Теория 1. Сопряжённые распределения и экспоненциальный класс распределений.

Якаков Дмитрий Спартакович, 417 группа

$\boxed{\text{Задание 3}}$

$\mathrm{Pareto}(x|a,b) = \frac{b a^b}{x^{b+1}} I[x \geq a]$, a - фикс.

$\mathrm{Pareto}(x|a,b) = \frac{b a^b}{x} I[x \geq a] \cdot e^{-b \log x} =$

$= \frac{f(x)}{g(\theta)} \exp(\theta^T u(x))$,

где $f(x) = \frac{1}{x} I[x \geq a]$;

$g(\theta) = \frac{1}{b a^b}$;

$\theta = \begin{pmatrix} a \\ b \end{pmatrix}$;

$u(x) = \begin{pmatrix} u_1(x) \\ u_2(x) \end{pmatrix} = \begin{pmatrix} 0 \\ -\log x \end{pmatrix}$.

$\mathbb{E} \log x = \mathbb{E} - u_2(x) = -\frac{\partial}{\partial b} \log g(\theta) =$

$= \frac{\partial}{\partial b} \log[b a^b] = \frac{1}{b a^b}[a^b + b a^b \log a] = \frac{1}{b}[1 + b \log a] =$

$= \frac{1}{b} + \log a$

Ответ: $\mathbb{E} \log x = \frac{1}{b} + \log a$.

$\boxed{\text{Задание 1}}$

а) $\theta_{ML} = \arg\max\limits_{\theta} p(x_1, \dots, x_N|\theta) = \arg\max\limits_{\theta} \prod\limits_{i=1}^{N} p(x_i|\theta) =$

$= \arg\max\limits_{\theta} \prod\limits_{i=1}^{N} \frac{1}{\theta} I[0 \leq x_i \leq \theta] = \arg\max\limits_{\theta} \frac{1}{\theta^n} I[x_{(0)} \geq 0] \cdot$

$\cdot I[x_{(N)} \le \theta] = x_{(N)}$,

где $x_{(0)} = \min_i x_i$, $x_{(N)} = \max_i x_i$

б) Сопряжённым распределением к $U[0,\theta]$ будет распределение Парето ~~$p(\theta) = Pareto(\theta|a,b)$~~

$$p(\theta) = \text{Pareto}(\theta|a,b) = \frac{ba^b}{\theta^{b+1}} I[\theta \ge a].$$

Покажем это:

$$p(\theta|X) = \frac{1}{Z}\left(\prod_{i=1}^{N} \frac{1}{\theta} I[0 \le x_i \le \theta]\right) \cdot \frac{ba^b}{\theta^{b+1}} I[\theta \ge a] =$$

$$= \frac{1}{\theta^{b+N+1}} I[\theta \ge \max\{x_{(N)}, a\}] \cdot \underbrace{\frac{1}{Z} \cdot ba^b \cdot I[x_{(0)} \ge 0]}_{C} =$$

$$= C \cdot \frac{1}{\theta^{b+N+1}} I[\theta \ge \max\{x_{(N)}, a\}]$$

Нормировочную константу C найдём из след. усл-я:

$$\int_{\max\{x_{(N)},a\}}^{+\infty} C \cdot \frac{1}{\theta^{b+N+1}} d\theta = C \cdot \frac{1}{\theta^{b+N}} \cdot \frac{1}{-(b+N)} \Bigg|_{\max\{x_{(N)},a\}}^{+\infty} =$$

$$= C \cdot \frac{1}{(\max\{x_{(N)},a\})^{b+N}} \cdot \frac{1}{b+N} = 1 \Rightarrow C = [\max\{x_{(N)},a\}]^{b+N} \cdot (b+N)$$

$\Rightarrow$ апостериорное распределение $p(\theta / x_1, \ldots, x_N)$ имеет вид:

$$p(\theta|X) = \frac{(b+N)\cdot[\max\{x_{(N)},a\}]^{b+N}}{\theta^{b+N+1}} I[\theta \ge \max\{x_{(N)},a\}] =$$

$$= \{a' = \max\{x_{(N)},a\},\ b' = b+N\} = \frac{b' a'^{b'}}{\theta^{b'+1}} \cdot I[\theta \ge a'] =$$

$= \text{Pareto}(\theta | a', b')$ $\Rightarrow$ исходное предположение об априорном распределении было верным в) Найдём статистики апостериорного распределения: мат. ожидание, медиану и моду.

$\theta \sim Pareto(X, a', b')$

- $\mathbb{E}\theta = \int\limits_{a'}^{+\infty} \frac{b' a'^{b'}}{\theta^{b'+1}} \cdot \theta \, d\theta = \int\limits_{a'}^{+\infty} \frac{1}{\theta^{b'}} \cdot b' a'^{b'} \, d\theta = b' a'^{b'} \cdot \frac{1}{\theta^{b'-1}} \cdot$

$\cdot \frac{1}{-b'+1} \Big|_{a'}^{+\infty} = b' a'^{b'} \cdot \frac{1}{a'^{b'-1}} \cdot \frac{1}{b'-1} = \frac{b' a'}{b'-1}$,

причём мат. ожидание $\exists$ при $b'-1>0$, $b'>1$

- $\int\limits_{a'}^{med\theta} \frac{b' a'^{b'}}{\theta^{b'+1}} d\theta = b' a'^{b'} \frac{1}{\theta^{b'}} \cdot \frac{1}{-b'} \Big|_{a'}^{med\theta} = b' a'^{b'} \frac{1}{(med\theta)^{b'}} \cdot \frac{1}{-b'} +$

$+ 1 = \frac{1}{2} \Rightarrow (med\theta)^{b'} = 2a'^{b'} \Rightarrow med\theta = \sqrt[b']{2} \cdot a'$

- Мода - это такое значение, на котором дост-ся максимум пл-ти распределения.

$Pareto(\theta | X, a', b') = \frac{b' a'^{b'}}{\theta^{b'+1}} I[\theta \geq a']$

ф-ция явл-ся убывающей $\Rightarrow mod\theta = a'$

<u>Ответ:</u> 1) $\theta_{ML} = x_{(N)}$;

2) $p(\theta) = Pareto(\theta | a, b)$;

3) $p(\theta | X) = Pareto(\theta | X, a', b')$,

где $a' = \max\{x_{(N)}, a\}$, $b' = b + N$;

4) $\left.\begin{aligned} \mathbb{E}\theta &= \frac{1}{b'-1} b' a' \ (b'>1), \\ med\theta &= \sqrt[b']{2} \cdot a', \\ mod\theta &= a' \end{aligned}\right\} \theta \sim p(\theta | X)$

Задание 2

- ∃ в городе всего ℓ маршрутов автобусов. Тогда вер-ть встретить автобус с номером i ~~[illegible]~~ равняется $\frac{1}{\ell}$, при условии, что мы не знаем никакие дополнительные данные $\Rightarrow$ номера маршрутов распределены равномерно

Т.к. ℓ – неизв. параметр, то номера $\sim U[0, \theta]$

- Чтобы использовать байесовский вывод, положим в качестве априорного распределение на θ сопряжённое к равномерному распр-ю – распр-е Парето, т.е. $p(\theta) = \mathrm{Pareto}(\theta | a, b)$
- Поскольку ∃ как минимум 1 автобусный маршрут, то логично положить $a = 1$.

Чтобы оценить параметр b, воспользуемся данными из интернета. Возьмём город Краснодар, в котором 100 маршрутов, и пусть Краснодар меньше около 10% городов России. Тогда:

$$P(\theta=1)+\dots+P(\theta=100) = \Big\{ P(\theta=i) = \int_i^{i+1} p(\theta)\, d\theta -$$

таким образом применим непр. распределение к дискретным автобусам. $\Big\} =$

$$= \int_1^{101} \frac{b\, 1^b}{\theta^{b+1}}\, d\theta = b \cdot \frac{1}{\theta^b} \cdot \frac{1}{-b} \Big|_1^{101} = -\frac{1}{101^b} + 1 = 0.9$$

$$\Rightarrow 101^b = \frac{1}{0.1} \Rightarrow b = \log_{10} 101 \approx 2$$

Итого, $a = 1$, $b = 2$.

• Наиболее адекватной статистикой апостериорного распределения будет медиана, так как:

1) она, в отличие от мат. ожидания, устойчива к выбросам (к примеру, маленькие города, где мало маршрутов, и большие города, где много маршрутов);

2) она, в отличие от моды, учитывает параметр „b" из распределения (см. задание 1).

• Найдём оценки на кол-во автобусных маршрутов

$\boxed{a)}$ увидели автобус с номером 100:

$$\left.\begin{aligned} a' &= \max\{x_{(n)}, a\} = \max\{100, 1\} = 100 \\ b' &= b + N = 2 + 1 = 3 \end{aligned}\right\} \Rightarrow$$

$$\operatorname{med}\Theta = \sqrt[3]{2} \cdot 100 \approx 126$$

$\boxed{б)}$ увидели ещё автобус с номером 50:

$$\left.\begin{aligned} a' &= \max\{x_{(n)}, a\} = \max\{100, 1\} = 100 \\ b' &= b + N = 2 + 2 = 4 \end{aligned}\right\} \Rightarrow$$

$$\text{med}\,\theta = \sqrt[4]{2} \cdot 100 \approx 119$$

$\boxed{в}$ увидели ещё автобус с номером 150:

$$\left.\begin{aligned} a' &= \max\{x_{(n)}, a\} = \max\{150, 1\} = 150 \\ b' &= b + N = 2 + 3 = 5 \end{aligned}\right\} \Rightarrow$$

$$\text{med}\,\theta = \sqrt[5]{2} \cdot 150 \approx 172$$