Теория вероятностей и математическая статистика

7 декабря 2015 г.

Содержание

Функция правдоподобия и ее свойства

В отличии от метода наименьших квадратов, метод наибольшего правдоподобия не требует от вероятностного распределения существования конечных абсолютных моментов какого-либо порядка, но предполагает существование плотности вероятности и ее гладкость по параметрам распределения в точке максимума. Например, для оценки показателя степени полиномиального хвоста можно использовать часть эмпирических данных $\{x_k\}_1^K \subset \{\xi_n^a\}_1^N$, его правый конец с отброшенными редкими максимальными значениями. Оставшиеся точки сортируются в порядке возрастания (строится вариационный ряд) и показатель степени a оценивается по формуле *метода максимального правдоподобия* (maximum likelihood method)

$$a \approx \theta_N = 1 + N\left(\sum_1^N \ln \frac{x_n}{x_1}\right)^{-1}, \quad x_n < x_{n+1}, \tag{1}$$

который обсуждается в настоящей лекции.

Если независимо распределенные случайные точки $\{x_n\}$ имеют общее распределение $p(x|\theta)$, зависящее от неизвестного параметра θ, то N-точечная плотность вероятности такого набора равна

$$p_N(x_1, \ldots, x_N|\theta) \stackrel{\text{def}}{=} \prod_1^N p(x_n|\theta).$$

Аргумент максимума этой функции по θ характеризует *наиболее вероятное* событие. Положение экстремума не изменяется при любом монотонном преобразовании этой функции, поэтому можно перейти к *логарифмической функции правдоподобия*

$$\mathcal{L}(\mathbf{x}|\theta) \stackrel{\text{def}}{=} \ln p_N(\mathbf{x}|\theta), \quad \mathbf{x} \stackrel{\text{def}}{=} \{x_n\}_1^N \text{ -выборочные значения.} \qquad (2)$$

Для отыскания экстремума используется необходимое условие $\theta_N^*: \ \partial_\theta \mathcal{L}(\mathbf{x}|\theta) = \partial_\theta p(\mathbf{x}|\theta)/p(\mathbf{x}|\theta) = 0$. Вторая производная $\partial_\theta^2 \mathcal{L}(\mathbf{x}|\theta^*)$ отрицательна, и чем больше ее модуль, тем быстрее убывает плотность $p_N(\mathbf{x}|\theta)$ при отклонении от θ^*: т.е. $|\partial_\theta^2 \mathcal{L}(\mathbf{x}|\theta^*)|$ характеризует *локализацию* экстремума функции $\theta\mathcal{L}(\mathbf{x}|\theta)$ в точке θ^*.

Информацией Фишера называется величина

$$I(\theta) = -\mathbb{E}\,\partial_\theta^2 \mathcal{L}(x|\theta), \tag{3}$$

являющаяся в силу ЗБЧ пределом выборочных средних при $N \to \infty$:

$$\frac{1}{N}\partial_\theta^2 \mathcal{L}(\mathbf{x}|\theta) = \frac{1}{N}\sum_{n=1}^{N} \partial_\theta^2 \mathcal{L}(x_n|\theta) \to \mathbb{E}\,\partial_\theta^2 \mathcal{L}(x|\theta) = -I(\theta).$$

Для многомерных параметров эту роль играет определитель матрицы вторых производных функции правдоподобия, характеризующая локализацию экстремума по совокупности параметров, а также производные $\partial_{\theta-k}^2 \mathcal{L}(x|\theta)$, характеризующие локализацию экстремума по каждому параметру θ_k в отдельности.
Производная функции правдоподобия возникает при дифференцировании матожиданий, зависящих от оцениваемого параметра:

$$\boxed{\partial_\theta\, \mathbb{E}\, f(\xi) = \partial_\theta \int dx\, f(x)\, p(x|\theta) = \int dx\, f(x)\, \partial_\theta p(x|\theta) = \mathbb{E}\, f(\xi)\, \partial_\theta \mathcal{L}(\xi|\theta)}$$

Нетрудно видеть, что функция правдоподобия объединения множеств независимых эмпирических данных равна сумме логарифмических функций правдоподобия этих множеств.

Предположим, что θ –истинное значение неизвестного параметра распределения. Поскольку $\partial_\theta \mathcal{L}(\mathbf{x}|\theta) = 0$ в точке θ^*_N и, согласно закону больших чисел или ЦПТ, $N^{-1}\partial^2_\theta \mathcal{L}(\mathbf{x}|\theta) \to \mathbb{E}\,\partial^2_\theta \mathcal{L}(x|\theta)$ при $N \to \infty$, то, разлагая $\mathcal{L}(x|\theta)$ в ряд Тейлора в окрестности точки θ^*_N, получим

$$p(\mathbf{x}|\theta) = e^{\mathcal{L}(\mathbf{x}|\theta)} \approx e^{\mathcal{L}(\mathbf{x}|\theta^*_N) + \frac{(\theta-\theta^*_N)^2}{2}\partial^2_\theta \mathcal{L}(\mathbf{x}|\theta^*_N)} \approx e^{\mathcal{L}(\mathbf{x}|\theta^*_N) - N\frac{(\theta-\theta^*_N)^2}{2I(\theta)^{-1}}},$$

$$\partial^2_\theta \mathcal{L}(\mathbf{x}|\theta^*_N) \approx N(\mathbb{E}\,\partial^2_\theta \mathcal{L}(x|\theta^*_N)) \approx -N\,I(\theta).$$

Это наблюдение объясняет роль информации Фишера и мотивирует гипотезу о предельном распределении погрешности оценок ММП:

$$P\left(\sqrt{\frac{N}{\sigma^2_N}}\,(\theta-\theta^*_N) \in B\right) \to \frac{1}{\sqrt{2\pi}}\int_B e^{-\frac{x^2}{2}}\,dx, \quad \sigma^2_N \stackrel{\text{def}}{=} \left(\frac{|\partial^2_\theta \mathcal{L}(\mathbf{x}|\theta^*_N)|}{N}\right)^{-1} \approx I^{-1}(\theta), \tag{4}$$

то есть $\theta - \theta^*_N = O\left(N^{-\frac{1}{2}}\right)$, где $\theta^*_N = \operatorname{argmax} \mathcal{L}(\mathbf{x}|\theta)$.

Вернемся к примеру. Если д. с. в. $x \in [x_{\min}, \infty)$ имеет распределение Парето с неизвестным показателем θ

$$p(x|\theta) = \frac{\theta - 1}{x_{\min}} \left(\frac{x_{\min}}{x} \right)^{\theta}, \; p_N(x_1, \ldots, x_N|\theta) = \prod_1^N \frac{\theta - 1}{x_{\min}} \left(\frac{x_{\min}}{x_n} \right)^{\theta},$$

то с учетом нормировки функция правдоподобия $\mathcal{L}$ равна

$$\mathcal{L}(\mathbf{x}|\theta) = N \ln(\theta - 1) - N \ln x_{\min} - \theta \sum_1^N \ln \frac{x_n}{x_{\min}}.$$

Точку экстремума этой функции находим из условия

$$0 = \partial \mathcal{L}/\partial\theta = \frac{N}{\theta - 1} - \sum_1^N \ln \frac{x_n}{x_{\min}} \Rightarrow \boxed{\theta_N^* = 1 + N \left(\sum_1^N \ln \frac{x_n}{x_1} \right)^{-1}}$$

Нетрудно видеть, что $\partial^2 \mathcal{L}/\partial\theta^2 = -N(\theta - 1)^{-2} < 0$, так что точка θ_N^* –выборочная оценка аргумента максимума. В качестве точки $x_{\min}$ естественно использовать оценку $x_{\min} \approx x_1$, $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_N$.

Теорема 1

Пусть $\{x_1, \ldots, x_N\}$ *-упорядоченная по возрастанию выборка* $\{\xi_1, \ldots, \xi_N\}$ *из распределения Парето* $P(x_{\min}, \theta)$. *Выборочное среднее*

$$N^{-1} \sum_n \ln \frac{x_n}{x_1} = (\theta_N^* - 1)^{-1} \approx (\theta - 1)^{-1}$$

дает состоятельную оценку показателя θ, *а точка* x_1 *является состоятельной оценкой для* $x_{\min}$.

Доказательство. Минимум i. i. d. r. v. $\{\xi_n\}$ имеет плотность распределения

$$P\{x_1^{(N)} \stackrel{\text{def}}{=} \min(\xi_1, \ldots, \xi_N) \in dx\} = dP(\xi > x)^N = -d\left(\frac{x_{\min}}{x}\right)^{N(\theta-1)}.$$

Обезразмеривая интеграл с помощью замены $\frac{x_{\min}}{x} \to y$ и интегрируя по частям логарифм отношения $\min(\xi_1, \ldots, \xi_N)/x_{\min}$, получаем

$$\begin{aligned} \mathbb{E} \ln \frac{x_1^{(N)}}{x_{\min}} &= -\int_{x_{\min}}^{\infty} d\left(\frac{x_{\min}}{x}\right)^{N(\theta-1)} \ln \frac{x}{x_{\min}} = -\int_1^{\infty} d\left(\frac{1}{y}\right)^{N(\theta-1)} \ln y \\ &= \int_1^{\infty} \left(\frac{1}{y}\right)^{N(\theta-1)} \frac{dy}{y} = \frac{1}{N(\theta-1)} \to 0, \quad N \to \infty. \end{aligned} \tag{5}$$

Докажите, что имеют место следующие выражения для плотностей распределения минимального и максимального выборочного значений:

Упражнение 1

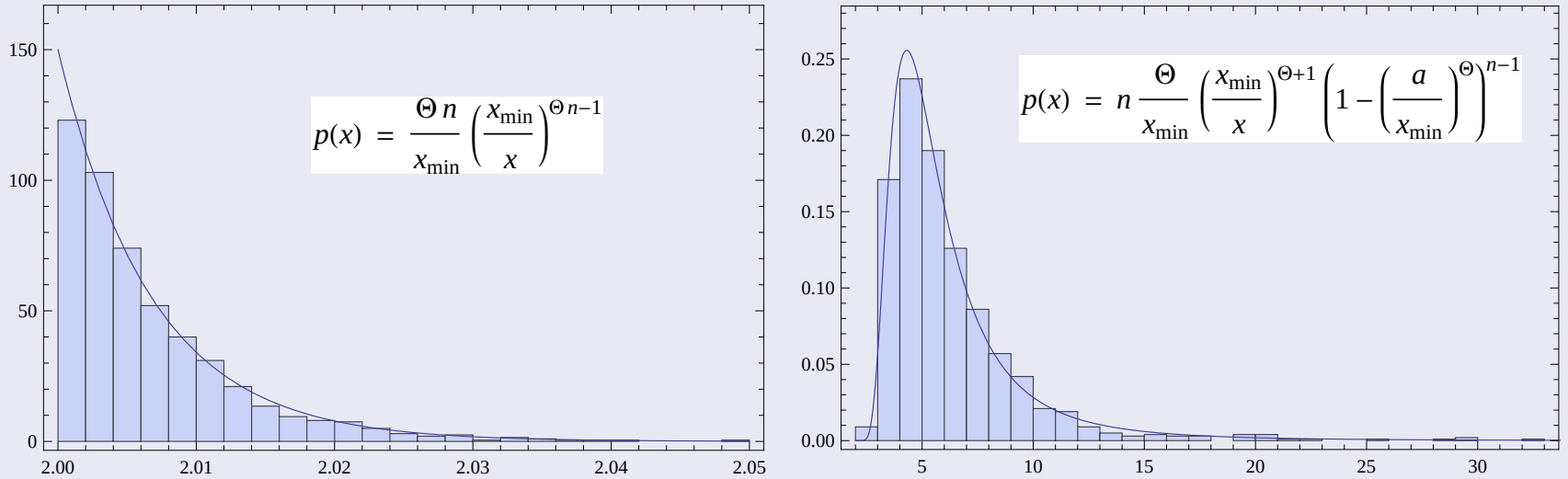

Рис. 1: Гистограммы и функции распределения оценок $P(\min x_n > x)$ (слева) и $P(\max x_n > x)$ (справа) по $N = 1000$ выборкам размера $n = 13$ из распределения Парето $p_{2,3}(x)$: $x_{\min} = 2$, $\theta = 3$, $\Theta = \theta + 1$.

Покажем, что из (5) следует *состоятельность* оценки $x_1 = x_1^{(N)} = \min\{\xi_1, \ldots, \xi_N\}$ для точки x_{min}, то есть $P(x_1^{(N)} - x_{min} \geq \varepsilon) \to 0$ при $N \to \infty$. Действительно,

$$\mathbb{E} \ln \frac{x_1^{(N)}}{x_{min}} > \mathbb{E} \ln \frac{x_1^{(N)}}{x_{min}} I_{\{x_1 - x_{min} \geq \varepsilon\}}(x_1)$$
$$= \mathbb{E} \ln \left(1 + \frac{x_1^{(N)} - x_{min}}{x_{min}}\right) I_{\{x_1 - x_{min} \geq \varepsilon\}}(x_1)$$
$$> \ln \left(1 + \frac{\varepsilon}{x_{min}}\right) \mathbb{E}\, I_{\{x_1 - x_{min} \geq \varepsilon\}}(x_1).$$

Поэтому при $N \to \infty$ из (5) вытекает условие состоятельности:

$$P(x_1^{(N)} - x_{min} \geq \varepsilon) = \mathbb{E}\, I_{\{x_1 - x_{min} \geq \varepsilon\}}(x_1)$$
$$\leq \ln \left(1 + \frac{\varepsilon}{x_{min}}\right)^{-1} \mathbb{E} \ln \frac{x_1^{(N)}}{x_{min}} = \ln \left(1 + \frac{\varepsilon}{x_{min}}\right)^{-1} \frac{1}{N(\theta - 1)} \to 0.$$

Используя (5), покажем, что $\mathbb{E}\frac{1}{N}\sum_n \ln\frac{x_n}{x_1} = (\theta-1)^{-1} + O(N^{-1})$:

$$\mathbb{E}\,\frac{1}{N}\sum_n \ln\frac{x_n}{x_1} \equiv \mathbb{E}\,\frac{1}{N}\sum_n \ln\frac{x_n}{x_{\min}} + \mathbb{E}\,\ln\frac{x_{\min}}{x_1} =$$

$$= \mathbb{E}\,\frac{1}{N}\sum_n \ln\frac{x_n}{x_{\min}} + O(N^{-1}).$$

С другой стороны, $\mathbb{E}\frac{1}{N}\sum_n \ln\frac{x_n}{x_{\min}}$ – несмещенная оценка параметра $(\theta-1)^{-1}$. Действительно, для переменной $y = \frac{x_{\min}}{x}$ имеем

$$\mathbb{E}\frac{1}{N}\sum_{n=1}^{N} \ln\frac{x_n}{x_{\min}} = \int_{x_{\min}}^{\infty} d\left(\frac{x_{\min}}{x}\right)^{\theta-1} \ln\frac{x}{x_{\min}} =$$

$$= \int_1^{\infty} dy^{-\theta+1}\ln y = \int_1^{\infty} y^{-\theta}dy = \frac{1}{\theta-1}.$$

Теорема доказана. □

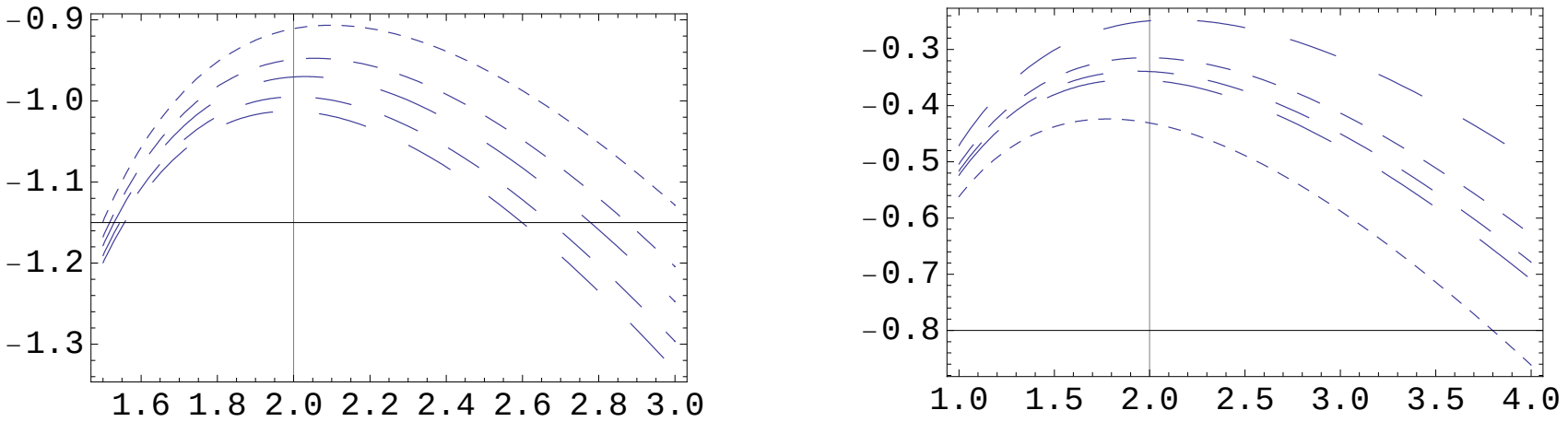

Рис. 2: Графики функции правдоподобия $\mathcal{L}_N(\mathbf{x}|\theta)/N$ для случайных выборок из распределений Парето (слева) ($\Omega = [1, \infty)$, $p(x) = O(x^{-\theta})$, $\theta = 2$) и Пуассона $p(x|\theta) = \theta e^{-\theta x}$, $x \in R_+$, $\theta = 2$ (справа). По оси x меняется θ. С увеличением объема выборки $N \in \{100, \ldots, 500\}$ длина пунктира увеличивается. Положение максимума является случайной величиной, которая сходится к $\theta = 2$ при $N \to \infty$

Упражнение 2

Показать, что если ξ имеет распределение Парето на $(x_{\min}, \infty)$, то дисперсия случайной величины $\log \frac{\xi}{x_{\min}}$ равна $\sigma^2 = (\theta - 1)^{-2}$.

С помощью ММП несложно получить совместную оценку среднего и дисперсии нормального распределения. В этом случае

$$\mathcal{L}(\mathbf{x}|\mu,\sigma) = -N\ln\sigma - \sum_1^N \frac{(x_n-\mu)^2}{2\sigma^2},$$

$$\mu_N^*:\ 0 = \partial_\mu \mathcal{L}(\mathbf{x}|\mu,\sigma) \Rightarrow \sum_1^N (x_n-\mu) = 0 \Rightarrow \mu_N^* = \frac{1}{N}\sum_n x_n,$$

$$\sigma_N^*:\ 0 = \partial_\sigma \mathcal{L}(\mathbf{x}|\mu,\sigma) \Rightarrow \sigma_N^2 = \frac{1}{N}\sum_1^N (x_n-\mu_N^*)^2.$$

Как нам уже известно, такая оценка среднего является несмещенной, а оценка дисперсии смещена (несмещенная оценка имеет вид $\widetilde{\sigma}_N^2 \stackrel{\text{def}}{=} \frac{1}{N-1}\sum_1^N (x_n-\mu_N^*)^2$), но состоятельна, то есть $\sigma_N \xrightarrow{P} \sigma$.

Доказательство сходимости по вероятности выглядит следующим образом:

$$P(|\sigma_N - \sigma| \geq \varepsilon) = P\left(\left|\frac{N-1}{N}(\widetilde{\sigma}_N - \sigma) - \frac{1}{N}\sigma\right| \geq \varepsilon\right)$$
$$\geq P\left(\left|\frac{N-1}{N}(\widetilde{\sigma}_N - \sigma)\right| - \frac{1}{N}\sigma \geq \varepsilon\right)$$
$$= P\left(\left|\widetilde{\sigma}_N - \sigma\right| \geq \varepsilon + \frac{\sigma+\varepsilon}{N-1}\right) \to 0 \quad \text{при} \quad N \to \infty$$

так как несмещенная оценка $\widetilde{\sigma}^2_N = \frac{1}{N-1}\sum_1^N (x_n - \mu^*_N)^2$ состоятельна. Действительно, $\widetilde{\sigma}_N \to \sigma$ по закону больших чисел и поэтому

$$P\left(\left|\widetilde{\sigma}_N - \sigma\right| \geq \varepsilon + \frac{\delta+\varepsilon}{N-1}\right) \geq \frac{\mathbb{E}(\widetilde{\sigma}_N - \sigma)^2}{\varepsilon + \frac{\delta+\varepsilon}{N-1}} = \frac{\mathbb{E}\widetilde{\sigma}^2_N - \sigma^2}{\varepsilon + \frac{\delta+\varepsilon}{N-1}} \to 0 \quad \text{при} \quad N \to \infty$$

□

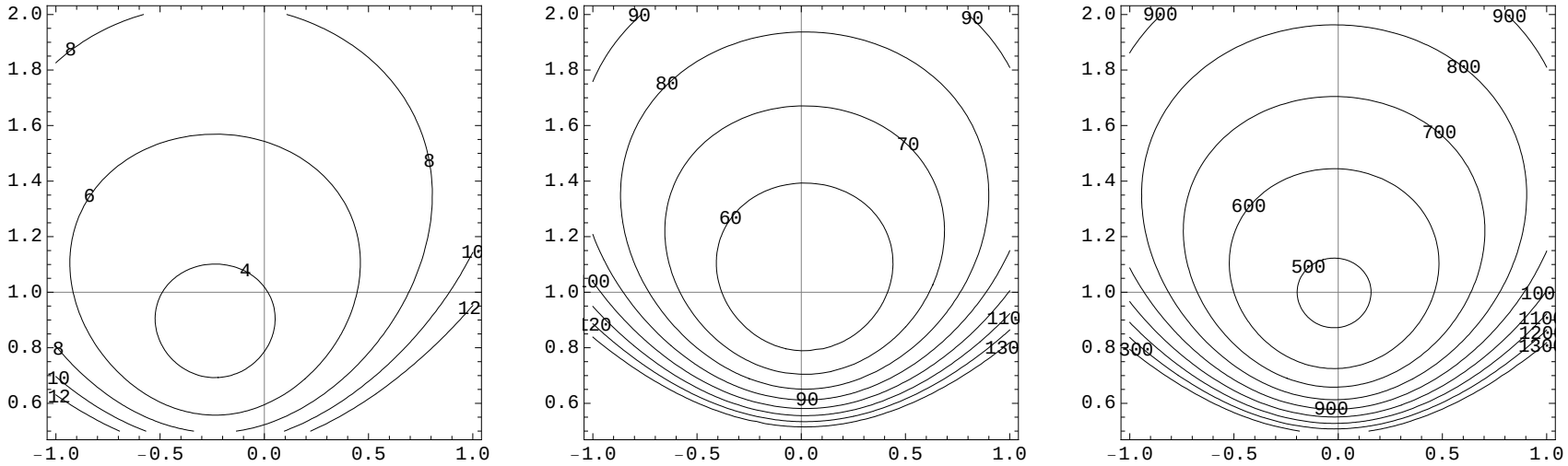

Рис. 3: На рисунках изображены линии уровня логарифмической функции правдоподобия $-\mathcal{L}(\mathbf{x}|\mu,\sigma)$, $\mathbf{x} = \{x_1, \ldots, x_N\}$ стандартного нормального распределения, построенные по 10, 100 и 1000 случайных точек в переменных μ, σ. По оси абсцисс меняется среднее значение, по оси ординат - дисперсия. Правильная локализация экстремума $\mu = 0$, $\sigma = 1$ заметно улучшается (см. правый график) при увеличении объема выборки до $N = 1000$.

Рассмотрим применение ММП для определения параметра экспоненциального распределения $p(\tau|\theta) = \theta e^{-\theta\tau}$, $\tau > 0$. Пусть $\tau = \{\tau_n\}_1^N$ – экспоненциально распределенные i. i. d. r. v. с неизвестным θ, которым соответствует функция правдоподобия

$$\mathcal{L}(\tau|\theta) = \ln\left(\theta^N e^{-\theta\sum_1^N \tau_n}\right) = N\ln\theta - \theta\sum_1^N \tau_n.$$

Используя ММП, находим

$$\theta_N^*: \ \partial_\theta \mathcal{L}(\tau|\theta) = 0 \Rightarrow \theta_N^* = \left(\frac{1}{N}\sum_{n=1}^N \tau_n\right)^{-1}.$$

Упражнение 3

Доказать, что оценка θ_N^* интенсивности экспоненциальной д. с. в. – несмещенная, а ее информация Фишера равна $I(\theta) = \theta^{-2}$.

Информация Фишера и неравенство Рао–Крамера

Информация Фишера используется для характеризации достоверности оценок. Она является аналогом обратной дисперсии в центральной предельной теореме, но для ее существования не требуется, чтобы дисперсия была конечна.

Лемма 1

Пусть x – д. с. в. с распределением $p(x|\theta)$, зависящим от θ, и $T(\mathbf{x}) = T(x_1, \ldots, x_N)$ – некоторая статистика (функция от случайной выборки, не зависящая от параметра явно). Тогда

1) $\mathbb{E}_\theta\, \partial_\theta \mathcal{L}(\mathbf{x}|\theta) = 0$;
2) $\partial_\theta \mathbb{E}_\theta\, T(\mathbf{x}) = \mathbb{E}_\theta\, T(\mathbf{x}) \partial_\theta \mathcal{L}(\mathbf{x}|\theta)$;
3) *Информация Фишера* $I(\theta) \stackrel{\text{def}}{=} -\mathbb{E}_\theta\, \partial_\theta^2 \mathcal{L}(x|\theta)$, *связанная с наблюдением случайной величины x, может быть записана в одной из трех эквивалентных форм:*

$$NI(\theta) = -\mathbb{E}_\theta\, \partial_\theta^2 \mathcal{L}(\mathbf{x}|\theta) = \mathbb{E}_\theta \left(\partial_\theta \mathcal{L}(\mathbf{x}|\theta)\right)^2 = \mathbb{D}_\theta\, \partial_\theta\, \mathcal{L}(\mathbf{x}|\theta),\ \mathbf{x} = \{x_1, \ldots, x_N\}.$$

Доказательство. Утверждение (1) вытекает из формулы интегрирования по частям:

$$\mathbb{E}_\theta\, \partial_\theta \mathcal{L}(\mathbf{x}|\theta) = \sum_n \int_{X^N} \frac{\partial_\theta p(x_n|\theta)}{p(x_n|\theta)} \prod_{k=1}^{N} p(x_k|\theta)\, dx =$$

$$= \sum_n \partial_\theta \int_X p(x_n|\theta)\, dx_n = \sum_n \partial_\theta 1 = 0.$$

Формула (2) доказывается с помощью дифференцирования под знаком интеграла:

$$\partial_\theta \mathbb{E}_\theta\, T(\mathbf{x}) = \partial_\theta \int_{X^N} T(\mathbf{x}) \prod_{1}^{N} p(x_n|\theta)\, d^N x =$$

$$= \int_{X^N} T(\mathbf{x}) \sum_n \frac{\partial_\theta p(x_n|\theta)}{p(x_n|\theta)} \prod_{n=1}^{N} p(x_n|\theta)\, d^N x = \mathbb{E}_\theta\, T(\mathbf{x}) \partial_\theta \mathcal{L}(\mathbf{x}|\theta).$$

(3) вытекает из нормировки плотности вероятности $\int_X p(x|\theta)dx = 1$. В случае $N = 1$, $\mathcal{L}(x|\theta) = \ln p(x|\theta)$, имеем $\partial_\theta p(x|\theta) = p(x|\theta)\partial_\theta \mathcal{L}(x|\theta)$ поэтому

$$
\begin{aligned}
0 &= \partial_\theta^2 \int_X p(x|\theta)dx = \partial_\theta \int_X \partial_\theta \mathcal{L}(x|\theta)\, p(x|\theta)dx \\
&= \int_X \partial_\theta \mathcal{L}(x|\theta) \frac{\partial_\theta p(x|\theta)}{p(x|\theta)}\, p(x|\theta)dx + \int_X \partial_\theta^2 \mathcal{L}(x|\theta)\, p(x|\theta)dx \\
&= \int_X \left(\partial_\theta \mathcal{L}(x|\theta)\right)^2 p(x|\theta)dx + \int_X \partial_\theta^2 \mathcal{L}(x|\theta)\, p(x|\theta)dx \\
&= \mathbb{E}_\theta \left(\partial_\theta \mathcal{L}(\,\cdot\,|\theta)\right)^2 + \mathbb{E}_\theta\, \partial_\theta^2 \mathcal{L}(\,\cdot\,|\theta) = \mathbb{D}_\theta\, \partial_\theta \mathcal{L}(\,\cdot\,|\theta) + \mathbb{E}_\theta\, \partial_\theta^2 \mathcal{L}(\,\cdot\,|\theta), \qquad (6)
\end{aligned}
$$

поскольку $\mathbb{E}_\theta\, \partial_\theta \mathcal{L}(\,\cdot\,|\theta) = 0$ в силу (1), где $\mathbb{E}_\theta\, \partial_\theta^2 \mathcal{L}(\,\cdot\,|\theta) = -I(\theta)$. Таким образом, доказано, что

$$
I(\theta) = \mathbb{D}_\theta\, \partial_\theta \mathcal{L}(\,\cdot\,|\theta) = \mathbb{E}_\theta \left(\partial_\theta \mathcal{L}(\,\cdot\,|\theta)\right)^2. \qquad (7)
$$

Если $N > 1$, то для независимых д. с. в. $\{x_n\}$ равенство (6) выполняется в силу линейности операции дифференцирования

$$0 = \partial_\theta^2 \int_X \sum_{n=1}^{N} p(x_n|\theta)dx = \partial_\theta \int_X \sum_n \left(\partial_\theta \mathcal{L}(x_n|\theta)\right) p(x_n|\theta)dx$$

$$= \sum_n \mathbb{E}_\theta \left(\left(\partial_\theta \mathcal{L}(x_n,\theta)\right)^2 + \partial_\theta^2 \mathcal{L}(x_n,|\theta) \right) = \mathbb{D}_\theta\, \partial_\theta \mathcal{L}(x_n,\theta) + \mathbb{E}_\theta\, \partial_\theta^2 \mathcal{L}(x_n,\theta),$$

Поэтому

$$\mathbb{E}_\theta \left(\partial_\theta \mathcal{L}(\mathbf{x}|\theta)\right)^2 = \left(\sum_{k\neq n} + \sum_{k=n}\right) \int_{X^N} \frac{\partial_\theta p(x_k|\theta)}{p(x_k|\theta)} \frac{\partial_\theta p(x_n|\theta)}{p(x_n|\theta)} \prod_{m=1}^{N} p(x_m|\theta)\, dx$$

$$= N(N-1)\left(\int_X \partial_\theta p(x_1|\theta)\, dx_1\right)^2 + N \int_X \left(\frac{\partial_\theta p(x_1|\theta)}{p(x_1|\theta)}\right)^2 p(x_1|\theta)\, dx_1$$

$$= N\, \mathbb{E}\left(\partial_\theta \mathcal{L}(\cdot|\theta)\right)^2 = N\, I(\theta), \qquad (8)$$

так как $\int \partial_\theta p(x|\theta)dx = \partial_\theta \int p(x|\theta)dx = \partial_\theta\, 1 = 0$. Отсюда следует утверждение (3) в многомерном случае. □

Упражнение 4

Вычислить информацию Фишера для распределения Парето $P(x_{\min}, \theta)$ *с плотностью*

$$p(x|x_{\min}, \theta) = \frac{\theta}{x_{\min}} \left(\frac{x_{\min}}{x} \right)^{\theta+1} I_{[x_{\min}, \infty)}(x).$$

Упражнение 5

Убедитесь, что информация Фишера для оценки параметров основных вероятностных распределений вычисляется по формулам, указанным в следующей таблице:

	$N(\theta, \sigma)$	$N(\mu, \theta)$	$\Gamma(\alpha, \theta)$	$C(\theta, 1)$	$\Pi(\theta)$	$Bi(n, \theta)$	$P(1, \theta)$
$I(\theta)$	$\frac{1}{\sigma^2}$	$\frac{2}{\theta^2}$	$\frac{\alpha}{\theta^2}$	$\frac{1}{2}$	$\frac{1}{\theta}$	$\frac{n}{\theta(1-\theta)}$	$\frac{1}{\theta^2}$

Таблица 1: Информация Фишера $I(\theta)$ для основных распределений

Неравенство Рао–Крамера

Статистикой (в узком смысле этого слова) называется любая функция $T : X \to Y$ выборочных данных $\mathbf{x} = \{x_1, \dots x_N\}$. Если статистика используется для оценки параметра θ распределения д. с. в. x_n, то функция T называется *несмещенной статистикой*, если $\mathbb{E}_\theta T(\mathbf{x}) = \theta$. Разность $b(\theta) = \mathbb{E}_\theta T(\mathbf{x}) - \theta$ называется *смещением*:

$$\mathbb{E}_\theta \big(T(\mathbf{x}) - \theta\big)^2 = \mathbb{D}_\theta T + b^2(\theta),$$

где $\mathbb{D}_\theta T$ – *дисперсия статистики* T. Если оценивается функция $\tau(\theta)$ от параметра θ, то определения не меняются: условие несмещенности имеет вид $\mathbb{E}_\theta T(\mathbf{x}) = \tau(\theta)$, а *смещение* по определению равно $b(\theta) = \mathbb{E}_\theta T(\mathbf{x}) - \tau(\theta)$. Если смещения нет, то

$$\boxed{\tau(\theta) = \mathbb{E}_\theta T(\mathbf{x}), \quad \mathbb{D}_\theta T = \mathbb{E}_\theta \big(T(\mathbf{x}) - \tau(\theta)\big)^2}$$

Неравенство Рао–Крамера устанавливает нижнюю границу для дисперсии смещенной статистики. Статистика называется *эффективной*, если ее дисперсия достигает нижней границы.

Теорема 2

Если $b(\theta)$, $\tau(\theta)$ – гладкие функции, то

$$\mathbb{D}_\theta T(\mathbf{x}) \geq \frac{(\tau'(\theta) + b'(\theta))^2}{NI(\theta)}. \tag{9}$$

Если $b(\theta) = 0$, то эффективность статистики $T(x)$ достигается, если $T(\mathbf{x}) - \tau(\theta) = a(\theta)\partial_\theta \mathcal{L}(\mathbf{x}|\theta)$, где $a(\theta)$ – некоторая функция. Это условие выполнено, если

$$p(x|\theta) = e^{A(\theta)B(x)+C(\theta)+E(x)},$$

$$T(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} B(x_n),\ \tau(\theta) = -\frac{C'(\theta)}{A'(\theta)}. \tag{10}$$

При этом $T(\mathbf{x}) - \tau(\theta) = a(\theta)\,\partial_\theta \mathcal{L}(\mathbf{x}|\theta)$, где $a(\theta) = \frac{1}{NA'(\theta)}$.

Доказательство. Докажем первую часть теоремы. Равенство

$$\partial_\theta \mathbb{E}_\theta T(\mathbf{x}) = b'(\theta) + \tau'(\theta)$$

выполнено по условию теоремы. С другой стороны,

$$\partial_\theta \mathbb{E}_\theta T(\mathbf{x}) = \mathbb{E}_\theta\, T(\mathbf{x}) \partial_\theta \mathcal{L}(\mathbf{x}|\theta) = \mathbb{E}_\theta\, (T(\mathbf{x}) - \mathbb{E}_\theta\, T(\mathbf{x})) \partial_\theta \mathcal{L}(\mathbf{x}|\theta)$$

согласно утверждениям (2) и (1) леммы 1 соответственно. Применяя неравенство Коши $\mathbb{E}\,|(a,b)_{\mathcal{H}}| \leq \sqrt{\mathbb{E}\,||a||^2_{\mathcal{H}}\, \mathbb{E}\,||b||^2_{\mathcal{H}}}$ в случае $\mathcal{H} = L_2(\mathbb{R}, dp(x))$, $a = T(\mathbf{x}) - \mathbb{E}_\theta\, T(\mathbf{x})$, $b = \partial_\theta \mathcal{L}(\mathbf{x}|\theta)$, получаем

$$b'(\theta) + \tau'(\theta) = \mathbb{E}_\theta\, T(\mathbf{x})\, \partial_\theta \mathcal{L}(\mathbf{x}|\theta) \leq \sqrt{\mathbb{D}\, T(\mathbf{x})\, \mathbb{D}\, \partial_\theta \mathcal{L}(\mathbf{x}|\theta)},$$

где $\mathbb{D}\, \partial_\theta \mathcal{L}(\mathbf{x}|\theta) = NI(\theta)$ в силу третьего утверждения леммы 1. Отсюда следует неравенство Рао–Крамера:

$$\boxed{\mathbb{D} T(\mathbf{x}) \geq \frac{(b'(\theta) + \tau'(\theta))^2}{NI(\theta)}}$$

Первая часть теоремы доказана.

Докажем достаточность второго условия теоремы. В случае $b(\theta) = 0$, неравенство (9) переходит в $\mathbb{D}_\theta T(\mathbf{x}) \geq \frac{(\tau'(\theta))^2}{NI(\theta)}$. Из второго и первого утверждений леммы 1 следует, что

$$\tau'(\theta) = \partial_\theta \mathbb{E}_\theta\, T(\mathbf{x}) = \mathbb{E}_\theta T(\mathbf{x})\, \partial_\theta \mathcal{L}(\mathbf{x}|\theta) = \mathbb{E}_\theta (T(\mathbf{x}) - \tau(\theta))\, \partial_\theta \mathcal{L}(\mathbf{x}|\theta).$$

С учетом связи $T(\mathbf{x}) - \tau(\theta) = a(\theta)\partial_\theta \mathcal{L}(\mathbf{x}|\theta)$, правая часть в последнем равенстве имеет две эквивалентные формы:

$$\tau'(\theta) = \mathbb{E}_\theta (T(\mathbf{x}) - \tau(\theta))\, \partial_\theta \mathcal{L}(\mathbf{x}|\theta) = a(\theta)\mathbb{E}_\theta (\partial_\theta \mathcal{L}(\mathbf{x}|\theta))^2 = a(\theta)^{-1}\mathbb{D}_\theta T(\mathbf{x}).$$

Перемножим последние два равенства:

$$(\tau'(\theta))^2 = \mathbb{E}_\theta (\partial_\theta \mathcal{L}(\mathbf{x}|\theta))^2\, \mathbb{D}_\theta T(\mathbf{x}) = NI(\theta)\, \mathbb{D}_\theta T(\mathbf{x})$$

Используя это равенство получаем оценку дисперсии снизу:

$$\boxed{\mathbb{D}_\theta T(\mathbf{x}) = \frac{(\tau'(\theta))^2}{NI(\theta)}, \text{ если } b(\theta) = 0 \quad \text{и } T(\mathbf{x}) - \tau(\theta) = a(\theta)\partial_\theta \mathcal{L}(\mathbf{x}|\theta)}$$

Для завершения доказательства убедимся, что из (10) следует

$$T(\mathbf{x}) - \tau(\theta) = \frac{1}{NA'(\theta)}\,\partial_\theta \mathcal{L}(\mathbf{x}|\theta). \tag{11}$$

Достаточно рассмотреть случай $N = 1$ (см. n. (3) леммы 1). Для функции $\mathcal{L}(x|\theta) = A(\theta)B(x) + C(\theta) + E(x)$ при выборе $T(x) = B(x)$, $\tau(\theta) = -\frac{C'(\theta)}{A'(\theta)}$ имеем

$$T(x) - \tau(\theta) = B(x) + \frac{C'(\theta)}{A'(\theta)} = \frac{1}{A'(\theta)}\left(A'(\theta)B(x) + C'(\theta)\right)$$
$$= \frac{1}{A'(\theta)}\,\partial_\theta \mathcal{L}(x|\theta).$$

Поэтому $T(x) - \tau(\theta) = a(\theta)\partial_\theta \mathcal{L}(x|\theta)$ при $N = 1$, $a(\theta) = \frac{1}{A'(\theta)}$ и оценка Рао–Крамера достигает нижней границы. □

Доказательство необходимости условий (10) можно найти в книге *Крамер Г. Математические методы статистики.* – М.: Мир, 1975. Следующее утверждение вытекает из ЦПТ.

Следствие 1

Если $\tau(\theta) = \theta$ *и статистика* T *несмещенная, т.е.* $b(\theta) = 0$, *то* $I(\theta)\,\mathbb{D}_\theta T(x) \geq 1$. *Если выполнено* (10), *то оптимальная оценка* θ^*_N *параметра* θ *определяется из условия* $\partial_\theta \mathcal{L}(x|\theta^*_N) = 0$ *или*

$$T(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_n B(x_n) = -\frac{C'(\theta^*_N)}{A'(\theta^*_N)} \stackrel{\text{def}}{=} \tau(\theta^*_N).$$

При этом

$$\lim_{N\to\infty} \mathbb{P}\left(\frac{\tau(\theta^*_N) - \tau(\theta)}{\sqrt{\mathbb{D}_\theta T(\mathbf{x})}} \in B\right) = \frac{1}{\sqrt{2\pi}}\int_B e^{-\frac{x^2}{2}}\,dx,$$

$P(x)$	$N(\theta,\sigma)$	$N(\mu,\theta)$	$\Gamma(\alpha,\theta)$	$\Pi(\theta)$	$Bi(n,\theta)$	$P(1,\theta)$
$\tau(\theta)$	θ	θ^2	$\alpha\theta$	θ	θn	θ^{-1}
$I(\theta)$	σ^{-2}	$2\theta^{-2}$	$\alpha\theta^{-2}$	θ^{-1}	$n/\theta(1-\theta)$	θ^{-2}
$\mathbb{D}_\theta T$	σ^2	$2\theta^4$	$\alpha\theta^2$	θ	$\theta(1-\theta)/n$	θ^{-2}
$B(x)$	x	$(x-\mu)^2$	x	x	x/n	$\ln x$

Таблица 2: Таблица значений $\tau(\theta)$, $B(x)$ и $\mathbb{D}_\theta T(\mathbf{x})$ для основных распределений

Упражнение 6

Учитывая, что $p(x|\theta) = e^{A(\theta)B(x)+C(\theta)+E(x)}$ *и* $T(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} B(x_n)$, *вычислите* $\tau(\theta) = \mathbb{E}_\theta T(\mathbf{x})$, $I(\theta)$, $B(x)$ *и* $\mathbb{D}_\theta T(x) = \frac{(\tau'(\theta))^2}{I(\theta)}$ *для распределений, указанных в таблице 2.*

Оптимальные статистики

Статистика, имеющая минимальную дисперсию, называется *оптимальной*. Предположим, что для оценки параметра θ используется несмещенная статистика $T(\mathbf{x})$. Далее мы рассмотрим примеры и свойства оптимальных статистик.

Пример 1

Пусть x – бернуллиевская с. в., принимающая значения $\{0, 1\}$ с вероятностью θ и $1 - \theta$ и $\mathbf{x} = \{x_1, \ldots, x_N\}$ – серия независимых испытаний. Выборочное среднее

$$T(\mathbf{x}) \stackrel{\text{def}}{=} \langle x \rangle_N = \frac{1}{N} \sum_1^N x_n \tag{12}$$

является несмещенной оценкой параметра $\theta = \mathbb{E}\, x$.

Ее дисперсия равна $\mathbb{D}_\theta T(\mathbf{x}) = \frac{\theta(1-\theta)}{N}$. Точно так же свойством несмещенности $\mathbb{E}\, T(\mathbf{x}) = \theta$ обладают статистики

$$\widetilde{T}(\mathbf{a}, \mathbf{x},) = \frac{1}{N}\sum_1^N a_n x_n, \quad a_n \in \mathbb{C}, \quad \sum_n a_n = N,$$

$$\mathbb{D}_\theta \widetilde{T}(\mathbf{a}, \mathbf{x},) = \frac{\theta(1-\theta)}{N^2}\sum_n a_n^2 \leq \frac{a^2\theta(1-\theta)}{N},$$

где $a = \max|a_n|$. Таким образом, существует континуум несмещенных статистик и возникает задача построения и характеризации оптимальных.

Покажем, как решается задача для распределения Бернулли.

Лемма 2

Статистика (12) *является оптимальной в классе несмещенных статистик.*

Доказательство. Покажем, что $\mathbb{D}_\theta \hat{T}(\mathbf{x}) \geq \frac{\theta(1-\theta)}{N}$ для любой другой статистики $\hat{T}$. Рассмотрим функцию правдоподобия и ее производную для распределения Бернулли $P_\theta(x) = \theta^x(1-\theta)^{1-x}$, где $x \in \{0,1\}$:

$$\mathcal{L}(\mathbf{x}|\theta) = \ln\theta \sum_n x_n + \ln(1-\theta)\sum_n (1-x_n),$$

$$\partial_\theta \mathcal{L}(\mathbf{x}|\theta) = \frac{1}{\theta}\sum_n x_n - \frac{1}{1-\theta}\sum_n (1-x_n) = \frac{\sum_n (x_n-\theta)}{\theta(1-\theta)}.$$

Напомним, что $\mathbb{E}x_n = \theta$, $\mathbb{D}x_n = p(1-p)$. Поэтому $\mathbb{E}\partial_\theta \mathcal{L}(\mathbf{x}|\theta) = 0$, $I(\theta) = \mathbb{E}_\theta(\partial_\theta L(\mathbf{x}|\theta))^2 = \frac{N}{\theta(1-\theta)}$.

Заметим, что $\mathbb{E}\,\hat{T}(x) = \theta$ по определению для любой несмещенной статистики $\hat{T}(\mathbf{x})$, Как следует из леммы 1, пп. 1–2 и условия несмещенности

$$\mathbb{E}_\theta\,\partial_\theta \mathcal{L}(\mathbf{x}|\theta) = 0, \quad \mathbb{E}_\theta\,\hat{T}(\mathbf{x})\partial_\theta \mathcal{L}(\mathbf{x}|\theta) = \partial_\theta \mathbb{E}_\theta\,\hat{T}(x) = \partial_\theta \theta = 1.$$

Поэтому $\mathbb{E}_\theta(\hat{T}(\mathbf{x}) - \theta)\,\partial_\theta \mathcal{L}(\mathbf{x}|\theta) = 1$, так как параметр θ неслучайный. Теперь из неравенства Коши–Буняковского–Шварца $\mathbb{E}\,|a\,b| \le \sqrt{\mathbb{E}\,|a|^2\,\mathbb{E}\,|b|^2}$ или (неравенства Рао–Крамера (9) без смещения) получаем требуемое неравенство для $a = \hat{T}(\mathbf{x}) - \theta$, $b = \partial_\theta \mathcal{L}(\mathbf{x}|\theta)$, $\mathbb{D}_\theta\,\hat{T}(\mathbf{x}) = \mathbb{E}_\theta(\hat{T}(\mathbf{x}) - \theta)^2$:

$$\mathbb{D}_\theta\,\hat{T}(\mathbf{x}) \ge \frac{1}{\mathbb{E}_\theta(\partial_\theta \mathcal{L}(\mathbf{x}|\theta))^2} = \frac{1}{I(\theta)} = \frac{\theta(1-\theta)}{N} = \mathbb{D}_\theta\,T(\mathbf{x}).$$

Лемма доказана. □

В заключение приведем два примера оптимальных статистик.

Пример 2

Для равномерного распределения на $[0, \theta]$ *оптимальной состоятельной оценкой параметра* θ *является* $\max_n x_n$.

Пример 3

Для отрицательного распределения Бернулли $\overline{Bi}(M, \theta)$ *оптимальной несмещенной оценкой параметра* θ *является*

$$\tau_N = \frac{MN}{\sum_{n=1}^{N} x_n + NM}.$$

Упражнение 7

Убедитесь в справедливости утверждений примеров 2 и 3.