Введение в математическую статистику. Оценивание характеристик распределения.

Леонид Иосипой

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

9 июня 2021

- Повторение
- Оценивание параметров распределения (продолжение)
- Оценивание характеристик распределения

1. Задача оценивания параметров распределения.

В статистике данные часто рассматриваются как реализация выборки из некоторого распределения, известного с точностью до одного или нескольких параметров. Будем оценивать параметры по выборке:

- $x_1, \ldots, x_n$ — это известные числа, реализация выборки (независимые и одинаково распределенные случайные величины) из функции распределения $F_\theta(u)$;
- $\theta \in \Theta$ — это неизвестный параметр, Θ — диапазон его возможных значений, а $\{F_\theta(u), \theta \in \Theta\}$ — семейство функций распределения;
- задача состоит в том, чтобы оценить (восстановить) θ по реализации выборки $x_1, \ldots, x_n$ наиболее точно.

1. Задача оценивания параметров распределения.

Оценка $\widehat{\theta}(x_1, \ldots, x_n)$ — это функция от n переменных. Подставляя в оценку $\widehat{\theta}$ реализацию выборки $x_1, \ldots, x_n$, мы получим число — оценку неизвестного параметра θ.

2. Свойства оценок.

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

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

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

2. Свойства оценок.

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

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

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).

3. Методы получения оценок.

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

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

3. Методы получения оценок.

(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 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

Так как распределение известно с точностью до θ:

- будем обозначать через $\mathbb{P}_\theta(X = u)$ вероятность принять какое-то значение в дискретном случае;
- будем обозначать плотность распределения через $f_\theta(u)$ в непрерывном случае.

Метод максимального правдоподобия

Введем величину:

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

Функцией правдоподобия называется величина:

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

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

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

Это и будет оценкой максимального правдоподобия.

Метод максимального правдоподобия

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.

Оценивание параметров распределения

Задача

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

Оценивание параметров распределения

Решение. Найдем сначала функцию правдоподобия:

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

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

Найдем логарифм функции правдоподобия:

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

Оценивание параметров распределения

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 выборки из экспоненциального распределения $\mathrm{Exp}(\theta)$ с неизвестным параметром $\theta > 0$. Оценить θ с помощью метода максимального правдоподобия.

Оценивание параметров распределения

Решение. Найдем сначала функцию правдоподобия:

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

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

где мы воспользовались тем, что $x_i \geq 0$ для всех $i = 1, \ldots, n$.

Перейдем к логарифму функции правдоподобия:

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

Оценивание параметров распределения

Приравняем производную к нулю:

$$\frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0.$$

Получим следующую оценку:

$$\widehat{\theta}(x_1, \ldots, x_n) = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}.$$

Свойства оценок

Какими свойствами обладают оценки, полученные методом моментов?

При некоторых условиях на регулярность модели:

- Возможная смещённость
- Состоятельность

Свойства оценок

Какими свойствами обладают оценки, полученные методом максимального правдоподобия?

При некоторых условиях на регулярность модели:

- Возможная смещённость
- Состоятельность
- Асимптотическая эффективность

  Это означает, что дисперсия при $n \to \infty$ является наименьшей возможной среди многих других оценок.

Оценивание характеристик распределения

Теперь допустим нам дана реализация выборки $x_1, \ldots, x_n$ из некоторого распределения, о котором мы ничего не знаем.

Как оценить характеристики этого распределения?

Оценивание характеристик распределения

Начнем с характеристик, которые называются «средними».

Теоретическое среднее	Выборочное среднее
Математическое ожидание: $$\mathbb{E}[X]$$	Выборочное среднее: $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

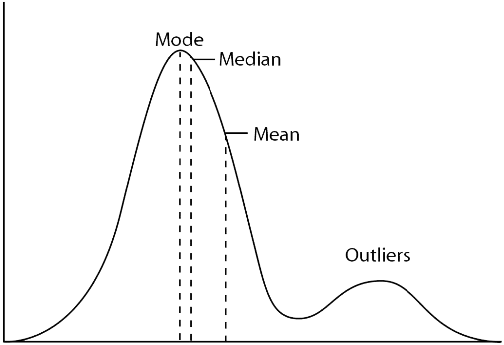
Оценивание характеристик распределения

Теоретическое среднее	Выборочное среднее
Теоретическая медиана: $$x_{1/2},$$ которая определяется как решение уравнения $$F(x_{1/2}) = \frac{1}{2},$$ где $F(u)$ — функция распределения.	Выборочная медиана: $$\mathrm{MED} = \begin{cases} x_{(k+1)}, & n = 2k+1, \\ (x_{(k)} + x_{(k+1)})/2, & n = 2k. \end{cases}$$ Здесь $$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$ это так называемый вариационный ряд, состоящий из упорядоченных по возрастанию элементов реализации выборки $x_1, \ldots, x_n$.

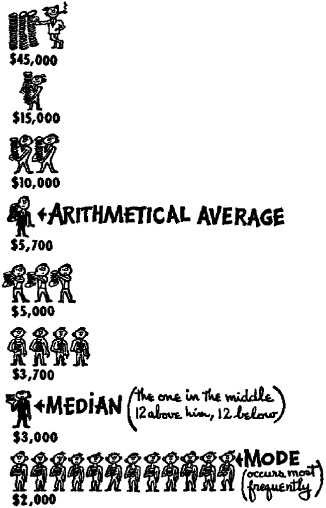
Оценивание характеристик распределения

Теоретическое среднее	Выборочное среднее
Теоретическая мода: ▶ В дискретном случае — значение, которое принимаются с наибольшей вероятностью. ▶ В непрерывном случае — точка максимума функции плотности.	Выборочная мода: ▶ В дискретном случае — самое распространенное значение реализации выборки. ▶ В непрерывном случае — нет.

Оценка среднего



Оценка среднего



Оценивание характеристик распределения

Как оценивать другие характеристики?

Допустим, мы хотим оценить $\mathbb{E}[g(X)]$, где X — случайная величина, из распределения которой получена выборка, а $g : \mathbb{R} \to \mathbb{R}$ — некоторая (известная) функция.

Это можно сделать с помощью оценки Монте-Карло:

$$\frac{1}{n} \sum_{i=1}^{n} g(x_i).$$

Оценивание характеристик распределения

Примеры:

1. Математическое ожидание:

$$\mathbb{E}[X] \approx \frac{1}{n} \sum_{i=1}^{n} x_i.$$

2. Моменты старшего порядка: для $k > 1$

$$\mathbb{E}[X^k] \approx \frac{1}{n} \sum_{i=1}^{n} x_i^k.$$

3. Более сложные функции. Например:

$$\mathbb{E}[X^3 \sin(X) \log(X)] \approx \frac{1}{n} \sum_{i=1}^{n} x_i^3 \sin(x_i) \log(x_i).$$

Оценивание характеристик распределения

Оценки Монте-Карло являются хорошими: они обладают несмещенностью и состоятельностью.

1. Несмещенность:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right] = \frac{1}{n}\Big(\mathbb{E}[g(X_1)] + \ldots \mathbb{E}[g(X_n)]\Big) = \mathbb{E}[g(X)].$$

2. Состоятельность: согласно закону больших чисел

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \quad \xrightarrow{\mathbb{P}} \quad \mathbb{E}[g(X)].$$

Оценивание характеристик распределения

А как на основе реализация выборки $x_1, \ldots, x_n$ из некоторого распределения X оценить дисперсию $\mathrm{Var}(X)$? Или какой-то другой центральный момент?

Оценивание характеристик распределения

Если бы математическое ожидание $\mathbb{E}[X]$ было бы известным, можно было бы воспользоваться оценкой Монте-Карло:

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mathbb{E}[X]\right)^2.$$

Но что делать, если $\mathbb{E}[X]$ неизвестно?

Оценивание характеристик распределения

Plug-in principle 1: если оценка некоторой характеристики требует знания каких-то других неизвестных характеристик, то можно попробовать подставить в оценку вместо неизвестных характеристик их оценки.

При этом, естественно, нет никаких гарантий, что у полученной оценки несмещенность и состоятельность сохранятся.

Оценивание характеристик распределения

Обозначим оценку для математического ожидания через $\overline{x}$,

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

Подставим ее в оценку для дисперсии, которая была выше:

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

Данная оценка будет состоятельной, но смещенной: она оценивает не $\mathrm{Var}(X)$, а $\frac{n-1}{n}\mathrm{Var}(X)$.

Оценка дисперсии

Действительно, используя свойства мат. ожидания, получаем:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{1}{n^2}\sum_{i,j=1}^{n}X_iX_j\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \frac{1}{n^2}\sum_{i\neq j}^{n}\mathbb{E}[X_i]\mathbb{E}[X_j]$$

$$= \mathbb{E}[X^2] - \frac{1}{n}\mathbb{E}[X^2] - \frac{n-1}{n}(\mathbb{E}X)^2$$

$$= \frac{n-1}{n}\mathrm{Var}(X).$$

Оценивание характеристик распределения

Смещение можно исправить, умножив оценку на дробь, обратную к $\frac{n-1}{n}$. Получим несмещенную оценку дисперсии S^2:

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

Оценивание характеристик распределения

Plug-in principle 2: если необходимо оценить какую-то функцию от нескольких неизвестных характеристик, то можно подставить оценки характеристик в эту функцию.

Опять же, никаких гарантий, что мы получим «хорошую» оценку, нет. Построенные оценки, скорее всего, придется корректировать, как и в случае с дисперсией.

Оценивание характеристик распределения

Например, оценкой стандартного отклонения по смещённой дисперсии является:

$$\widehat{\sigma}_b = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

А по несмещённой дисперсии:

$$\widehat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Обе оценки являются смещенными.

Оценивание характеристик распределения

Оценка коэффициента асимметрии

$$\gamma_1 = \frac{\mathbb{E}[(X - \mathbb{E}X)^3]}{(\mathrm{Var}(X))^{3/2}}$$

такая

$$\widehat{\gamma}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right)^{3/2}}.$$

Эта оценка тоже является смещенной.

Оценивание характеристик распределения

Оценка коэффициента эксцесса

$$\gamma_2 = \frac{\mathbb{E}[(X - \mathbb{E}X)^4]}{(\mathrm{Var}(X))^2}$$

такая

$$\widehat{\gamma}_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right)^2}.$$

Эта оценка тоже является смещенной.

Спасибо за внимание!