Математическая статистика. ДЗ 13.

Прохоров Юрий, 776

Задача 1

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — простая выборка из распределения с плотностью $f_\theta(x)$. Построить оптимальную оценку параметра θ, если

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

(b) $f_\theta(x) = \begin{cases} \dfrac{1}{2\theta} e^{-\sqrt{x/\theta}} & , \ x \geq 0 \\ 0 & , \ x < 0 \end{cases}$

Решение:

(a) Данная модель не является регулярной, поэтому будем строить эффективную оценку следующим алгоритмом:

1. Найдем полную достаточную статистику $T(\mathbf{X})$.
2. Найдем какую-нибудь простую несмещенную оценку $\widehat{\theta}_1(\mathbf{X})$.
3. Согласно теореме Рао-Блеквелла-Колмогорова, построим оптимальную оценку

$$\widehat{\theta}(\mathbf{X}) = \mathbb{E}\big[\widehat{\theta}_1(\mathbf{X}) \bigm| T(\mathbf{X})\big]$$

Будем действовать по этому алгоритму.

1. Функция правдоподобия выборки:

$$L(\mathbf{X} \,|\, \theta) = \prod_{i=1}^{n} \mathbb{I}\{X_i \geq \theta\} e^{-(X_i - \theta)} = \underbrace{\mathbb{I}\{X_{(1)} \geq \theta\} \cdot e^{n\theta}}_{g(X_{(1)}, \theta)} \cdot \underbrace{e^{-\sum_{i=1}^n X_i}}_{h(\mathbf{X})}$$

   По критерию факторизации, статистика

$$T(\mathbf{X}) = X_{(1)} = \min_{i=\overline{1,n}} X_i$$

   является **достаточной статистикой**.

2. Найдем ее распределение. Воспользуемся следующим свойством экспоненциального распределения:

$$\left.\begin{matrix} \xi_1, \ldots, \xi_n \text{ — независимые} \\ \xi_i \sim \mathrm{Exp}(\lambda_i) \end{matrix}\right\} \quad \Longrightarrow \quad \xi_{(1)} = \min_{i=\overline{1,n}} \xi_i \sim \mathrm{Exp}(\lambda_1 + \ldots + \lambda_n)$$

   У нас $X_i = \theta + \xi_i$, где $\xi_i \sim \mathrm{Exp}(1)$ — i.i.d.. Поэтому $X_{(1)} = \theta + \xi_{(1)}$, где $\xi_{(1)} \sim \mathrm{Exp}(n)$, поэтому

$$f_{X_{(1)}}(x) = \begin{cases} ne^{-n(x-\theta)} & , \ x \geq \theta, \\ 0 & , \ x < \theta. \end{cases}$$

3. Проверим полноту этой статистики. Пусть выполнено, что

$$\mathbb{E}_\theta\big[g(X_{(1)})\big] = \int\limits_{\theta}^{+\infty} g(x)\, ne^{-n(x-\theta)}\, dx = 0, \qquad \forall \theta \in \mathbb{R}$$

$$\int_{\theta}^{+\infty} g(x)\, e^{-nx}\, dx = 0, \qquad \forall \theta \in \mathbb{R}$$

Дифференцируем это равенство по θ:

$$-g(\theta)e^{-n\theta} = 0 \qquad \Longrightarrow \qquad g(\theta) = 0, \qquad \forall \theta \in \mathbb{R}$$

Таким образом, статистика $X_{(1)}$ является **полной достаточной статистикой**.

4. Далее, найдем какую-нибудь несмещенную оценку. Заметим, что если $\xi \sim \mathrm{Exp}(\lambda)$, то $\mathbb{E}\xi = \frac{1}{\lambda}$. Тогда

$$\mathbb{E}_\theta X_1 = \theta + 1 \qquad \Longrightarrow \qquad \widehat{\theta}_1 = X_1 - 1 \quad — \quad \text{несмещенная оценка}$$

По теореме Рао-Блеквелла-Колмогорова, оценка

$$\widehat{\theta} = \mathbb{E}_\theta \big[\widehat{\theta}_1 \,\big|\, X_{(1)}\big] = \mathbb{E}_\theta \big[X_1 \,\big|\, X_{(1)}\big] - 1$$

является оптимальной.

Вычислим данное условное матожидание.

- Функция распределения X_i:

$$F_{X_i}(x) = \mathbb{P}_\theta\{X_i < x\} = \max(0, 1 - e^{\theta - x})$$

- Функция распределения $X_{(1)}$:

$$F_{X_{(1)}}(t) = \mathbb{P}_\theta\{X_{(1)} < t\} = \max(0, 1 - e^{n(\theta - t)})$$

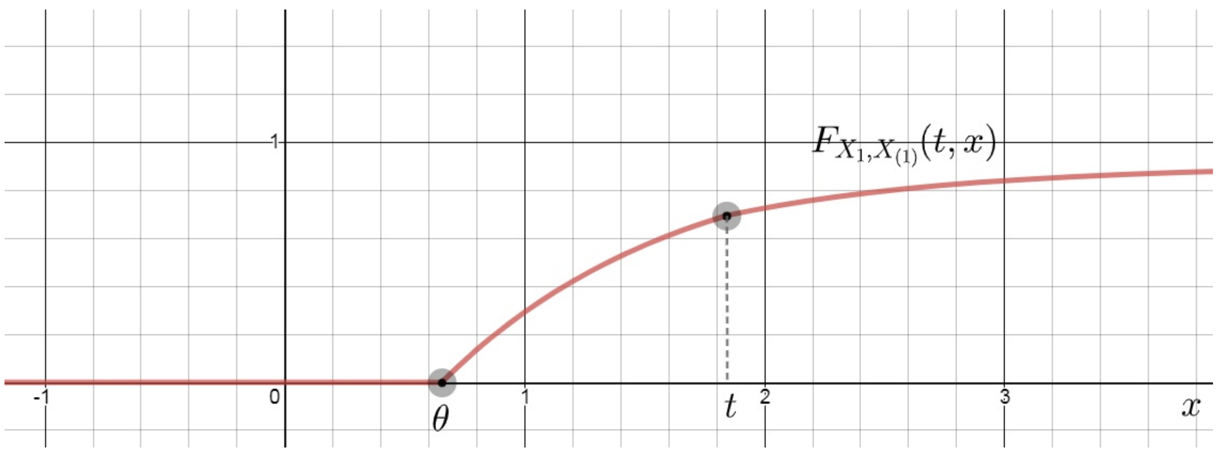
- Функция совместного распределения X_1 и $X_{(1)}$:

$$\begin{aligned} F_{X_1, X_{(1)}}(x,t) &= \mathbb{P}_\theta\{X_1 < x,\, X_{(1)} < t\} = \mathbb{P}_\theta\{X_1 < x\} - \mathbb{P}_\theta\{X_1 < x,\, X_{(1)} \geq t\} = \\ &= \mathbb{P}_\theta\{X_1 < x\} - \mathbb{P}_\theta\{t \leq X_1 < x\} \cdot \mathbb{P}_\theta\{X_2 \geq t\} \cdots \mathbb{P}_\theta\{X_n \geq t\} \end{aligned}$$

Легко видеть, что при $x \leq \theta$ или $t \leq \theta$ эта вероятность равна 0. В остальных случаях легко вычислить:

$$F_{X_1, X_{(1)}}(x,t) = \begin{cases} 1 - e^{n(\theta - t)} - e^{\theta - x}\big(1 - e^{(n-1)(\theta - t)}\big) & , \qquad \theta < t < x, \\ 1 - e^{\theta - x} & , \qquad \theta < x \leq t, \\ 0 & , \qquad x \leq \theta \text{ или t} \leq \theta. \end{cases}$$

Вот так ведет себя эта функция при фиксированном $t > \theta$.



Важно, что эта функция непрерывна, хотя ее частная производная по x разрывна в точках $x = 0$ и $x = t$.

- Функция условного распределения:

$$F_{X_1\,|\,X_{(1)}}(x\,|\,t) = \mathbb{P}_\theta\{X_1 < x \,\big|\, X_{(1)} = t\}$$

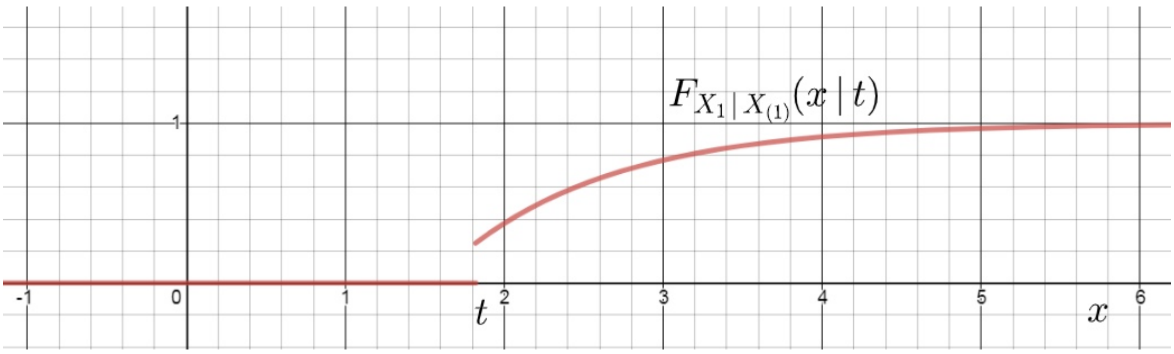
При $t \le \theta$ она неопределена, при $x \le \theta$ и при $t \ge x$ она равна 0. При $\theta < t < x$:

$$F_{X_1\,|\,X_{(1)}}(x\,|\,t) = \mathbb{P}_\theta\{X_1 < x \,\big|\, X_{(1)} = t\} = \lim_{\Delta\to 0} \mathbb{P}_\theta\big\{X_1 < x \,\big|\, t \le X_{(1)} \le t+\Delta\big\}$$

При достаточно малом $\Delta > 0$:

$$\mathbb{P}_\theta\big\{X_1 < x \,\big|\, t \le X_{(1)} \le t+\Delta\big\} = \frac{\mathbb{P}_\theta\big\{X_1 < x,\, t \le X_{(1)} \le t+\Delta\big\}}{\mathbb{P}_\theta\big\{t \le X_{(1)} \le t+\Delta\big\}} = \frac{F_{X_1,X_{(1)}}(x,t+\Delta) - F_{X_1,X_{(1)}}(x,t)}{F_{X_{(1)}}(t+\Delta) - F_{X_{(1)}}(t)} =$$
$$= \frac{e^{n(\theta-t)}\big(1-e^{-n\Delta}\big) - e^{\theta-x}e^{(n-1)(\theta-t)}\big(1-e^{-(n-1)\Delta}\big)}{e^{n(\theta-t)}\big(1-e^{-n\Delta}\big)} =$$
$$= 1 - \frac{e^{\theta-x}}{e^{\theta-t}} \cdot \frac{1-e^{-(n-1)\Delta}}{1-e^{-n\Delta}} = 1 - e^{t-x}\frac{(n-1)\Delta + o(\Delta)}{n\Delta + o(\Delta)} =$$
$$= 1 - e^{t-x}\frac{n-1+o(1)}{n+o(1)} \xrightarrow{\Delta\to+0} 1 - \frac{n-1}{n}e^{t-x}$$

При фиксированном $t > \theta$ она имеет вид:



В точке t эта функция имеет разрыв с 0 до $\frac{1}{n}$. Это соответствует тому, что

$$\mathbb{P}\{X_1 = t\,|\,X_{(1)} = t\} = \frac{1}{n}$$

Это логично, потому что все X_i независимы и одинаково непрерывно распределены, поэтому вероятность X_1 быть наименьшим значением есть $\frac{1}{n}$.

- Вычислим матожидание при фиксированном значении $X_{(1)}$:

$$m(t) = \mathbb{E}_\theta\big[X_1 \,\big|\, X_{(1)} = t\big] = \int\limits_\theta^{+\infty} x\,dF_{X_1\,|\,X_{(1)}}(x\,|\,t) = t\cdot\mathbb{P}\{X_1 = t\,|\,X_{(1)} = t\} + \int\limits_t^{+\infty} x\,F'_{X_1\,|\,X_{(1)}}(x\,|\,t)\,dx =$$
$$= t\cdot\frac{1}{n} + \frac{n-1}{n}\underbrace{\int\limits_t^{+\infty} xe^{t-x}\,dx}_{t+1} = \frac{t}{n} + \frac{(n-1)(t+1)}{n} = t + 1 - \frac{1}{n}$$

Данную цепочку равенств можно проинтерпретировать:

$$\mathbb{E}_\theta\big[X_1 \,\big|\, X_{(1)} = t\big] = \frac{1}{n}\cdot t + \frac{n-1}{n}\int\limits_t^{+\infty} xe^{t-x}\,dx$$

Здесь с вероятностью $\frac{1}{n}$ оказывается, что X_1 — минимальный элемент выборки. Соответствующее его значение равно t. С вероятностью $\frac{n-1}{n}$ оказывается, что X_1 имеет распределение $\mathrm{Exp}(1)$, смещенное на t.

- Искомое условное матожидание:

$$\mathbb{E}_\theta\big[X_1 \,\big|\, X_{(1)}\big] = m(X_{(1)}) = X_{(1)} + 1 - \frac{1}{n}$$

Итого, оптимальная оценка:

$$\widehat{\theta} = X_{(1)} - \frac{1}{n}$$

Важно заметить, что необязательно было проводить все эти громоздкие вычисления, а можно было воспользоваться теоремой Лемана-Шеффе:

Теорема Лемана-Шеффе: (Lehmann-Scheffé)

> Пусть $T(\mathbf{X})$ — полная достаточная статистика, и функция $g(\cdot)$ такая, что
>
> $$\mathbb{E}_\theta\big[g(T(\mathbf{X}))\big] = \theta, \qquad \forall\theta$$
>
> Тогда $\widehat{\theta} = g(T(\mathbf{X}))$ — оптимальная оценка параметра θ.

В этом случае функцию $g(t) = t - \frac{1}{n}$ легко угадать.

(b) Данная модель является регулярной, поэтому оптимальную оценку можно искать следующим способом:

1. Построим оценку максимального правдоподобия $\widehat{\theta}^{\text{ОМП}}$.
2. Проверим, является ли это эффективной оценкой.
3. Если эффективная оценка существует, то она является как оценкой максимального правдоподобия, так и оптимальной оценкой.

Будем действовать по этому алгоритму.

1. Логарифм функция правдоподобия:

$$\ln L(\mathbf{X}\,|\,\theta) = \sum_{i=1}^{n} \ln\left(\frac{1}{2\theta} e^{-\sqrt{X_i/\theta}}\right) = -n\ln(2\theta) - \frac{1}{\sqrt{\theta}}\sum_{i=1}^{n}\sqrt{X_i}$$

$$\frac{d\ln L}{d\theta} = -\frac{n}{\theta} + \frac{1}{2\theta^{3/2}}\sum_{i=1}^{n}\sqrt{X_i} = 0 \qquad \Longrightarrow \qquad \widehat{\theta}^{\text{ОМП}} = \left(\frac{1}{2n}\sum_{i=1}^{n}\sqrt{X_i}\right)^2$$

2. Проверим эффективность, то есть что случайные величины $\frac{d\ln L}{d\theta}$ и $\big(\widehat{\theta} - \theta\big)$ пропорциональны.

$$\frac{d\ln L}{d\theta} = \frac{1}{2\theta^{3/2}}\left(\sum_{i=1}^{n}\sqrt{X_i} - 2n\sqrt{\theta}\right)$$

   Пропорциональности нет, поэтому ОМП — неэффективная оценка. Про оптимальность ничего сказать не можем.

3. Построим достаточную статистику $T(\mathbf{X})$ с помощью критерия факторизации:

$$L(\mathbf{X}\,|\,\theta) = \frac{1}{(2\theta)^n}\exp\left\{-\frac{1}{\sqrt{\theta}}\sum_{i=1}^{n}\sqrt{X_i}\right\} \qquad \Longrightarrow \qquad T(\mathbf{X}) = \sum_{i=1}^{n}\sqrt{X_i}$$

   Найдем распределение $\sqrt{X}$. При $x \geq 0$:

$$f_{\sqrt{X}}(x) = \frac{d}{dx}F_{\sqrt{X}}(x) = \frac{d}{dx}\mathbb{P}\{\sqrt{X} < x\} = \frac{d}{dx}\mathbb{P}\{X < x^2\} = \frac{d}{dx}F_X(x^2) = f_X(x^2)\cdot 2x = \frac{x}{\theta}e^{-x/\sqrt{\theta}}$$

Итого получаем, что

$$f_{\sqrt{X}}(x) = \begin{cases} \frac{x}{\theta} \cdot e^{-x/\sqrt{\theta}} & , \ x \geq 0 \\ 0 & , \ x < 0 \end{cases} \qquad \Longrightarrow \qquad \sqrt{X_i} \sim \text{Gamma}\left(2, \sqrt{\theta}\right)$$

Если $\xi \sim \text{Gamma}(k, \lambda)$, то плотность распределения

$$f_\xi(x) = \frac{x^{k-1}}{\Gamma(k)\lambda^k} \cdot e^{-x/\lambda}, \qquad x \geq 0$$

Воспользуемся свойством аддитивности гамма-распределения: если $\xi_i \sim \Gamma(k_i, \lambda)$ — независимые, то

$$\xi_1 + \ldots + \xi_n \sim \Gamma\left(\sum_{i=1}^n k_i, \, \lambda\right)$$

Таким образом, достаточная статистика имеет распределение

$$T(\mathbf{X}) = \sum_{i=1}^n \sqrt{X_i} \sim \text{Gamma}\big(2n, \sqrt{\theta}\big), \qquad f_T(x) = \begin{cases} \dfrac{x^{2n-1}}{(2n-1)!\,\theta^n} \cdot e^{-x/\sqrt{\theta}} & , \ x \geq 0 \\ 0 & , \ x < 0 \end{cases}$$

4. Проверяем полноту этой статистики. Пусть для $\forall \theta > 0$:

$$\mathbb{E}_\theta\big[g(T(\mathbf{X}))\big] = \int\limits_0^{+\infty} g(x)\, \frac{x^{2n-1}}{(2n-1)!\,\theta^n} \cdot e^{-x/\sqrt{\theta}}\, dx = 0$$

Обозначим $y = \frac{1}{\sqrt{\theta}} > 0$. Тогда получаем, что

$$\int\limits_0^{+\infty} g(x)\, x^{2n-1}\, e^{-yx}\, dx = 0, \qquad \forall y > 0$$

Заметим, что слева написано преобразование Лапласа функции $f(x) = g(x)x^{2n-1}$:

$$F(y) = \mathcal{L}[f](y) \overset{\text{def}}{=} \int\limits_0^{+\infty} f(x)\, e^{-yx}\, dx = 0, \qquad \forall y > 0$$

Из единственности преобразования Лапласа получаем, что $f(x) \equiv 0$, и, значит, $g(x) \equiv 0$. Таким образом, $T(\mathbf{X})$ — **полная достаточная статистика**.

5. Построим оптимальную оценку с помощью теоремы Лемана-Шеффе. Найдем такую функцию $g(t)$, что

$$\mathbb{E}_\theta\big[g(T(\mathbf{X}))\big] = \theta, \qquad \forall \theta > 0$$

Известно, что матожидание распределения $\text{Gamma}(k, \lambda)$ равно $k\lambda$, а дисперсия — $k\lambda^2$. Поэтому можно вычислить второй момент:

$$\mathbb{E}_\theta\big(T(\mathbf{X})\big)^2 = \mathbb{V}_\theta T(\mathbf{X}) + \big(\mathbb{E}_\theta T(\mathbf{X})\big)^2 = 2n\theta + \big(2n\sqrt{\theta}\big)^2 = 2n(1+2n)\theta$$

Мы рассматриваем именно второй момент, а не саму дисперсию, потому что искомая функция $g(x)$ не должна зависеть от параметра θ.

Тогда в качестве функции $g(x)$ подойдет:

$$g(x) = \frac{x^2}{2n(1+2n)}$$

Тогда оптимальная оценка параметра:

$$\widehat{\theta} = g(T(\mathbf{X})) = \frac{1}{2n(1+2n)} \left(\sum_{i=1}^n \sqrt{X_i}\right)^2$$

Заметим, что оптимальная оценка является несмещенной, в отличие от оценки максимального правдоподобия:

$$\widehat{\theta}^{\text{опт.}} = \frac{1}{4n^2+2n} \left(\sum_{i=1}^n \sqrt{X_i}\right)^2, \qquad \widehat{\theta}^{\text{ОМП}} = \frac{1}{4n^2} \left(\sum_{i=1}^n \sqrt{X_i}\right)^2$$

Тем не менее, оценка максимального правдоподобия является асимптотически оптимальной.