Числовые характеристики распределений. Метод максимального правдоподобия.

Леонид Иосипой

Курс «Вероятностные модели и статистика»
Центр непрерывного образования, ВШЭ

8 апреля 2021

- Повторение
- Числовые характеристики случайных величин
- Задача поиска больных
- Метод максимального правдоподобия

Повторение

1. Независимость.

   Случайные величины X и Y называются независимыми тогда и только тогда, когда для любых (борелевских) подмножеств $A, B \subset \mathbb{R}$ выполняется равенство

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B).$$

Повторение

2. Числовые характеристики случайных величин.

1) Математическое ожидание.

Для дискретной случайной величины X с таблицей распределения

X	a_1	a_2	a_3	$\ldots$
$\mathbb{P}$	p_1	p_2	p_3	$\ldots$

математическим ожиданием называется число $\mathbb{E}[X]$, которое вычисляется по формуле

$$\mathbb{E}[X] = \sum_i a_i \cdot \mathbb{P}(X = a_i) = \sum_i a_i p_i.$$

Повторение

2. Числовые характеристики случайных величин.

1) Математическое ожидание.

Для непрерывной случайной величины X с плотностью $f(u)$ математическим ожиданием называется число $\mathbb{E}[X]$, которое вычисляется по формуле

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} u \cdot f(u) du.$$

Повторение

2. Числовые характеристики случайных величин.

1) Математическое ожидание.

Оба определения согласованы согласно правилам перехода от дискретного случая к непрерывному:

$$\sum_i \longleftrightarrow \int_{-\infty}^{+\infty}, \qquad a_i \longleftrightarrow u \qquad \text{и} \qquad p_i \longleftrightarrow f(u)du.$$

$$\sum_i a_i \cdot p_i \longleftrightarrow \int_{-\infty}^{+\infty} u \cdot f(u)du.$$

Повторение

2. Числовые характеристики случайных величин.

1) Математическое ожидание.

(E1) $\mathbb{E}[c] = c$ для любого $c \in \mathbb{R}$;

(E2) $\mathbb{E}[cX] = c\mathbb{E}[X]$ для любого $c \in \mathbb{R}$;

(E3) $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ для любых X и Y;

(E4) $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$, если X и Y независимы;

(E5) Если $X \geq 0$, то $\mathbb{E}[X] \geq 0$.

Повторение

2. Числовые характеристики случайных величин.

1) Математическое ожидание.

Как найти $\mathbb{E}[g(X)]$ для произвольной функции $g : \mathbb{R} \to \mathbb{R}$?

- Если X имеет дискретное распределение и принимает значения a_i с вероятностями p_i, то

$$\mathbb{E}[g(X)] = \sum_i g(a_i) p_i.$$

- Если X имеет непрерывное распределение с плотностью распределения $f(u)$, то

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(u) f(u) du.$$

Повторение

2. Числовые характеристики случайных величин.

2) Дисперсия.

Дисперсией случайной величины X называется число $\mathrm{Var}(X)$, которое вычисляется по формуле

$$\mathrm{Var}(X) = \mathbb{E}\Big[(X - \mathbb{E}X)^2\Big] = \mathbb{E}\big[X^2\big] - \big(\mathbb{E}X\big)^2.$$

Это определение и для дискретных, и для непрерывных распределений. Но обратите внимание, что мат. ожидание считается в этих случаях по-разному.

Повторение

2. Числовые характеристики случайных величин.

2) Дисперсия.

(V1) $\mathrm{Var}(c) = 0$ для любого $c \in \mathbb{R}$.

(V2) $\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$ для любого $c \in \mathbb{R}$.

(V3) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, если X и Y независимы.

(V4) $\mathrm{Var}(X + c) = \mathrm{Var}(X)$ для любого $c \in \mathbb{R}$.

(V5) $\mathrm{Var}(X) \geq 0$.

Повторение

2. Числовые характеристики случайных величин.

- если $X \sim \mathbf{B}_p$, то $\mathbb{E}[X] = p$, $\mathrm{Var}(X) = p(1-p)$;
- если $X \sim \mathbf{B}_{n,p}$, то $\mathbb{E}[X] = np$, $\mathrm{Var}(X) = np(1-p)$;
- если X равномерно распределен на $[a, b]$, то $\mathbb{E}[X] = (a+b)/2$, $\mathrm{Var}(X) = (b-a)^2/12$;
- если $X \sim \mathcal{N}(a, \sigma^2)$, то $\mathbb{E}[X] = a$, $\mathrm{Var}(X) = \sigma^2$.

Числовые характеристики случайных величин

2. Дисперсия.

Среднеквадратическое отклонение (или стандартное отклонение) — это квадратный корень из дисперсии случайной величины:

$$\sigma = \sqrt{\mathrm{Var}(X)}.$$

Разделив случайную величину X на σ, мы получим случайную величину с дисперсией $\mathrm{Var}(X/\sigma) = 1$. Данная операция называется нормировкой.

Числовые характеристики случайных величин

3. Моменты старших порядков.

- $\mathbb{E}[X^k]$ — k-ый момент X;
- $\mathbb{E}[|X|^k]$ — абсолютный k-ый момент X;
- $\mathbb{E}[(X - \mathbb{E}X)^k]$ — центральный k-ый момент X;
- $\mathbb{E}[|X - \mathbb{E}X|^k]$ — абсолютный центральный k-ый момент X.

Все вышеприведенные моменты характеризуют распределение случайной величины. Они часто появляются в задачах концентрации (о них мы поговорим позже).

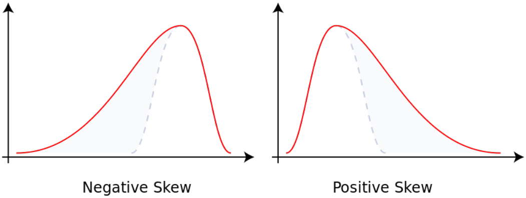
Числовые характеристики случайных величин

4. Коэффициент асимметрии (Skewness).

Коэффициент асимметрии — величина, характеризующая асимметрию распределения данной случайной величины.

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

γ_1 — нормированный центральный момент порядка 3.



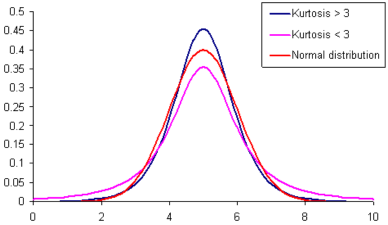
Числовые характеристики случайных величин

4. Коэффициент эксцесса (Kurtosis).

Коэффициент эксцесса — мера остроты пика плотности распределения случайной величины.

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

γ_2 — нормированный центральный момент порядка 4.



Задача поиска больных

Предыстория. Во время Второй мировой войны всех призывников в армию США подвергали медицинскому обследованию. Реакция Вассермана позволяет обнаруживать в крови больных сифилисом определенные антитела.

Р. Дорфманом была предложена простая методика, на основе которой необходимое для выявления всех больных число проверок удалось уменьшить в 5 раз!

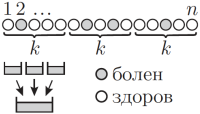
Задача поиска больных

Вероятностная модель. Предположим, что:

- вероятность обнаружения антител p одна и та же для всех n обследуемых; для определенности пусть $p = 0.01$.
- результаты анализов для различных людей независимы (то есть моделью является последовательность из n независимых бернуллиевских случайных величин с вероятностью «успеха» p).

Задача поиска больных

Методика. Смешиваются пробы крови k человек и анализируется полученная смесь.



Пусть X_j — количество проверок в j-й группе, $j = 1, \dots, n/k$.

$$X_j = \begin{cases} 1 & \text{с вероятностью } (1-p)^k \quad \text{(все } k \text{ человек здоровы)}, \\ k+1 & \text{с вероятностью } 1-(1-p)^k \quad \text{(есть больные)}. \end{cases}$$

Задача поиска больных

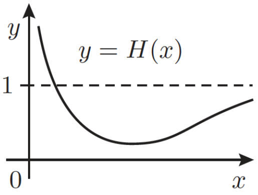
Будем подбирать оптимальное k, минимизирующее общее ожидаемое количество проверок $Z = X_1 + \ldots + X_{n/k}$.

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}[X_1 + \ldots + X_{n/k}] \\
&= \frac{n}{k}\Big(1 \cdot (1-p)^k + (k+1) \cdot (1 - (1-p)^k)\Big) \\
&= \frac{n}{k}\Big(k + 1 - k(1-p)^k\Big) \\
&= n\Big(1 + 1/k - (1-p)^k\Big).
\end{aligned}$$

Положим $H(x) = 1 + 1/x - (1-p)^x$ при $x > 0$.

Задача поиска больных

Необходимо найти минимум $H(x) = 1 + 1/x - (1-p)^x$, $x > 0$.



Уравнение $H'(x) = 0$ будет следующим:

$$1/x^2 + (1-p)^x \ln(1-p) = 0.$$

Оно не решается! :(

Задача поиска больных

Заменим функцию $H(x)$ на ее аппроксимацию (рядом Тейлора) первого порядка

$$H(x) \approx 1 + 1/x - (1 - px) = 1/x + px,$$

имеющую точку минимума $x_{\min} = 1/\sqrt{p}$.

Пусть $p = 0.01$, тогда $k_{\min} = 1/\sqrt{p} = 10$.

Получим для $k = k_{\min}$:

$$\begin{aligned}\mathbb{E}[Z] &\approx n\Big(1 + 1/k_{\min} - (1 - k_{\min} p)\Big) \\ &= n\left(\frac{1}{10} + \frac{1}{10}\right) \\ &= \frac{n}{5}.\end{aligned}$$

Метод максимального правдоподобия

Перейдем теперь к методам оценки неизвестных параметров.

Допустим, что у нас есть реализация выборки из некоторого распределения, известного с точностью до одного или нескольких параметров.

Как на основе этих данных оценить значения параметров?

Метод максимального правдоподобия

Известно много различных методов построения оценок: метод моментов, метод максимального правдоподобия, метод спейсингов и т.д.

Мы изучим всего один из них — метод максимального правдоподобия. Он является наиболее важным, популярным и интуитивно понятным методом.

Метод максимального правдоподобия

В теории оценивания неизвестные параметры обычно обозначаются через $\theta_1, \theta_2, \ldots, \theta_d$.

Чтобы упростить формулировки и обозначения, мы будем считать, что неизвестный параметр многомерный:

$$\theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \Theta \subset \mathbb{R}^d.$$

Метод максимального правдоподобия

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

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 $x_1, \dots, x_n$ из некоторого распределения с неизвестным (многомерным) параметром θ.

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

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 выборки из экспоненциального распределения с неизвестным параметром интенсивности $\theta > 0$, плотность распределения которого равна

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

Оценить θ с помощью метода максимального правдоподобия.

Метод максимального правдоподобия

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:

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

Метод максимального правдоподобия

Приравняем производную к нулю:

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

Получим следующую оценку:

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

Спасибо за внимание!