Лекции по математической статистике

Шабанов Д.А., конспектировал Алексей Хачиянц

Содержание

1 Лекция 1

1.1 Ради чего мы собрались?

Допустим, что у нас есть какое-то наблюдение $\boldsymbol{X}$, то есть какой-то случайный вектор (хотя в асимптотических случаях будем считать, что $\boldsymbol{X}$ бесконечномерен). Далее, распределение вектора $\boldsymbol{X}$ считается неизвестным. Основная задача математической статистики состоит в том, чтобы *выдать обоснованное мнение* о распределении $\boldsymbol{X}$. Рассмотрим несколько основных подзадач:

(a) Допустим, что мы знаем, что искомое распределение P принадлежит какому-то классу распределений $\mathcal{P}$. Мы хотим проверить, можно ли выделить более узкий подкласс распределений $\mathcal{P}_0 \subset \mathcal{P}$ такой, что $\mathsf{P} \in \mathcal{P}_0$. Данный вид задач называется *проверкой гипотез*.

Пример: допустим, что нам известно, что случайный вектор X имеет нормальное распределение и мы хотим проверить гипотезу о том, что X имеет нормальное распределение с нулевым матожиданием.

(b) Теперь предположим, что распределение P пришло из параметрического семейства: $\mathsf{P} \in \{\mathsf{P}_\theta \colon \theta \in \Theta\}$ и мы хотим оценить истинное значение параметра θ_0, то есть построить оценку истинного значения $\hat{\theta}_0$. Такие задачи называются *точечным оцениванием* (если мы оцениваем каким-то значением) или же *интервальным оцениванием* (если мы в качестве оценки предлагаем какую-то область).

(c) Пусть наблюдение содержит пары: $\boldsymbol{X} = ((X_1, Y_1), \ldots, (X_n, Y_n))$. Возникает вопрос: можно ли сказать, что Y_i и X_i независимы?

Пример: Допустим, что мы собираем данные о цвете волос и цвете глаз у людей и пытаемся проверить, есть ли зависимость между ними.

(d) Допустим, что наблюдение разбилось на две части: $\boldsymbol{X} = (Y_1, \ldots, Y_m) \sqcup (Z_1, \ldots, Z_n)$. Вопрос таков: можно ли сказать, что Y_i и Z_j равны по распределению? Другими словами, можно ли сказать, что в наблюдении действительно будут одинаково распределённые случайные величины?

1.2 Точечное оценивание

Начнём с точечного оценивания. Пусть $\boldsymbol{X}$ — выборка (случайный вектор с независимыми и одинаково распределёнными компонентами) из неизвестного параметрически заданного распределения $\mathsf{P} \in \{\mathsf{P}_\theta \colon \theta \in \Theta\}$, причём параметром является набор из k действительных чисел: $\Theta \subseteq \mathbb{R}^k$, $k \geqslant 1$. По выборке мы должны каким-то образом оценить истинное значение параметра θ. Для этого строятся *оценки*. Дадим определение:

Определение 1. *Оценка* — это борелевская[1] функция $T \colon \mathcal{X} \mapsto \Theta$, где $\mathcal{X}$ — выборочное пространство. Проще говоря, это функция от выборки.

Для оценок нет единого обозначения. Если оценивается параметр θ, то оценку обозначают через $\hat{\theta}(\boldsymbol{X})$, $\theta^*(\boldsymbol{X})$ или же $\tilde{\theta}(\boldsymbol{X})$ (и так далее, но суть ясна).

Мы научились делать какие-то предсказания насчёт значения θ. Но как проверить адекватность предсказания? Для этого нужно выделить какие-то полезные свойства оценок и проверять их. Выпишем четыре основных свойства.

Определение 2. Пусть $\hat{\theta}(\boldsymbol{X})$ — это оценка параметра θ. Тогда будем называть оценку $\hat{\theta}$ *несмещённой*, если для любого $\theta \in \Theta$

$$\mathsf{E}_\theta[\hat{\theta}(\boldsymbol{X})] = \theta.$$

В данном случае E_θ означает, что при взятии математического ожидания мы предполагаем, что выборка $\boldsymbol{X}$ взята из распределения P_θ.

Определение 3. Пусть размер выборки $\boldsymbol{X}$ увеличивается и $\hat{\theta}_n(\boldsymbol{X}) = \hat{\theta}_n(X_1, \ldots, X_n)$. Тогда последовательность оценок $\hat{\theta}_n(\boldsymbol{X})$ будет называться *состоятельной* оценкой параметра θ, если для любого $\theta \in \Theta$

$$\hat{\theta}_n(\boldsymbol{X}) \xrightarrow{\mathsf{P}_\theta} \theta.$$

Понятие *сильно состоятельной* оценки совпадает с определением состоятельной оценки с тем отличием, что в ней сходимость по вероятности заменяется на сходимость почти наверное:

$$\hat{\theta}_n(\boldsymbol{X}) \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \theta.$$

[1] Напомню, что борелевская функция — это отображение такое, что для него полный прообраз борелевского множества будет борелевским множеством.

Определение 4. Пусть размер выборки $\boldsymbol{X}$ увеличивается и $\hat{\theta}_n(\boldsymbol{X}) = \hat{\theta}_n(X_1, \ldots, X_n)$. Тогда последовательность оценок $\hat{\theta}_n(\boldsymbol{X})$ будет называться *асимптотически нормальной* оценкой параметра θ, если для любого $\theta \in \Theta$

$$\sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta) \xrightarrow{d_\theta} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\theta)),$$

где $\boldsymbol{\Sigma}(\theta)$ — это *асимптотическая дисперсия* $\hat{\theta}_n(\boldsymbol{X})$.

Теперь приведём общую идею того, как можно строить оценки. Допустим, что мы смогли найти функционал G такой, что для всех $\theta \in \Theta$ $G(\mathsf{P}_\theta) = \theta$. Тогда «хорошей» оценкой можно считать $\hat{\theta}(\boldsymbol{X}) = G(\mathsf{P}_n^*)$, где P_n^* — это *эмпирическое распределение*:[2]

$$\mathsf{P}_n^*(B) = \frac{1}{n}\sum_{k=1}^{n}[x_i \in B]$$

На эмпирическое распределение можно смотреть, как на дискретное распределение, равновероятно сосредоточенное в точках выборки $\boldsymbol{X}$. Теперь рассмотрим несколько примеров:

(a) Допустим, что функционал G устроен по следующему правилу:

$$G(\mathsf{P}_\theta) = \int_{\mathcal{X}} g(x)\mathsf{P}_\theta(\mathrm{d}x).$$

Это интеграл Лебега по мере P_θ от функции g. Людям, не знакомым с ним, можно читать его следующим образом: если P_θ — это абсолютно непрерывная вероятностная мера с плотностью p_θ, то

$$\int_{\mathcal{X}} g(x)\mathsf{P}_\theta(\mathrm{d}x) = \int_{\mathcal{X}} g(x)p_\theta(x)\,\mathrm{d}x.$$

Если же P_θ есть дискретная вероятностная мера, то

$$\int_{\mathcal{X}} g(x)\mathsf{P}_\theta(\mathrm{d}x) = \sum_{x\in\mathcal{X}} g(x)\mathsf{P}_\theta(\{x\}).$$

В таком случае оценка будет иметь вид

$$G(\mathsf{P}_n^*) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \equiv \overline{g(\boldsymbol{X})}.$$

$\overline{g(\boldsymbol{X})}$ обычно называют *выборочным средним*.

(b) Теперь скажем, что оценка имеет вид линейной комбинации функций от порядковых статистик (такие оценки обычно называют *L-оценками*):

$$\hat{\theta}_n(\boldsymbol{X}) = \sum_{i=1}^{n} \alpha_i\varphi_i(X_{(i)}).$$

На всякий случай напомним определение порядковой статистики. Пусть $\omega \in \Omega$ и $x_k = X_k(\omega)$. Далее, перенумеруем последовательность так, чтобы она шла по возрастанию: $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$. Эту последовательность называют *вариационным рядом*. Случайную величину $X_{(k)}(\omega) = x_{(k)}$ называют *k-й порядковой статистикой*. Сразу же заметим, что $X_{(1)} = \min\{X_1, \ldots, X_n\}$ и $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

В качестве примера возьмём G такой, что $G(\mathsf{P}_\theta)$ будет равен α-квантилю P_θ. Напомню, что α-квантиль c_α функции распределения F равен минимальному значению x, в котором $F(x) \geqslant \alpha$:

$$c_\alpha = \min\{x\colon F(x) \geqslant \alpha\}.$$

Тогда *выборочным α-квантилем* назовём

$$G(\mathsf{P}_n^*) = \begin{cases} X_{(n\alpha)}, & n\alpha \text{ целое} \\ X_{(\lceil n\alpha\rceil)}, & n\alpha \text{ не целое} \end{cases}$$

[2] Здесь и в дальнейшем для обозначения индикатора будем использовать *нотацию Айверсона*: $[P] = 1$, если P правда и 0 иначе.

(c) Теперь скажем, что функционал G устроен следующим образом:

$$G(\mathsf{Q}) = \arg\min_{\theta\in\Theta} \int_{\mathcal{X}} \psi(x,\theta)\mathsf{Q}(\mathrm{d}x).$$

Такие оценки называют *M-оценками*. Тогда

$$G(\mathsf{P}_n^*) = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{k=1}^{n} \psi(X_k,\theta).$$

В качетсве примера M-оценки можно взять оценку максимального правдоподобия: в ней $\psi(x,\theta) = -\log p_\theta(x)$.

Попробуем разобраться, в каких задачах можно построить «хорошие» оценки в том смысле, что они будут обладать свойствами, введённых выше.

Кто может определять распределение? Первая мысль, которая приходит в голову — моменты. Из этого мы сразу получаем так называемый *метод моментов*. Но этот метод плох: у него есть много недостатков, а плюс один — на него можно давать задачи на контрольной. Основная сложность состоит в том, что нужно уметь считать моменты, как функции от параметра, что может быть совсем нетривиально или же невозможно (особенно в случае, когда распределение далеко не табличное).

Для примера рассмотрим распределение Коши $\mathrm{C}(\theta,1)$. Обычно его задают плотностью $f(x;\theta,1)$, которая равна

$$f(x;\theta,1) = \frac{1}{\pi(1+(x-\theta)^2)}.$$

Если взять случайную величину с таким распределением, то у неё не будет матожидания, дисперсии да и всех моментов выше первого. Конечно, можно посчитать матожидание логарифма (в таком случае интеграл сойдётся), но выразить её, как явную функцию от θ, не получится. Небольшое примечание: в данном распределении θ выступает в качестве медианы.

Второй подход предлагает использовать квантили. Очевидно, что этот подход универсален, так как по любой функции распределения можно построить квантили. Но у этого метода есть сложность — квантили считать ещё сложнее, чем моменты. Так что этот метод в основном подходит только для определения «сдвига»: если плотность симметрична относительно какой-то точки, то эта точка служит медианой, а медиану на практике считать не так уж и сложно.

Ещё одна идея завязывается на M-оценках. Описать её явно не так уж и просто, но общую философию можно сказать так: «мы живём в наиболее вероятном мире». Допустим, мы посчитали функционал и определили θ при фиксированном $\boldsymbol{X}$, при котором он максимален (или минимален). Тогда эту θ берут в качестве настоящей, так как мы считаем, что она отражает действительность.

1.3 Метод моментов

Пусть $\boldsymbol{X}$ — выборка из распределения $\mathsf{P} \in \{\mathsf{P}_\theta \colon \theta \in \Theta\}$, причём $\Theta \subseteq \mathbb{R}^k$. Далее, возьмём *пробные функции* $g_1(x),\dots,g_k(x)$ такие, что вектор

$$m(\theta) = (\mathsf{E}_\theta[g_1(X_1)],\dots,\mathsf{E}_\theta[g_k(X_1)])$$

задаёт биекцию между Θ и $m(\Theta)$. Тогда оценкой $\hat\theta(\boldsymbol{X})$ по методу моментов с пробными функциями $g_1(x),\dots,g_k(x)$ называется решение системы уравнений (относительно θ)

$$\begin{cases} m_1(\theta) = \overline{g_1(\boldsymbol{X})} \\ \dots \\ m_k(\theta) = \overline{g_k(\boldsymbol{X})} \end{cases}$$

Здесь $m_j(\theta) = \mathsf{E}_\theta[g_j(X_1)]$. Так как $m(\theta)$ задаёт биекцию, то

$$\hat\theta(\boldsymbol{X}) = m^{-1}(\overline{g_1(\boldsymbol{X})},\dots,\overline{g_k(\boldsymbol{X})}).$$

Обычно в качестве пробных функций берут $x, x^2, \dots, x^k$.

Оказывается, что с теоретической точки зрения оценки с помощью метода моментов не так уж и плохи.

Лемма. *Пусть $\hat\theta_n(\boldsymbol{X})$ — это оценка θ по методу моментов. Тогда*

(a) *Если m^{-1} непрерывна, то $\hat{\theta}_n(\boldsymbol{X})$ есть сильно состоятельная оценка для θ.*

(b) *Если m^{-1} дифференцируема и $\mathsf{E}_\theta[g_j^2(X_1)] < +\infty$ для всех $j = 1, \ldots, k$, то $\hat{\theta}_n(\boldsymbol{X})$ есть асимптотически нормальная оценка θ.*

Доказательство. Для начала заметим, что по усиленному закону больших чисел

$$\overline{g_j(\boldsymbol{X})} = \frac{1}{n}\sum_{k=1}^{n} g_j(X_k) \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \mathsf{E}_\theta[g_j(X_1)].$$

Следовательно, по теореме о наследовании сходимости

$$\hat{\theta}_n(\boldsymbol{X}) = m^{-1}(\overline{g_1(\boldsymbol{X})}, \ldots, \overline{g_k(\boldsymbol{X})}) \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} m^{-1}(m_1(\theta), \ldots, m_k(\theta)) = m^{-1}(m(\theta)) = \theta.$$

Для доказательства асимптотической нормальности заметим следующее. Введём следующие векторы: $\boldsymbol{Y}_i = (g_1(X_i), \ldots, g_k(X_i))$. Далее, так как элементы выборки независимы и одинаково распределены, то $\boldsymbol{Y}_i$ тоже независимы и одинаково распределены. Так как мы предполагаем, что $\mathsf{E}_\theta[g_j^2(X_1)] < +\infty$, то матрица ковариаций $\boldsymbol{\Sigma}$ невырождена. Тогда мы можем применить многомерную центральную предельную теорему:

$$\sqrt{n}\left(\frac{\boldsymbol{Y}_1 + \ldots + \boldsymbol{Y}_n}{n} - \mathsf{E}_\theta[\boldsymbol{Y}_1]\right) \xrightarrow{d_\theta} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}(\theta))$$

Для дальнейших рассуждений нужна следующая

Лемма (о наследовании асимптотической сходимости)**.** *Пусть $\xi_i \in \mathbb{R}^m$ — последовательность случайных векторов, для которых существует вектор $\boldsymbol{a} \in \mathbb{R}^n$ и матрица $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ такие, что*

$$\sqrt{n}(\xi_n - \boldsymbol{a}) \xrightarrow{d_\theta} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Далее, пусть $H\colon \mathbb{R}^m \mapsto \mathbb{R}^n$ — это дифференцируемая функция. Тогда

$$\sqrt{n}(H(\xi_n) - H(\boldsymbol{a})) \xrightarrow{d_\theta} H'(\boldsymbol{a})\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

где $H'(\boldsymbol{a})$ — это матрица Якоби функции H в точке $\boldsymbol{a}$.

Теперь применим эту лемму, положив $H = m^{-1}$:

$$\sqrt{n}\left(m^{-1}\left(\frac{\boldsymbol{Y}_1 + \ldots + \boldsymbol{Y}_n}{n}\right) - m^{-1}(\mathsf{E}_\theta[\boldsymbol{Y}_1])\right) = \sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta) \xrightarrow{d_\theta} (m^{-1})'(m(\theta))\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}(\theta)).$$

Осталось заметить, что $(m^{-1})'(m(\theta))\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}(\theta)) = \mathcal{N}(\boldsymbol{0}, \mathbf{A}(\theta))$ для некоторой матрицы $\mathbf{A}(\theta)$. □

1.4 Выборочные квантили

Теперь приступим ко второму методу описания функции распределения — через квантили. Оказывается, что у выборочных квантилей есть очень хорошее свойство.

Теорема 1. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из распределения $F(x)$ с плотностью $f(x)$, а z_p — p-квантиль распределения $F(x)$. Далее, пусть $f(x)$ непрерывно дифференцируема в окрестности z_p и $f(z_p) > 0$. Тогда выборочный квантиль есть асимптотически нормальная оценка теоретического квантиля:*

$$\sqrt{n}(X_{[np]+1} - z_p) \xrightarrow{d_\theta} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(z_p)}\right).$$

Доказательство. Начнём с того, что посчитаем плотность k-й порядковой статистики. Для этого заметим, что если $X_{(k)} \leqslant x$, то хотя бы k элементов выборки не больше x. Тогда, если $F_k(x)$ — это функция распределения k-й статистики, то

$$F_k(x) = \mathsf{P}(X_{(k)} \leqslant x) = \sum_{s=k}^{n} \binom{n}{s} F^s(x)(1 - F(x))^{n-s}.$$

Теперь посчитаем плотность $f_k(x)$ функции распределения $F_k(x)$, продифференцировав её:

$$f_k(x) = \sum_{s=k}^{n} sf(x)\binom{n}{s}F^{s-1}(x)(1 - F(x))^{n-s} - \sum_{s=k}^{n-1}(n-s)f(x)\binom{n}{s}F^s(x)(1 - F(x))^{n-s-1}.$$

Осталось упростить эту сумму. Для этого заметим, что

$$s\binom{n}{s} = \frac{n!}{(s-1)!(n-s)!} = n\binom{n-1}{s-1}, \quad (n-s)\binom{n}{s} = \frac{n!}{s!(n-s-1)!} = n\binom{n-1}{s}.$$

Тогда, заменив индекс суммирования в первой сумме на $t = s-1$, получим следующее:

$$f_k(x) = \sum_{t=k-1}^{n-1} nf(x)\binom{n-1}{t}F^t(x)(1-F(x))^{n-t-1} - \sum_{t=k}^{n-1} nf(x)\binom{n-1}{t}F^t(x)(1-F(x))^{n-t-1}.$$

Но тогда

$$f_k(x) = nf(x)\binom{n-1}{k-1}F^{k-1}(x)(1-F(x))^{n-k}.$$

Теперь покажем, что если мы возьмём последовательность случайных величин T_n, построенных по правилу

$$T_n = \frac{f(z_p)\sqrt{n}}{\sqrt{p(1-p)}}(X_{(k)} - z_p), \text{ где } k = [np]+1,$$

то $T_n \xrightarrow{d_\theta} \mathcal{N}(0,1)$. Для этого покажем один промежуточный факт: если ξ — случайная величина из распределения с плотностью f_ξ, а $\eta = a\xi + b$, где $a > 0$ и b — константы, то η имеет плотность f_η, равную

$$f_\eta(x) = \frac{1}{a}f_\xi\left(\frac{x-b}{a}\right).$$

Это несложно понять, если заметить следующее:

$$F_\eta(x) = \mathsf{P}(\eta \leqslant x) = \mathsf{P}(a\xi + b \leqslant x) = \mathsf{P}\left(\xi \leqslant \frac{x-b}{a}\right) = F_\xi\left(\frac{x-b}{a}\right).$$

Дифференцируя по x, мы получим желаемое. Воспользуемся этим. Пусть $q_n(x)$ — плотность T_n. Тогда

$$q_n(x) = \frac{\sqrt{p(1-p)}}{f(z_p)\sqrt{n}}f_k\left(z_p + \frac{x\sqrt{p(1-p)}}{f(z_p)\sqrt{n}}\right).$$

Для удобства введём следующее обозначение:

$$t_n = z_p + \frac{x\sqrt{p(1-p)}}{f(z_p)\sqrt{n}}.$$

Тогда

$$q_n(x) = n\binom{n-1}{k-1}F(t_n)^{k-1}f(t_n)(1-F(t_n))^{n-k}\sqrt{\frac{p(1-p)}{n}}\frac{1}{f(z_p)}.$$

Теперь сделаем финт ушами и скажем, что

$$q_n(x) = A_1(n)A_2(n)A_3(n),$$

где

$$A_1(n) = \frac{f(t_n)}{f(z_p)},$$
$$A_2(n) = n\binom{n-1}{k-1}\sqrt{\frac{p(1-p)}{n}}p^{k-1}(1-p)^{n-k},$$
$$A_3(n) = \left(\frac{F(t_n)}{p}\right)^{k-1}\left(\frac{1-F(t_n)}{1-p}\right)^{n-k}.$$

Найдём пределы всех трёх выражений при $n \to \infty$:

1. В первом всё очень просто: достаточно заметить, что $t_n \to z_p$ при $n \to \infty$ и $f(x)$ непрерывна в окрестности z_p. Тогда
$$\frac{f(t_n)}{f(z_p)} \xrightarrow[n\to\infty]{} 1 \implies \lim_{n\to\infty} A_1(n) = 1.$$

2. Для этого пункта заметим, что

$$A_2(n) = k\binom{n}{k}\sqrt{\frac{p(1-p)}{n}}p^{k-1}(1-p)^{n-k}.$$

Далее, воспользуемся формулой Стирлинга:

$$A_2(n) \sim k\frac{\sqrt{2\pi n}(\frac{n}{e})^n}{\sqrt{2\pi k}(\frac{k}{e})^k\sqrt{2\pi(n-k)}(\frac{n-k}{e})^k}\sqrt{\frac{p(1-p)}{n}}p^{k-1}(1-p)^{n-k}.$$

Теперь вспомним, что $k = [np] + 1$. Но из этого следует, что $k \sim np$ и $n - k \sim n(1-p)$. Тогда

$$A_2(n) \sim np\frac{\sqrt{2\pi n}(\frac{n}{e})^n}{\sqrt{2\pi np}(\frac{k}{e})^k\sqrt{2\pi n(1-p)}(\frac{n-k}{e})^k}\sqrt{\frac{p(1-p)}{n}}p^{k-1}(1-p)^{n-k}.$$

Упростим:

$$A_2(n) \sim \frac{1}{\sqrt{2\pi}}\left(\frac{np}{k}\right)^k\left(\frac{n(1-p)}{n-k}\right)^{n-k}.$$

Теперь докажем, что $A_2(n) \to 1/\sqrt{2\pi}$ при $n \to \infty$. Заметим, что

$$A_2(n) \sim \frac{1}{\sqrt{2\pi}}\exp\left\{k\ln\frac{np}{k} + (n-k)\ln\frac{n(1-p)}{n-k}\right\}.$$

Далее, k отличается от np не более, чем на 1. Тогда логарифмы можно разложить в ряд Тейлора в нуле:

$$\ln\frac{np}{k} = \frac{np}{k} - 1 + O\left(\frac{1}{k^2}\right) = \frac{np-k}{k} + O\left(\frac{1}{n^2}\right),$$
$$\ln\frac{n(1-p)}{n-k} = \frac{n(1-p)}{n-k} - 1 + O\left(\frac{1}{(n-k)^2}\right) = \frac{k-np}{n-k} + O\left(\frac{1}{n^2}\right).$$

Следовательно,

$$A_2(n) \sim \frac{1}{\sqrt{2\pi}}\exp\left\{np - k + k - np + O\left(\frac{1}{n^2}\right)\right\} \xrightarrow[n\to\infty]{} \frac{1}{\sqrt{2\pi}}.$$

3. Для этого пункта нужно заметить, что $F(t_n) \to F(z_p) = p$ при $n \to \infty$. Из этого можно сделать вывод, что

$$\left(\frac{F(t_n)}{p}\right)^{k-1} \sim \left(\frac{F(t_n)}{p}\right)^k.$$

Далее, заметим, что

$$A_3(n) \sim \exp\left\{k\ln\frac{F(t_n)}{p} + (n-k)\ln\frac{1-F(t_n)}{1-p}\right\}.$$

Теперь разложим $F(t_n)$ в ряд Тейлора в точке z_p, пользуясь непрерывной дифференцируемостью $f(x)$:

$$F(t_n) = F(z_p) + (t_n - z_p)f(z_p) + \frac{1}{2}(t_n - z_p)^2 f'(z_p) + o((t_n - z_p)^2)$$
$$= p + \sqrt{\frac{p(1-p)}{n}}x + \frac{f'(z_p)}{f^2(z_p)}\frac{p(1-p)}{2n}x^2 + o((t_n - z_p)^2).$$

Тогда

$$\frac{F(t_n)}{p} = 1 + \sqrt{\frac{1-p}{np}}x + \frac{f'(z_p)}{f^2(z_p)}\frac{1-p}{2n}x^2 + o\left(\frac{1}{n}\right)$$

Следовательно, если взять от этого логарифм, то его можно разложить в ряд Тейлора:

$$\ln\frac{F(t_n)}{p} = \sqrt{\frac{1-p}{np}}x + \frac{f'(z_p)}{f^2(z_p)}\frac{1-p}{2n}x^2 + o\left(\frac{1}{n}\right) - \frac{1}{2}\frac{1-p}{np}x^2 + o\left(\frac{1}{n}\right).$$

Тогда

$$k\ln\frac{F(t_n)}{p} = \sqrt{\frac{1-p}{np}}xk + \frac{f'(z_p)}{f^2(z_p)}\frac{k(1-p)}{2n}x^2 - \frac{1}{2}\frac{1-p}{p}\frac{k}{n}x^2 + o(1).$$

Аналогичными расуждениями можно показать, что

$$(n-k)\ln\frac{1-F(t_n)}{1-p} = -\sqrt{\frac{p}{n(1-p)}}x(n-k) - \frac{f'(z_p)}{f^2(z_p)}\frac{p(n-k)}{2n}x^2 - \frac{1}{2}\frac{p}{1-p}\frac{n-k}{n}x^2 + o(1).$$

Теперь воспользуемся тем, что $|k-np| \leqslant 1$. Это означает, что $k = np + O(1/n)$. Тогда

$$\begin{aligned} k\ln\frac{F(t_n)}{p} &= \sqrt{\frac{1-p}{np}}x\left(np + O\left(\frac{1}{n}\right)\right) + \frac{1}{2}\frac{f'(z_p)}{f^2(z_p)}\frac{1-p}{n}x^2\left(np + O\left(\frac{1}{n}\right)\right) \\ &\quad - \frac{1}{2}\frac{1-p}{p}\frac{x^2}{n}\left(np + O\left(\frac{1}{n}\right)\right) + o(1) \\ &= x\sqrt{np(1-p)} + O\left(\frac{1}{\sqrt{n}}\right) + \frac{p(1-p)}{2}\frac{f'(z_p)}{f^2(z_p)}x^2 + O\left(\frac{1}{n^2}\right) \\ &\quad - \frac{1-p}{2}x^2 + O\left(\frac{1}{n^2}\right) + o(1) \\ &= x\sqrt{np(1-p)} + \frac{p(1-p)}{2}\frac{f'(z_p)}{f^2(z_p)}x^2 - \frac{1-p}{2}x^2 + o(1). \end{aligned}$$

Аналогично поступая со вторым слагаемым, получим, что

$$(n-k)\ln\frac{1-F(t_n)}{1-p} = -x\sqrt{np(1-p)} - \frac{p(1-p)}{2}\frac{f'(z_p)}{f^2(z_p)}x^2 - \frac{p}{2}x^2 + o(1).$$

Следовательно,

$$A_3(n) \sim \exp\left\{-\frac{x^2}{2} + o(1)\right\} \xrightarrow[n\to\infty]{} \exp\left\{-\frac{x^2}{2}\right\}.$$

В итоге мы получили, что для всех x

$$\lim_{n\to\infty} q_n(x) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}.$$

Это означает, что $q_n(x)$ будет равномерно сходится к плотности $\mathcal{N}(0,1)$ на любом компакте. Из равномерной сходимости на отрезке $[a,b]$ следует, что

$$\lim_{n\to\infty}(F_{T_n}(b) - F_{T_n}(a)) = \lim_{n\to\infty}\int_a^b q_n(x)\,\mathrm{d}x = \Phi(b) - \Phi(a).$$

Теперь нужно доказать, что $F_{T_n}(x) \to \Phi(x)$ для всех x, где $\Phi(x)$ — это стандартное нормальное распределение. Заметим, что для любого $a < b$

$$\overline{\lim_{n\to\infty}}\,|F_{T_n}(b) - \Phi(b)| \leqslant \overline{\lim_{n\to\infty}}\,|F_{T_n}(b) - F_{T_n}(a) + \Phi(b) - \Phi(a)| + \overline{\lim_{n\to\infty}}\,|F_{T_n}(a) - \Phi(a)|$$

Далее, первый предел уходит в ноль. Тогда

$$\overline{\lim_{n\to\infty}}\,|F_{T_n}(b) - \Phi(b)| \leqslant \overline{\lim_{n\to\infty}}\,|F_{T_n}(a) - \Phi(a)|.$$

Однако

$$\overline{\lim_{n\to\infty}}\,|F_{T_n}(a) - \Phi(a)| \leqslant \overline{\lim_{n\to\infty}}\,(F_{T_n}(a) + \Phi(a)) = \Phi(a) + \overline{\lim_{n\to\infty}}\,F_{T_n}(a).$$

Далее, снова сделаем финт ушами, пользуясь тем, что $F_{T_n}(a) < 1$:

$$\Phi(a) + \overline{\lim_{n\to\infty}}\,F_{T_n}(a) \leqslant \Phi(a) + \overline{\lim_{n\to\infty}}\,(F_{T_n}(a) - F_{T_n}(-a) + 1) = \Phi(a) + \Phi(a) - \Phi(-a) + 1.$$

Но эту сумму можно сделать сколь угодно малой, устремив a к $-\infty$. Следовательно, $T_n \xrightarrow{d_\theta} \mathcal{N}(0,1)$ и мы получаем желаемое. □

Посмотрим на частный случай квантили, а именно на медиану. Стоит заметить, что выборочную медиану определяют не как выборочный 1/2-квантиль:

Определение 5. Выборочной медианой $\hat{\mu}$ выборки $X_1, \ldots, X_n$ называют величину

$$\hat{\mu}(\boldsymbol{X}) = \begin{cases} X_{(k+1)}, & n = 2k+1 \\ (X_{(k)} + X_{(k+1)})/2, & n = 2k \end{cases}$$

Оказывается, что для такого определения свойство асимптотической нормальности тоже выполняется:

Упражнение. В условиях теоремы про асимптотическую нормальность выборочной квантили

$$\sqrt{n}(\hat{\mu}(\boldsymbol{X}) - z_{1/2}) \xrightarrow{d_\theta} \mathcal{N}\left(0, \frac{1}{4f^2(z_{1/2})}\right).$$

В качестве примера рассмотрим выборку $\boldsymbol{X}$ из распределения Коши $\mathrm{C}(\theta, 1)$. У данного распределения всё очень плохо с моментами, но при этом асимптотическая нормальность выборочной медианы выполняется:

$$\sqrt{n}(\hat{\mu}(\boldsymbol{X}) - \theta) \xrightarrow{d_\theta} \mathcal{N}\left(0, \frac{\pi^2}{4}\right).$$

2 Лекция 2

2.1 Сравнение оценок

Начнём с такой темы, как сравнение оценок. Очевидно, что оценок можно придумать навалом, можно и много оценок с хорошими свойствами наподобие состоятельности или асимптотической нормальности и прочими. Вопрос: кто из них лучше и как сравнивать их?

Пусть $\hat{\theta}(\boldsymbol{X})$ и $\theta^*(\boldsymbol{X})$ — оценки неизвестного параметра θ с «хорошими» свойствами. Грубо говоря, если оценка не состоятельна, то она нам не интересна глобально. Возникает вопрос: как сравнивать оценки $\hat{\theta}(\boldsymbol{X})$ и $\theta^*(\boldsymbol{X})$?

Базовый принцип сравнения завязан на так называемых функциях потерь и риска.

Определение 6. Функцию $\rho(x, y) \geqslant 0$ называют *функцией потерь*.

Простыми словами, функция потерь считает, сколько мы потеряем, если вместо x подставим y. Данная функция обычно должна иметь некоторые хорошие свойства (если она метрика — то это вообще прекрасно, но это не обязательно).

Приведём несколько примеров:

(a) Самый базовый пример — это квадратичная функция потерь: $\rho(x, y) = (x - y)^2$ $(x, y \in \mathbb{R})$.

(b) $\rho(x, y) = |x - y|$ — L_1-метрика.

(c) $\rho(\boldsymbol{x}, \boldsymbol{y}) = \langle \mathbf{A}(\boldsymbol{x} - \boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$, где $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{A} \succ 0$ (это означает, что матрица положительно определена).

Определение 7. *Функцией риска* оценки $\hat{\theta}(\boldsymbol{X})$ параметра θ называется средний размер потерь нашей оценки:

$$R(\theta, \hat{\theta}(\boldsymbol{X})) = \mathsf{E}_\theta[\rho(\theta, \hat{\theta}(\boldsymbol{X}))].$$

Как мы можем сравнивать оценки с помощью функции риска? Для этого есть аж четыре подхода.

1. Самый банальный подход называется *равномерным подходом*. Суть его крайне проста: просто сравниваем функции риска.

   Определение 8. Пусть $\hat{\theta}(\boldsymbol{X})$ и $\theta^*(\boldsymbol{X})$ — оценки θ. Будем говорить, что оценка $\hat{\theta}(\boldsymbol{X})$ лучше, чем оценка $\theta^*(\boldsymbol{X})$ (в равномерном смысле), если для всех $\theta \in \Theta$ функция риска для $\hat{\theta}(\boldsymbol{X})$ не превосходит функцию риска для $\theta^*(\boldsymbol{X})$:

   $$R(\theta, \hat{\theta}(\boldsymbol{X})) \leqslant R(\theta, \theta^*(\boldsymbol{X})),$$

   причём существует параметр $\theta_0 \in \Theta$ такой, что в нём неравенство будет строгим: $R(\theta_0, \hat{\theta}(\boldsymbol{X})) < R(\theta_0, \theta^*(\boldsymbol{X}))$.

   Как правило, если мы взяли хорошую функцию потерь (наподобие квадратичной), то равество функций риска везде означает равенство оценок, так что последнее требование не совсем обязательно.

   Что теперь напрашивется? Поиск лучшей оценки. Будем называть оценку $\hat{\theta}(\boldsymbol{X})$ наилучшей, если она лучше любой другой в равномерном подходе. Однако поиск наилучшей оценки — это достаточно бессмысленная задача, так как мы её не найдём. Рассмотрим следующий пример. Пусть $\rho(x, y) = (x - y)^2$ — квадратичная функция потерь, параметр $\theta \in [0, 1]$, а оценка возвращает тождественный ноль: $\hat{\theta}(\boldsymbol{X}) \equiv 0$. Тогда функция риска будет равна

   $$R(\theta, \hat{\theta}(\boldsymbol{X})) = \mathsf{E}_\theta[(\theta - \hat{\theta}(\boldsymbol{X}))^2] = \theta^2.$$

Однако, если $\theta = 0$, то $R(\theta, \hat{\theta}(\boldsymbol{X})) = 0$. Аналогично, если оценка возвращает не 0, а какое-то $\theta_0 \in [0, 1]$, то $R(\theta_0, \hat{\theta}(\boldsymbol{X})) = 0$. Но тогда если бы у нас была наилучшая оценка $\theta^*(\boldsymbol{X})$, то у неё должен быть тождественно нулевой риск: $R(\theta, \theta^*(\boldsymbol{X})) \equiv 0$. Тем самым мы получаем, что $\hat{\theta}(\boldsymbol{X}) = \theta$. Следовательно, наилучшей оценки в равномерном смысле просто нет.

Как тогда быть? Можно сказать, что будем смотреть не на все оценки, а только лишь на какой-то класс. Тогда можно ввести следующее определение:

Определение 9. Оценка $\hat{\theta}(\boldsymbol{X})$ называется наилучшей оценкой в классе оценок $\mathcal{K}$, если она лучше, чем любая другая оценка $\theta^*(\boldsymbol{X}) \in \mathcal{K}$. Если при подсчёте функции риска используется квадратичная функция потерь, то оценку $\hat{\theta}(\boldsymbol{X})$ называют *оптимальной* в классе оценок $\mathcal{K}$.

В качестве примера класса оценок можно сказать, что $\mathcal{K}$ — это все несмещённые оценки $\tau(\theta)$.

2. *Байесовский подход.* Смысл этого подхода состоит в следующем: равномерный подход пытается считать, что все параметры в каком-то смымле равноправны и мы должны иметь неравенство для каждого $\theta \in \Theta$. Байесовский подход утверждает, что параметры не равноправны и некоторые параметры более вероятны, чем другие и мы пытаемя рассматривать параметр θ как случайную величину. Тогда нам важно выполнение неравенства на главных значениях, а на маловероятных уже без разницы. Формализуем это:

   Определение 10. Пусть Q — некоторое вероятностное распределение на Θ с плотностью $q(t)$, а $\hat{\theta}(\boldsymbol{X})$ и $\theta^*(\boldsymbol{X})$ есть оценки θ. Тогда будем говорить, что оценка $\hat{\theta}(\boldsymbol{X})$ лучше оценки $\theta^*(\boldsymbol{X})$, если

   $$R_{\mathsf{Q}}(\hat{\theta}(\boldsymbol{X})) < R_{\mathsf{Q}}(\theta^*(\boldsymbol{X})), \text{ где } R_{\mathsf{Q}}(\hat{\theta}(\boldsymbol{X})) = \int_{\Theta} R(\theta, \hat{\theta}(\boldsymbol{X}))q(\theta)\,\mathrm{d}\theta.$$

   Далее, будем называть оценку $\hat{\theta}(\boldsymbol{X})$ *наилучшей*, если для неё R_{Q} минимально:

   $$R_{\mathsf{Q}}(\hat{\theta}(\boldsymbol{X})) = \inf_{\theta^*(\boldsymbol{X})} R_{\mathsf{Q}}(\theta^*(\boldsymbol{X})).$$

   Стоит заметить, что наилучшая оценка зависит от выбора распределения Q, то есть байесовский подход зависит от выбора распределения. По сути, на Q можно смотреть, как на априорное представление о том, какие значения могут принимать параметры. Грубо говоря, мы можем пытаться делать какие-то такие вещи: если какая-то оценка оказалась ближе к истине, чем другие, но их мы откидывать не хотим, тогда просто возьмём их с какими-то весами.

   В качестве классического примера можно взять задачу про однорукого бандита.[3] Допустим, что есть игральный автомат, у которого несколько ручек — для красоты будем считать, что ручек две, и вероятность выигрыша у одной ручки больше, чем у другой. Наша задача — оптимизировать стратегию среднего выигрыша. Самый простой подход состоит в следующем: прокрутим обе ручки по 100 раз и посмотрим, с какой ручки больше выигрыша (пусть вышло так, что с первой). Равномерный подход говорит, что первая ручка лучше и достаточно крутить только её. Байесовский же подход утверждает, что могло случиться так, что вторая ручка лучше, но мы её просто не использовали слишком малое количество раз и нам просто не повезло — если бы взяли миллион раз, то было бы лучше. Поэтому стратегия байесовского подхода будет в таком стиле: десять раз крутим первую ручку и один раз вторую.

3. *Минимаксный подход.* Этот подход достаточно тесно связан с байесовским подходом. Он апеллирует к следующему: мы вообще не хотим больших значений функций потерь, то есть если она мала для всех значений $\theta \in \Theta$, то нас это устраивает.

   Определение 11. Пусть $\hat{\theta}(\boldsymbol{X})$ и $\theta^*(\boldsymbol{X})$ — оценки θ. Будем говорить, что оценка $\hat{\theta}(\boldsymbol{X})$ лучше, чем оценка $\theta^*(\boldsymbol{X})$ (в минимаксном смысле), если

   $$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}(\boldsymbol{X})) < \sup_{\theta \in \Theta} R(\theta, \theta^*(\boldsymbol{X})).$$

   Оценку $\hat{\theta}(\boldsymbol{X})$ будем называть наилучшей, если для неё выполнено следующее равенство:

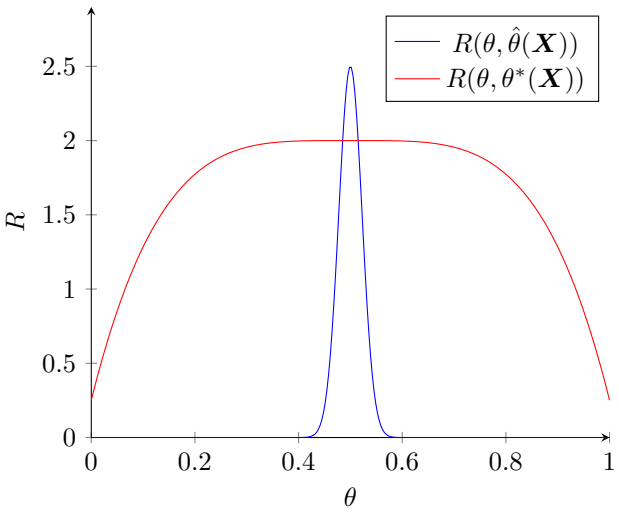
   $$\sup_{\theta \in \Theta} R(\theta, \hat{\theta}(\boldsymbol{X})) = \inf_{\theta^*(\boldsymbol{X})} \sup_{\theta \in \Theta} R(\theta, \theta^*(\boldsymbol{X})).$$

[3] На самом деле это задача так называемого *последовательного анализа.*

Пусть у нас есть две оценки $\hat{\theta}(\boldsymbol{X})$ и $\theta^*(\boldsymbol{X})$ для параметра $\theta \in [0,1]$ со следующими функциями риска (см. рисунок справа). Что можно сказать про эти оценки?

- Если в байесовском подходе сказать, что мы берём равномерное распределение, то есть $\mathsf{Q} = \mathrm{U}(0,1)$, то оценка $\hat{\theta}(\boldsymbol{X})$ будет лучше оценки $\theta^*(\boldsymbol{X})$, так как у неё меньше площадь под кривой.
- Теперь рассмотрим минимаксный подход. В нём оценка $\theta^*(\boldsymbol{X})$ будет лучше оценки $\hat{\theta}(\boldsymbol{X})$, так как у неё меньше максимальное значение.
- В равномерном же подходе эти оценки несравнимы, так как есть и участки, где $R(\theta, \hat{\theta}(\boldsymbol{X})) > R(\theta, \theta^*(\boldsymbol{X}))$, так и участки, где $R(\theta, \hat{\theta}(\boldsymbol{X})) < R(\theta, \theta^*(\boldsymbol{X}))$.

Функции риска для $\hat{\theta}(\boldsymbol{X})$ и $\theta^*(\boldsymbol{X})$



Стоит заметить, что если в байесовском подходе взять распределение, сосредоточенное в центре, то оценка $\theta^*(\boldsymbol{X})$ будет лучше, чем $\hat{\theta}(\boldsymbol{X})$.

Всё это сравнение оценок приводит к понятию *допустимой оценки*. Суть в том, что равномерный подход наиболее сильный — если оценка лучше в равномерном подходе, то она будет лучше и в байесовском, и в минимаксном подходе. Поэтому имеет смысл смотреть только те оценки, для которых нет более хороших оценок в равномерном смысле, так как иначе можно просто взять более хорошую.

Определение 12. Будем называть оценку $\hat{\theta}(\boldsymbol{X})$ *допустимой*, если не существует оценки $\theta^*(\boldsymbol{X})$ такой, что $\theta^*(\boldsymbol{X})$ лучше, чем $\hat{\theta}(\boldsymbol{X})$ в равномерном подходе.

Есть ещё один подход для сравнения оценок, но он принципиально отличается от предыдущих тем, что он не использует функции риска и он работает только для асимптотически нормальных оценок.

4. *Асимптотический подход.* Он завязан на определении асимптотической дисперсии.

   Определение 13. Пусть $\hat{\theta}_n(\boldsymbol{X})$ и $\theta^*_n(\boldsymbol{X})$ — асимптотически нормальные оценки параметра θ с асимптотическими дисперсиями $\sigma_1^2(\theta)$ и $\sigma_2^2(\theta)$ соответственно. В таком случае будем говорить, что оценка $\hat{\theta}_n(\boldsymbol{X})$ лучше, чем оценка $\theta^*_n(\boldsymbol{X})$, если для всех $\theta \in \Theta$ $\sigma_1^2(\theta) \leqslant \sigma_2^2(\theta)$.

   Интуитивно на это можно смотреть, как на равномерное ограничение предельного среднеквадратичного уклонения. Почему это может быть важно? С помощью асимптотической дисперсии можно строить доверительные интервалы, и при её уменьшении будет уменьшаться длина интервала, что хорошо.

   Для примера попробуем сравнить какие-нибудь оценки.

   Задача 1. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из $\mathcal{N}(\theta, 1)$, $\overline{\boldsymbol{X}}$ — выборочное среднее, а $\hat{\mu}(\boldsymbol{X})$ — выборочная медиана. Сравните эти оценки в асимптотическом подходе.*

   Решение. Для начала сразу же уточним, что для нормального распределения математическое ожидание является медианой, поэтому оценивается одна и та же величина и задача действительно осмысленна. Далее, согласно центральной предельной теореме выборочне среднее является асимптотически нормальной оценкой:

   $$\sqrt{n}(\overline{\boldsymbol{X}} - \theta) \xrightarrow{d_\theta} \mathcal{N}(0, \mathsf{D}_\theta[X_1]) = \mathcal{N}(0,1).$$

   Заметим, что плотность $f(x)$ нормального распределения непрерывно дифференцируема и положительна. Следовательно, можно воспользоваться теоремой об асимптотической нормальности выборочной медианы:

   $$\sqrt{n}(\hat{\mu}(\boldsymbol{X}) - \theta) \xrightarrow[n\to\infty]{d_\theta} \mathcal{N}\left(0, \frac{1}{4f^2(\theta)}\right) = \mathcal{N}\left(0, \frac{\pi}{2}\right).$$

   Но тогда выборочная медиана хуже в асимптотическом подходе, чем выборочное среднее. И это неспроста. □

   Этот факт кажется не самым очевидным, так как известно, что медиана более устойчива к выбросам в выборке.

2.2 Неравенство Рао-Крамера

Будем считать, что мы работаем в классе $\mathcal{K}$ несмещённых оценок $\tau(\theta)$. Поставим следующую задачу: как найти оптимальную оценку в этом классе? Другими словами, нам нужно решить следующую задачу: если $\hat{\theta}(\boldsymbol{X})$ — это несмещённая оценка $\tau(\theta)$, то нам нужно равномерно по всем $\theta \in \Theta$ минимизровать $\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \tau(\theta))^2] = \mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})]$.

Оказывается, что если наложить так называемые *условия регулярности*, то можно предоставить нижнюю оценку для дисперсии оценки. Сформулируем эти условия:

1. Пусть множество параметров Θ — это открытый интервал на $\mathbb{R}$.

2. Далее, пусть параметрическое семейство распределений $\{\mathsf{P}_\theta \colon \theta \in \Theta\}$ является доминируемым с плотностью $p_\theta(x)$. Немного разъясним это требование.

   Определение 14. Параметрическое семейство распределений называется *доминируемым*, если все распределения в нём принадлежат к одному типу (абсолютно непрерывному или дискретному).

   Если все распределения абсолютно непрерывны, то под плотностью $p_\theta(x)$ подразумевается обычная плотность вероятностного распределения P_θ. Если же все распределения дискретны, то $p_\theta(x) = \mathsf{P}_\theta(\{x\})$.

   В таком случае удобно ввести понятие *правдоподобия*. Это просто плотность выборки $\boldsymbol{X}$ и она равна

   $$p_\theta(\boldsymbol{X}) = \prod_{i=1}^{n} p_\theta(X_i).$$

3. Пусть $A = \{\boldsymbol{x} \colon p_\theta(\boldsymbol{x}) > 0\}$. Тогда A не зависит от θ.

Сделаем небольшое отступление: для того, чтобы доказывать некоторые утверждения одновременно и для дискретных, и для абсолютно непрерывных распределений, скажем, что[4]

$$\int_A p_\theta(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) = \begin{cases} \int_A p_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, & \mathsf{P}_\theta \text{ абсолютно непрерывна} \\ \sum_{x \in A} p_\theta(\boldsymbol{x}), & \mathsf{P}_\theta \text{ дискретна} \end{cases}$$

4. Для любой статистики $S(\boldsymbol{X})$ такой, что $\mathsf{E}_\theta[S^2(\boldsymbol{X})] < +\infty$, есть возможность дифференцирования под знаком интеграла:

   $$\frac{\partial}{\partial\theta}\,\mathsf{E}_\theta[S(\boldsymbol{X})] = \mathsf{E}_\theta[S(\boldsymbol{X})U_\theta(\boldsymbol{X})], \text{ где } U_\theta(\boldsymbol{X}) = \frac{\partial}{\partial\theta}\ln p_\theta(\boldsymbol{X}).$$

   Теперь покажем, почему это то же самое, что и дифференцирование под знаком интеграла:

   $$\begin{aligned} \frac{\partial}{\partial\theta}\,\mathsf{E}_\theta[S(\boldsymbol{X})] &= \frac{\partial}{\partial\theta}\int_A S(\boldsymbol{x})p_\theta(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) = \int_A S(\boldsymbol{x})\frac{\partial p_\theta(\boldsymbol{x})}{\partial\theta}\mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_A S(\boldsymbol{x})\left(\frac{1}{p_\theta(\boldsymbol{x})}\frac{\partial p_\theta(\boldsymbol{x})}{\partial\theta}\right)p_\theta(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_A S(\boldsymbol{x})\frac{\partial \ln p_\theta(\boldsymbol{x})}{\partial\theta}p_\theta(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) = \mathsf{E}_\theta[S(\boldsymbol{X})U_\theta(\boldsymbol{X})]. \end{aligned}$$

   Функцию $U_\theta(\boldsymbol{X})$ называют *вкладом* наблюдения $\boldsymbol{X}$.

5. Величина

   $$I_{\boldsymbol{X}}(\theta) = \mathsf{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln p_\theta(\boldsymbol{X})\right)^2\right],$$

   называемая *информацией* (по Фишеру) выборки $\boldsymbol{X}$, должна быть положительна и конечна для любых $\theta \in \Theta$.

Теперь можно сформулировать теорему, которая будет давать оценку снизу.

Теорема 2 (Неравенство Рао-Крамéра)**.** *Пусть τ — это дифференцируемая функция от параметра. Далее, пусть $\hat{\theta}(\boldsymbol{X})$ — это несмещённая оценка $\tau(\theta)$ такая, что $\mathsf{E}_\theta[\hat{\theta}^2(\boldsymbol{X})] < +\infty$ и выполнены условия регулярности. Тогда для всех $\theta \in \Theta$*

$$\mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})] \geqslant \frac{(\tau'(\theta))^2}{I_{\boldsymbol{X}}(\theta)}.$$

[4] В принципе, это связано с ремаркой из первой лекции, ибо здесь под μ подразумевается мера Лебега.

Доказательство. Воспользуемся четвёртым условием регулярности, подставив в неё $S(\boldsymbol{X}) \equiv 1$:

$$\frac{\partial}{\partial\theta}\,\mathsf{E}_\theta[S(\boldsymbol{X})] = \frac{\partial}{\partial\theta}\,\mathsf{E}_\theta[1] = 0,$$
$$\mathsf{E}_\theta[S(\boldsymbol{X})U_\theta(\boldsymbol{X})] = \mathsf{E}_\theta[U_\theta(\boldsymbol{X})].$$

Следовательно, $\mathsf{E}_\theta[U_\theta(\boldsymbol{X})] = 0$.

Далее, снова воспользуемся четвёртым свойством регулярности, подставив в него $S(\boldsymbol{X}) = \hat{\theta}(\boldsymbol{X})$:

$$\frac{\partial}{\partial\theta}\,\mathsf{E}_\theta[S(\boldsymbol{X})] = \frac{\partial}{\partial\theta}\,\mathsf{E}_\theta[\hat{\theta}(\boldsymbol{X})] = \tau'(\theta),$$
$$\mathsf{E}_\theta[S(\boldsymbol{X})U_\theta(\boldsymbol{X})] = \mathsf{E}_\theta[\hat{\theta}(\boldsymbol{X})U_\theta(\boldsymbol{X})] = \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \tau(\theta))U_\theta(\boldsymbol{X})]$$

Тем самым $\tau'(\theta) = \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \tau(\theta))U_\theta(\boldsymbol{X})]$. Теперь вспомним неравенство Коши-Буняковского-Шварца: для любых случайных величин X, Y с конечными вторыми моментами

$$\mathsf{E}^2[XY] \leqslant \mathsf{E}[X^2]\,\mathsf{E}[Y^2].$$

Тогда согласно этому неравенству

$$(\tau'(\theta))^2 \leqslant \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \tau(\theta))^2]\,\mathsf{E}_\theta[U_\theta^2(\boldsymbol{X})] = \mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})]I_{\boldsymbol{X}}(\theta).$$

Отсюда получаем желаемое:

$$\mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})] \geqslant \frac{(\tau'(\theta))^2}{I_{\boldsymbol{X}}(\theta)}.$$ □

А можно ли вообще достичь эту оценку? В некоторых случаях можно.

Определение 15. Пусть $\hat{\theta}(\boldsymbol{X})$ — несмещённая оценка для $\tau(\theta)$. Будем называть её *эффективной*, если для неё достигается равенство в неравенстве Рао-Крамера, то есть для всех $\theta \in \Theta$

$$\mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})] = \frac{(\tau'(\theta))^2}{I_{\boldsymbol{X}}(\theta)}.$$

Оказывается, для проверки оценки на эффективность даже есть критерий.

Теорема 3 (Критерий эффективности)**.** *В условиях неравенства Рао-Крамера $\hat{\theta}(\boldsymbol{X})$ будет эффективной оценкой $\tau(\theta)$ тогда и только тогда, когда*

$$\hat{\theta}(\boldsymbol{X}) - \tau(\theta) = c(\theta)U_\theta(\boldsymbol{X}),\ \textit{где}\ c(\theta) = \frac{\tau'(\theta)}{I_{\boldsymbol{X}}(\theta)}.$$

Доказательство. Как известно, в неравенстве Коши-Буняковского-Шварца равенство достигается тогда и только тогда, когда между случайными величинами есть линейная зависимость. Следовательно, $\hat{\theta}(\boldsymbol{X})$ будет эффективной оценкой для $\tau(\theta)$ только в том случае, если существуют функции $c(\theta)$ и $a(\theta)$ такие, что

$$\hat{\theta}(\boldsymbol{X}) - \tau(\theta) = c(\theta)U_\theta(\boldsymbol{X}) + a(\theta).$$

Теперь посчитаем эти функции.

- Для начала возьмём матожидание:

  $$\mathsf{E}_\theta[\hat{\theta}(\boldsymbol{X}) - \tau(\theta)] = \mathsf{E}_\theta[c(\theta)U_\theta(\boldsymbol{X}) + a(\theta)].$$

  Тогда

  $$\tau(\theta) - \tau(\theta) = c(\theta)\cdot 0 + a(\theta) \implies a(\theta) = 0.$$

- Теперь помножим обе части равенства на $U_\theta(\boldsymbol{X})$ и возьмём матожидание:

  $$\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \tau(\theta))U_\theta(\boldsymbol{X})] = \mathsf{E}_\theta[c(\theta)U_\theta^2(\boldsymbol{X})].$$

  Но тогда

  $$\tau'(\theta) = c(\theta)I_{\boldsymbol{X}}(\theta) \implies c(\theta) = \frac{\tau'(\theta)}{I_{\boldsymbol{X}}(\theta)}.$$

Тем самым получаем желаемое. □

Теперь попробуем применить этот критерий.

Задача 2. *Пусть* $\boldsymbol{X} = (X_1, \ldots, X_n)$ — *выборка из распределения Бернулли* $\mathrm{Bin}(1, \theta)$, *где* $\theta \in (0, 1)$. *Найти эффективную оценку* θ *и информацию* $I_{\boldsymbol{X}}(\theta)$.

Решение. Воспользуемся критерием эффективности. Для этого нужно посчитать вклад. Начнём с того, что запишем плотность выборки.

Небольшое отступление: если нам известно, что случайная величина ξ принимает значения $x_1, \ldots, x_m$ с вероятностями $p_1, \ldots, p_m$, то плотность можно записать следующим образом:

$$p_\xi(x) = \prod_{i=1}^{m} p_i^{[x=x_i]}.$$

Воспользуемся этим и запишем правдоподобие:

$$p_\theta(\boldsymbol{X}) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} \theta^{[X_i=1]}(1-\theta)^{[X_i=0]} = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}$$

Теперь несложно посчитать правдоподобие:

$$\begin{aligned} \ln p_\theta(\boldsymbol{X}) &= \sum_{k=1}^{n} X_i \ln\theta + (1 - X_i)\ln(1-\theta) \\ U_\theta(\boldsymbol{X}) &= \frac{X_1 + \ldots + X_n}{\theta} - \frac{n - X_1 - \ldots - X_n}{1-\theta} \\ &= n\left(\frac{\overline{\boldsymbol{X}}}{\theta} - \frac{1-\overline{\boldsymbol{X}}}{1-\theta}\right) = \frac{n}{\theta(1-\theta)}(\overline{\boldsymbol{X}} - \theta). \end{aligned}$$

Отсюда видно, что $\overline{\boldsymbol{X}}$ является эффективной оценкой θ и информация Фишера равна

$$I_{\boldsymbol{X}}(\theta) = \frac{n}{\theta(1-\theta)}.$$ □

2.3 Информация Фишера

Начнём с того, что обобщим понятие информации по Фишеру:

Определение 16. Пусть $S(\boldsymbol{X})$ — это некоторая статистика с плотностью $g_\theta(x)$. Тогда *информацией Фишера* статистики $S(\boldsymbol{X})$ называется

$$I_S(\theta) = \mathsf{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln g_\theta(S(\boldsymbol{X}))\right)^2\right].$$

У информации Фишера есть несколько свойств, которые оправдывают её название.

1. Допустим, что распределение $S(\boldsymbol{X})$ не зависит от θ. Тогда $I_S(\theta) = 0$.

Доказательство. Это достаточно очевидное утверждение, так как в таком случае $\ln g_\theta(S(\boldsymbol{X}))$ не будет зависеть от θ и при дифференцировании по нему обратится в 0, а матожидание нуля есть ноль. □

2. Пусть $S(\boldsymbol{X})$ и $T(\boldsymbol{X})$ — независимые статистики для всех $\theta \in \Theta$ и выполнены условия регулярности. Тогда
$$I_{(S,T)}(\theta) = I_S(\theta) + I_T(\theta).$$

Доказательство. Пусть статистики $S(\boldsymbol{X})$ и $T(\boldsymbol{X})$ имеют плотности $f_\theta(s)$ и $g_\theta(t)$ соответственно. Тогда случайный вектор $(S(\boldsymbol{X}), T(\boldsymbol{X}))$ имеет совместную плотность $h_\theta(s,t) = f_\theta(s)g_\theta(t)$. Следовательно,

$$\ln h_t(s,t) = \ln f_\theta(s) + \ln g_\theta(t).$$

Из этого можно сделать вывод, что

$$\mathsf{E}_\theta\left[\frac{\partial}{\partial\theta}\ln h_\theta(S(\boldsymbol{X}), T(\boldsymbol{X}))\right] = \mathsf{E}_\theta\left[\frac{\partial}{\partial\theta}\ln f_\theta(S(\boldsymbol{X}))\right] + \mathsf{E}_\theta\left[\frac{\partial}{\partial\theta}\ln g_\theta(T(\boldsymbol{X}))\right] = 0.$$

Но тогда

$$\begin{aligned} I_{(S,T)}(\theta) &= \mathsf{D}_\theta\left[\frac{\partial}{\partial\theta}\ln h_\theta(S(\boldsymbol{X}), T(\boldsymbol{X}))\right] \\ &= \mathsf{D}_\theta\left[\frac{\partial}{\partial\theta}\ln f_\theta(S(\boldsymbol{X}))\right] + \mathsf{D}_\theta\left[\frac{\partial}{\partial\theta}\ln g_\theta(T(\boldsymbol{X}))\right] = I_S(\theta) + I_T(\theta). \end{aligned}$$ □

У этого свойства есть достаточно приятное следствие. Пусть

$$i(\theta) = \mathsf{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \ln p_\theta(X_1) \right)^2 \right].$$

Тогда

$$I_{\boldsymbol{X}}(\theta) = \sum_{i=1}^{n} I_{X_i}(\theta) = n I_{X_1}(\theta) = n i(\theta).$$

3. Для любой статистики $S(\boldsymbol{X})$ $I_S(\theta) \leqslant I_{\boldsymbol{X}}(\theta)$.

Доказательство. Для начала покажем, что если у статистики $S(\boldsymbol{X})$ есть плотность $g_\theta(s)$, то

$$\mathsf{E}_\theta \left[\frac{\partial}{\partial \theta} \ln p_\theta(\boldsymbol{X}) \,\middle|\, S(\boldsymbol{X}) \right] = \frac{\partial}{\partial \theta} \ln g_\theta(S(\boldsymbol{X})).$$

Будем делать это по определению условного математического ожидания. Начнём с того, что функция справа является борелевской функцией от $S(\boldsymbol{X})$. Осталось проверить интегральное свойство: для любого борелевского множества B

$$\mathsf{E}_\theta \left[\frac{\partial}{\partial \theta} \ln p_\theta(\boldsymbol{X}) [S(\boldsymbol{X}) \in B] \right] = \mathsf{E}_\theta \left[\frac{\partial}{\partial \theta} \ln g_\theta(S(\boldsymbol{X})) [S(\boldsymbol{X}) \in B] \right].$$

Заметим, что

$$\begin{aligned} \mathsf{E}_\theta \left[\frac{\partial}{\partial \theta} \ln p_\theta(\boldsymbol{X}) [S(\boldsymbol{X}) \in B] \right] &= \int_A \frac{\partial \ln p_\theta(\boldsymbol{x})}{\partial \theta} p_\theta(\boldsymbol{x}) [S(\boldsymbol{x}) \in B] \mu(\mathrm{d}\boldsymbol{x}) = \int_A \frac{\partial p_\theta(\boldsymbol{x})}{\partial \theta} [S(\boldsymbol{x}) \in B] \mu(\mathrm{d}\boldsymbol{x}) \\ &= \frac{\partial}{\partial \theta} \int_A p_\theta(\boldsymbol{x}) [S(\boldsymbol{x}) \in B] \mu(\mathrm{d}\boldsymbol{x}) = \frac{\partial}{\partial \theta} \mathsf{P}_\theta(S(\boldsymbol{X}) \in B) \end{aligned}$$

Но, с другой стороны

$$\begin{aligned} \frac{\partial}{\partial \theta} \mathsf{P}_\theta(S(\boldsymbol{X}) \in B) &= \frac{\partial}{\partial \theta} \int_B g_\theta(\boldsymbol{s}) \mu'(\mathrm{d}\boldsymbol{s}) = \int_B \frac{\partial \ln p_\theta(\boldsymbol{s})}{\partial \theta} g_\theta(\boldsymbol{s}) \mu'(\mathrm{d}\boldsymbol{s}) \\ &= \mathsf{E}_\theta \left[\frac{\partial}{\partial \theta} \ln g_\theta(S(\boldsymbol{X})) [S(\boldsymbol{X}) \in B] \right]. \end{aligned}$$

Теперь будем смотреть на следующую величину:

$$M = \mathsf{E}_\theta \left[\frac{\partial \ln p_\theta(\boldsymbol{X})}{\partial \theta} \frac{\partial \ln g_\theta(S(\boldsymbol{X}))}{\partial \theta} \right].$$

Согласно неравенству Коши-Буняковского-Шварца

$$M^2 \leqslant \mathsf{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \ln p_\theta(\boldsymbol{X}) \right)^2 \right] \mathsf{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \ln g_\theta(S(\boldsymbol{X})) \right)^2 \right] = I_{\boldsymbol{X}}(\theta) I_S(\theta).$$

Теперь докажем, что $M = I_S(\theta)$. Для этого заметим, что

$$\mathsf{E}_\theta \left[\frac{\partial \ln p_\theta(\boldsymbol{X})}{\partial \theta} \frac{\partial \ln g_\theta(S(\boldsymbol{X}))}{\partial \theta} \right] = \mathsf{E}_\theta \left[\mathsf{E}_\theta \left[\frac{\partial \ln p_\theta(\boldsymbol{X})}{\partial \theta} \frac{\partial \ln g_\theta(S(\boldsymbol{X}))}{\partial \theta} \,\middle|\, S(\boldsymbol{X}) \right] \right].$$

Далее, функцию от условия можно вынести из условного математического ожидания. Тогда

$$\begin{aligned} \mathsf{E}_\theta \left[\frac{\partial \ln p_\theta(\boldsymbol{X})}{\partial \theta} \frac{\partial \ln g_\theta(S(\boldsymbol{X}))}{\partial \theta} \right] &= \mathsf{E}_\theta \left[\frac{\partial \ln g_\theta(S(\boldsymbol{X}))}{\partial \theta} \mathsf{E}_\theta \left[\frac{\partial \ln p_\theta(\boldsymbol{X})}{\partial \theta} \,\middle|\, S(\boldsymbol{X}) \right] \right] \\ &= \mathsf{E}_\theta \left[\left(\frac{\partial \ln g_\theta(S(\boldsymbol{X}))}{\partial \theta} \right)^2 \right] = I_{\boldsymbol{S}}(\theta). \end{aligned}$$

Тогда

$$I_S^2(\theta) \leqslant I_{\boldsymbol{X}}(\theta) I_S(\theta) \implies I_S(\theta) \leqslant I_{\boldsymbol{X}}(\theta). \qquad \square$$

4. Оценка $S(\boldsymbol{X})$ будет достаточной тогда и только тогда, когда для всех $\theta \in \Theta$ $I_S(\theta) = I_{\boldsymbol{X}}(\theta)$.

Доказательство. Это свойство будет доказано позднее. □

Рассмотрим пример.

Задача 3. *Пусть* $\boldsymbol{X} = (X_1, \dots, X_n)$ *— выборка из экспоненциального распределения* $\mathrm{Exp}(\theta)$. *Далее, пусть* $X_{(1)} = \min_{1 \leqslant i \leqslant n} X_i$. *Найти* $I_{X_{(1)}}(\theta)$ *и* $I_{\boldsymbol{X}}(\theta)$.

Решение. Начнём с того, что найдём распределение $X_{(1)}$. Для этого заметим, что

$$\mathsf{P}_\theta(X_{(1)} \leqslant x) = 1 - \mathsf{P}_\theta(X_{(1)} \geqslant x) = 1 - \prod_{i=1}^{n} \mathsf{P}_\theta(X_i \geqslant x) = 1 - e^{-n\theta x}.$$

Следовательно, $X_{(1)} \sim \mathrm{Exp}(n\theta)$.

Теперь можно считать информации Фишера. Начнём с $I_{\boldsymbol{X}}(\theta)$. Вспомним, что $I_{\boldsymbol{X}}(\theta) = ni(\theta)$. Тогда

$$\frac{\partial}{\partial \theta} \ln p_\theta(X_1) = \frac{\partial}{\partial \theta}(\ln \theta - \theta X_1) = \frac{1}{\theta} - X_1.$$

Тогда

$$i(\theta) = \mathsf{E}_\theta \left[\left(X_1 - \frac{1}{\theta} \right)^2 \right] = \mathsf{D}_\theta[X_1] = \frac{1}{\theta^2} \implies I_{\boldsymbol{X}}(\theta) = \frac{n}{\theta^2}.$$

Теперь посчитаем $I_{X_{(1)}}(\theta)$ аналогичным образом. Заметим, что

$$\frac{\partial}{\partial \theta} \ln p_\theta(X_{(1)}) = \frac{\partial}{\partial \theta}(\ln n + \ln \theta - n\theta X_1) = \frac{1}{\theta} - nX_1.$$

Тогда

$$I_{X_1}(\theta) = \mathsf{E}_\theta \left[\left(nX_{(1)} - \frac{1}{\theta} \right)^2 \right] = n^2 \, \mathsf{D}_\theta[X_{(1)}] = \frac{1}{\theta^2}.$$ □

2.4 Многомерное неравенство Рао-Крамера

У неравенства Рао-Крамера есть многомерный аналог. Но для его доказательства нужно одно неравенство.

Теорема 4 (Матричное неравенство Коши-Буняковского-Шварца)**.** *Пусть* $\boldsymbol{\Psi}$ *и* $\mathbf{H}$ *— случайные матрицы одного и того же размера и матрица* $\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top]$ *обратима. Тогда*[5]

$$\mathsf{E}_\theta[\boldsymbol{\Psi}\boldsymbol{\Psi}^\top] \succcurlyeq \mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top](\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top])^{-1} \, \mathsf{E}_\theta[\mathbf{H}\boldsymbol{\Psi}^\top],$$

причём равенство достигается тогда и только тогда, когда $\boldsymbol{\Psi} = \mathbf{Z}\mathbf{H}$, *где*

$$\mathbf{Z} = \mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top](\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top])^{-1}.$$

Доказательство. Для начала заметим, что для любой матрицы $\mathbf{A}$ матрица $\mathbf{A}\mathbf{A}^\top \succcurlyeq 0$. Тогда для любой матрицы $\mathbf{Z}$ выполнено следующее:

$$(\boldsymbol{\Psi} - \mathbf{Z}\mathbf{H})(\boldsymbol{\Psi} - \mathbf{Z}\mathbf{H})^\top \succcurlyeq 0$$

Возьмём матожидание:

$$\mathsf{E}_\theta[(\boldsymbol{\Psi} - \mathbf{Z}\mathbf{H})(\boldsymbol{\Psi} - \mathbf{Z}\mathbf{H})^\top] \succcurlyeq 0$$

Теперь раскроем матожидание по линейности:

$$\mathsf{E}_\theta[\boldsymbol{\Psi}\boldsymbol{\Psi}^\top] - \mathbf{Z}\,\mathsf{E}_\theta[\mathbf{H}\boldsymbol{\Psi}^\top] - \mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top]\mathbf{Z}^\top + \mathbf{Z}\,\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top]\mathbf{Z}^\top \succcurlyeq 0$$

Возьмём $\mathbf{Z} = \mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top](\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top])^{-1}$ и попробуем упростить выражение:

$$\begin{aligned}
\mathbf{Z}\,\mathsf{E}_\theta[\mathbf{H}\boldsymbol{\Psi}^\top] &= \mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top](\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top])^{-1} \, \mathsf{E}_\theta[\mathbf{H}\boldsymbol{\Psi}^\top], \\
\mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top]\mathbf{Z}^\top &= \mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top](\mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top](\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top])^{-1})^\top \\
&= \mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top]((\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top])^{-1})^\top(\mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top])^\top \\
&= \mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top](\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top])^{-1} \, \mathsf{E}_\theta[\mathbf{H}\boldsymbol{\Psi}^\top], \\
\mathbf{Z}\,\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top]\mathbf{Z}^\top &= \mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top](\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top])^{-1} \, \mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top](\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top])^{-1} \, \mathsf{E}_\theta[\mathbf{H}\boldsymbol{\Psi}^\top] \\
&= \mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top](\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top])^{-1} \, \mathsf{E}_\theta[\mathbf{H}\boldsymbol{\Psi}^\top].
\end{aligned}$$

[5] Запись $\mathbf{A} \succcurlyeq \mathbf{B}$ означает, что матрица $\mathbf{A} - \mathbf{B}$ неотрицательно определена.

Следовательно,

$$\mathsf{E}_\theta[\boldsymbol{\Psi}\boldsymbol{\Psi}^\top] - \mathsf{E}_\theta[\boldsymbol{\Psi}\mathbf{H}^\top](\mathsf{E}_\theta[\mathbf{H}\mathbf{H}^\top])^{-1}\,\mathsf{E}_\theta[\mathbf{H}\boldsymbol{\Psi}^\top] \succcurlyeq 0.$$

Для критерия равенства нужно заметить, что

$$\mathsf{E}_\theta[(\boldsymbol{\Psi} - \mathbf{Z}\mathbf{H})(\boldsymbol{\Psi} - \mathbf{Z}\mathbf{H})^\top] = \mathbf{0} \iff \boldsymbol{\Psi} = \mathbf{Z}\mathbf{H}. \qquad \square$$

Дело осталось за малым.

Теорема 5 (Многомерное неравенство Рао-Крамера)**.** *Пусть пространство параметров $\Theta \subseteq \mathbb{R}^k$, $k > 1$. Далее, пусть $\tau\colon \Theta \mapsto \mathbb{R}^k$ — это некоторая дифференцируемая по $\boldsymbol{\theta}$ функция и её якобиан равен*

$$\tau'(\boldsymbol{\theta}) = \left\| \frac{\partial(\tau(\boldsymbol{\theta}))_i}{\partial \boldsymbol{\theta}_j} \right\|_{i,j=1}^{k} \in \mathbb{R}^{k\times k}.$$

Далее, $\hat\theta(\boldsymbol{X}) \in \mathbb{R}^k$ — это несмещённая оценка для $\tau(\theta)$ с матрицей ковариаций

$$\mathsf{D}_{\boldsymbol{\theta}}[\hat\theta(\boldsymbol{X})] = \mathsf{E}_{\boldsymbol{\theta}}[(\hat\theta(\boldsymbol{X}) - \tau(\boldsymbol{\theta}))(\hat\theta(\boldsymbol{X}) - \tau(\boldsymbol{\theta}))^\top]$$

Пусть $I_{\boldsymbol{X}}(\theta)$ — информация Фишера:

$$I_{\boldsymbol{X}}(\boldsymbol{\theta}) = \left\| \mathsf{E}_{\boldsymbol{\theta}} \left[\frac{\partial \ln p_\theta(\boldsymbol{X})}{\partial \boldsymbol{\theta}_i} \frac{\partial \ln p_\theta(\boldsymbol{X})}{\partial \boldsymbol{\theta}_j} \right] \right\|_{i,j=1}^{k} \in \mathbb{R}^{k\times k}$$

Другими словами, это матрица ковариаций вклада

$$U_{\boldsymbol{\theta}}(\boldsymbol{X}) = \left(\frac{\partial}{\partial \boldsymbol{\theta}_1} \ln p_\theta(\boldsymbol{X}), \ldots, \frac{\partial}{\partial \boldsymbol{\theta}_k} \ln p_\theta(\boldsymbol{X}) \right)^\top$$

Многомерное неравенство Рао-Крамера утверждает, что в условиях регулярности

$$\mathsf{D}_{\boldsymbol{\theta}}[\hat\theta(\boldsymbol{X})] \succcurlyeq \tau'(\boldsymbol{\theta}) I_{\boldsymbol{X}}^{-1}(\boldsymbol{\theta})(\tau'(\boldsymbol{\theta}))^\top.$$

Доказательство. Рассуждая ровно так же, как и в одномерном случае, получаем, что $\mathsf{E}_{\boldsymbol{\theta}}[U_{\boldsymbol{\theta}}(\boldsymbol{X})] = \mathbf{0}$ и

$$\tau'(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}[\hat\theta(\boldsymbol{X})(U_{\boldsymbol{\theta}}(\boldsymbol{X}))^\top] = \mathsf{E}_{\boldsymbol{\theta}}[(\hat\theta(\boldsymbol{X}) - \tau(\boldsymbol{\theta}))(U_{\boldsymbol{\theta}}(\boldsymbol{X}))^\top].$$

Далее, по матричному неравенству Коши-Буняковского-Шварца получаем, что

$$\mathsf{D}_{\boldsymbol{\theta}}[\hat\theta(\boldsymbol{X})] \succcurlyeq \tau'(\boldsymbol{\theta}) I_{\boldsymbol{X}}^{-1}(\boldsymbol{\theta})(\tau'(\boldsymbol{\theta}))^\top. \qquad \square$$

3 Лекция 3

3.1 Оценки максимального правдоподобия

Сегодня мы будем обсуждать то, что преобразило статистику — метод максимального правдоподобия. Начнём с того, что введём понятие правдоподобия.

Определение 17. Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — наблюдение с неизвестным распределением $\mathsf{P} \in \{\mathsf{P}_\theta, \theta \in \Theta\}$, где $\{\mathsf{P}_\theta, \theta \in \Theta\}$ есть доминируемое семейство с плотностью $p_\theta(\boldsymbol{x})$.[6] Тогда *функцией правдоподобия* называется случайная величина $f_\theta(\boldsymbol{X}) = p_\theta(\boldsymbol{X})$.

Примечание. Если $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка, то плотность случайного вектора разбивается в произведение плотностей координат:

$$f_\theta(\boldsymbol{X}) = p_\theta(\boldsymbol{X}) = \prod_{i=1}^{n} p_\theta(X_i).$$

Теперь можно ввести и сам метод.

Определение 18. Оценкой параметра θ по методу максимального правдоподобия, или же *оценкой максимального правдоподобия*, называется

$$\hat\theta(\boldsymbol{X}) = \arg\max_{\theta\in\Theta} f_\theta(\boldsymbol{X}).$$

[6] Плотность понимается в обобщённом смысле — за подробностями обращайтесь ко второй лекции.

Данное определение уже накладывает несколько ограничений: как минимум, то, что максимум существует и он единственен. Но в дальнейшем будем считать, что они выполнены. Философия этого метода уже не так очевидна. Она состоит в том, что «мы живём в наиболее вероятном мире».

Допустим, что мы рассматриваем схему Бернулли и у нас выпало много нулей и мало единиц. Мы думаем: «Наверное, это неспроста!» Тогда, наверное, так и должно быть на самом деле — действительно, в дискретном случае функция правдоподобия есть вероятность того, что выпадает данный набор. Далее, мы подбираем θ такое, что данный набор наблюдений наиболее вероятен — коли уж он выпал, то истинным значением должно быть только то, в котором он выпадает с наибольшей вероятностью.

Рассмотрим пару примеров нахождения оценки максимального правдоподобия, для того, чтобы понять, что вообще происходит.

Задача 4. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из равномерного распределения $\mathrm{U}(0, \theta)$, $\theta > 0$. Найти оценку максимального правдоподобия параметра θ.*

Решение. Начнём с того, что распишем функцию правдоподобия:

$$f_\theta(\boldsymbol{X}) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} \frac{1}{\theta}[0 \leqslant X_i \leqslant \theta] = \frac{1}{\theta^n}[0 \leqslant X_{(1)} \leqslant X_{(n)} \leqslant \theta].$$

Теперь нам нужно максимизировать её, как функцию от θ. Заметим, что θ^{-n} есть монотонно убывающая функция, поэтому нужно взять минимальное θ такое, что функция правдоподобия не обратится в ноль. Но тогда $\hat{\theta}(\boldsymbol{X}) = X_{(n)}$. □

Задача 5. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из нормального распределения $\mathcal{N}(\mu, \sigma^2)$. Найдите оценку максимального правдоподобия параметра $\boldsymbol{\theta} = (\mu, \sigma^2)$.*

Решение. Опять же, начнём с того, что распишем функцию правдоподобия:

$$f_{\boldsymbol{\theta}}(\boldsymbol{X}) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(X_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(X_i - \mu)^2}{2\sigma^2}\right\} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right\}.$$

Нам нужно максимизировать её одновременно по μ и по θ. Но работать с экспонентой достаточно неудобно, поэтому прологарифмируем её. Это не поменяет решения, так как логарифм биективно переводит $\mathbb{R}_{++}$ в $\mathbb{R}$:

$$\ln f_{\boldsymbol{\theta}}(\boldsymbol{X}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Эта функция дифференцируема, поэтому можно достаточно легко найти всех претендентов на точки экстремума приравниваем производной к нулю:

$$0 = \frac{\partial}{\partial\mu}\ln f_{\boldsymbol{\theta}}(\boldsymbol{X}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu),$$
$$0 = \frac{\partial}{\partial\sigma^2}\ln f_{\boldsymbol{\theta}}(\boldsymbol{X}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Отсюда несложно получить, что претендентами на точки экстремума будут

$$\hat{\mu}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{\boldsymbol{X}},$$
$$\hat{\sigma}^2(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = S^2.$$

Но данная точка действительно будет являться максимумом, что несложно проверить по значениям производных. Так как она есть единственный максимум, то это глобальный максимум и $\hat{\theta}(\boldsymbol{X}) = (\overline{\boldsymbol{X}}, S^2)$. □

Примечание. Несложно показать, что S^2 — смещённая оценка. Но оценка максимального правдоподобия не обязательно несмещённая. Вообще, несмещённость — достаточно сильное свойство (и получить из оценки несмещённую может быть весьма нетривиально) в том плане, что мы хотим равенство матожиданий. Но у ОМП есть так называемая *асимптотическая несмещённость*: предел матожиданий будет именно таким, каким и должен быть.

У оценок максимального правдоподобия и функции правдоподобия есть достаточно интересные свойства. Но они требуют некоторых условий регулярности. Будем постепенно формулировать их и доказывать свойства.

(R1) Параметрическое семейство распределений $\{ \mathsf{P}_\theta \mid \theta \in \Theta \}$ — это доминируемое семейство с плотностью $p_\theta(x)$ и *различимыми распределениями*, то есть $\mathsf{P}_{\theta_0} = \mathsf{P}_{\theta_1}$ почти везде тогда и только тогда, когда $\theta_0 = \theta_1$.

(R2) $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка растущего размера из неизвестного распределения $\mathsf{P} \in \{ \mathsf{P}_\theta \mid \theta \in \Theta \}$.

(R3) $A = \{ x \colon p_\theta(x) > 0 \}$ не зависит от θ.

Теорема 6 (Экстремальное свойство правдоподобия)**.** *В условиях регулярности (R1)—(R3) для всех различных θ_0, $\theta_1 \in \Theta$*

$$\lim_{n\to\infty} \mathsf{P}_{\theta_0}(f_{\theta_0}(\boldsymbol{X}) > f_{\theta_1}(\boldsymbol{X})) = 1.$$

Доказательство. Будем считать, что мы будем работаем в A. Посмотрим, при каких условиях выполняется событие $f_{\theta_0}(\boldsymbol{X}) > f_{\theta_1}(\boldsymbol{X})$. Для этого прологарифмируем и преобразуем выражение:

$$\ln \frac{f_{\theta_1}(\boldsymbol{X})}{f_{\theta_0}(\boldsymbol{X})} < 0 \implies \frac{1}{n}\sum_{i=1}^{n} \ln \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)} < 0.$$

По усиленному закону больших чисел

$$\frac{1}{n}\sum_{i=1}^{n} \ln \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)} \xrightarrow{\mathsf{P}_{\theta_0}\text{-п.н.}} \mathsf{E}_{\theta_0}\left[\ln \frac{p_{\theta_1}(X_1)}{p_{\theta_0}(X_1)}\right].$$

Теперь докажем, что[7]

$$\mathsf{E}_{\theta_0}\left[\ln \frac{p_{\theta_1}(X_1)}{p_{\theta_0}(X_1)}\right] < 0.$$

Воспользуемся неравенством Йенсена:

$$\mathsf{E}_{\theta_0}\left[\ln \frac{p_{\theta_1}(X_1)}{p_{\theta_0}(X_1)}\right] \leqslant \ln \mathsf{E}_{\theta_0}\left[\frac{p_{\theta_1}(X_1)}{p_{\theta_0}(X_1)}\right] = \ln \int_A \frac{p_{\theta_1}(X_1)}{p_{\theta_0}(X_1)} p_{\theta_0}(X_1)\mu(\mathrm{d}x) = \ln \mathsf{E}_{\theta_1}[1] = 0.$$

Но почему оно не равно нулю? Предположим, что это так:

$$\mathsf{E}_{\theta_0}\left[\ln \frac{p_{\theta_1}(X_1)}{p_{\theta_0}(X_1)}\right] = 0.$$

Но в таком случае можно воспользоваться критерием равенства для неравенства Йенсена: $\phi(\mathsf{E}[\xi]) = \mathsf{E}[\phi(\xi)]$ тогда и только тогда, когда ϕ линейна почти везде. Но $\ln(x)$ нелинейна. Тогда получаем, что аргумент должен быть равен единице почти везде: $\mu(\{ x \colon p_{\theta_0}(x) = p_{\theta_1}(x) \}) = 1$. Но это означает, что $\theta_0 = \theta_1$, что противоречит условию.

В итоге получаем, что

$$\lim_{n\to\infty} \mathsf{P}_{\theta_0}(f_{\theta_0}(\boldsymbol{X}) > f_{\theta_1}(\boldsymbol{X})) = \mathsf{P}_{\theta_0}\left(\mathsf{E}_{\theta_0}\left[\ln \frac{p_{\theta_1}(X_1)}{p_{\theta_0}(X_1)}\right] < 0\right) = 1. \qquad \square$$

Следствие (Состоятельность оценки максимального правдоподобия)**.** Если Θ конечно, что оценка максимального правдоподобия состоятельна.

Доказательство. Пусть $\hat{\theta}_n(\boldsymbol{X})$ — это оценка максимального правдоподобия. Тогда по экстремальному свойству правдоподобия для любого $\theta_0 \in \Theta$

$$\lim_{n\to\infty} \mathsf{P}_{\theta_0}(\hat{\theta}_n(\boldsymbol{X}) = \theta_0) = \lim_{n\to\infty} \mathsf{P}_\theta(\forall \theta \neq \theta_0 f_{\theta_0}(\boldsymbol{X}) > f_\theta(\boldsymbol{X})) = 1. \qquad \square$$

Введём ещё два условия регулярности:

(R4) Θ есть открытый интервал на $\mathbb{R}$.

[7] Небольшое примечание: если добавить минус к этому матожиданию, то получим широко известную *дивергенцию Кульбака-Лейблера*. По сути, мы доказываем то, что она неотрицательна и то, что она равна нулю тогда и только тогда, когда плотности равны почти везде.

(R5) $p_\theta(x)$ непрерывно дифференцируема по θ для всех $x \in A$.

Теперь можно доказать хорошее свойство, плотно связанное с оценками максимального правдоподобия.

Теорема 7 (Состоятельность решения уравнения правдоподобия)**.** *В условиях регулярности (R1)—(R5)* уравнение правдоподобия

$$\frac{\partial}{\partial\theta}\ln f_\theta(\boldsymbol{X}) = 0$$

с вероятностью, стремящейся к 1, имеет решение, которое сходится по вероятности к истинному значению параметра.

Доказательство. Пусть θ_0 есть истинное значение параметра. Возьмём $\delta > 0$ такое, что $[\theta_0-\delta, \theta_0+\delta] \subset \Theta$ (это возможно из-за открытости Θ). Далее, введём следующее событие:

$$A_n = \{\, f_{\theta_0}(\boldsymbol{X}) > f_{\theta_0+\delta}(\boldsymbol{X}), f_{\theta_0}(\boldsymbol{X}) > f_{\theta_0-\delta}(\boldsymbol{X})\,\}$$

Тогда по экстремальному свойству правдоподобия

$$\lim_{n\to\infty} \mathsf{P}_{\theta_0}(A_n) = 1.$$

Что можно сказать, если выполнено A_n? Так как по (R6) производная логарифма функции правдоподобия непрерывна по θ, то на отрезке $[\theta_0 - \delta, \theta_0 + \delta]$ будет точка, в которой возрастание заменяется убыванием. Следовательно, на нём будет хотя бы один корень уравнения правдоподобия. Допустим, что на этом отрезке есть несколько корней (не обязательно конечное число). Пусть $\tilde\theta(\boldsymbol{X})$ — ближайший к θ_0 корень. Это возможно, так как предел корней тоже является корнем (так как производная непрерывна). Тогда оказывается, что для всех $\varepsilon > 0$

$$\lim_{n\to\infty} \mathsf{P}_{\theta_0}(|\tilde\theta(\boldsymbol{X}) - \theta_0| \leqslant \varepsilon) = 1.$$

Почему это так? Зафиксируем ε и заметим, что рассуждения выше легальны для $\delta = \varepsilon$. Следовательно, с вероятностью, стремящейся к 1, на отрезке $[\theta_0 - \varepsilon, \theta_0 + \varepsilon]$ будет корень. Но $\tilde\theta(\boldsymbol{X})$ — ближайший к θ_0 корень. Тогда он лежит в этом отрезке. □

Вот доказали мы эту теорему. Но она больше напоминает решето, из которого вытекают проблемы. Почему?

1. Например, корней уравнения правдоподобия может быть несколько. В доказательстве мы выбираем ближайший из них к истинному значению. Но как его выбрать?
2. Даже если мы его найдём, то он зависит от истинного значения, то есть вообще не является оценкой.
3. Почему $\tilde\theta(\boldsymbol{X})$ есть точка максимума? Мы сказали, что это корень уравнения, но он не обязательно даёт максимум — может оказаться, что это точка минимума или же точка перегиба (если корней несколько).
4. Корень существует не всегда, а только с большой вероятностью.

Впрочем, если сказать, что уравнение правдоподобия имеет только один корень, то всё достаточно неплохо. Четвёртый и второй вопросы отпадают сразу же, да и первый тоже. Остался третий, но он тоже исправляется:

Теорема 8. *Если для всех $\boldsymbol{X} = (X_1, \ldots, X_n)$ уравнение правдоподобия имеет единственный корень $\hat\theta(\boldsymbol{X})$, то с вероятностью, стремящейся к 1, $\hat\theta(\boldsymbol{X})$ будет оценкой максимального правдоподобия и онценка максимального правдоподобия будет состоятельной.*

Доказательство. По сути, доказательство повторяет то, что было сказано выше. Пусть θ_0 — истинное значение параметра. Снова возьмём $\delta > 0$ такое, что $[\theta_0 - \delta, \theta_0 + \delta] \subset \Theta$ и заметим, что если

$$\lim_{n\to\infty} \mathsf{P}_{\theta_0}(A_n) = 1, \text{ где } A_n = \{\, f_{\theta_0}(\boldsymbol{X}) > f_{\theta_0+\delta}(\boldsymbol{X}), f_{\theta_0}(\boldsymbol{X}) > f_{\theta_0-\delta}(\boldsymbol{X})\,\}.$$

Однако если выполнено A_n, то внутри $[\theta_0 - \delta, \theta_0 + \delta]$ есть точка локального максимума. Как известно, в ней производная равна нулю, и, следовательно, она будет корнем уравнения правдоподобия. Но тогда эта точка есть $\hat\theta(\boldsymbol{X})$. Далее, это должен быть глобальный максимум, так как иначе найдётся локальный минимум, что противоречит единственности корня. Следовательно, если выполнено A_n, то $\hat\theta(\boldsymbol{X})$ есть оценка максимального правдоподобия. Тогда

$$\lim_{n\to\infty} \mathsf{P}_{\theta_0}(\hat\theta(\boldsymbol{X}) = \text{ОМП}) = 1.$$

Но, как известно, $\hat{\theta}(\boldsymbol{X})$ есть ближайший к θ_0 корень. Тогда для всех $\varepsilon > 0$

$$\lim_{n\to\infty} \mathsf{P}_{\theta_0}(|\hat{\theta}(\boldsymbol{X}) - \theta_0| \leqslant \varepsilon) = 1.$$

Тем самым получаем, что и ОМП будет состоятельной оценкой параметра θ. □

Поехали дальше. Эти условия регулярности уже далеко не так очевидны на первый взгляд.

(R6) $p_\theta(x)$ трижды непрерывно дифференцируема по θ для всех $x \in A$.

(R7) Интеграл

$$\int_A p_\theta(x)\mu(\mathrm{d}x)$$

можно дважды дифференцировать под знаком интеграла.

(R8) Для всех $\theta \in \Theta$

$$0 < i(\theta) = \mathsf{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln p_\theta(X_1)\right)^2\right] < +\infty.$$

(R9) Для любого $\theta_0 \in \Theta$ существует $\delta > 0$ и функция $M(x)$ такая, что для всех $\theta \in [\theta_0 - \delta, \theta_0 + \delta]$

$$\left|\frac{\partial^3}{\partial\theta^3}\ln p_\theta(x)\right| \leqslant M(x), \text{ причём } \mathsf{E}_{\theta_0}[M(X_1)] < +\infty.$$

Теорема 9. *В условиях регулярности (R1)—(R9) любая состоятельная последовательность* $\{\,\hat{\theta}_n(\boldsymbol{X}) \mid n \in \mathbb{N}\,\}$ *корней уравнения правдоподобия удовлетворяет свойству асимптотической нормальности: для всех* $\theta_0 \in \Theta$

$$\sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta_0) \xrightarrow{d_{\theta_0}} \mathcal{N}\left(0, \frac{1}{i(\theta_0)}\right).$$

Доказательство. Обозначим $\mathcal{L}(\boldsymbol{X},\theta) = \ln f_\theta(\boldsymbol{X})$ и будем говорить, что $\mathcal{L}^{(n)}(\boldsymbol{X},\theta)$ есть n-я частная производная $\mathcal{L}(\boldsymbol{X},\theta)$ по θ. Далее, пусть θ_0 есть истинное значение параметра, то есть $\hat{\theta}_n(\boldsymbol{X})$ сходится к θ_0 по вероятности P_{θ_0}. Разложим $\mathcal{L}'(\boldsymbol{X},\theta)$ в ряд Тейлора в точке θ_0:

$$\mathcal{L}'(\boldsymbol{X},\theta) = \mathcal{L}'(\boldsymbol{X},\theta_0) + \mathcal{L}''(\boldsymbol{X},\theta_0)(\theta - \theta_0) + \frac{1}{2}\mathcal{L}'''(\boldsymbol{X},\tilde{\theta})(\theta - \theta_0)^2,$$

где $\tilde{\theta}$ находится между θ и θ_0. Теперь подставим $\theta = \hat{\theta}_n(\boldsymbol{X})$:

$$\mathcal{L}'(\boldsymbol{X},\hat{\theta}_n(\boldsymbol{X})) = \mathcal{L}'(\boldsymbol{X},\theta_0) + \mathcal{L}''(\boldsymbol{X},\theta_0)(\hat{\theta}_n(\boldsymbol{X}) - \theta_0) + \frac{1}{2}\mathcal{L}'''(\boldsymbol{X},\tilde{\theta}_n)(\hat{\theta}_n(\boldsymbol{X}) - \theta_0)^2,$$

где $\tilde{\theta}_n$ находится между $\hat{\theta}_n(\boldsymbol{X})$ и θ_0. Теперь вспомним, что $\hat{\theta}_n(\boldsymbol{X})$ есть корень уравнения правдоподобия. Тогда $\mathcal{L}'(\boldsymbol{X},\hat{\theta}_n(\boldsymbol{X})) = 0$. Теперь выразим $\sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta_0)$ из данного равенства следующим образом:

$$\sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta_0) = \frac{-\sqrt{n}\mathcal{L}'(\boldsymbol{X},\theta_0)}{\mathcal{L}''(\boldsymbol{X},\theta_0) + \frac{1}{2}\mathcal{L}'''(\boldsymbol{X},\tilde{\theta})(\hat{\theta}_n(\boldsymbol{X}) - \theta_0)} = \frac{-\frac{1}{\sqrt{n}}\mathcal{L}'(\boldsymbol{X},\theta_0)}{\frac{1}{n}\mathcal{L}''(\boldsymbol{X},\theta_0) + \frac{1}{2n}\mathcal{L}'''(\boldsymbol{X},\tilde{\theta})(\hat{\theta}_n(\boldsymbol{X}) - \theta_0)}$$

Теперь рассмотрим это выражение по частям.

- Начнём с числителя. Заметим, что

$$-\frac{1}{\sqrt{n}}\mathcal{L}'(\boldsymbol{X},\theta_0) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\theta}\ln p_\theta(X_i)\bigg|_{\theta=\theta_0} = -\sqrt{n}\sum_{i=1}^n U_{\theta_0}(X_i)$$

  Как известно, возможность дифференцирования под знаком интеграла показывает, что $\mathsf{E}_{\theta_0}[U_{\theta_0}(X_1)] = 0$. Тогда по центральной предельной теореме

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^n U_{\theta_0}(X_i) \xrightarrow{d_{\theta_0}} -\mathcal{N}\left(0, \mathsf{D}_{\theta_0}[U_{\theta_0}(X_1)]\right) = -\mathcal{N}(0, i(\theta_0)).$$

- По усиленному закону больших чисел

$$\frac{1}{n}\mathcal{L}''(\boldsymbol{X},\theta_0) \xrightarrow{\mathsf{P}_{\theta_0}\text{-п.н.}} \mathsf{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\ln p_\theta(X_1)\Big|_{\theta=\theta_0}\right].$$

Докажем, что

$$-\mathsf{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln p_\theta(X_1)\right] = i(\theta).$$

Доказательство. Заметим, что

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\ln p_\theta(x) &= \frac{\partial}{\partial\theta}\left(\frac{1}{p_\theta(x)}\frac{\partial p_\theta(x)}{\partial\theta}\right) = -\frac{1}{(p_\theta(x))^2}\left(\frac{\partial p_\theta(x)}{\partial\theta}\right)^2 + \frac{1}{p_\theta(x)}\frac{\partial^2 p_\theta(x)}{\partial\theta^2}\\ &= -\left(\frac{\partial}{\partial\theta}\ln p_\theta(x)\right)^2 + \frac{1}{p_\theta(x)}\frac{\partial^2 p_\theta(x)}{\partial\theta^2}.\end{aligned}$$

Теперь возьмём матожидание и воспользуемся тем, что мы можем два раза дифференцировать под знаком интеграла:

$$\begin{aligned}\mathsf{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln p_\theta(X_1)\right] &= \int_A \frac{\partial^2\ln p_\theta(x)}{\partial\theta^2}p_\theta(x)\mu(\mathrm{d}x)\\ &= -\int_A\left(\frac{\partial\ln p_\theta(x)}{\partial\theta}\right)^2 p_\theta(x)\mu(\mathrm{d}x) + \int_A\frac{\partial^2 p_\theta(x)}{\partial\theta^2}\mu(\mathrm{d}x)\\ &= -i(\theta) + \frac{\partial^2}{\partial\theta^2}\int_A p_\theta(x)\mu(\mathrm{d}x) = -i(\theta)\end{aligned}$$ □

Отсюда получаем, что

$$\frac{1}{n}\mathcal{L}''(\boldsymbol{X},\theta_0) \xrightarrow{\mathsf{P}_{\theta_0}\text{-п.н.}} -i(\theta).$$

- Теперь перейдём к третьему члену и покажем, что он стремится к нулю по вероятности. Заметим, что по условию $\hat\theta(\boldsymbol{X})$ сходится к θ_0 по вероятности. Из этого можно сделать вывод, что $\tilde\theta_n$ тоже стремится к θ_0 по вероятности. Далее, по (R9)

$$\left|\frac{1}{n}\mathcal{L}'''(\boldsymbol{X},\theta)\right| \leqslant \frac{1}{n}\sum_{i=1}^n M(X_i),$$

что имеет конечное матожидание. Тогда получаем произведение ограниченной случайной величины на нечто, что сходится к нулю по вероятности. Это сходится к нулю по вероятности.

Комбинируя вышесказанное, по лемме Слуцкого получаем, что

$$\sqrt{n}(\hat\theta_n(\boldsymbol{X}) - \theta_0) \xrightarrow{d_{\theta_0}} \frac{1}{i(\theta_0)}\mathcal{N}(0, i(\theta_0)) = \mathcal{N}\left(0, \frac{1}{i(\theta_0)}\right).$$ □

И получаем приятное с практической точки зрения свойство:

Следствие. Если в условиях теоремы для всех $\boldsymbol{X} = (X_1,\ldots,X_n)$ существует единственное решение уравнения правдоподобия, то оно является оценкой максимального правдоподобия, причём ОМП будет асимптотически нормальной оценкой параметра θ с асимптотической дисперсией $i^{-1}(\theta)$.

Оказывается, что можно предложить нижнюю границу не только для обычной дисперсии (что даёт неравенство Рао-Крамера), но и для асимптотической дисперсии. Этот результат называется *теоремой Бахадура*. Сформулируем его:

Теорема 10 (Бахадур)**.** *Если в условиях регулярности (R1)—(R9) оценка $\hat\theta(\boldsymbol{X})$ является асимптотически нормальной оценкой параметра θ с асимптотической дисперсией $\sigma^2(\theta)$, то $\sigma^2(\theta) \geqslant i^{-1}(\theta)$ почти везде (то есть неравенство нарушается только на множестве лебеговой меры 0).*

Ну и сразу же пример, в котором ограничение выполнено не везде, но почти везде.

Задача 6. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из нормального распределения $\mathcal{N}(\theta, 1)$. Введём следующую оценку параметра θ:*

$$\hat{\theta}_n(\boldsymbol{X}) = \begin{cases} \overline{\boldsymbol{X}}, & |\overline{\boldsymbol{X}}| \geqslant n^{-1/4} \\ \overline{\boldsymbol{X}}/2, & |\overline{\boldsymbol{X}}| < n^{-1/4} \end{cases}$$

Найдите асимптотическую дисперсию $\sigma^2(\theta)$ и сравните её с обратной информацией Фишера $i^{-1}(\theta)$ одного элемента.

Решение. Как известно, усиленный закон больших чисел имеет скорость сходимости порядка $\mathcal{O}(n^{-1/2})$ с вероятностью, стремящейся к 1. Тогда если $\theta \neq 0$, то

$$\lim_{n\to\infty} \mathsf{P}_\theta(\hat{\theta}_n(\boldsymbol{X}) = \overline{\boldsymbol{X}}) = 1$$

Теперь рассмотрим распределение $\sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta)$. Для этого воспользуемся формулой полной вероятности:

$$\begin{aligned} \mathsf{P}(\sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta) \leqslant x) &= \mathsf{P}(\sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta) \leqslant x \mid \hat{\theta}_n(\boldsymbol{X}) = \overline{\boldsymbol{X}})\mathsf{P}(\hat{\theta}_n(\boldsymbol{X}) = \overline{\boldsymbol{X}}) \\ &\quad + \mathsf{P}(\sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta) \leqslant x \mid \hat{\theta}_n(\boldsymbol{X}) = \overline{\boldsymbol{X}}/2)\mathsf{P}(\hat{\theta}_n(\boldsymbol{X}) = \overline{\boldsymbol{X}}/2) \\ &= \mathsf{P}(\sqrt{n}(\overline{\boldsymbol{X}} - \theta) \leqslant x)\mathsf{P}(\hat{\theta}_n(\boldsymbol{X}) = \overline{\boldsymbol{X}}) \\ &\quad + \mathsf{P}(\sqrt{n}(\overline{\boldsymbol{X}}/2 - \theta) \leqslant x)\mathsf{P}(\hat{\theta}_n(\boldsymbol{X}) = \overline{\boldsymbol{X}}/2) \end{aligned}$$

Теперь устремим n к бесконечности. Согласно результату выше и центральной предельной теореме получаем, что

$$\lim_{n\to\infty} \mathsf{P}(\sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta) \leqslant x) = \mathsf{P}(\mathcal{N}(0,1) \leqslant x) = \Phi(x) \implies \sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta) \xrightarrow{d_\theta} \mathcal{N}(0,1).$$

Далее скажем, что пусть $\theta = 0$. В таком случае утверждение о скорости сходимости усиленного закона больших чисел говорит, что

$$\lim_{n\to\infty} \mathsf{P}_\theta(\hat{\theta}_n(\boldsymbol{X}) = \overline{\boldsymbol{X}}/2) = 1.$$

Применяя формулу полной вероятности, получаем, что

$$\mathsf{P}(\sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta) \leqslant x) = \mathsf{P}(\sqrt{n}\,\overline{\boldsymbol{X}} \leqslant x)\mathsf{P}(\hat{\theta}_n(\boldsymbol{X}) = \overline{\boldsymbol{X}}) + \mathsf{P}(\sqrt{n}\,\overline{\boldsymbol{X}} \leqslant 2x)\mathsf{P}(\hat{\theta}_n(\boldsymbol{X}) = \overline{\boldsymbol{X}}/2).$$

Следовательно,

$$\lim_{n\to\infty} \mathsf{P}(\sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta) \leqslant x) = \Phi(2x) \implies \sqrt{n}(\hat{\theta}_n(\boldsymbol{X}) - \theta) \xrightarrow{d_\theta} \mathcal{N}(0, 1/4).$$

Однако информация Фишера равна

$$i(\theta) = \mathsf{E}_\theta[(X_1 - \theta)^2] = \mathsf{D}_\theta[X_1] = 1.$$

Тогда мы получаем, что везде, кроме нуля, $\sigma^2(\theta) \geqslant i^{-1}(\theta)$. □

4 Лекция 4

4.1 Теорема Бахадура и следствия из неё

На прошлой лекции мы сформулировали теорему Бахадура. Сейчас мы попытаемся доказать её. Но для начала потребуем выполнение следующих условий регулярности:

(R1) $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка растущего размера из неизвестного распределения $\mathsf{P} \in \{\mathsf{P}_\theta, \theta \in \Theta\}$, где $\{\mathsf{P}_\theta, \theta \in \Theta\}$ — доминируемое семейство с плотностью $p_\theta(x)$ по мере μ.

(R2) Будем считать, что θ есть одномерный параметр и Θ есть открытый интервал на $\mathbb{R}$.

(R3) Множество $A = \{x\colon p_\theta(x) > 0\}$ не зависит от θ.

(R4) Функция $\ell(\theta, x) = \ln p_\theta(x)$ дважды непрерывно дифференцируема по θ для всех $x \in A$. Для удобства будем считать, что

$$\ell^{(n)}(\theta, x) = \frac{\partial^n}{\partial \theta^n}\ell(\theta, x).$$

(R5) Для всех $\theta \in \Theta$ $\mathsf{E}_\theta[\ell'(\theta, X_1)] = 0$ и $0 < \mathsf{E}_\theta[(\ell'(\theta, X_1))^2] = i(\theta) < +\infty$. Стоит заметить, что $-\mathsf{E}_\theta[\ell''(\theta, X_1)] = i(\theta)$.

(R6) Для всех $\theta_0 \in \Theta$ существует $\delta > 0$ и функция $M(x)$ такая, что для всех $\theta \in [\theta_0 - \delta, \theta_0 + \delta]$ $|\ell'(\theta, x)| \leqslant M(x)$ и $\mathsf{E}_\theta[M(X_1)] < +\infty$.

Теорема 11. *Пусть в условиях регулярности (R1)—(R6) для последовательности оценок $t_n(\boldsymbol{X})$ параметра θ выполнено условие асимптотической нормальности: для всех $\theta \in \Theta$*

$$\sqrt{n}(t_n(\boldsymbol{X}) - \theta) \xrightarrow{d_\theta} \mathcal{N}(0, \sigma^2(\theta)).$$

Далее, зафиксируем $\theta_0 \in \Theta$ и положим $\theta_n = \theta_0 + n^{-1/2}$. Если

$$\varlimsup_{n\to\infty} \mathsf{P}_{\theta_n}(t_n(\boldsymbol{X}) \leqslant \theta_n) \leqslant \frac{1}{2},$$

то $\sigma^2(\theta_0) \geqslant i^{-1}(\theta)$.

Для доказательства данной теоремы нужно сперва доказать несколько вспомогательных утверждений и ввести несколько обозначений. Начнём с того, что введём *логарифмическую функцию правдоподобия*:

$$\mathcal{L}(\theta, \boldsymbol{X}) = \ln f_\theta(\boldsymbol{X}) = \sum_{i=1}^n \ell(\theta, X_i).$$

Далее введём следующую случайную величину:

$$T_n = \frac{1}{\sqrt{i(\theta_0)}} \left(\mathcal{L}(\theta_n, \boldsymbol{X}) - \mathcal{L}(\theta_0, \boldsymbol{X}) + \frac{i(\theta_0)}{2} \right).$$

Теперь покажем, что для случайной величины T_n выполнено одно интересное свойство:

Лемма. *При фиксированном θ_0 последовательность случайных величин T_n сходится по распределению к стандартному нормальному распределению: $T_n \xrightarrow{d_{\theta_0}} \mathcal{N}(0, 1)$.*

Доказательство. Воспользуемся дважды непрерывной дифференцируемостью $\ell(\theta, x)$ и разложим $\mathcal{L}(\theta_n, \boldsymbol{X})$ по θ в окрестности θ_0:

$$\begin{aligned}\mathcal{L}(\theta_n, \boldsymbol{X}) &= \mathcal{L}(\theta_0, \boldsymbol{X}) + \mathcal{L}'(\theta_0, \boldsymbol{X})(\theta_n - \theta_0) + \frac{1}{2}\mathcal{L}''(\tilde{\theta}_n, \boldsymbol{X})(\theta_n - \theta_0)^2 \\ &= \mathcal{L}(\theta_0, \boldsymbol{X}) + \frac{1}{\sqrt{n}}\mathcal{L}'(\theta_0, \boldsymbol{X}) + \frac{1}{2n}\mathcal{L}''(\tilde{\theta}_n, \boldsymbol{X}),\end{aligned}$$

где $\tilde{\theta}_n \in (\theta_0, \theta_n)$. Далее, введём следующую случайную величину:

$$\xi_n = \frac{1}{2n} \left(\mathcal{L}''(\tilde{\theta}_n, \boldsymbol{X}) - \mathcal{L}''(\theta_0, \boldsymbol{X}) \right).$$

Докажем, что эта последовательность стремится к нулю почти наверное: $\xi_n \xrightarrow{\mathsf{P}_{\theta_0}\text{-п.н.}} 0$. Для этого введём следующую функцию:

$$A(x, \delta) = \sup_{\theta \in [\theta_0 - \delta, \theta_0 + \delta]} |\ell''(\theta, x) - \ell''(\theta_0, x)|.$$

Далее, положим $m(\delta) = \mathsf{E}_{\theta_0}[A(X_1, \delta)]$. Заметим, что если $\delta_1 > \delta_2$, то $A(x, \delta_1) \geqslant A(x, \delta_2)$ и при $\delta \to 0$ $A(x, \delta) \to 0$. Тогда $A(x, \delta)$ монотонно убывает к нулю при $\delta \to 0$. Далее, воспользуемся свойством регулярности (R6):

$$A(x, \delta) = \sup_{\theta \in [\theta_0 - \delta, \theta_0 + \delta]} |\ell''(\theta, x) - \ell''(\theta_0, x)| \leqslant \sup_{\theta \in [\theta_0 - \delta, \theta_0 + \delta]} 2M(x) = 2M(x).$$

Тем самым можно применить теорему Лебега о мажорируемой сходимости. Из неё следует, что $m(\delta) \to 0$ при $\delta \to 0$. Далее, заметим, что если взять $\delta > n^{-1/2}$, то

$$|\xi_n| = \frac{1}{2n} \left| \sum_{i=1}^n (\ell''(\tilde{\theta}_n, X_i) - \ell''(\theta_0, X_i)) \right| \leqslant \frac{1}{2n} \sum_{i=1}^n |\ell''(\tilde{\theta}_n, X_i) - \ell''(\theta_0, X_i)| \leqslant \frac{1}{2n} \sum_{i=1}^n A(X_i, \delta).$$

Однако по усиленному закону больших чисел

$$\frac{1}{2n}\sum_{i=1}^{n} A(X_i,\delta) \xrightarrow{\mathsf{P}_{\theta_0}\text{-п.н.}} \frac{1}{2}m(\delta).$$

Из вышесказанного следует, что для любого $\delta > 0$

$$\overline{\lim_{n\to\infty}}\,|\xi_n| \leqslant \frac{1}{2}m(\delta) \quad \mathsf{P}_{\theta_0}\text{-п.н.}$$

Следовательно, $\xi_n \xrightarrow{\mathsf{P}_{\theta_0}\text{-п.н.}} 0$.

Теперь преобразуем T_n следующим образом:

$$\begin{aligned} T_n &= \frac{1}{\sqrt{i(\theta_0)}}\left(\mathcal{L}(\theta_n,\boldsymbol{X}) - \mathcal{L}(\theta_0,\boldsymbol{X}) + \frac{i(\theta_0)}{2}\right) \\ &= \frac{1}{\sqrt{i(\theta_0)}}\left(\mathcal{L}(\theta_0,\boldsymbol{X}) + \frac{1}{\sqrt{n}}\mathcal{L}'(\theta_0,\boldsymbol{X}) + \frac{1}{2n}\mathcal{L}''(\tilde{\theta}_n,\boldsymbol{X}) - \mathcal{L}(\theta_0,\boldsymbol{X}) + \frac{i(\theta_0)}{2}\right) \\ &= \frac{1}{\sqrt{i(\theta_0)}}\left(\frac{1}{\sqrt{n}}\mathcal{L}'(\theta_0,\boldsymbol{X}) + \xi_n + \frac{1}{2n}\mathcal{L}''(\theta_0,\boldsymbol{X}) + \frac{i(\theta_0)}{2}\right). \end{aligned}$$

Осталось показать, что это стремится по распределению туда, куда нужно.

- Начнём с члена $n^{-1/2}\mathcal{L}'(\theta_0,\boldsymbol{X})$. Заметим, что из условия регулярности (R5) и центральной предельной теоремы следует, что

$$\frac{1}{\sqrt{n}}\mathcal{L}'(\theta_0,\boldsymbol{X}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'(\theta_0,X_i) \xrightarrow{d_{\theta_0}} \mathcal{N}(0,i(\theta_0)).$$

- Теперь рассмотрим член $(2n)^{-1}\mathcal{L}'(\theta_0,\boldsymbol{X})$. По усиленному закону больших чисел

$$\frac{1}{2n}\mathcal{L}''(\theta_0,\boldsymbol{X}) = \frac{1}{2n}\sum_{i=1}^{n}\ell''(\theta_0,X_i) \xrightarrow{\mathsf{P}_{\theta_0}\text{-п.н.}} \frac{1}{2}\,\mathsf{E}_{\theta_0}[\ell''(\theta_0,X_1)] = -\frac{i(\theta_0)}{2}.$$

- Из пунктов выше следует, что

$$\xi_n + \frac{1}{2n}\mathcal{L}''(\theta_0,\boldsymbol{X}) + \frac{i(\theta_0)}{2} \xrightarrow{\mathsf{P}_{\theta_0}\text{-п.н.}} 0.$$

В итоге по лемме Слуцкого получаем, что

$$T_n \xrightarrow{d_{\theta_0}} \frac{1}{\sqrt{i(\theta_0)}}\mathcal{N}(0,i(\theta_0)) = \mathcal{N}(0,1). \qquad \square$$

Для доказательства следующей леммы нам понадобится достаточно известный факт, связанный с равномерной интегрируемостью.[8] Но перед этим докажем одно простое утверждение:

Лемма (Абсолютная непрерывность математического ожидания)**.** *Пусть ξ — интегрируемая случайная величина. Тогда для любого $\varepsilon > 0$ существует $\delta > 0$ такое, что для любого события F такого, что $\mathsf{P}(F) < \delta$, выполнено $\mathsf{E}[|\xi|[F]] < \varepsilon$.*

Доказательство. Пусть это не так. Тогда существует $\varepsilon > 0$ и последовательность событий F_n такая, что $\mathsf{P}(F_n) < 2^{-n}$ и $\mathsf{E}[|\xi|[F]] \geqslant \varepsilon$. Но $|\xi|[F_n] \to 0$ почти наверное и $|\xi|[F_n] \leqslant |\xi|$. Следовательно, по теореме Лебега о мажорируемой сходимости $\varepsilon \leqslant \mathsf{E}[|\xi|[F_n]] \to 0$. Противоречие. $\square$

Теорема 12. *Пусть ξ_n — последовательность неотрицательных интегрируемых случайных величин таких, что ξ_n сходится к ξ по распределению. Тогда $\mathsf{E}[\xi_n] \to \mathsf{E}[\xi]$ тогда и только тогда, когда случайные величины ξ_n равномерно интегрируемы, то есть*

$$\lim_{c\to+\infty}\sup_{n\in\mathbb{N}}\mathsf{E}\,[\xi_n[\xi_n \geqslant c]] = 0.$$

[8] Доказательство этого факта — дополнительный материал, который на лекции был оставлен в качестве упражнения.

Доказательство. Начнём с доказательства в прямую сторону: то есть докажем, что из сходимости матожиданий следует равномерная интегрируемость. Заметим, что $\mathsf{E}[|\xi_n - \xi|] \to 0$, так как случайные величины неотрицательны (и, следовательно, сходятся в L_1). Теперь нам нужно показать, что для любого $\varepsilon > 0$ можно подобрать $c > 0$ такое, что для всех $n \in \mathbb{N}$ $\mathsf{E}[\xi_n[\xi_n \geqslant c]] < 2\varepsilon$.

Хотелось бы сказать, что про теореме Лебега о мажорируемой сходимости несложно подобрать c для каждой случайной величины ξ_n по отдельности, после чего взять супремум. Но он не обязательно конечен. Поэтому будем действовать так: пусть для всех $n > N$ $\mathsf{E}[|\xi_n - \xi|] < \varepsilon$. Для всех $n \leqslant N$ подберём c_n такие, что $\mathsf{E}[\xi_n[\xi_n \geqslant c_n]] < 2\varepsilon$, после чего возьмём максимальное из них. Поэтому конечное число членов спереди нас не пугает и достаточно рассмотреть только случай $n > N$. Заметим, что

$$\mathsf{E}[\xi_n[\xi_n > c]] \leqslant \mathsf{E}[|\xi_n - \xi|[\xi_n > c]] + \mathsf{E}[\xi[\xi_n > c]].$$

С первым интегралом всё понятно: $\mathsf{E}[|\xi_n - \xi|[\xi_n \geqslant c]] \leqslant \mathsf{E}[|\xi_n - \xi|] < \varepsilon$. Для второго заметим, что

$$\mathsf{P}(\xi_n \geqslant c) \leqslant \frac{\mathsf{E}[\xi_n]}{c} \leqslant \frac{\sup_{n>N} \mathsf{E}[\xi_n]}{c} \leqslant \frac{\sup_{n>N}(\mathsf{E}[\xi] + \mathsf{E}[|\xi_n - \xi|])}{c} \leqslant \frac{\mathsf{E}[\xi] + \varepsilon}{c} \xrightarrow[c\to\infty]{} 0.$$

Следовательно, можно найти c такое, что можно воспользоваться абсолютной непрерывностью матожидания и получить, что $\mathsf{E}[\xi[\xi_n \geqslant c]] < \varepsilon$. Тем самым получается равномерная интегрируемость.

Теперь докажем в другую сторону. Пусть последовательность ξ_n равномерно интегрируема. Введём следующую функцию:

$$f_c(x) = \begin{cases} x, & x < c \\ 0, & x \geqslant c+1 \\ c - c(x-c), & c \leqslant x < c+1 \end{cases}$$

По сути, это отсечённая функция $y = x$, дополненная линейным куском до непрерывности. Далее, заметим, что по неравенству треугольника

$$\begin{aligned} \mathsf{E}[|\xi_n - \xi|] &= \mathsf{E}[|(\xi_n - f_c(\xi_n)) - (\xi - f_c(\xi)) + (f_c(\xi_n) - f_c(\xi))|] \\ &\leqslant \mathsf{E}[|\xi_n - f_c(\xi_n)|] + \mathsf{E}[|\xi - f_c(\xi)|] + \mathsf{E}[|f_c(\xi_n) - f_c(\xi)|] \end{aligned}$$

Теперь зафиксируем произвольный $\varepsilon > 0$ и будем ограничивать все члены по отдельности:

- Начнём с того, что при $x \geqslant c$ $f_c(x) \leqslant x$. Поэтому $\mathsf{E}[|\xi_n - f_c(\xi_n)|] \leqslant \mathsf{E}[\xi_n[\xi_n > c]]$. Но равномерная интегрируемость позволяет выбрать $c_1 = c_1(\varepsilon)$ такое, что $\mathsf{E}[\xi_n[\xi_n > c]] < \varepsilon/3$ для всех n.
- Со вторым членом всё аналогично: выбираем $c_2 = c_2(\varepsilon)$ такое, что $\mathsf{E}[\xi[\xi > c]] < \varepsilon/3$ для всех n.
- Для третьего достаточно вспомнить, что ξ_n сходятся к ξ по распределению и альтернативное определение сходимости по распределению. Из него следует, что $\mathsf{E}[|f_c(\xi_n) - f_c(\xi)|] < \varepsilon/3$ для любых c и достаточно больших n.

Комбинируя эти три оценки, получим сходимость в L_1, из которой следует сходимость матожиданий. □

Теперь можно вернуться к нашей теореме.

Лемма. *Пусть $\Phi(x)$ — функция распределения стандартной нормальной случайной величины. Тогда для всех $y \in \mathbb{R}$*

$$\lim_{n\to\infty} \mathsf{P}_{\theta_n}(T_n \geqslant y) = 1 - \Phi(y - \sqrt{i(\theta)}).$$

Доказательство. Для удобства будем писать, что T_n есть функция от выборки: $T_n = T_n(\boldsymbol{X})$. Теперь распишем $\mathsf{P}_{\theta_n}(T_n < y)$:

$$\begin{aligned} \mathsf{P}_{\theta_n}(T_n(\boldsymbol{X}) \geqslant y) &= \int_{\boldsymbol{x}:\ T_n(\boldsymbol{x})<y} p_{\theta_n}(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) = \int_{\boldsymbol{x}:\ T_n(\boldsymbol{x})<y} \frac{p_{\theta_n}(\boldsymbol{x})}{p_{\theta_0}(\boldsymbol{x})} p_{\theta_0}(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_{\boldsymbol{x}:\ T_n(\boldsymbol{x})<y} \exp\left\{\mathcal{L}(\theta_n, \boldsymbol{X}) - \mathcal{L}(\theta_0, \boldsymbol{X})\right\} p_{\theta_0}(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_{\boldsymbol{x}:\ T_n(\boldsymbol{x})<y} \exp\left\{\sqrt{i(\theta_0)}T_n(\boldsymbol{x}) - \frac{i(\theta)}{2}\right\} p_{\theta_0}(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) \\ &= e^{-i(\theta)/2}\, \mathsf{E}_{\theta_0}\left[e^{T_n(\boldsymbol{X})\sqrt{i(\theta_0)}}[T_n(\boldsymbol{X}) < y]\right]. \end{aligned}$$

Теперь воспользуемся тем, что $T_n(\boldsymbol{X}) \xrightarrow{d_{\theta_0}} \xi \sim \mathcal{N}(0,1)$ и ограниченностью функции внутри матожидания (что даёт равномерную интегрируемость). Тогда можно сказать, что

$$\mathsf{E}_{\theta_0}\left[e^{T_n(\boldsymbol{X})\sqrt{i(\theta_0)}}[T_n(\boldsymbol{X}) < y]\right] \to \mathsf{E}\left[e^{\xi\sqrt{i(\theta_0)}}[\xi < y]\right].$$

Это матожидание достаточно легко считается:

$$\begin{aligned}\mathsf{E}\left[e^{\xi\sqrt{i(\theta_0)}}[\xi < y]\right] &= \int_{-\infty}^{y} e^{z\sqrt{i(\theta_0)}}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}\mathrm{d}z \\ &= \int_{-\infty}^{y}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{z^2}{2} + z\sqrt{i(\theta_0)} - \frac{i(\theta_0)}{2} + \frac{i(\theta_0)}{2}\right\}\mathrm{d}z \\ &= e^{i(\theta_0)/2}\int_{-\infty}^{y}\frac{1}{\sqrt{2\pi}}e^{-(z-\sqrt{i(\theta_0)})^2/2}\mathrm{d}z = e^{i(\theta_0)/2}\Phi(y - \sqrt{i(\theta_0)}).\end{aligned}$$

Тем самым получаем, что

$$\lim_{n\to\infty}\mathsf{P}_{\theta_n}(T_n(\boldsymbol{X}) < y) = \Phi(y - \sqrt{i(\theta_0)}).$$ □

Теперь можно приступить к доказательству теоремы, сформулированной ещё пару страниц назад.

Доказательство. Зафиксируем $y > \sqrt{i(\theta_0)}$ и введём два события: $D_n = \{t_n > \theta_n\}$ и $S_n = \{T_n \geqslant y\}$. Сразу же воспользуемся леммой выше и скажем, что

$$\lim_{n\to\infty}\mathsf{P}_{\theta_n}(S_n) = 1 - \Phi(y - \sqrt{i(\theta_0)}) < \frac{1}{2}.$$

Теперь заметим следующее:

$$\varliminf_{n\to\infty}\mathsf{P}_{\theta_n}(t_n(\boldsymbol{X}) \leqslant \theta_n) = \varliminf_{n\to\infty}(1 - \mathsf{P}_{\theta_n}(t_n(\boldsymbol{X}) > \theta_n)) = 1 - \varlimsup_{n\to\infty}\mathsf{P}_{\theta_n}(t_n(\boldsymbol{X}) > \theta_n).$$

Следовательно, согласно условию

$$1 - \varlimsup_{n\to\infty}\mathsf{P}_{\theta_n}(t_n(\boldsymbol{X}) > \theta_n) \leqslant \frac{1}{2} \implies \varlimsup_{n\to\infty}\mathsf{P}_{\theta_n}(t_n(\boldsymbol{X}) > \theta_n) \geqslant \frac{1}{2}.$$

Из этого следует, что существует подпоследовательность индексов $\{n_k \mid k \in \mathbb{N}\}$ такая, что

$$\lim_{k\to\infty}\mathsf{P}_{\theta_{n_k}}(D_{n_k}) \geqslant \frac{1}{2} > \lim_{k\to\infty}\mathsf{P}_{\theta_{n_k}}(S_{n_k}).$$

Это означает, что при достаточно больших k $\mathsf{P}_{\theta_{n_k}}(D_{n_k}) > \mathsf{P}_{\theta_{n_k}}(S_{n_k})$. Теперь докажем ещё одну лемму:[9]

Лемма. *Для всех достаточно больших k $\mathsf{P}_{\theta_0}(D_{n_k}) > \mathsf{P}_{\theta_0}(S_{n_k})$.*

Доказательство. Пусть $n = n_k$ и k достаточно велико для того, чтобы выполнялось условие $\mathsf{P}_{\theta_{n_k}}(D_{n_k}) > \mathsf{P}_{\theta_{n_k}}(S_{n_k})$. Заметим, что событие S_n можно переписать следующим образом:

$$\begin{aligned}S_n = \{T_n(\boldsymbol{X}) \geqslant y\} &= \left\{\frac{1}{\sqrt{i(\theta_0)}}\left(\mathcal{L}(\theta_n, \boldsymbol{X}) - \mathcal{L}(\theta_0, \boldsymbol{X}) + \frac{i(\theta_0)}{2}\right) > y\right\} \\ &= \{\mathcal{L}(\theta_n, \boldsymbol{X}) - \mathcal{L}(\theta_0, \boldsymbol{X}) > y'\} = \{p_{\theta_n}(\boldsymbol{X}) > \lambda p_{\theta_0}(\boldsymbol{X})\}.\end{aligned}$$

Теперь рассмотрим следующее неравенство:

$$(p_{\theta_n}(\boldsymbol{x}) - \lambda p_{\theta_0}(\boldsymbol{x}))[\boldsymbol{x} \in S_n] \geqslant (p_{\theta_n}(\boldsymbol{x}) - \lambda p_{\theta_0}(\boldsymbol{x}))[\boldsymbol{x} \in D_n].$$

Почему оно действительно выполняется?

- Пусть $\boldsymbol{x} \in S_n$. Тогда $[\boldsymbol{x} \in S_n] = 1$ и $[\boldsymbol{x} \in D_n] \leqslant 1$, что выполнено всегда.
- Теперь пусть $\boldsymbol{x} \notin S_n$. Тогда $[\boldsymbol{x} \in S_n]$ и $p_{\theta_n}(\boldsymbol{x}) - \lambda p_{\theta_0}(\boldsymbol{x}) < 0$. Так как индикатор принимает значения 0 или 1, то получается, что неположительное число не больше нуля, что похоже на правду.

[9] Лемм много не бывает.

Проинтегрируем по мере μ:

$$\mathsf{P}_{\theta_n}(S_n) - \lambda \mathsf{P}_{\theta_0}(S_n) \geqslant \mathsf{P}_{\theta_n}(D_n) - \lambda \mathsf{P}_{\theta_0}(D_n)$$

Простой перегруппировкой мы получаем желаемое:

$$\lambda(\mathsf{P}_{\theta_0}(D_n) - \mathsf{P}_{\theta_0}(S_n)) \geqslant \mathsf{P}_{\theta_n}(D_n) - \mathsf{P}_{\theta_n}(S_n) > 0 \implies \mathsf{P}_{\theta_0}(D_n) > \mathsf{P}_{\theta_0}(S_n). \qquad \square$$

Теперь покажем, что $\mathsf{P}_{\theta_0}(D_n)$ и $\mathsf{P}_{\theta_0}(S_n)$ на самом деле имеют пределы. По лемме про сходимость T_n по распределению к $\mathcal{N}(0,1)$

$$\lim_{n\to\infty} \mathsf{P}_{\theta_0}(S_n) = 1 - \Phi(y).$$

Далее,

$$\begin{aligned}\lim_{n\to\infty} \mathsf{P}_{\theta_0}(D_n) &= \lim_{n\to\infty} \mathsf{P}_{\theta_0}\left(t_n(\boldsymbol{X}) \geqslant \theta_0 + \frac{1}{\sqrt{n}}\right) = \lim_{n\to\infty} \mathsf{P}_{\theta_0}\left(\sqrt{n}(t_n(\boldsymbol{X}) - \theta_0) \geqslant 1\right) \\ &= \lim_{n\to\infty} \mathsf{P}_{\theta_0}\left(\frac{\sqrt{n}(t_n(\boldsymbol{X}) - \theta_0)}{\sigma(\theta_0)} \geqslant \frac{1}{\sigma(\theta_0)}\right).\end{aligned}$$

Однако по центральной предельной теореме левая часть неравенства сходится по распределению к $\mathcal{N}(0,1)$. Следовательно, предел равен $1 - \Phi(\sigma^{-1}(\theta_0))$. Теперь воспользуеся результатом выше и скажем, что

$$\lim_{n\to\infty} \mathsf{P}_{\theta_0}(S_n) \leqslant \lim_{n\to\infty} \mathsf{P}_{\theta_0}(D_n).$$

Отсюда получаем, что

$$1 - \Phi(y) \leqslant 1 - \Phi\left(\frac{1}{\sigma(\theta_0)}\right) \implies \Phi\left(\frac{1}{\sigma(\theta_0)}\right) \leqslant \Phi(y) \implies \frac{1}{\sigma(\theta_0)} \leqslant y \implies \sigma^2(\theta_0) \geqslant \frac{1}{y^2}.$$

Однако мы брали произвольное $y > \sqrt{i(\theta_0)}$. Тогда $\sigma^2(\theta_0) \geqslant i^{-1}(\theta_0)$, что и требовалось доказать. $\square$

Теперь перейдём к самой теореме Бахадура.

Теорема 13. *Пусть в условиях регулярности (R1)—(R6) для последовательности оценок $t_n(\boldsymbol{X})$ параметра θ выполнено условие асимптотической нормальности: для всех $\theta \in \Theta$*

$$\sqrt{n}(t_n(\boldsymbol{X}) - \theta) \xrightarrow{d_\theta} \mathcal{N}(0, \sigma^2(\theta)).$$

Тогда для всех $\theta_0 \in \Theta$, кроме множества лебеговой меры 0, выполнено следующее условие:

$$\varlimsup_{n\to\infty} \mathsf{P}_{\theta_n}(t_n(\boldsymbol{X}) \leqslant \theta_n) \leqslant \frac{1}{2}.$$

Доказательство. Рассмотрим следующую функцию:

$$f_n(\theta) = \left|\mathsf{P}_\theta(t_n(\boldsymbol{X}) \leqslant \theta) - \frac{1}{2}\right| = \left|\mathsf{P}_\theta(\sqrt{n}(t_n(\boldsymbol{X}) - \theta) \leqslant 0) - \frac{1}{2}\right|.$$

Тогда по центральной предельной теореме для любого фиксированного $\theta \in \Theta$ $f_n(\theta) \to 0$ при $n \to \infty$. Стоит заметить, что $f_n(\theta) \in [0, 1/2]$.

Теперь введём две вещи: функцию $g_n(\theta) = f_n(\theta + n^{-1/2})$ и колмогоровскую тройку $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathsf{Q})$, где Q — вероятностная мера стандартного нормального распределения. На этой тройке введём случайную величину ξ, действующую по правилу $\xi(x) = x$. Посчитаем матожидание $g_n(\xi)$:

$$\mathsf{E}[g_n(\xi)] = \int_{-\infty}^{+\infty} g_n(x) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x = \int_{-\infty}^{+\infty} f_n\left(x + \frac{1}{\sqrt{n}}\right) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x.$$

Сделаем замену $y = x + n^{-1/2}$. Тогда

$$\begin{aligned}\mathsf{E}[g_n(\xi)] &= \int_{-\infty}^{+\infty} f_n(y) \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(y^2 - \frac{2y}{\sqrt{n}} + \frac{1}{n}\right)\right\} \mathrm{d}y \\ &= \int_{-\infty}^{+\infty} f_n(y) \exp\left\{\frac{y}{\sqrt{n}} - \frac{1}{2n}\right\} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \mathrm{d}y \\ &= \mathsf{E}\left[f_n(\xi) \exp\left\{\frac{\xi}{\sqrt{n}} - \frac{1}{2n}\right\}\right].\end{aligned}$$

Однако

$$f_n(\xi)\exp\left\{\frac{\xi}{\sqrt{n}}-\frac{1}{2n}\right\}\xrightarrow{\mathsf{Q}\text{-п.н.}}0 \text{ и } f_n(\xi)\exp\left\{\frac{\xi}{\sqrt{n}}-\frac{1}{2n}\right\}\leqslant\max(e^{\xi},1).$$

По теореме Лебега о мажорируемой сходимости получаем, что $\mathsf{E}[g_n(\xi)]\to 0$. Так как $g_n(x)\in[0,1/2]$, то это означает, что $g_n(\xi)$ сходится к нулю в L_1. Тогда $g_n(\xi)$ сходится по вероятностной мере Q к нулю. Но тогда можно выбрать подпоследовательность индексов $\{n_k, k\in\mathbb{N}\}$ такую, что $g_{n_k}(\xi)\to 0$ Q-п.н.

Из этого можно сделать вывод, что существует борелевское множество $B\in\mathcal{B}(\mathbb{R})$ такое, что $\mathsf{Q}(B)=0$ и для всех $x\notin B$ $g_{n_k}(x)\to 0$. Так как мы использовали в качестве вероятностной меры вероятностную меру стандартного нормального распределения, то это означает, что множество B имеет лебегову меру 0. Теперь заметим, что для всех $\theta\in\Theta\setminus B$

$$g_{n_k}(\theta)\to 0 \implies P_{\theta_{n_k}}(t_{n_k}(\boldsymbol{X})\leqslant\theta_{n_k})\to\frac{1}{2}\implies\varlimsup_{n\to\infty}\mathsf{P}_{\theta_n}(t_n(\boldsymbol{X})\leqslant\theta_n)\leqslant\frac{1}{2}.$$ □

На этом доказательство теоремы Бахадура заканчивается. Теперь можно и следствия посмотреть. Но для начала введём одно определение:

Определение 19. Пусть $\hat{\theta}_n(\boldsymbol{X})$ — асимптотически нормальная оценка параметра θ с асимптотической дисперсией $\sigma^2(\theta)$. Если $\sigma^2(\theta)=i^{-1}(\theta)$, то будем называть оценку $\hat{\theta}_n(\boldsymbol{X})$ *асимптотически эффективной.*

Следствие. В условиях регулярности оценка максимального правдоподобия асимптотически эффективна. Поэтому считается, что оценка максимального правдоподобия — это «почти» наилучшая оценка в асимптотическом плане.

Оказывается, что у оценки максимального правдоподобия есть ещё одно интересное свойство.

Теорема 14 (Эффективность оценки максимального правдоподобия)**.** *Если в условиях неравенства Рао-Крамера $\hat{\theta}(\boldsymbol{X})$ является эффективной оценкой θ, то $\hat{\theta}(\boldsymbol{X})$ есть оценка максимального правдоподобия.*

Доказательство. Воспользуемся критерием эффективности оценки:

$$\hat{\theta}(\boldsymbol{X})-\theta=\frac{1}{I_{\boldsymbol{X}}(\theta)}\frac{\partial}{\partial\theta}\ln p_\theta(\boldsymbol{X})=\frac{\mathcal{L}(\theta,\boldsymbol{X})}{I_{\boldsymbol{X}}(\theta)}.$$

Так как информация Фишера положительна, то $\mathcal{L}(\theta,\boldsymbol{X})$ имеет тот же знак, что и $\hat{\theta}(\boldsymbol{X})-\theta$: при $\theta<\hat{\theta}(\boldsymbol{X})$ $\mathcal{L}(\theta,\boldsymbol{X})>0$ и наоборот. Тогда получаем, что $\hat{\theta}(\boldsymbol{X})$ — это единственная точка максимума $f_\theta(\boldsymbol{X})$ (как функции от θ), то есть $\hat{\theta}(\boldsymbol{X})$ есть оценка максимального правдоподобия. □

На этом с асимптотическим подходом заканчиваем.

4.2 Байесовский подход

Ранее мы вводили так называемый байесовский подход к сравнению оценок. В нём задача поиска наилучшей оценки параметра θ вводилась следующим образом:

$$\hat{\theta}(\boldsymbol{X})=\arg\min_{\theta^*(\boldsymbol{X})}R_{\mathsf{Q}}(\theta^*,\theta)=\arg\min_{\theta^*(\boldsymbol{X})}\int_\Theta R(\theta^*(\boldsymbol{X}),t)q(t)\lambda(\mathrm{d}t),$$

где Q — вероятностная мера на Θ с плотностью $q(t)$ по мере λ. У данной плотности есть своё название:

Определение 20. Плотность $q(t)$ называют *априорной плотностью* параметра θ.

Далее, будем считать, что $\boldsymbol{X}=(X_1,\ldots,X_n)$ — это выборка из неизвестного распределения $\mathsf{P}\in\{\mathsf{P}_\theta\mid\theta\in\Theta\}$, где $\{\mathsf{P}_\theta\mid\theta\in\Theta\}$ есть доминируемое семейство с плотностью $p_\theta(x)$ по мере μ. Тогда можно ввести следующую штуку:

Определение 21. *Апостериорной плотностью* параметра θ называют

$$q(t\mid\boldsymbol{x})=\frac{q(t)p_t(\boldsymbol{x})}{\int_\Theta q(\tau)p_\tau(\boldsymbol{x})\lambda(\mathrm{d}\tau)}.$$

На данный момент будем считать, что мы работаем с квадратичной функцией потерь: $R(\hat{\theta}(\boldsymbol{X}),\theta)=\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X})-\theta)^2]$. Можно ли получить решение с такой функцией потерь в общем случае? Можно.

Определение 22. *Байесовской оценкой* параметра θ будем называть

$$\hat{\theta}_{\mathsf{Q}}(\boldsymbol{X}) = \int_{\Theta} tq(t \mid \boldsymbol{X})\lambda(\mathrm{d}t).$$

На следующей лекции мы докажем, что она является наилучшей оценкой в байесовском подходе. Для этого нам понадобится следующая теорема:

Теорема 15 (о наилучшем квадратичном прогнозе)**.** *Пусть ξ и η — случайные величины. Тогда*

$$\mathsf{E}[\xi \mid \eta] = \arg\min_{f(\eta)} \mathsf{E}[(\xi - f(\eta))^2].$$

Доказательство. Добавим и вычтем $\mathsf{E}[\xi \mid \eta]$, после чего раскроем:

$$\begin{aligned}\mathsf{E}[(\xi - f(\eta))^2] &= \mathsf{E}[(\xi - \mathsf{E}[\xi \mid \eta] + \mathsf{E}[\xi \mid \eta] - f(\eta))^2] \\ &= \mathsf{E}[(\xi - \mathsf{E}[\xi \mid \eta])^2] + 2\,\mathsf{E}[(\xi - \mathsf{E}[\xi \mid \eta])(\mathsf{E}[\xi \mid \eta] - f(\eta))] + \mathsf{E}[(\mathsf{E}[\xi \mid \eta] - f(\eta))^2].\end{aligned}$$

Первое слагаемое никак не зависит от f. Докажем, что второе слагаемое равно нулю. Для этого воспользуемся формулой полной вероятности и тем, что функцию от условия можно вынести за условное матожидание:

$$\begin{aligned}\mathsf{E}[(\xi - \mathsf{E}[\xi \mid \eta])(\mathsf{E}[\xi \mid \eta] - f(\eta))] &= \mathsf{E}[\mathsf{E}[(\xi - \mathsf{E}[\xi \mid \eta])(\mathsf{E}[\xi \mid \eta] - f(\eta))] \mid \eta] \\ &= \mathsf{E}[(\mathsf{E}[\xi \mid \eta] - f(\eta))\,\mathsf{E}[(\xi - \mathsf{E}[\xi \mid \eta]) \mid \eta]] \\ &= \mathsf{E}[(\mathsf{E}[\xi \mid \eta] - f(\eta))(\mathsf{E}[\xi \mid \eta] - \mathsf{E}[\xi \mid \eta])] = 0.\end{aligned}$$

Тем самым минимум достигается при $f(\eta) = \mathsf{E}[\xi \mid \eta]$. □

5 Лекция 5

5.1 Байесовские оценки

Напомню, что в байесовском подходе задача нахождения наилучшей оценки параметра θ формулируется так:

$$\hat{\theta}(\boldsymbol{X}) = \arg\min_{\theta^*(\boldsymbol{X})} \int_{\Theta} R(\theta^*(\boldsymbol{X}), t)q(t)\lambda(\mathrm{d}t),$$

где $q(t)$ — плотность вероятностного распределения Q на Θ по мере λ. Оказывается, что при некоторых условиях можно получить явное выражение для наилучшей оценки. Для доказательства этого факта сделаем несколько предположений:

- Мы будем использовать квадратичную функцию потерь: $R(\hat{\theta}(\boldsymbol{X}, \theta)) = \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta)^2]$.
- $\boldsymbol{X}$ имеет неизвестное распределение $\mathsf{P} \in \{\mathsf{P}_\theta \mid \theta \in \Theta\}$, где $\{\mathsf{P}_\theta \mid \theta \in \Theta\}$ — доминируемое семейство распределений с плотностью $p_\theta(x)$ по мере μ.
- Будем считать, что параметр есть число: $\Theta \in \mathbb{R}$. Далее, зафиксируем вероятностное распределение Q на Θ с плотностью $q(t)$ по мере λ.

Теперь докажем следующую теорему:

Теорема 16 (о байесовской оценке)**.** *Байесовская оценка $\hat{\theta}_{\mathsf{Q}}(\boldsymbol{X})$ — это наилучшая оценка параметра θ в байесовском подходе.*

Доказательство. Введём следующую функцию: $f(t, \boldsymbol{x}) = q(t)p_t(\boldsymbol{x})$. Заметим, что

$$\int_{\Theta\times\mathcal{X}} f(t, \boldsymbol{x})\lambda(\mathrm{d}t)\mu(\mathrm{d}\boldsymbol{x}) = \int_{\Theta} q(t) \left(\int_{\mathcal{X}} p_t(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) \right) \lambda(\mathrm{d}t) = \int_{\Theta} q(t)\lambda(\mathrm{d}t) = 1.$$

Тогда можно считать, что $f(t, \boldsymbol{x})$ есть плотность некоторого вероятностного распределения $\tilde{\mathsf{P}}$ на $\Theta \times \mathcal{X}$ по мере $\lambda \otimes \mu$. Далее, посмотрим на вектор $(\theta, \boldsymbol{X})$, как на случайный вектор на колмогоровской тройке $(\Theta \times \mathcal{X}, \mathcal{F}, \tilde{\mathsf{P}})$, где $\mathcal{F}$ — соответствующая сигма-алгебра, действующий по следующему правилу: $(\theta, \boldsymbol{X})(t, \boldsymbol{x}) = (t, \boldsymbol{x})$. Тогда легко заметить следующее:

$$\begin{aligned}\int_{\Theta} R(\hat{\theta}(\boldsymbol{X}), t)q(t)\lambda(\mathrm{d}t) &= \int_{\Theta} \mathsf{E}_t[(\hat{\theta}(\boldsymbol{X}) - t)^2]q(t)\lambda(\mathrm{d}t) \\ &= \int_{\Theta} \left(\int_{\mathcal{X}} (\hat{\theta}(\boldsymbol{x}) - t)^2 p_t(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) \right) q(t)\lambda(\mathrm{d}t) \\ &= \int_{\Theta\times\mathcal{X}} (\hat{\theta}(\boldsymbol{x}) - t)^2 f(t, \boldsymbol{x})\lambda(\mathrm{d}t)\mu(\mathrm{d}\boldsymbol{x}).\end{aligned}$$

Но это есть ни что иное, как математическое ожидание $(\hat\theta(\boldsymbol{X}) - t)^2$ по вероятностной мере $\tilde{\mathsf{P}}$. Будем обозначать его через $\tilde{\mathsf{E}}[(\hat\theta(\boldsymbol{X}) - t)^2]$. Тогда задача поиска наилучшей оценки сводится к следующей:

$$\tilde{\mathsf{E}}[(\hat\theta(\boldsymbol{X}) - t)^2] \to \min_{\hat\theta(\boldsymbol{X})}.$$

Теорема о наилучшем квадратичном прогнозе говорит, что ответ есть $\tilde{\mathsf{E}}[\theta \mid \boldsymbol{X}]$. Теперь осталось показать, что $\hat\theta_{\mathsf{Q}}(\boldsymbol{X}) = \tilde{\mathsf{E}}[\theta \mid \boldsymbol{X}]$. Для этого заметим, что

- $q(t)$ есть плотность θ.
- Выборка $\boldsymbol{X}$ имеет плотность
$$g(\boldsymbol{x}) = \int_\Theta f(t, \boldsymbol{x})\lambda(\mathrm{d}t).$$
- Далее, условная плотность $\boldsymbol{X}$ относительно θ равна $f(t,x)/q(t) = p_t(x)$.[10]

Отсюда несложно получить условную плотность θ относительно $\boldsymbol{X}$. Она равна

$$\frac{f(t,\boldsymbol{x})}{g(\boldsymbol{x})} = \frac{q(t)p_t(\boldsymbol{x})}{\int_\Theta q(\tau)p_\tau(\boldsymbol{x})\lambda(\mathrm{d}\tau)} = q(t \mid \boldsymbol{x}).$$

Тогда

$$\tilde{\mathsf{E}}[\theta \mid \boldsymbol{X} = \boldsymbol{x}] = \int_\Theta tq(t \mid \boldsymbol{x})\lambda(\mathrm{d}t) = \hat\theta_{\mathsf{Q}}(\boldsymbol{x}) \implies \hat\theta_{\mathsf{Q}}(\boldsymbol{X}) = \tilde{\mathsf{E}}[\theta \mid \boldsymbol{X}]. \qquad \square$$

Примечание. Стоит заметить, что считать условную плотность на самом деле не обязательно, так как байесовскую оценку можно представить в виде отношения интегралов по совместной плотности:

$$\hat\theta_{\mathsf{Q}}(\boldsymbol{X}) = \frac{\int_\Theta tf(t,\boldsymbol{X})\lambda(\mathrm{d}t)}{\int_\Theta f(t,\boldsymbol{X})\lambda(\mathrm{d}t)}.$$

Теперь рассмотрим какой-нибудь пример.

Задача 7. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из равномерного распределения $\mathrm{U}(0,\theta)$, $\theta > 0$. Найдите байесовскую оценку параметра θ, если априорная плотность распределения равна*

(a) $q(t) = [t \in [0,1]]$,

(b) $q(t) = t^{-2}[t \geqslant 1]$.

Решение. Начнём с того, что посчитаем плотность выборки:

$$p_\theta(\boldsymbol{X}) = \frac{1}{\theta^n}[0 \leqslant X_{(1)} \leqslant X_{(n)} \leqslant \theta] = \theta^{-n}[\theta \geqslant X_{(n)}].$$

(a) Совместная плотность равна $f(t,\boldsymbol{X}) = q(t)p_t(\boldsymbol{X})$. Так как $X_{(n)} \in [0,1]$ почти наверное, то можно написать следующее:

$$f(t,\boldsymbol{X}) = t^{-n}[t \in [0,1], t \geqslant X_{(n)}] = t^{-n}[t \in [X_{(n)}, 1]].$$

Тогда

$$\int_\Theta tf(t,\boldsymbol{X})\lambda(\mathrm{d}t) = \int_{X_{(n)}}^1 t^{1-n}\mathrm{d}t = \frac{x^{2-n}}{2-n}\bigg|_{X_{(n)}}^1 = \frac{1}{n-2}(X_{(n)}^{2-n} - 1)$$

$$\int_\Theta f(t,\boldsymbol{X})\lambda(\mathrm{d}t) = \int_{X_{(n)}}^1 t^{-n}\mathrm{d}t = \frac{x^{1-n}}{1-n}\bigg|_{X_{(n)}}^1 = \frac{1}{n-1}(X_{(n)}^{1-n} - 1)$$

Отсюда получаем, что байесовской оценкой в таком случае будет

$$\hat\theta_{\mathsf{Q}}(\boldsymbol{X}) = \frac{n-1}{n-2}\frac{X_{(n)}^{2-n} - 1}{X_{(n)}^{1-n} - 1} = \frac{n-1}{n-2}\frac{X_{(n)} - X_{(n)}^{n-1}}{1 - X_{(n)}^{n-1}}.$$

Стоит заметить, что при больших n $\hat\theta_{\mathsf{Q}}(\boldsymbol{X}) \sim X_{(n)}$.

[10] Из-за этого её часто обозначают $p(\boldsymbol{X} \mid \theta)$.

(b) В данном случае совместная плотность устроена немного по-другому:

$$f(t, \boldsymbol{X}) = t^{-(n+2)}[t \geqslant 1, t \geqslant X_{(n)}] = t^{-(n+2)}[t \geqslant \max(1, X_{(n)})].$$

Тогда

$$\int_\Theta t f(t, \boldsymbol{X})\lambda(\mathrm{d}t) = \int_{\max(1,X_{(n)})}^{+\infty} t^{-n-1}\mathrm{d}t = -\frac{x^{-n}}{n}\Big|_{\max(1,X_{(n)})}^{+\infty} = \frac{1}{n}(\max(1, X_{(n)}))^{-n}$$

$$\int_\Theta f(t, \boldsymbol{X})\lambda(\mathrm{d}t) = \int_{\max(1,X_{(n)})}^{+\infty} t^{-n-2}\mathrm{d}t = -\frac{x^{-n-1}}{n+1}\Big|_{\max(1,X_{(n)})}^{+\infty} = \frac{1}{n+1}(\max(1, X_{(n)}))^{-n-1}$$

Отсюда получаем, что

$$\hat{\theta}_{\mathsf{Q}}(\boldsymbol{X}) = \frac{n+1}{n}\max(1, X_{(n)}).$$ □

5.2 Минимаксные оценки

Помимо байесовского подхода к сравнению оценок вводился так называемый минимаксный подход. В нём задача поиска наилучшей оценки ставится следующим образом:

$$\hat{\theta}(\boldsymbol{X}) = \arg\min_{\theta^*(\boldsymbol{X})} \sup_{\theta\in\Theta} R(\theta^*(\boldsymbol{X}), \theta).$$

Сначала совершенно непонятно, можно ли вообще предъявить какой-то критерий минимаксности. Но он есть. Сделаем те же предположения, что делались для байесовских оценок:

- Мы будем использовать квадратичную функцию потерь: $R(\hat{\theta}(\boldsymbol{X}, \theta)) = \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta)^2]$.
- $\boldsymbol{X}$ имеет неизвестное распределение $\mathsf{P} \in \{\mathsf{P}_\theta \mid \theta \in \Theta\}$, где $\{\mathsf{P}_\theta \mid \theta \in \Theta\}$ — доминируемое семейство распределений с плотностью $p_\theta(x)$ по мере μ.
- Будем считать, что параметр есть число: $\Theta \in \mathbb{R}$. Далее, зафиксируем вероятностное распределение Q на Θ с плотностью $q(t)$ по мере λ.

Лемма (Достаточное условие минимаксности оценки)**.** *Пусть $\hat{\theta}(\boldsymbol{X})$ — это оценка параметра θ такое, что существует вероятностное распределение Q на Θ с плотностью $q(t)$ по мере λ, соответствующее следующему условию: для всех $\theta \in \Theta$*

$$\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta)^2] \leqslant \int_\Theta \mathsf{E}_t[(\hat{\theta}_{\mathsf{Q}}(\boldsymbol{X}) - t)^2]q(t)\lambda(\mathrm{d}t).$$

Тогда $\hat{\theta}(\boldsymbol{X})$ есть минимаксная оценка, то есть это наилучшая оценка в минимаксном подходе.

Доказательство. Возьмём произвольную оценку $\theta^*(\boldsymbol{X})$. Заметим, что

$$\int_\Theta \mathsf{E}_t[(\theta^*(\boldsymbol{X}) - t)^2]q(t)\lambda(\mathrm{d}t) \leqslant \int_\Theta \sup_{\theta\in\Theta} \mathsf{E}_\theta[(\theta^*(\boldsymbol{X}) - \theta)^2]q(t)\lambda(\mathrm{d}t) = \sup_{\theta\in\Theta} \mathsf{E}_\theta[(\theta^*(\boldsymbol{X}) - \theta)^2].$$

Однако данный интеграл можно ограничить снизу интегралом для байесовской оценки:

$$\int_\Theta \mathsf{E}_t[(\hat{\theta}_{\mathsf{Q}}(\boldsymbol{X}) - t)^2]q(t)\lambda(\mathrm{d}t) \leqslant \int_\Theta \mathsf{E}_t[(\theta^*(\boldsymbol{X}) - t)^2]q(t)\lambda(\mathrm{d}t)$$

Тем самым получаем цепочку неравенств, которая показывает минимаксность:

$$\sup_{\theta\in\Theta} \mathsf{E}_\theta[(\theta^*(\boldsymbol{X}) - \theta)^2] \geqslant \int_\Theta \mathsf{E}_t[(\hat{\theta}_{\mathsf{Q}}(\boldsymbol{X}) - t)^2]q(t)\lambda(\mathrm{d}t) \geqslant \sup_{\theta\in\Theta} \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta)^2].$$ □

Из данного утверждения можно сделать очень интересное следствие:

Следствие. Пусть для оценки $\hat{\theta}(\boldsymbol{X})$ существует такое вероятностное распределение Q на Θ с плотностью $q(t)$ по мере λ со следующими условиями:

(a) Существует множество $\Psi \subseteq \Theta$ такое, что $\mathsf{Q}(\Psi) = 1$ и для всех $\theta \in \Psi$ $\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta)] = c = \mathrm{const}$.

(b) Для всех $\theta \notin \Psi$ $\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta)] \leqslant c$

(c) Оценка является байесовской оценкой для распределения Q: $\hat{\theta}(\boldsymbol{X}) = \hat{\theta}_{\mathsf{Q}}(\boldsymbol{X})$

Тогда $\hat{\theta}(\boldsymbol{X})$ есть минимаксная оценка.

Идея доказательства состоит в том, что

$$\sup_{\theta\in\Theta} \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta)^2] = c = \int_\Theta \mathsf{E}_t[(\hat{\theta}_{\mathsf{Q}}(\boldsymbol{X}) - t)^2]q(t)\lambda(\mathrm{d}t).$$

Теперь сделаем лирическое отступление. Допустим, что мы возьмём другое распределение $\tilde{\mathsf{Q}}$. Что тогда можно сказать про отношение следующих интегралов:

$$\int_\Theta \mathsf{E}_t[(\hat{\theta}_{\tilde{\mathsf{Q}}}(\boldsymbol{X}) - t)^2]\tilde{q}(t)\lambda(\mathrm{d}t) \text{ и } \int_\Theta \mathsf{E}_t[(\hat{\theta}_{\mathsf{Q}}(\boldsymbol{X}) - t)^2]q(t)\lambda(\mathrm{d}t)?$$

На самом деле первый интеграл не больше второго. Действительно, $\hat{\theta}_{\tilde{\mathsf{Q}}}(\boldsymbol{X})$ есть байесовская оценка для распределения $\tilde{\mathsf{Q}}$. Тогда

$$\begin{aligned}\int_\Theta \mathsf{E}_t[(\hat{\theta}_{\tilde{\mathsf{Q}}}(\boldsymbol{X}) - t)^2]\tilde{q}(t)\lambda(\mathrm{d}t) &\leqslant \int_\Theta \mathsf{E}_t[(\hat{\theta}_{\mathsf{Q}}(\boldsymbol{X}) - t)^2]\tilde{q}(t)\lambda(\mathrm{d}t)\\ &\leqslant \sup_{\theta\in\Theta} \mathsf{E}_\theta[(\hat{\theta}_{\mathsf{Q}}(\boldsymbol{X}) - \theta)^2] = \int_\Theta \mathsf{E}_t[(\hat{\theta}_{\mathsf{Q}}(\boldsymbol{X}) - t)^2]q(t)\lambda(\mathrm{d}t).\end{aligned}$$

То есть распределение Q таково, что для него значение интеграла максимально. У такого распределения есть название:

Определение 23. Вероятностное распределение Q на Θ называется *наихудшим априорным распределением.*

Теперь попробуем посчитать минимаксную оценку.

Задача 8. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из распределения Бернулли* $\mathrm{Bern}(\theta)$. *Найти минимаксную оценку для θ.*

Решение. Казалось бы, совершенно непонятно, как подступаться к этой задаче, даже с учётом достаточного условия выше. Начнём с того, что попробуем найти оценку, для которой функция потерь не зависит от параметра. Для этого поэкспериментируем с обычными оценками. Начнём с великолепной оценки — с выборочного среднего $\overline{\boldsymbol{X}}$. Для неё

$$\mathsf{E}_\theta[(\overline{\boldsymbol{X}} - \theta)^2] = \mathsf{D}_\theta[\overline{\boldsymbol{X}}] = \frac{\theta(1-\theta)}{n}.$$

Как видно, для неё функция потерь зависит от параметра. Сделаем грязный трюк и будем искать оценку в виде $\hat{\theta}(\boldsymbol{X}) = a\overline{\boldsymbol{X}} + b$, где $a > 0$ и b — некоторые константы. Тогда

$$\begin{aligned}\mathsf{E}_\theta[(a\overline{\boldsymbol{X}} + b - \theta)^2] &= \mathsf{E}_\theta[(a(\overline{\boldsymbol{X}} - \theta) + (b - (1-a)\theta))^2]\\ &= a^2\,\mathsf{E}_\theta[(\overline{\boldsymbol{X}} - \theta)^2] + (b - (1-a)\theta)^2\\ &= \frac{a^2(\theta - \theta^2)}{n} + b^2 - 2b\theta(1-a) + \theta^2(1-a)^2\\ &= \theta^2\left(-\frac{a^2}{n} + (1-a)^2\right) + \theta\left(\frac{a^2}{n} - 2b(1-a)\right) + b^2.\end{aligned}$$

Занулим коэффициенты перед θ и θ^2. Тогда получается следующая система уравнений:

$$\begin{cases} n(1-a)^2 = a^2 \\ 2nb(1-a) = a^2 \end{cases} \quad\Longrightarrow\quad \begin{cases} \sqrt{n} - \sqrt{n}a = a \\ 2b = 1 - a \end{cases}$$

Тем самым получаем, что

$$a = \frac{\sqrt{n}}{\sqrt{n}+1}, \quad b = \frac{1}{2(\sqrt{n}+1)}.$$

Отсюда получаем, что в качестве кандидата можно рассматривать оценку

$$\hat{\theta}(\boldsymbol{X}) = \frac{\sqrt{n}}{\sqrt{n}+1}\overline{\boldsymbol{X}} + \frac{1}{2(\sqrt{n}+1)}.$$

Для неё функция риска будет равна

$$\mathsf{E}_\theta[(\hat\theta(\boldsymbol{X})-\theta)^2]=\frac{1}{4(\sqrt{n}+1)^2}.$$

Теперь попробуем найти распределение Q такое, что для него $\hat\theta(\boldsymbol{X})$ будет байесовской оценкой. Для этого посмотрим на функцию правдоподобия выборки:

$$p_\theta(\boldsymbol{X})=\prod_{i=1}^n p_\theta(X_i)=\prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i}=\theta^{n\overline{\boldsymbol{X}}}(1-\theta)^{n(1-\overline{\boldsymbol{X}})}.$$

Теперь будем смотреть на неё, как на функцию от θ. Ничего не напоминает? Правильно, это напоминает бета-распределение. Возьмём в качестве априорного распределения Q бета-распределение $\mathrm{B}(\alpha,\beta)$:

$$q(t)=\frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha,\beta)}.$$

Тогда совместная плотность распределения равна

$$q(\theta)p_\theta(\boldsymbol{X})=\frac{\theta^{n\overline{\boldsymbol{X}}+\alpha-1}(1-\theta)^{n(1-\overline{\boldsymbol{X}})+\beta-1}}{B(\alpha,\beta)}.$$

Так как знаменатель при подсчёте будет только зависеть от $\boldsymbol{X}$, то можно смело сказать, что $q(t\,|\,\boldsymbol{X})$ равно плотности бета-распределения $\mathrm{B}(n\overline{\boldsymbol{X}}+\alpha,n(1-\overline{\boldsymbol{X}})+\beta)$.[11] Теперь найдём математическое ожидание случайной величины $\xi\sim\mathrm{B}(\alpha,\beta)$:

$$\mathsf{E}[\xi]=\frac{1}{B(\alpha,\beta)}\int_0^1 t^\alpha(1-t)^{\beta-1}\mathrm{d}t=\frac{B(\alpha+1,\beta)}{B(\alpha,\beta)}=\frac{\Gamma(\alpha+1)\Gamma(\beta)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+1)}=\frac{\alpha}{\alpha+\beta}.$$

Тогда

$$\hat\theta_{\mathsf{Q}}(\boldsymbol{X})=\frac{n\overline{\boldsymbol{X}}+\alpha}{n+\alpha+\beta}.$$

Теперь мы хотим подобрать параметры α и β так, чтобы байесовская оценка $\hat\theta_{\mathsf{Q}}(\boldsymbol{X})$ была равна нашей оценке $\hat\theta(\boldsymbol{X})$. Для этого заметим, что

$$\hat\theta(\boldsymbol{X})=\frac{2\sqrt{n}\overline{\boldsymbol{X}}+1}{2\sqrt{n}+2}=\frac{n\overline{\boldsymbol{X}}+\sqrt{n}/2}{n+n\sqrt{2}}.$$

Тогда $\alpha=\beta=\sqrt{n}/2$. Следовательно, $\mathsf{Q}=\mathrm{B}(\sqrt{n}/2,\sqrt{n}/2)$ есть наихудшее априорное распределение и $\hat\theta(\boldsymbol{X})$ есть минимаксная оценка. □

Иногда бывают случаи, когда наихудшее распределение найти не получается. В таких случаях спасает следующая лемма:

Лемма. *Пусть $\hat\theta(\boldsymbol{X})$ — оценка параметра θ такая, что существует последовательность вероятностных мер $\{\mathsf{Q}_k,k\in\mathbb{N}\}$ на Θ с плотностями $q_k(t)$ по соответствующим мерам λ_k, удовлетворяющая следующему условию: для всех $\theta\in\Theta$*

$$\mathsf{E}_\theta[(\hat\theta(\boldsymbol{X})-\theta)^2]\leqslant\varlimsup_{k\to\infty}\int_\Theta \mathsf{E}_t[(\hat\theta_{\mathsf{Q}_k}(\boldsymbol{X})-t)^2]q_k(t)\lambda_k(\mathrm{d}t).$$

Тогда $\hat\theta(\boldsymbol{X})$ есть минимаксная оценка.

Доказательство. Доказательство почти дословно повторяет доказательство в случае одной. Зафиксируем произвольную оценку $\theta^*(\boldsymbol{X})$. Далее, для любого натурального k

$$\int_\Theta \mathsf{E}_t[(\theta^*(\boldsymbol{X})-t)^2]q_k(t)\lambda_k(\mathrm{d}t)\leqslant\int_\Theta\sup_{\theta\in\Theta}\mathsf{E}_\theta[(\theta^*(\boldsymbol{X})-\theta)^2]q_k(t)\lambda_k(\mathrm{d}t)=\sup_{\theta\in\Theta}\mathsf{E}_\theta[(\theta^*(\boldsymbol{X})-\theta)^2].$$

Однако данный интеграл можно ограничить снизу интегралом для байесовской оценки:

$$\int_\Theta \mathsf{E}_t[(\hat\theta_{\mathsf{Q}_k}(\boldsymbol{X})-t)^2]q_k(t)\lambda_k(\mathrm{d}t)\leqslant\int_\Theta \mathsf{E}_t[(\theta^*(\boldsymbol{X})-t)^2]q_k(t)\lambda_k(\mathrm{d}t)$$

Следовательно, получаем желаемое:

$$\sup_{\theta\in\Theta}\mathsf{E}_\theta[(\theta^*(\boldsymbol{X})-\theta)^2]\geqslant\varlimsup_{k\to\infty}\int_\Theta \mathsf{E}_t[(\hat\theta_{\mathsf{Q}_k}(\boldsymbol{X})-t)^2]q_k(t)\lambda_k(\mathrm{d}t)\geqslant\sup_{\theta\in\Theta}\mathsf{E}_\theta[(\hat\theta(\boldsymbol{X})-\theta)^2].$$ □

[11] Небольшое примечание: когда априорное $q(t)$ и апостериорное $q(t\,|\,\boldsymbol{X})$ распределения параметров относятся к одному типу, но могут различаться параметрами, то говорят, что распределение $q(t\,|\,\boldsymbol{X})$ есть *сопряженное* распределению $p_t(\boldsymbol{X})$. В данном случае было показано, что сопряжённым для распределения Бернулли является бета-распределение.

Задача 9. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из нормального распределения $\mathcal{N}(\theta, 1)$. Найти минимаксную оценку для θ.*

Решение. Для начала заметим, что оценка $\overline{\boldsymbol{X}}$ подходит в качестве претендента на минимаксность, так как функция потерь не зависит от параметра.

$$\mathsf{E}_\theta[(\overline{\boldsymbol{X}} - \theta)^2] = \mathsf{D}_\theta[\overline{\boldsymbol{X}}] = \frac{1}{n}.$$

Далее посмотрим на функцию правдоподобия для выборки:

$$p_\theta(\boldsymbol{X}) = \prod_{i=1}^n p_\theta(X_i) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^n (X_i - \theta)^2\right\}.$$

Если посмотреть на неё, как на функцию от θ, то выражение будет напоминать плотность нормального распределения. Скажем, что $\mathsf{Q}_k = \mathcal{N}(0, \sigma_k^2)$, где σ_k^2 есть некоторое положительное число. Тогда

$$\begin{aligned} q(t) &= \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2\sigma_k^2}\right\}, \\ q(t)p_t(\boldsymbol{X}) &= \frac{1}{\sigma_k^2(2\pi)^{(n+1)/2}} \exp\left\{-\frac{1}{2}\left(\frac{t^2}{\sigma_k^2} + nt^2 - 2tn\overline{\boldsymbol{X}} + \sum_{i=1}^n X_i^2\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(n + \frac{1}{\sigma^2}\right)\left(t - \frac{n}{n + \sigma_k^{-2}}\overline{\boldsymbol{X}}\right)^2\right\} \end{aligned}$$

Отсюда можно сделать вывод, что[12]

$$q(t \mid \boldsymbol{X}) \sim \mathcal{N}\left(\frac{n}{n + \sigma_k^{-2}}\overline{\boldsymbol{X}}, \frac{1}{n + \sigma_k^{-2}}\right).$$

Тогда

$$\hat{\theta}_{\mathsf{Q}_k}(\boldsymbol{X}) = \frac{n}{n + \sigma_k^{-2}}\overline{\boldsymbol{X}}.$$

Великолепно, мы взяли какую-то последовательность вероятностных распределений. Но можно ли сказать, что будет выполнено условие леммы? Оказывается, что можно. Заметим, что

$$\begin{aligned} \int_\Theta \mathsf{E}_t[(\hat{\theta}_{\mathsf{Q}_k}(\boldsymbol{X}) - t)^2] q_k(t)\lambda_k(\mathrm{d}t) &= \int_{\mathbb{R}^n}\int_\Theta (\hat{\theta}_{\mathsf{Q}_k}(\boldsymbol{x}) - t)^2 q_k(t)p_t(\boldsymbol{x})\lambda_k(\mathrm{d}t)\mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_{\mathbb{R}^n}\int_\Theta (t - \hat{\theta}_{\mathsf{Q}_k}(\boldsymbol{x}))^2 q(t \mid \boldsymbol{x})g(\boldsymbol{x})\lambda_k(\mathrm{d}t)\mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_{\mathbb{R}^n}\left(\int_\Theta (t - \hat{\theta}_{\mathsf{Q}_k}(\boldsymbol{x}))^2 q(t \mid \boldsymbol{x})\lambda_k(\mathrm{d}t)\right) g(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_{\mathbb{R}^n} \frac{1}{n + \sigma_k^{-2}} g(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) = \frac{1}{n + \sigma_k^{-2}}. \end{aligned}$$

Тогда для минимаксности $\overline{\boldsymbol{X}}$ должно быть выполнено следующее условие:

$$\frac{1}{n} \leqslant \varlimsup_{k\to\infty} \frac{1}{n + \sigma_k^{-2}}.$$

Но для этого достаточно взять последовательность σ_k^2, которая будет стремиться к $+\infty$ (например, $\sigma_k^2 = k$). □

6 Лекция 6

На этой лекции разбирались задачи из первого домашнего задания. Здесь мы оставим только формулировки.

Задача 10. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из равномерного распределения на отрезке $[0, \theta]$. Проверьте на несмещенность, состоятельность, сильную состоятельность и асимптотическую нормальность следующие оценки параметра θ: $2\overline{\boldsymbol{X}}$, $\overline{\boldsymbol{X}} + X_{(n)}/2$, $(n+1)X_{(1)}$, $X_{(1)} + X_{(n)}$, $\frac{n+1}{n}X_{(n)}$.*

[12] Сопряжённое для нормального с известной дисперсией есть нормальное.

Задача 11. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из гамма-распределения с параметрами (α, λ). Предложите асимптотически нормальную оценку $\alpha > 0$ и вычислите ее асимптотическую дисперсию, если*

(a) λ известно;

(b) λ тоже неизвестно.

Задача 12. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из равномерного распределения на отрезке $[0, \theta]$. Сравните следующие оценки параметра θ в равномерном подходе с квадратичной функцией потерь:*

1. $\hat{\theta}_1(\boldsymbol{X}) = 2\overline{\boldsymbol{X}}$;
2. $\hat{\theta}_2(\boldsymbol{X}) = X_{(1)} + X_{(n)}$;
3. $\hat{\theta}_3(\boldsymbol{X}) = \frac{n+1}{n} X_{(n)}$.

Задача 13. *Пусть $\theta_1^*(\boldsymbol{X})$ и $\theta_2^*(\boldsymbol{X})$ — две "почти наилучшие" оценки параметра θ в среднеквадратичном походе (т.е. каждая из них не хуже любой другой оценки), имеющие одинаковые математические ожидания. Докажите, что тогда для любого θ они совпадают почти наверное, т.е. $\theta_1^*(\boldsymbol{X}) = \theta_2^*(\boldsymbol{X})$ P_θ-п.н.*

Задача 14. *$\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из распределения $\mathcal{N}(\theta, 1)$, $\theta > 0$. Сравните в равномерном подходе относительно квадратичной функции потерь оценки $\overline{\boldsymbol{X}}$ и $\max(0, \overline{\boldsymbol{X}})$.*

Задача 15. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из гамма-распределения с плотностью*

$$p_\theta(x) = \frac{2^\theta}{\Gamma(\theta)} x^{\theta-1} e^{-2x} [x \geqslant 0].$$

где $\theta > 0$ — неизвестный параметр. Для каких функций $\tau(\theta)$ существует эффективная оценка? Найдите информацию Фишера $i(\theta)$ одного элемента выборки.

Задача 16. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из нормального распределения с параметрами (μ, σ^2). Найдите эффективную оценку*

(a) параметра μ, если σ известно;

(b) параметра σ^2, если μ известно.

Вычислите информацию Фишера одного наблюдения в обоих случаях. Найдите информационную матрицу в случае, когда оба параметра μ и σ^2 неизвестны.

Задача 17. *Пусть $\boldsymbol{X}$ — наблюдение из "регулярного" семейства $\{ \mathsf{P}_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Theta \}$, $\Theta \subseteq \mathbb{R}^k$, $k > 1$. Докажите, что если $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ — эффективная оценка $\boldsymbol{\theta}$, то она является оценкой максимального правдоподобия для $\boldsymbol{\theta}$.*

Задача 18. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из распределения с плотностью*

$$p_{\boldsymbol{\theta}}(x) = \frac{1}{\alpha} \exp\left\{ -\frac{x - \beta}{\alpha} \right\} [x \geqslant \beta],$$

где $\boldsymbol{\theta} = (\alpha, \beta)$, $\alpha > 0$ — двумерный параметр. Найдите для $\boldsymbol{\theta}$ оценку максимального правдоподобия. Докажите, что полученная для α оценка $\hat{\alpha}_n(\boldsymbol{X})$ является асимптотически нормальной, и найдите ее асимптотическую дисперсию.

7 Лекция 7

7.1 Достаточные статистики и оптимальные оценки

Ранее мы рассмотрели асимптотический, байесовский и минимаксный подходы. Теперь осталось рассмотреть самый сильный подход — равномерный. Напомню, что в нём поиск наилучшей оценки параметра θ устроен так:

$$\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta)^2] \to \min_{\hat{\theta}(\boldsymbol{X})} \text{ равномерно по всем } \theta \in \Theta.$$

Однако, как обсуждалось ранее, если не сузить класс оценок, то эта задача будет бессмысленной (так как риск будет должен быть равен тождественному нулю). Поэтому скажем, что мы работаем в классе *несмещённых* оценок. В таком случае $\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta)^2] = \mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})]$. Такая переформулировка сразу приводит нас к определнию оптимальной оценки.

Определение 24. Пусть $\hat{\theta}(\boldsymbol{X})$ — несмещённая оценка параметра $\theta \in \mathbb{R}$. Будем говорить, что $\hat{\theta}(\boldsymbol{X})$ есть *оптимальная* оценка, если у неё равномерно наименьшая дисперсия, то есть для любой другой несмещённой оценки $\theta^*(\boldsymbol{X})$

$$\mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})] \leqslant \mathsf{D}_\theta[\theta^*(\boldsymbol{X})] \text{ для всех } \theta \in \Theta.$$

Данное определение дословно переносится на многомерный случай, но там неравенство на дисперсии записывается следующим образом:

$$\mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})] \preccurlyeq \mathsf{D}_\theta[\theta^*(\boldsymbol{X})] \iff \mathsf{D}_\theta[\theta^*(\boldsymbol{X})] - \mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})] \text{ неотрицательно определена.}$$

Определение 25. Пусть $\{\,\mathsf{P}_\theta \mid \theta \in \Theta\,\}$ — некоторое семейство распределений, а $\boldsymbol{X}$ — наблюдение с неизвестным распределением $\mathsf{P} \in \{\,\mathsf{P}_\theta \mid \theta \in \Theta\,\}$. Будем называть статистику $S(\boldsymbol{X})$ *достаточной*, если существует вариант условного распределения $\mathsf{P}_\theta(\boldsymbol{X} \in B \mid S(\boldsymbol{X}) = s)$, который не зависит от параметра θ, то есть существует измеримая функция $\mathsf{P}(B, s)$ такая, что

$$\mathsf{P}_\theta(\boldsymbol{X} \in B \mid S(\boldsymbol{X}) = s) = \mathsf{P}(B, s) \text{ п.н. по распределению } S(\boldsymbol{X}).$$

Теорема 17 (Колмогоров-Блэкуэлл-Рао)**.** *Пусть $\hat{\theta}(\boldsymbol{X})$ — некоторая несмещённая оценка $\tau(\theta) \in \mathbb{R}$ с конечным матожиданием: $\mathsf{E}_\theta[\hat{\theta}(\boldsymbol{X})] < +\infty$ для всех $\theta \in \Theta$. Далее, пусть $S(\boldsymbol{X})$ — это достаточная статистика для семейства распределений $\{\,\mathsf{P}_\theta \mid \theta \in \Theta\,\}$. Тогда*

- *Пусть $\theta^*(\boldsymbol{X}) = \mathsf{E}_\theta[\hat{\theta}(\boldsymbol{X}) \mid S(\boldsymbol{X})]$. Тогда $\theta^*(\boldsymbol{X})$ есть несмещённая оценка $\tau(\theta)$.*
- *Для всех $\theta \in \Theta$ $\mathsf{D}_\theta[\theta^*(\boldsymbol{X})] \leqslant \mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})]$.*
- *Равенство в неравенстве выше для всех $\theta \in \Theta$ достигается тогда и только тогда, когда $\hat{\theta}(\boldsymbol{X})$ является $S(\boldsymbol{X})$-измеримой функцией.*

Доказательство. Несмещённость $\theta^*(\boldsymbol{X})$ сразу же следует из формулы полной вероятности:

$$\mathsf{E}_\theta[\theta^*(\boldsymbol{X})] = \mathsf{E}_\theta[\mathsf{E}_\theta[\hat{\theta}(\boldsymbol{X}) \mid S(\boldsymbol{X})]] = \mathsf{E}_\theta[\hat{\theta}(\boldsymbol{X})] = \tau(\theta).$$

Менее очевидным фактом является то, что это действительно оценка, то есть то, что $\theta^*(\boldsymbol{X})$ не зависит от θ. Однако $\theta^*(\boldsymbol{X})$ есть интеграл $\hat{\theta}(\boldsymbol{X})$ по условному распределению $\boldsymbol{X}$ относительно $S(\boldsymbol{X})$. Но ни оно, ни функция $\hat{\theta}(\boldsymbol{x})$ не зависят от θ. Тем самым получаем желаемое.

Теперь покажем, что ограничение на дисперсии действительно выполнено. Для этого заметим, что функция $h(x) = (x - \tau(\theta))^2$ выпукла. Тогда можно воспользоваться неравенством Йенсена:

$$(\theta^*(\boldsymbol{X}) - \tau(\theta))^2 = (\mathsf{E}_\theta[\hat{\theta}(\boldsymbol{X}) \mid S(\boldsymbol{X})] - \tau(\theta))^2 \leqslant \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \tau(\theta))^2 \mid S(\boldsymbol{X})].$$

Возьмём матожидание с обеих сторон:

$$\mathsf{D}_\theta[\theta^*(\boldsymbol{X})] \leqslant \mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})].$$

Осталось понять, когда будет выполняться равенство в данном неравенстве. Для этого заметим, что

$$\begin{aligned}
\mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})] - \mathsf{D}_\theta[\theta^*(\boldsymbol{X})] &= \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \tau(\theta))^2] - \mathsf{E}_\theta[(\theta^*(\boldsymbol{X}) - \tau(\theta))^2] \\
&= \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}))^2] - \mathsf{E}_\theta[(\theta^*(\boldsymbol{X}))^2] \\
&= \mathsf{E}_\theta[\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}))^2 - (\theta^*(\boldsymbol{X}))^2 \mid S(\boldsymbol{X})]] \\
&= \mathsf{E}_\theta[\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}))^2 \mid S(\boldsymbol{X})] - (\theta^*(\boldsymbol{X}))^2].
\end{aligned}$$

Выражение внутри матожидания напоминает условную дисперсию. Действительно, по линейности условного математического ожидания

$$\begin{aligned}
\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta^*(\boldsymbol{X}))^2 \mid S(\boldsymbol{X})] &= \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}))^2 - 2\hat{\theta}(\boldsymbol{X})\theta^*(\boldsymbol{X}) + (\theta^*(\boldsymbol{X}))^2 \mid S(\boldsymbol{X})] \\
&= \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}))^2 \mid S(\boldsymbol{X})] - 2\theta^*(\boldsymbol{X})\,\mathsf{E}_\theta[\hat{\theta}(\boldsymbol{X}) \mid S(\boldsymbol{X})] + (\theta^*(\boldsymbol{X}))^2 \\
&= \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}))^2 \mid S(\boldsymbol{X})] - (\theta^*(\boldsymbol{X}))^2.
\end{aligned}$$

Тогда

$$\mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})] - \mathsf{D}_\theta[\theta^*(\boldsymbol{X})] = \mathsf{E}_\theta[\mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta^*(\boldsymbol{X}))^2 \mid S(\boldsymbol{X})]] = \mathsf{E}_\theta[(\hat{\theta}(\boldsymbol{X}) - \theta^*(\boldsymbol{X}))^2].$$

Но данное матожидание равно нулю тогда и только тогда, когда $\hat{\theta}(\boldsymbol{X}) = \theta^*(\boldsymbol{X})$ P_θ-п.н. для всех $\theta \in \Theta$. $\square$

Следствие. Теорема Колмогорова-Блэкуэлла-Рао верна и в многомерном случае.

Доказательство. Пусть $\tau(\theta) \in \mathbb{R}^k$. Возьмём произвольный ненулевой вектор $\boldsymbol{a} \in \mathbb{R}^k$. Тогда $\hat{\theta}_{\boldsymbol{a}}(\boldsymbol{X})$ есть несмещённая оценка $\langle \tau(\theta), \boldsymbol{a} \rangle$. Следовательно, по одномерному случаю

$$\mathsf{D}_\theta[\langle \theta^*(\boldsymbol{X}), \boldsymbol{a} \rangle] \leqslant \mathsf{D}_\theta[\langle \hat{\theta}(\boldsymbol{X}), \boldsymbol{a} \rangle].$$

Но это означает, что

$$\boldsymbol{a}^\top \mathsf{D}_\theta[\theta^*(\boldsymbol{X})]\boldsymbol{a} \leqslant \boldsymbol{a}^\top \mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})]\boldsymbol{a}.$$

Тогда $\mathsf{D}_\theta[\theta^*(\boldsymbol{X})] \preccurlyeq \mathsf{D}_\theta[\hat{\theta}(\boldsymbol{X})]$. Критерий равенства доказывается аналогично одномерному случаю. □

Следствие. Пусть в условиях теоремы Колмогорова-Блэкуэлла-Рао для $\tau(\theta)$ существует единственная $S(\boldsymbol{X})$-измеримая несмещённая оценка $\theta^*(\boldsymbol{X})$. Тогда $\theta^*(\boldsymbol{X})$ — оптимальная оценка $\tau(\theta)$.

Доказательство. Пусть оценка $\theta^*(\boldsymbol{X})$ не оптимальна. Тогда найдётся оценка $\hat{\theta}(\boldsymbol{X})$, которая будет лучше, то есть её дисперсия меньше. Но тогда по теореме Колмогорова-Блэкуэлла-Рао $\theta^{**}(\boldsymbol{X}) = \mathsf{E}_\theta[\hat{\theta}(\boldsymbol{X}) \mid S(\boldsymbol{X})]$ будет не хуже, чем $\hat{\theta}(\boldsymbol{X})$. Однако и $\theta^*(\boldsymbol{X})$, и $\theta^{**}(\boldsymbol{X})$ $S(\boldsymbol{X})$-измеримы. Следовательно, они равны и $\theta^*(\boldsymbol{X})$ не хуже, чем $\hat{\theta}(\boldsymbol{X})$. Противоречие с тем, что $\hat{\theta}(\boldsymbol{X})$ лучше, чем $\theta^*(\boldsymbol{X})$. Следовательно, $\theta^*(\boldsymbol{X})$ оптимальна. □

Определение 26. Статистика $S(\boldsymbol{X})$ называется *полной* для семейства $\{ \mathsf{P}_\theta \mid \theta \in \Theta \}$, если из того, что

$$\mathsf{E}_\theta[f(S(\boldsymbol{X}))] = 0 \text{ для любого } \theta \in \Theta$$

следует, что $f(S(\boldsymbol{X})) = 0$ P_θ-п.н. для всех $\theta \in \Theta$.

Теорема 18 (об оптимальной оценке)**.** *Пусть $S(\boldsymbol{X})$ — полная достаточная статистика для семейства $\{ \mathsf{P}_\theta \mid \theta \in \Theta \}$. Тогда если $\phi(S(\boldsymbol{X}))$ есть несмещённая оценка $\tau(\theta)$, то она будет оптимальной оценкой.*

Доказательство. Пусть $\psi(S(\boldsymbol{X}))$ — другая несмещённая оценка $\tau(\theta)$. Тогда для любого $\theta \in \Theta$

$$\mathsf{E}_\theta[\phi(S(\boldsymbol{X})) - \psi(S(\boldsymbol{X}))] = 0.$$

Но тогда по полноте $\phi(S(\boldsymbol{X})) = \psi(S(\boldsymbol{X}))$ P_θ-п.н. Отсюда получаем, что $\phi(S(\boldsymbol{X}))$ есть единственная $S(\boldsymbol{X})$-измеримая несмещённая оценка $\tau(\theta)$. Следовательно, она оптимальна. □

Тем самым, если есть полная достаточная статистика, то можно легко находить оптимальные оценки для $\tau(\theta)$, решая *уравнение несмещённости*

$$\mathsf{E}_\theta[\phi(S(\boldsymbol{X}))] = \tau(\theta).$$

Для примера разберём одну задачу.

Задача 19. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из распределения Бернулли* $\mathrm{Bern}(\theta)$, $\theta \in (0, 1)$. *Найти полную и достаточную статистику.*

Доказательство. Ранее мы не вводили никаких методов поиска полных и достаточных статистик, поэтому придётся угадывать. Начнём с самого очевидного кандидата — выборочного среднего $\overline{\boldsymbol{X}}$, а точнее, с суммы элементов $S(\boldsymbol{X}) = \sum_{i=1}^n X_i$. Проверим, что она полна. Так как распределение Бернулли дискретно, то достаточно проверить, что для любого $\boldsymbol{x} \in \{ 0, 1 \}^n$ условная вероятность $\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x} \mid S(\boldsymbol{X}) = k)$ не зависит от θ. Раскроем по определению:

$$\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x} \mid S(\boldsymbol{X}) = k) = \frac{\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}, S(\boldsymbol{X}) = k)}{\mathsf{P}_\theta(S(\boldsymbol{X}) = k)}$$

Теперь возникает два случая. Если $\sum_{i=1}^n x_i \neq k$, то события в вероятности в числителе несовместны и $Pr_\theta(\boldsymbol{X} = \boldsymbol{x} \mid S(\boldsymbol{X}) = k) = 0$. Иначе же событие $S(\boldsymbol{X}) = k$ включается в событие $\boldsymbol{X} = \boldsymbol{x}$ и

$$\begin{aligned}\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x} \mid S(\boldsymbol{X}) = k) &= \frac{\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}}{\mathsf{P}_\theta(S(\boldsymbol{X}) = k)} \mathbf{1}_{\sum_{i=1}^n x_i = k} \\ &= \frac{\theta^k (1-\theta)^{n-k}}{\binom{n}{k} \theta^k (1-\theta)^{n-k}} \mathbf{1}_{\sum_{i=1}^n x_i = k} = \binom{n}{k}^{-1} \mathbf{1}_{\sum_{i=1}^n x_i = k}.\end{aligned}$$

Тем самым получаем, что условная вероятность действительно не зависит от θ и $S(\boldsymbol{X})$ будет достаточной статистикой. Теперь проверим полноту. Для этого возьмём произвольную функцию f и посчитаем матожидание $\mathsf{E}_\theta[f(S(\boldsymbol{X}))]$ по определению, пользуясь тем, что $S(\boldsymbol{X}) \sim \mathrm{Bin}(n, \theta)$:

$$\mathsf{E}_\theta[f(S(\boldsymbol{X}))] = \sum_{k=0}^n f(k) \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Далее допустим, что это матожидание равно нулю для любого $\theta \in (0,1)$. Но это многочлен от θ степени n, то есть у него не может быть континуум корней, если он не есть тождественный ноль. Но есть одна проблема: отсюда не следует, что $f(k) = 0$ для всех $k \in \{1, 2, \ldots, n\}$. Для этого преобразуем многочлен:

$$\sum_{k=0}^{n} f(k)\binom{n}{k}\theta^k(1-\theta)^{n-k} = (1-\theta)^n \sum_{k=0}^{n} f(k)\binom{n}{k}\left(\frac{\theta}{1-\theta}\right)^k.$$

Так как $1 - \theta > 0$, то получаем, что для всех $k \in \{1, 2, \ldots, n\}$ $f(k)\binom{n}{k} = 0$. Следовательно, $f(k) = 0$ для всех $k \in \{1, 2, \ldots, n\}$ и $f(S(\boldsymbol{X})) = 0$ P_θ-п.н. для всех $\theta \in (0,1)$. Тогда $S(\boldsymbol{X})$ есть полная статистика. □

Следствие. $\overline{\boldsymbol{X}}$ есть оптимальная оценка θ в схеме Бернулли.

7.2 Критерий факторизации Неймана-Фишера

В примере выше нам сильно повезло, что мы сразу угадали полную достаточную статистику. Но как быть в общем случае? То есть возникает два относительно фундаментальных вопроса:

- Как находить достаточную статистику?
- Как проверять достаточные статистики на полноту?

На первый вопрос даёт ответ следующая теорема.

Теорема 19 (Критерий факторизации Неймана-Фишера)**.** *Пусть $\{\mathsf{P}_\theta \mid \theta \in \Theta\}$ — это доминируемое семейство с плотностью $p_\theta(x)$. Тогда статистика $S(\boldsymbol{X})$ будет достаточной для данного семейства тогда и только тогда, когда существует представление $p_\theta(x)$ в виде*

$$p_\theta(x) = h(x)\psi_\theta(S(x)),$$

где ψ_θ и h — некоторые неотрицательные борелевские функции.

Доказательство для дискретного случая. Для начала покажем, что из того, что $S(\boldsymbol{X})$ есть достаточная статистика, следует нужное представление. Для этого заметим, что

$$\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) = \mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x} \mid S(\boldsymbol{X}) = S(\boldsymbol{x}))\mathsf{P}_\theta(S(\boldsymbol{X}) = S(\boldsymbol{x})).$$

Первый множитель в произведении не зависит от θ по достаточности $S(\boldsymbol{X})$, а второй зависит только от θ и $S(\boldsymbol{X})$. Тогда

$$\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) = h(\boldsymbol{x})\psi_\theta(S(\boldsymbol{x})),$$

где $h(\boldsymbol{x}) = \mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x} \mid S(\boldsymbol{X}) = S(\boldsymbol{x}))$, а $\psi_\theta(S(\boldsymbol{x})) = \mathsf{P}_\theta(S(\boldsymbol{X}) = S(\boldsymbol{x}))$.

Теперь предположим, что существует представление $\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x})$ в виде $h(\boldsymbol{x})\psi_\theta(S(\boldsymbol{x}))$. Тогда

$$\begin{aligned}
\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x} \mid S(\boldsymbol{X}) = s) &= \frac{\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}, S(\boldsymbol{X}) = s)}{\mathsf{P}_\theta(S(\boldsymbol{X}) = s)} = \frac{\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x})}{\mathsf{P}_\theta(S(\boldsymbol{X}) = s)}\mathbf{1}_{S(\boldsymbol{x})=s} \\
&= \frac{\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x})}{\sum_{\boldsymbol{y}:S(\boldsymbol{y})=S(\boldsymbol{x})} \mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{y})}\mathbf{1}_{S(\boldsymbol{x})=s} \\
&= \frac{h(\boldsymbol{x})\psi_\theta(S(\boldsymbol{x}))}{\sum_{\boldsymbol{y}:S(\boldsymbol{y})=S(\boldsymbol{x})} h(\boldsymbol{y})\psi_\theta(S(\boldsymbol{y}))}\mathbf{1}_{S(\boldsymbol{x})=s} \\
&= \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{y}:S(\boldsymbol{y})=S(\boldsymbol{x})} h(\boldsymbol{y})}\mathbf{1}_{S(\boldsymbol{x})=s}
\end{aligned}$$

Тем самым получаем, что условная вероятность не зависит от θ. Следовательно, $S(\boldsymbol{X})$ есть достаточная статистика. □

Заметьте, что это было доказательство для дискретного случая. Доказательство для непрерывного случая не такое тривиальное.[13] Для него требуется теорема о пересчёте условных математических ожиданий. Но для того, чтобы её сформулировать, нужно ввести несколько вещей.

Определение 27. Пусть $(\Omega, \mathcal{F}, \mathsf{P})$ — вероятностное пространство. Далее, пусть Q — это другая вероятностная мера на измеримом пространстве $(\Omega, \mathcal{F})$. Будем говорить, что вероятностная мера Q абсолютно непрерывна относительно вероятностой меры P, если для любого $A \in \mathcal{F}$ такого, что $\mathsf{P}(A) = 0$, выполнено, что $\mathsf{Q}(A) = 0$. Обозначение: $\mathsf{Q} \ll \mathsf{P}$.

[13] Владимир Васильевич, ну не аналогично они доказываются, не аналогично!

Пример. Пусть $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Далее, пусть $\mathsf{P} = \mathcal{N}(0,1)$, а $\mathsf{Q} = \mathrm{Exp}(1)$. Несложно понять, что $\mathsf{Q} \ll \mathsf{P}$. Однако в другую сторону это неверно, так как если взять $A = (-\infty, 0]$, то $\mathsf{Q}(A) = 0$, но $\mathsf{P}(A) = 1/2 \neq 0$.

Теорема 20 (Радон, Никодим)**.** *Пусть $(\Omega, \mathcal{F})$ есть измеримое пространство, а P и $\tilde{\mathsf{P}}$ — две вероятностные меры на нём, причём $\tilde{\mathsf{P}} \ll \mathsf{P}$. Тогда существует единственная P-п.н. случайная величина ξ такая, что для любого $A \in \mathcal{F}$*

$$\tilde{\mathsf{P}}(A) = \mathsf{E}[\xi \mathbf{1}_{\xi \in A}],$$

где E означает взятие математического ожидания по вероятностной мере P. Такую случайную величину называют производной Радона-Никодима *и обозначают*

$$\xi = \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}}.$$

Если кому интересно доказательство теоремы, то обращайтесь к учебнику по функциональному анализу (к тому же Колмогорову-Фомину).

Теперь докажем один достаточно простой факт.

Теорема 21. *Пусть ξ — это произвольная случайная величина, а $\tilde{\mathsf{P}} \ll \mathsf{P}$ — две вероятностные меры на измеримом пространстве $(\Omega, \mathcal{F})$. Тогда*

$$\tilde{\mathsf{E}}[\xi] = \mathsf{E}\left[\xi \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}}\right],$$

где $\tilde{\mathsf{E}}$ означает, что мы берём математическое ожидание по вероятностной мере $\tilde{\mathsf{P}}$.

Доказательство. Пусть $\xi(\omega) = \mathbf{1}_{\omega \in A}$, где $A \in \mathcal{F}$. Тогда по теореме Радона-Никодима

$$\tilde{\mathsf{E}}[\mathbf{1}_A] = \tilde{\mathsf{P}}(A) = \mathsf{E}\left[\mathbf{1}_A \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}}\right].$$

То есть условие теоремы выполнено для индикаторов. Отсюда по линейности математического ожидания несложно получить это утверждение для простых случайных величин. Далее возьмём последовательность простых случайных величин ξ_n, монотонно приближающих ξ. Пользуясь теоремой о монотонной сходимости, получим желаемое. □

Теорема 22 (Пересчёт условных математических ожиданий)**.** *Пусть $\tilde{\mathsf{P}} \ll \mathsf{P}$ — две вероятностные меры на измеримом пространстве $(\Omega, \mathcal{F})$. Далее, пусть ξ и η — случайные величины, причём $\mathsf{E}[|\xi|] < +\infty$. Тогда*

$$\tilde{\mathsf{E}}[\xi \mid \eta] = \frac{\mathsf{E}\left[\xi \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \,\middle|\, \eta\right]}{\mathsf{E}\left[\frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \,\middle|\, \eta\right]}$$

Доказательство. Покажем, что правая часть равенства удовлетворяет свойствам условного математического ожидания. Действительно, она есть отношение измеримых функций от η, то есть это измеримая функция от η. Далее проверим интегральное свойство: для произвольного борелевского множества B

$$\begin{aligned}
\tilde{\mathsf{E}}\left[\frac{\mathsf{E}\left[\xi \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \,\middle|\, \eta\right]}{\mathsf{E}\left[\frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \,\middle|\, \eta\right]} \mathbf{1}_{\eta \in B}\right] &= \mathsf{E}\left[\frac{\mathsf{E}\left[\xi \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \,\middle|\, \eta\right]}{\mathsf{E}\left[\frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \,\middle|\, \eta\right]} \mathbf{1}_{\eta \in B} \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}}\right] \\
&= \mathsf{E}\left[\mathsf{E}\left[\frac{\mathsf{E}\left[\xi \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \,\middle|\, \eta\right]}{\mathsf{E}\left[\frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \,\middle|\, \eta\right]} \mathbf{1}_{\eta \in B} \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \,\middle|\, \eta\right]\right] \\
&= \mathsf{E}\left[\mathsf{E}\left[\xi \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \,\middle|\, \eta\right] \mathbf{1}_{\eta \in B}\right] = \mathsf{E}\left[\xi \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \mathbf{1}_{\eta \in B}\right] = \tilde{\mathsf{E}}[\xi \mathbf{1}_{\eta \in B}].
\end{aligned}$$

Тем самым интегральное свойство выполнено и правая часть равенства действительно есть условное математическое ожидание. □

В доказательстве был опущен один момент: делить на ноль нельзя, а проверка на то, что знаменатель не ноль, была пропущена. Но несложно показать, что он действительно не ноль.

Задача 20. *Докажите, что*

$$\tilde{\mathsf{P}}\left(\mathsf{E}\left[\frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}} \,\middle|\, \eta\right] = 0\right) = 0.$$

Доказательство. Для удобства скажем, что $\xi = \frac{\mathrm{d}\tilde{\mathsf{P}}}{\mathrm{d}\mathsf{P}}$. Далее, по формуле полной вероятности:

$$\tilde{\mathsf{P}}(\mathsf{E}[\xi \mid \eta] = 0) = \tilde{\mathsf{E}}[\mathbf{1}_{\mathsf{E}[\xi \mid \eta]=0}] = \mathsf{E}[\xi \mathbf{1}_{\mathsf{E}[\xi \mid \eta]=0}] = \mathsf{E}[\mathsf{E}[\xi \mathbf{1}_{\mathsf{E}[\xi \mid \eta]=0} \mid \eta]] = \mathsf{E}[\mathsf{E}[\xi \mid \eta]\mathbf{1}_{\mathsf{E}[\xi \mid \eta]=0}] = 0. \qquad \square$$

8 Лекция 8

8.1 Критерий факторизации Неймана-Фишера: продолжение

На предыдущей лекции был сформулирован критерий достаточности статистики $S(\boldsymbol{X})$, называемый критерием факторизации Неймана-Фишера. Но он был доказан только в дискретном случае, так как для доказательства в непрерывном случае требовалась формула пересчёта условных матожиданий. Теперь она есть и доказана, поэтому можно приступить и к критерию.

Доказательство, данное ниже, будет корректно для всех случаев, но мы сконцентрируемся на непрерывном случае.

Доказательство. Зафиксируем некоторое $\theta' \in \Theta$ и для любого $\theta \neq \theta'$ введём следующую вероятностную меру:

$$\mathsf{P}_{\theta,\theta'} = \frac{\mathsf{P}_\theta + \mathsf{P}_{\theta'}}{2}.$$

Сразу же заметим, что $\mathsf{P}_\theta \ll \mathsf{P}_{\theta,\theta'}$ и $\mathsf{P}_{\theta'} \ll \mathsf{P}_{\theta,\theta'}$. Тогда по теореме Радона-Никодима существуют производные

$$\frac{\mathrm{d}\mathsf{P}_\theta}{\mathrm{d}\mathsf{P}_{\theta,\theta'}}(x) = \frac{2p_\theta(x)}{p_\theta(x) + p_{\theta'}(x)} \equiv f_\theta(x),$$
$$\frac{\mathrm{d}\mathsf{P}_{\theta'}}{\mathrm{d}\mathsf{P}_{\theta,\theta'}}(x) = \frac{2p_{\theta'}(x)}{p_\theta(x) + p_{\theta'}(x)} \equiv f'_\theta(x).$$

Теперь выразим плотность $p_\theta(x)$ через $f_\theta(x)$:

$$p_\theta(x) = p_{\theta'}(x)\frac{f_\theta(x)}{2 - f_\theta(x)}.$$

Если доказать, что $f_\theta(x)$ есть функция от $S(x)$ то мы сразу же получим, что $p_\theta(x) = h(x)\psi_\theta(S(x))$, где $h(x) = p_{\theta'}(x)$, а $\psi_\theta(S(x)) = f_\theta(x)/(2 - f_\theta(x))$. Это план доказательства в прямую сторону.

Какой же может быть план доказательства в обратную сторону? Пусть имеет место факторизация: $p_\theta(x) = h(x)\psi_\theta(S(x))$. Тогда

$$f_\theta(x) = \frac{2h(x)\psi_\theta(S(x))}{h(x)\psi_\theta(S(x)) + h(x)\psi_{\theta'}(S(x))} = \frac{2\psi_\theta(S(x))}{\psi_\theta(S(x)) + \psi_{\theta'}(S(x))}.$$

Следовательно, $f_\theta(x)$ есть функция от $S(x)$ и θ. После чего этим можно будет воспользоваться.

Для начала докажем в обратную сторону. Пусть имеет место факторизация. Тогда $f_\theta(x)$ и $f'_\theta(x)$ есть функции от $S(x)$. Воспользуемся формулой пересчёта условного математического ожидания $T(\boldsymbol{X})$ по $S(\boldsymbol{X})$:

$$\mathsf{E}_\theta[T(\boldsymbol{X}) \mid S(\boldsymbol{X})] = \frac{\mathsf{E}_{\theta,\theta'}[T(\boldsymbol{X})f_\theta(\boldsymbol{X}) \mid S(\boldsymbol{X})]}{\mathsf{E}_{\theta,\theta'}[f_\theta(\boldsymbol{X}) \mid S(\boldsymbol{X})]} = \frac{f_\theta(\boldsymbol{X})\,\mathsf{E}_{\theta,\theta'}[T(\boldsymbol{X}) \mid S(\boldsymbol{X})]}{f_\theta(\boldsymbol{X})}$$
$$= \mathsf{E}_{\theta,\theta'}[T(\boldsymbol{X}) \mid S(\boldsymbol{X})].$$

Аналогично получаем, что $\mathsf{E}_{\theta'}[T(\boldsymbol{X}) \mid S(\boldsymbol{X})] = \mathsf{E}_{\theta,\theta'}[T(\boldsymbol{X}) \mid S(\boldsymbol{X})]$. Следовательно, для любого $\theta \neq \theta'$ $\mathsf{E}_\theta[T(\boldsymbol{X}) \mid S(\boldsymbol{X})] = \mathsf{E}_{\theta'}[T(\boldsymbol{X}) \mid S(\boldsymbol{X})]$. Теперь возьмём $T(\boldsymbol{X}) = \mathbf{1}_{\boldsymbol{X}\in B}$, где $B \in \mathcal{B}(\mathbb{R}^n)$. Отсюда получаем, что условное распределение $\boldsymbol{X}$ относительно $S(\boldsymbol{X})$ не зависит от θ, но это и одначает достаточность статистики $S(\boldsymbol{X})$.

Приступим к доказательству в прямую сторону. Нам нужно доказать, что $f_\theta(x)$ есть $S(x)$-измеримая функция. Для этого покажем, что $\mathsf{E}_\theta[f_\theta(x) \mid S(\boldsymbol{X})] = f_\theta(\boldsymbol{X})$. В силу достаточности статистики $S(\boldsymbol{X})$ $\mathsf{E}_\theta[f_\theta(x) \mid S(\boldsymbol{X})]$ не зависит от θ и

$$\mathsf{E}_{\theta'}[f_\theta(x) \mid S(\boldsymbol{X})] = \mathsf{E}_\theta[f_\theta(x) \mid S(\boldsymbol{X})].$$

Покажем, что $\mathsf{E}_{\theta,\theta'}[f_\theta(x) \mid S(\boldsymbol{X})] = \mathsf{E}_\theta[f_\theta(x) \mid S(\boldsymbol{X})]$. Для этого проверим интегральное свойство. Зафиксируем произвольное борелевское множество $B \in \mathcal{B}(\mathbb{R}^k)$ (будем считать, что $S(\boldsymbol{X}) \in \mathbb{R}^k$) и распишем:

$$\mathsf{E}_{\theta,\theta'}[f_\theta(x)\mathbf{1}_{S(\boldsymbol{X})\in B}] = \frac{1}{2}\,\mathsf{E}_\theta[f_\theta(x)\mathbf{1}_{S(\boldsymbol{X})\in B}] + \frac{1}{2}\,\mathsf{E}_{\theta'}[f_\theta(x)\mathbf{1}_{S(\boldsymbol{X})\in B}].$$

Далее, по интегральному свойству

$$\begin{aligned}\mathsf{E}_{\theta,\theta'}[f_\theta(x)\mathbf{1}_{S(\boldsymbol{X})\in B}] &= \frac{1}{2}\,\mathsf{E}_\theta[\mathsf{E}_\theta[f_\theta(x)\,|\,S(\boldsymbol{X})]\mathbf{1}_{S(\boldsymbol{X})\in B}] + \frac{1}{2}\,\mathsf{E}_{\theta'}[\mathsf{E}_{\theta'}[f_\theta(x)\,|\,S(\boldsymbol{X})]\mathbf{1}_{S(\boldsymbol{X})\in B}]\\ &= \frac{1}{2}\,\mathsf{E}_\theta[\mathsf{E}_\theta[f_\theta(x)\,|\,S(\boldsymbol{X})]\mathbf{1}_{S(\boldsymbol{X})\in B}] + \frac{1}{2}\,\mathsf{E}_{\theta'}[\mathsf{E}_\theta[f_\theta(x)\,|\,S(\boldsymbol{X})]\mathbf{1}_{S(\boldsymbol{X})\in B}]\\ &= \mathsf{E}_{\theta,\theta'}[\mathsf{E}_\theta[f_\theta(x)\,|\,S(\boldsymbol{X})]\mathbf{1}_{S(\boldsymbol{X})\in B}].\end{aligned}$$

Тем самым получаем желаемое. Но вопрос: а зачем нам это нужно? Затем, чтобы применить формулу пересчёта условных математических ожиданий. Применим её к статистике $f_\theta(\boldsymbol{X})$:

$$\mathsf{E}_\theta[f_\theta(x)\,|\,S(\boldsymbol{X})] = \frac{\mathsf{E}_{\theta,\theta'}[f_\theta^2(x)\,|\,S(\boldsymbol{X})]}{\mathsf{E}_{\theta,\theta'}[f_\theta(x)\,|\,S(\boldsymbol{X})]}$$

Но тогда мы получаем интересную вещь:

$$\mathsf{E}_{\theta,\theta'}[f_\theta^2(x)\,|\,S(\boldsymbol{X})] = (\mathsf{E}_{\theta,\theta'}[f_\theta(x)\,|\,S(\boldsymbol{X})])^2.$$

Однако нам известно, что для любых случайных величин ξ и η $\mathsf{D}[\xi\,|\,\eta] = \mathsf{E}[(\xi - \mathsf{E}[\xi\,|\,\eta])^2\,|\,\eta] = \mathsf{E}[\xi^2\,|\,\eta] - (\mathsf{E}[\xi\,|\,\eta])^2$. Тогда

$$\mathsf{E}_{\theta,\theta'}[(f_\theta(x) - \mathsf{E}_{\theta,\theta'}[f_\theta(x)\,|\,S(\boldsymbol{X})])^2\,|\,S(\boldsymbol{X})] = 0.$$

Взяв матожидание $\mathsf{E}_{\theta,\theta'}$, получим, что $\mathsf{E}_{\theta,\theta'}[f_\theta(x)\,|\,S(\boldsymbol{X})] = f_\theta(x)$. Следовательно, $\mathsf{E}_\theta[f_\theta(x)\,|\,S(\boldsymbol{X})] = f_\theta(\boldsymbol{X})$ и получается искомое разложение $p_\theta(x)$, что и требовалось доказать. □

Попробуем применить этот критерий. Ранее мы находили оптимальную оценку для схемы Бернулли. Теперь найдём её для равномерного распределения.

Задача 21. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из равномерного распределения $\mathrm{U}[0,\theta]$, $\theta > 0$. Найти оптимальную оценку θ.*

Доказательство. План действий следующий: нужно найти достаточную статистику, проверить её на полноту, после чего решить уравнение несмещённости. Первое сделать сосем несложно, пользуясь критерием факторизации. Распишем плотность выборки:

$$p_\theta(\boldsymbol{X}) = \prod_{k=1}^n p_\theta(X_k) = \prod_{k=1}^n \frac{1}{\theta}\mathbf{1}_{0\leqslant X_k\leqslant\theta} = \frac{1}{\theta^n}\mathbf{1}_{0\leqslant X_{(1)}\leqslant X_{(n)}\leqslant\theta}.$$

Достаточной статистикой в данном случае будет являться $X_{(n)}$, так как в критерии факторизации можно взять $h(\boldsymbol{X}) = \mathbf{1}_{0\leqslant X_{(1)}\leqslant X_{(n)}}$, $\psi_\theta(S(\boldsymbol{X}) = \theta^{-n}\mathbf{1}_{X_{(n)}\leqslant\theta}$.

Небольшое лирическое отступление: а почему мы рассматриваем именно эту статистику? Достаточных статистик полно — как минимум, сама выборка будет достаточной. Но есть такая «рекомендация»: если параметр k–мерный, то статистику тоже подбирают k-мерной. Но так можно сделать не всегда. Например, если взять $\mathcal{N}(\mu,\sigma^2)$, то достаточной статистикой будет $(\sum_{k=1}^n X_k, \sum_{k=1}^n X_k^2)$. Однако, если взять распределение $\mathcal{N}(\theta^2,\theta^2)$, то связь между μ и σ^2 есть, а связь между $\sum_{k=1}^n X_k$ и $\sum_{k=1}^n X_k^2$ построить особо не получится. Тем самым получается, что для одномерного параметра θ статистика $(\sum_{k=1}^n X_k, \sum_{k=1}^n X_k^2)$ будет двумерной достаточной, но она не будет полной.

Вернёмся к задаче и покажем, что $X_{(n)}$ есть полная статистика. Для этого вспомним, что плотность максимума равна

$$p_{X_{(n)}}(x) = \frac{nx^{n-1}}{\theta^n}\mathbf{1}_{x\in[0,\theta]}.$$

Посчитаем матожидание произвольной функции $g(X_{(1)})$:

$$\mathsf{E}_\theta[g(X_{(n)})] = \int_0^\theta g(x)\frac{nx^{n-1}}{\theta^n}\mathrm{d}x.$$

Теперь предположим, что для всех $\theta > 0$ $\mathsf{E}_\theta[g(X_{(n)})] = 0$. Это равносильно тому, что

$$\int_0^\theta g(x)x^{n-1}\mathrm{d}x = 0.$$

Продифференцируем по θ:

$$g(\theta)\theta^{n-1} = 0 \implies g(\theta) = 0 \text{ для всех } \theta > 0.$$

Тогда $g(X_{(1)}) = 0$ P_θ-п.н. для всех $\theta > 0$ и $X_{(n)}$ есть полная достаточная статистика. Осталось решить уравнение несмещённости: $\mathsf{E}_\theta[\phi(X_{(1)})] = \theta$. Заметим, что это равносильно тому, что:

$$\frac{\theta^{n+1}}{n} = \int_0^\theta \phi(x) x^{n-1} \mathrm{d}x.$$

Продифференцируем по θ:

$$\frac{n+1}{n}\theta^n = \phi(\theta)\theta^{n-1} \implies \phi(\theta) = \frac{n+1}{n}\theta.$$

Следовательно, оптимальной оценкой θ является $(1 + n^{-1})X_{(n)}$. □

Что ещё бывает здесь полезного? Например, можно считать условные математические ожидания через полные достаточные статистики. Рассмотрим какой-нибудь плохой пример.

Задача 22. *Пусть* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *— выборка из стандартного равномерного распределения* $\mathrm{U}[0, 1]$. *Вычислить* $\mathsf{E}[\sqrt{X_1 + X_2} \,|\, X_{(n)}]$.

Доказательство. Как бы мы могли считать такую вещь? Можно было бы попытаться посчитать условную плотность, но никто не гарантирует её существования, да и функции очень плохие — вычисления будут отвратительными. Попробуем свести её к статистической задаче. Для этого скажем, что мы смотрим не на стандартное равномерное распределение, а на $\mathrm{U}[0, \theta]$, где $\theta > 0$ (потом просто подставим $\theta = 1$). В таком случае известно, что $X_{(n)}$ есть полная достаточная статистика в равномерной модели $\mathrm{U}[0, \theta]$. Тогда $\mathsf{E}_\theta[\sqrt{X_1 + X_2} \,|\, X_{(n)}]$ не зависит от θ, а зависит только от $X_{(n)}$. И чем же это является? Ответ такой: оптимальной оценкой своего математического ожидания, то есть

$$\mathsf{E}_\theta[\mathsf{E}_\theta[\sqrt{X_1 + X_2} \,|\, X_{(n)}]] = \mathsf{E}_\theta[\sqrt{X_1 + X_2}].$$

Осталось посчитать это матожидание:

$$\begin{aligned}\mathsf{E}_\theta[\sqrt{X_1 + X_2}] &= \int_0^\theta \int_0^\theta \sqrt{x+y}\frac{1}{\theta^2}\mathrm{d}x\mathrm{d}y = \frac{1}{\theta^2}\int_0^\theta \left(\int_0^\theta \sqrt{x+y}\mathrm{d}x\right)\mathrm{d}y \\ &= \frac{1}{\theta^2}\int_0^\theta \frac{2}{3}(y+\theta)^{3/2}\mathrm{d}y = \frac{1}{\theta^2}\frac{2}{3}\frac{2}{5}(2\theta)^{5/2} = \frac{16\sqrt{2}}{15}\sqrt{\theta}.\end{aligned}$$

Тогда получаем, что

$$\mathsf{E}_\theta[\sqrt{X_1 + X_2} \,|\, X_{(n)}] = \frac{16\sqrt{2}}{15}\sqrt{X_{(n)}}.$$ □

Этот метод позволяет быстро считать условные матожидания, но он работает только тогда, когда в условии стоит полная достаточная статистика. На этом с достаточностью заканчиваем.

8.2 Полнота

Если мы ищем достаточную статистику, то мы параллельно проверяем её на полноту. Но как это делать? В случае схемы Бернулли или же равномерного проверки были сделана честно, но в общем случае это достаточно трудно.

Это достаточно тяжёлый вопрос. Для достаточности есть достаточно общий критерий факторизации, который даёт ответ в большинстве разумных случаев. Для полноты же такого удобного критерия нет. Однако в некоторых хороших случаях можно вывести условия для полноты — например, если семейство распределений является экспоненциальным.

Определение 28. Пусть $\{\, \mathsf{P}_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Theta \,\}$ — семейство распределений с k-мерным параметром: $\Theta \subseteq \mathbb{R}^k$, $k \in \mathbb{N}$. Далее, пусть оно доминируемо с плотностью $p_{\boldsymbol{\theta}}(x)$ по мере μ. Если эта плотность представима в следующем виде:

$$p_{\boldsymbol{\theta}}(x) = h(x)\exp\left\{\sum_{i=1}^k u_i(x)a_j(\boldsymbol{\theta}) + v(\boldsymbol{\theta})\right\},$$

где h и u_i — это борелевские функции, то семейство распределений $\{\, \mathsf{P}_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Theta \,\}$ называется *экспоненциальным*.

Ранее мы уже видели примеры семейств распределений, не являющихся экспоненциальным — то же равномерное, так как индикатор в виде экспоненты не представить.

Теперь предоложим, что семейство является экспоненциальным. Чему будет равна достаточная статистика? Понятно, что она будет равна $\boldsymbol{S}(x) = (u_1(x), \ldots, u_k(x))$. Оказывается, что если функции $a_i(\boldsymbol{\theta})$ достаточно хороши, то $\boldsymbol{u}(x)$ будет и полной.

Теорема 23 (об экспоненциальном семействе)**.** *Пусть* $\{\,\mathsf{P}_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Theta\,\}$ *— экспоненциальное семейство распределений, а функция* $\boldsymbol{a}(\boldsymbol{\theta}) = (a_1(\boldsymbol{\theta}), \ldots, a_k(\boldsymbol{\theta}))$ *такова, что при пробегании* $\boldsymbol{\theta}$ *всего* Θ $\boldsymbol{a}(\boldsymbol{\theta})$ *зачерчивает* k*-мерный параллелепипед (или шар). Тогда* $\boldsymbol{S}(\boldsymbol{X}) = (u_1(\boldsymbol{X}), \ldots, u_k(\boldsymbol{X}))$ *будет полной достаточной статистикой для* $\boldsymbol{\theta}$*.*

Она обычно формулируется в таком виде, но есть вопрос: а важно ли то, что размерности параметра и статистики совпадают? В принципе, необязательно, но это достаточно разумное требование, так как иначе могут возникать не самые приятные подводные камни.

Доказательство данной теоремы требует две леммы, в доказательстве которых используется комплексный анализ. Первая из них описывает плотность $S(\boldsymbol{X})$, а вторая обычно называется теоремой единственности аналитической функции комплексного переменного. Начнём с плотности. Было бы неплохо сказать что-то про плотность $S(\boldsymbol{X})$, так как при рассуждении про полноту требуется распределение этой статистики.

Пусть

$$\psi_\theta(S(\boldsymbol{X})) = \exp\left\{\sum_{i=1}^{k} u_i(x)a_j(\boldsymbol{\theta}) + v(\boldsymbol{\theta})\right\},$$

где $\boldsymbol{S}(\boldsymbol{X}) = (u_1(\boldsymbol{X}), \ldots, u_k(\boldsymbol{X}))$, то есть $p_\theta(\boldsymbol{X}) = h(\boldsymbol{X})\psi_\theta(\boldsymbol{S}(\boldsymbol{X}))$.

Лемма. *Статистика* $\boldsymbol{S}(\boldsymbol{X})$ *имеет плотность* $\psi_\theta(\boldsymbol{s})$ *по мере*

$$\nu(B) = \int_{\boldsymbol{S}^{-1}(B)} h(\boldsymbol{x})\mathrm{d}x \text{ (если } p_\theta(x) \text{ абсолютно непрерывна).}$$

Доказательство. Для существования плотности нужно показать, что распределение статистики $\boldsymbol{S}(\boldsymbol{X})$ абсолютно непрерывно относительно меры ν. Пусть G_θ есть распределение $\boldsymbol{S}(\boldsymbol{X})$. Если $\nu(B) = 0$, то $h(\boldsymbol{x}) = 0$ для всех $\boldsymbol{x} \in \boldsymbol{S}^{-1}(B)$ (так как $h(x) \geqslant 0$). Тогда посмотрим на $\mathsf{P}_\theta(\boldsymbol{S}(\boldsymbol{X}) \in B)$. Заметим, что

$$\mathsf{P}_\theta(\boldsymbol{S}(\boldsymbol{X}) \in B) = \mathsf{E}_\theta[\mathbf{1}_{\boldsymbol{S}(\boldsymbol{X})\in B}] = \int_{\mathbb{R}^k} \mathbf{1}_{\boldsymbol{S}(\boldsymbol{x})\in B} p_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_{\boldsymbol{S}^{-1}(B)} h(\boldsymbol{x})\psi_\theta(\boldsymbol{S}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} = 0.$$

Следовательно, распределение G_θ абсолютно непрерывно относительно ν и по теореме Радона-Никодима существует производная $\frac{\mathrm{d}G_\theta}{\mathrm{d}\nu}$. Осталось проверить, что она равна $\psi_\theta(\boldsymbol{s})$. Снова распишем $\mathsf{P}_\theta(\boldsymbol{S}(\boldsymbol{X}) \in B)$:

$$\mathsf{P}_\theta(\boldsymbol{S}(\boldsymbol{X}) \in B) = \int_{\boldsymbol{S}^{-1}(B)} h(\boldsymbol{x})\psi_\theta(\boldsymbol{S}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}.$$

Теперь пересчитываем интеграл, делая замену $\boldsymbol{s} = \boldsymbol{S}(\boldsymbol{x})$:

$$\mathsf{P}_\theta(\boldsymbol{S}(\boldsymbol{X}) \in B) = \int_B \psi_\theta(\boldsymbol{s})\nu(\mathrm{d}\boldsymbol{s}). \qquad \square$$

Вторая лемма утверждает следующее:

Лемма (единственности аналитических функций)**.** *Пусть* G_1 *и* G_2 *— две меры Лебега такие, что существует параллелепипед* $D \subseteq \mathbb{R}^k$ *со следующим условием: для любого* $\boldsymbol{a} \in D$

$$\int_{\mathbb{R}^k} e^{\langle \boldsymbol{a},\boldsymbol{y}\rangle} G_1(\mathrm{d}\boldsymbol{y}) = \int_{\mathbb{R}^k} e^{\langle \boldsymbol{a},\boldsymbol{y}\rangle} G_2(\mathrm{d}\boldsymbol{y}) < \infty$$

Тогда $G_1 = G_2$*.*

Идея доказательства. <здесь была идея, но её я допишу позже> $\square$

Теперь докажем теорему об экспоненциальном семействе.

Доказательство. Достаточность статистики $\boldsymbol{S}(\boldsymbol{X})$ следует из критерия факторизации Неймана-Фишера. Покажем, что она будет полна. Пусть $\mathsf{E}_{\boldsymbol{\theta}}[f(\boldsymbol{S}(\boldsymbol{X}))] = 0$ для всех $\boldsymbol{\theta} \in \Theta$. Заметим, что по лемме о плотности статистики

$$\mathsf{E}_{\boldsymbol{\theta}}[f(\boldsymbol{S}(\boldsymbol{X}))] = \int_{\mathbb{R}^k} f(\boldsymbol{s})\psi_{\boldsymbol{\theta}}(\boldsymbol{s})\nu(\mathrm{d}\boldsymbol{s}) = 0.$$

Теперь введём две функции: $f^+(\boldsymbol{s}) = \max\{f(\boldsymbol{s}), 0\}$ и $f^-(\boldsymbol{s}) = \min\{-f(\boldsymbol{s}), 0\}$. Тогда для любого $\boldsymbol{\theta} \in \Theta$

$$\int_{\mathbb{R}^k} f^+(\boldsymbol{s})\psi_{\boldsymbol{\theta}}(\boldsymbol{s})\nu(\mathrm{d}\boldsymbol{s}) = \int_{\mathbb{R}^k} f^-(\boldsymbol{s})\psi_{\boldsymbol{\theta}}(\boldsymbol{s})\nu(\mathrm{d}\boldsymbol{s}).$$

Теперь вспомним, чему равно $\psi_\theta(\boldsymbol{s})$. Тогда

$$\int_{\mathbb{R}^k} f^+(\boldsymbol{s}) \exp\left\{\langle \boldsymbol{a}(\boldsymbol{\theta}), \boldsymbol{s}\rangle + v(\boldsymbol{\theta})\right\} \nu(\mathrm{d}\boldsymbol{s}) = \int_{\mathbb{R}^k} f^-(\boldsymbol{s}) \exp\left\{\langle \boldsymbol{a}(\boldsymbol{\theta}), \boldsymbol{s}\rangle + v(\boldsymbol{\theta})\right\} \nu(\mathrm{d}\boldsymbol{s})$$

Следовательно, существует некоторый параллелепипед $D \subseteq \mathbb{R}^k$ такой, что для любого $\boldsymbol{a} \in D$:

$$\int_{\mathbb{R}^k} e^{\langle \boldsymbol{a}, \boldsymbol{s}\rangle} f^+(\boldsymbol{s}) \nu(\mathrm{d}\boldsymbol{s}) = \int_{\mathbb{R}^k} e^{\langle \boldsymbol{a}, \boldsymbol{s}\rangle} f^-(\boldsymbol{s}) \nu(\mathrm{d}\boldsymbol{s}).$$

Но это есть ни что иное, как лемма о единственности аналитической функции. Тогда

$$f^+(\boldsymbol{s})\nu(\mathrm{d}\boldsymbol{s}) = f^-(\boldsymbol{s})\nu(\mathrm{d}\boldsymbol{s}) \implies f^+(\boldsymbol{s}) = f^-(\boldsymbol{s}) \text{ по мере } \nu.$$

Так как $\boldsymbol{S}(\boldsymbol{X})$ имеет плотность по мере ν, то $f(\boldsymbol{S}(\boldsymbol{X})) = 0$ $\mathsf{P}_{\boldsymbol{\theta}}$-п.н. для любого $\boldsymbol{\theta} \in \Theta$. Следовательно, $\boldsymbol{S}(\boldsymbol{X})$ полна. □

Теперь рассмотрим пример применения теоремы.

Задача 23. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из нормального распределения $\mathcal{N}(\mu, \sigma^2)$. Рассмотрим три случая:*

(a) Среднее μ неизвестно, дисперсия σ^2 известна: $\mu = \theta$, $\theta \in \mathbb{R}$;

(b) Среднее μ известно, дисперсия σ^2 неизвестна: $\sigma^2 = \theta$, $\theta \in \mathbb{R}_{++}$;

(c) Ни среднее μ, ни дисперсия σ^2 неизвестны: $\boldsymbol{\theta} = (\mu, \sigma^2)$, $\boldsymbol{\theta} \in \mathbb{R} \times \mathbb{R}_{++}$.

Найти оптимальные оценки параметра θ.

Доказательство. Для начала выпишем плотность выборки:

$$p_\theta(\boldsymbol{X}) = \prod_{k=1}^n p_\theta(X_k) = \prod_{k=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(X_k-\mu)^2}{2\sigma^2}\right\} = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^n (X_k-\mu)^2\right\}.$$

Теперь будем решать пункты по отдельности.

(a) Заметим, что

$$\sum_{k=1}^n (X_k-\mu)^2 = \sum_{k=1}^n X_k^2 - 2\mu\sum_{k=1}^n X_k + n\mu^2.$$

Тогда

$$p_\theta(\boldsymbol{X}) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^n X_k^2 + \frac{\mu}{\sigma^2}\sum_{k=1}^n X_k - \frac{n\mu^2}{2\sigma^2} - \frac{n}{2}\ln\sigma^2\right\}.$$

Следовательно, это экспоненциальное семейство, в котором $a_1(\mu) = \mu/(2\sigma^2)$, $u_1(\boldsymbol{X}) = \sum_{k=1}^n X_k$, $h(\boldsymbol{X}) = (2\pi)^{-n/2}\exp\left\{-(2\sigma)^{-1}\sum_{k=1}^n X_k^2\right\}$, а $v(\mu) = -n\mu^2/(2\sigma^2) - n\ln\sigma$. Так как θ пробегает всю прямую, то $a_1(\mu)$ тоже пробегает всю прямую, а, следовательно, и какой-то отрезок. Тогда $u_1(\boldsymbol{X})$ есть полная достаточная статистика. Теперь достаточно решить уравнение несмещённости, но его уже решали очень много раз — подходит выборочное среднее $\overline{\boldsymbol{X}}$, что и будет оптимальной оценкой среднего.

(b) В данном случае никакие преобразования не требуются. Это будет экспоненциальное семейство, в котором $a_1(\theta) = -1/(2\sigma^2)$, $u_1(\boldsymbol{X}) = \sum_{k=1}^n (X_k-\mu)^2$, $v(\theta) = -n\ln\sigma$, $h(\boldsymbol{X}) = (2\pi)^{-n/2}$. В данном случае $a_1(\theta)$ пробегает $(-\infty, 0)$, поэтому $u_1(\boldsymbol{X})$ есть полная достаточная статистика. Теперь нужно решить уравнение несмещённости. Для этого заметим, что

$$\mathsf{E}_\theta\left[\sum_{k=1}^n (X_k-\mu)^2\right] = n\,\mathsf{D}_\theta[X_1] = n\sigma^2.$$

Тогда оптимальной оценкой σ^2 будет $n^{-1}\sum_{k=1}^n (X_k-\mu)^2$.

(c) В данном случае запишем плотность следующим образом:

$$p_{\boldsymbol{\theta}}(\boldsymbol{X}) = \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{k=1}^{n} X_k^2 + \frac{\mu}{\sigma^2} \sum_{k=1}^{n} X_k - \frac{n\mu^2}{2\sigma^2} - \frac{n}{2} \ln \sigma^2 \right\}.$$

Тогда это будет экспоненциальное семейство распределений с $\boldsymbol{a}(\boldsymbol{\theta}) = (\mu/(2\sigma^2), -1/(2\sigma^2))$, $\boldsymbol{u}(\boldsymbol{X}) = (\sum_{k=1}^{n} X_k, \sum_{k=1}^{n} X_k^2)$, $v(\boldsymbol{\theta}) = -n\mu^2/(2\sigma^2) - n\ln\sigma$, $h(\boldsymbol{X}) = (2\pi)^{-n/2}$. Заметим, что $\boldsymbol{a}(\boldsymbol{\theta})$ пробегает $\mathbb{R} \times \mathbb{R}_{++}$. Тогда $\boldsymbol{u}(\boldsymbol{X})$ есть полная и достаточная статистика. Теперь надо решить уравнения несмещённости. Сразу же скажем, что $\mathsf{E}_{\boldsymbol{\theta}}[\overline{\boldsymbol{X}}] = \mu$, то есть выборочное среднее есть оптимальная оценка среднего. Далее,

$$\mathsf{E}_{\boldsymbol{\theta}} \left[\frac{1}{n} \sum_{k=1}^{n} X_k^2 \right] = \mu^2 + \sigma^2$$

Нужно избавиться от μ^2. Для этого посчитаем матожидание квадрата выборочно среднего, пользуясь тем, что $\overline{\boldsymbol{X}} \sim \mathcal{N}(\mu, \sigma^2/n)$:

$$\mathsf{E}_{\boldsymbol{\theta}}[(\overline{\boldsymbol{X}})^2] = \mu^2 + \frac{\sigma^2}{n}.$$

Поэтому оптимальной оценкой σ^2 будет

$$S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{\boldsymbol{X}})^2.$$ □

9 Лекция 9

9.1 Линейная регрессия

Сегодня мы разберём классическую задачу линейной регрессии. Обычно она формулируется следующим образом: допустим, что мы хотим узнать какую-то величину $\boldsymbol{\ell} \in \mathbb{R}^n$ и проводим измерения. Но реальные инструменты не могут получить абсолютную точность, поэтому замеры будут шумными:

$$\boldsymbol{X} = \boldsymbol{\ell} + \boldsymbol{\varepsilon},$$

где $\boldsymbol{X}$ — результаты измерений, а $\boldsymbol{\varepsilon}$ — ошибка измерения. Обычно предполагается, что $\boldsymbol{\varepsilon}$ есть гауссовский вектор, но сегодня мы будем рассматривать общий случай. Далее, на $\boldsymbol{\varepsilon}$ накладываются следующие ограничения:

$$\mathsf{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathsf{D}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n.$$

То есть аппарат для наблюдения достаточно хорош, чтобы в среднем скос был нулевой, а ошибки между собой некоррелируют и имеют одинаковую неизвестную дисперсию σ^2. В таком виде можно представить почти все реальные задачи.

В чём состоит линейность модели? В том, что мы кое-что знаем про $\boldsymbol{\ell}$: это не какой-то произвольный вектор, а вектор из нетривиального *известного* линейного подпространства L размерности $k < n$. Но что означает, что мы знаем L? То, что нам известен его некоторый базис $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k \in \mathbb{R}^n$. Составим из базисных векторов матрицу $\mathbf{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k) \in \mathbb{R}^{n \times k}$. Понятно, что в таком случае $\operatorname{rank} \mathbf{Z} = k$. Тогда $\boldsymbol{\ell}$ можно разложить в линейную комбинацию:

$$\boldsymbol{\ell} = \sum_{i=1}^{k} \theta_i \boldsymbol{z}_i = \mathbf{Z}\boldsymbol{\theta}, \text{ где } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k).$$

Смысл вектора $\boldsymbol{\theta}$ очевиден: это неизвестные координаты вектора $\boldsymbol{\ell}$ в базисе $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$. Тогда задача линейной регрессии переформулируется следующим образом: по наблюдению $\boldsymbol{X}$ нужно оценить $\boldsymbol{\theta}$ и σ^2.

Рассмотрим какие-нибудь примеры:

- Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из нормального распределения $\mathcal{N}(\mu, \sigma^2)$. В таком случае задачей линейной регрессии можно считать задачу оценки параметров. Для этого скажем, что

$$\boldsymbol{X} = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix} + \begin{pmatrix} X_1 - \mu \\ \vdots \\ X_n - \mu \end{pmatrix}.$$

  Далее заметим, что первый вектор лежит в одномерном линейном подпространстве с базисным вектором $(1, \ldots, 1)$, а второй вектор содержит независимые и одинаково распределённые случайные величины с нормальным распределением $\mathcal{N}(0, \sigma^2)$. Тем самым получаем задачу линейной регрессии.

- Теперь рассмотрим пример посложнее. Пусть тело движется по прямой с постоянной скоростью. Далее, в некоторые моменты времени $t_1, \ldots, t_n$ мы измеряем координату тела:

$$X_i = a + bt_i + \varepsilon_i,$$

где a — стартовая позиция, b — скорость движения тела и ε_i — погрешность измерения на i-й итерации. Задача состоит в оценке неизвестных a и b. Как привести эту задачу к задаче линейной регрессии? Для этого поймём, как будет выглядеть вектор $\boldsymbol{\ell}$:

$$\boldsymbol{\ell} = \begin{pmatrix} a + bt_1 \\ \vdots \\ a + bt_n \end{pmatrix} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

Тем самым разумно положить $\boldsymbol{\theta} = (a, b)$: всё неизвестное в задаче, кроме σ^2, заносится в $\boldsymbol{\theta}$. Далее, матрицу $\mathbf{Z}$ можно ввести так:

$$\mathbf{Z} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix}.$$

Если не все наблюдения были проведены в одно и то же время, то это настоящий двумерный базис и получается легальная задача линейной регрессии. Далее, можно провести аналогичные рассуждения, если рассматривать ракету, которая летит по параболе. В таком случае будет добавляться член квадратичной зависимости от времени ct_i^2, но суть от этого не изменится.

Как было сказано ранее, задача состоит в оценивании пары $(\boldsymbol{\theta}, \sigma^2)$. Будем использовать *метод наименьших квдаратов*. Вообще, по этому методу можно оценивать почти всё, что угодно, но в задаче линейной регрессии он даёт содержательные свойства. Сформулируем сам сетод:

Определение 29. Оценкой $\boldsymbol{\theta}$ по методу наименьших квадратов называется

$$\hat{\boldsymbol{\theta}}(\boldsymbol{X}) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^k} \|\boldsymbol{X} - \mathbf{Z}\boldsymbol{\theta}\|^2.$$

Геометрический смысл этого определения крайне прост: $\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ есть проекция $\boldsymbol{X}$ на L. Но тогда $\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ есть проекция $\boldsymbol{X}$ на ортогональное подпространство $L^\perp$. Отсюда можно получить замкнутую форму для $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$.

Лемма. *Оценка наименьших квадратов имеет следующий вид:*

$$\hat{\boldsymbol{\theta}}(\boldsymbol{X}) = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \boldsymbol{X}.$$

Перед этим заметим, что $\mathbf{Z}^\top \mathbf{Z} \in \mathbb{R}^{k \times k}$ является обратимой матрицей, так как столбцы $\mathbf{Z}$ являются базисными.

Доказательство. Как было сказано выше, $\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ есть проекция $\boldsymbol{X}$ на $L^\perp$. Следовательно, для любого $\boldsymbol{\theta} \in \mathbb{R}^k$ $\mathbf{Z}\boldsymbol{\theta}$ ортогонально $\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$:

$$\langle \boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X}), \mathbf{Z}\boldsymbol{\theta} \rangle = \boldsymbol{\theta}^\top \mathbf{Z}^\top (\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})) = 0.$$

Но второе равенсто выполнено для всех $\boldsymbol{\theta} \in \mathbb{R}^k$ тогда и только тогда, когда $\mathbf{Z}^\top (\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})) = \mathbf{0}$, так как иначе можно взять $\boldsymbol{\theta} = \mathbf{Z}^\top (\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})) \neq \mathbf{0}$ и получить $0 = \boldsymbol{\theta}^\top \boldsymbol{\theta} > 0$. Следовательно,

$$\mathbf{Z}^\top \boldsymbol{X} = \mathbf{Z}^\top \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X}) \implies \hat{\boldsymbol{\theta}}(\boldsymbol{X}) = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \boldsymbol{X}. \qquad \square$$

Вернёмся к примерам и попробуем посчитать оценки:

- В первом случае матрица $\mathbf{Z}$ равна $(1, 1, \ldots, 1)^\top \in \mathbb{R}^{n \times 1}$. Тогда $\mathbf{Z}^\top \mathbf{Z} = n$ и

$$\hat{\mu}(\boldsymbol{X}) = \frac{1}{n} \mathbf{Z}^\top \boldsymbol{X} = \sum_{i=1}^{n} X_i = \overline{\boldsymbol{X}}.$$

- Для начала посчитаем $\mathbf{Z}^\top\mathbf{Z}$:

$$\mathbf{Z}^\top\mathbf{Z} = \begin{pmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{pmatrix}.$$

Теперь обратим её:

$$(\mathbf{Z}^\top\mathbf{Z})^{-1} = \frac{1}{n\sum_{i=1}^n t_i^2 - (\sum_{i=1}^n t_i)^2} \begin{pmatrix} \sum_{i=1}^n t_i^2 & -\sum_{i=1}^n t_i \\ -\sum_{i=1}^n t_i & n \end{pmatrix}.$$

Далее, посчитаем $\mathbf{Z}^\top\boldsymbol{X}$:

$$\mathbf{Z}^\top\boldsymbol{X} = \begin{pmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i t_i \end{pmatrix}$$

Осталось помножить:

$$(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\boldsymbol{X} = \frac{1}{n\sum_{i=1}^n t_i^2 - (\sum_{i=1}^n t_i)^2} \begin{pmatrix} (\sum_{i=1}^n X_i)(\sum_{i=1}^n t_i^2) - (\sum_{i=1}^n t_i)(\sum_{i=1}^n X_i t_i) \\ n\sum_{i=1}^n X_i t_i - (\sum_{i=1}^n X_i)(\sum_{i=1}^n t_i). \end{pmatrix}$$

Эта формула хороша всем, кроме одного: вычислять ответ с её помощью может оказаться непозволительно дорого, так как вычисление подразумевает перемножение и обращение матриц: матрицу 2×2 обернуть достаточно быстро, а матрицу 10000×10000 уже долго. Поэтому чаще оценку наименьших квадратов ищут по определению, то есть решают задачу минимизации квадратичной функции.

Попытаемся понять, какие свойства могут быть у оценок методом наименьших квадратов. Начнём с самого простого: посчитаем матожидание и дисперсию.

Лемма. *Для оценки методом наименьших квадратов $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ выполнено следующее:*

$$\mathsf{E}[\hat{\boldsymbol{\theta}}(\boldsymbol{X})] = \boldsymbol{\theta}, \quad \mathsf{D}[\hat{\boldsymbol{\theta}}(\boldsymbol{X})] = \sigma^2(\mathbf{Z}^\top\mathbf{Z})^{-1}.$$

Доказательство. Для начала заметим, что матрица $\mathbf{Z}^\top\mathbf{Z}$ симметрична. Далее, $\boldsymbol{X} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, поэтому $\mathsf{E}[\boldsymbol{X}] = \mathbf{Z}\boldsymbol{\theta}$ и $\mathsf{D}[\boldsymbol{X}] = \sigma^2\mathbf{I}_n$. Тогда

$$\begin{aligned} \mathsf{E}[\hat{\boldsymbol{\theta}}(\boldsymbol{X})] &= \mathsf{E}[(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\boldsymbol{X}] = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\,\mathsf{E}[\boldsymbol{X}] = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Z}\boldsymbol{\theta} = \boldsymbol{\theta}, \\ \mathsf{D}[\hat{\boldsymbol{\theta}}(\boldsymbol{X})] &= \mathsf{D}[(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\boldsymbol{X}] = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\,\mathsf{D}[\boldsymbol{X}]\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1} = \sigma^2(\mathbf{Z}^\top\mathbf{Z})^{-1}. \end{aligned}$$

□

Можно ли что-нибудь ещё сказать про эту оценку? Асимптотические свойства рассматривать бесполезно, так как не понятно, что будет происходить с оценкой при увеличении размерности $\boldsymbol{\ell}$. Однако ту же оптимальность можно проверить. И оказывается, что она оптимальна, но в немного более узком классе, чем обычно.

Лемма. *Оценка методом наименьших квадратов $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ является оптимальной оценкой $\boldsymbol{\theta}$ в классе линейных несмещённых оценок, то есть оценок вида $\mathbf{B}\boldsymbol{X}$, где $\mathbf{B}$ — неслучайная матрица.*

Доказательство. Пусть $\boldsymbol{\theta}^*(\boldsymbol{X}) = \mathbf{B}\boldsymbol{X}$ — другая несмещённая оценка $\boldsymbol{\theta}$. Тогда для любого $\boldsymbol{\theta} \in \mathbb{R}^k$

$$\boldsymbol{\theta} = \mathsf{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}^*(\boldsymbol{X})] = \mathbf{B}\,\mathsf{E}_{\boldsymbol{\theta}}[\boldsymbol{X}] = \mathbf{B}\mathbf{Z}\boldsymbol{\theta}.$$

Далее, $\mathbf{B} \in \mathbb{R}^{k\times n}$, $\mathbf{Z} \in \mathbb{R}^{n\times k}$ и $\mathbf{B}\mathbf{Z} \in \mathbb{R}^{k\times k}$. Но тогда $\mathbf{B}\mathbf{Z} = \mathbf{I}_k$. Далее, заметим, что

$$\mathsf{D}_{\boldsymbol{\theta}}[\mathbf{B}\boldsymbol{X}] = \mathbf{B}\,\mathsf{D}_{\boldsymbol{\theta}}[\boldsymbol{X}]\mathbf{B}^\top = \sigma^2\mathbf{B}\mathbf{B}^\top.$$

В итоге нужно доказать, что $\sigma^2\mathbf{B}\mathbf{B}^\top \succcurlyeq \sigma^2(\mathbf{Z}^\top\mathbf{Z})^{-1}$. Для этого рассмотрим следующую дисперсию:

$$\begin{aligned} \mathsf{D}_{\boldsymbol{\theta}}[(\mathbf{B} - (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top)\boldsymbol{X}] &= (\mathbf{B} - (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top)\sigma^2\mathbf{I}_n(\mathbf{B}^\top - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}) \\ &= \sigma^2(\mathbf{B}\mathbf{B}^\top - \mathbf{B}\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1} - (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{B}^\top + (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}) \\ &= \sigma^2\mathbf{B}\mathbf{B}^\top - \sigma^2(\mathbf{Z}^\top\mathbf{Z})^{-1} \succcurlyeq \mathbf{0}. \end{aligned}$$

□

Отлично, мы получили наилучшую линейную оценку для $\boldsymbol{\theta}$. Но ещё есть неизвестный параметр σ^2, который оценить с помощью метода наименьших квадратов не получится. И что делать? Оказывается, это можно сделать следующим образом.

Лемма. *Пусть $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ — это оценка методом наименьших квадратов. Тогда*

$$\mathsf{E}_{\boldsymbol{\theta}}[\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2] = (n-k)\sigma^2.$$

Доказательство. Для начала заметим, что $\mathsf{E}_\theta[\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})] = \mathbf{0}$. Тогда

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\theta}}[\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2] &= \sum_{i=1}^{n} \mathsf{D}_\theta[X_i - (\mathbf{Z}\boldsymbol{\theta})_i] = \operatorname{tr} \mathsf{D}_{\boldsymbol{\theta}}[\boldsymbol{X} - \mathbf{Z}\boldsymbol{\theta}] = \operatorname{tr} \mathsf{D}_{\boldsymbol{\theta}}[(\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top)\boldsymbol{X}] \\ &= \sigma^2 \operatorname{tr}(\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top)(\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top) \\ &= \sigma^2 \operatorname{tr}(\mathbf{I}_n - 2\mathbf{Z}^\top(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z} + \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top) \\ &= \sigma^2 \operatorname{tr}(\mathbf{I}_n - \mathbf{Z}^\top(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}) = \sigma^2(n - \operatorname{tr}((\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{Z})) = \sigma^2(n-k).\end{aligned}$$

□

Тем самым получаем, что несмещённой оценкой σ^2 является

$$\hat{\sigma}^2(\boldsymbol{X}) = \frac{1}{n-k}\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2.$$

Напоследок рассмотрим простую задачу на линейную регрессию.

Задача 24. *Допустим, что есть два груза с массами a и b. Их взвесили по отдельности и вместе. Результаты записали в вектор (X_1, X_2, X_3). Найти оценку для (a, b) методом наименьших квадратов.*

Решение. Для начала сведём это к задаче линейной регрессии:

$$\underbrace{\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}}_{\boldsymbol{X}} = \begin{pmatrix} a + \varepsilon_1 \\ b + \varepsilon_2 \\ a + b + \varepsilon_3 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}}_{\mathbf{Z}} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\boldsymbol{\theta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}}_{\boldsymbol{\varepsilon}}.$$

Теперь можно считать по уже известной формуле. Для начала посчитаем $(\mathbf{Z}^\top\mathbf{Z})^{-1}$:

$$\mathbf{Z}^\top\mathbf{Z} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \implies (\mathbf{Z}^\top\mathbf{Z})^{-1} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

Далее,

$$\mathbf{Z}^\top\boldsymbol{X} = \begin{pmatrix} X_1 + X_3 \\ X_2 + X_3 \end{pmatrix}$$

Следовательно,

$$\hat{\boldsymbol{\theta}}(\boldsymbol{X}) = \frac{1}{3}\begin{pmatrix} 2X_1 - X_2 + X_3 \\ -X_1 + 2X_2 + X_3 \end{pmatrix} \implies \begin{cases} \hat{a}(\boldsymbol{X}) = (2X_1 - X_2 + X_3)/3 \\ \hat{b}(\boldsymbol{X}) = (-X_1 + 2X_2 + X_3)/3 \end{cases}$$

□

9.2 Гауссовская линейная регрессия

Далее в общем случае какие-либо предположения делать уже крайне сложно, поэтому начнём добавлять какие-то ограничения на $\boldsymbol{\varepsilon}$. Самое популярное — затребовать нормальность распределения $\boldsymbol{\varepsilon}$. В таком случае линейную регрессию называют гауссовской. В ней

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n) \implies \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\ell}, \sigma^2\mathbf{I}_n) = \mathcal{N}(\mathbf{Z}\boldsymbol{\theta}, \sigma^2\mathbf{I}_n).$$

Оказывается, что в таком случае полученная раннее оценка методом наименьших квадратов обладает крайне приятными свойствами.

Теорема 24. *В гауссовской линейной регрессионой модели пара $(\hat{\boldsymbol{\theta}}(\boldsymbol{X}), \|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2)$ является полной достаточной статистикой для $(\boldsymbol{\theta}, \sigma^2)$.*

Доказательство. Найдём плотность $\boldsymbol{X}$:

$$\begin{aligned} p_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{(2\pi)^{n/2}\sqrt{|\sigma^2\mathbf{I}_n|}} \exp\left\{-\frac{1}{2}(\boldsymbol{X} - \mathbf{Z}\boldsymbol{\theta})^\top(\sigma^2\mathbf{I}_n)^{-1}(\boldsymbol{X} - \mathbf{Z}\boldsymbol{\theta})\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{X} - \mathbf{Z}\boldsymbol{\theta}\|^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X}) + \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X}) - \mathbf{Z}\boldsymbol{\theta}\|^2\right\}. \end{aligned}$$

Далее нужно приглядеться к этой норме. Вспомним, что $\boldsymbol{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ есть проекция $\boldsymbol{X}$ на $L^\perp$, а $\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})-\mathbf{Z}\boldsymbol{\theta}$ лежит в L. Но тогда по теореме Пифагора

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\|\boldsymbol{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2}{2\sigma^2} - \frac{\|\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})-\mathbf{Z}\boldsymbol{\theta}\|^2}{\sigma^2}\right\}.$$

Отсюда уже видно то, что по критерию факторизации $(\hat{\boldsymbol{\theta}}(\boldsymbol{X}), \|\boldsymbol{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2)$ есть достаточная статистика. Но нам нужна полнота, поэтому нужно несколько преобразовать плотность и привести её к виду экспоненциального семейства. Для этого немного откатимся назад и раскроем:

$$\begin{aligned} p_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{X}-\mathbf{Z}\boldsymbol{\theta}\|^2\right\} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\|\boldsymbol{X}\|^2}{2\sigma^2} + \frac{\langle \mathbf{Z}^\top\boldsymbol{X}, \boldsymbol{\theta}\rangle}{\sigma^2} - \frac{\|\mathbf{Z}\boldsymbol{\theta}\|^2}{2\sigma^2}\right\} \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{X}\|^2 + \sum_{i=1}^{k}(\mathbf{Z}^\top\boldsymbol{X})_i\frac{\theta_i}{\sigma^2} - \frac{\|\mathbf{Z}\boldsymbol{\theta}\|^2}{2\sigma^2} - \frac{n}{2}\ln\sigma^2\right\}. \end{aligned}$$

Заметим, что при пробегании $\boldsymbol{\theta}\in\mathbb{R}^k$ и $\sigma^2>0$ вектор $(-1/(2\sigma^2), \theta_1/\sigma^2, \ldots, \theta_k/\sigma^2)$ заметает полупространство $\mathbb{R}^{k+1}$. Тем самым по теореме об экспоненциальном семействе статистика $(\mathbf{Z}^\top\boldsymbol{X}, \|\boldsymbol{X}\|^2)$ является полной и достаточной статистикой для пары $(\boldsymbol{\theta}, \sigma^2)$. Осталось доказать, что между этой статистикой и статистикой из условия теоремы есть биекция. Это мы докажем на следующей лекции. □

10 Лекция 10

Продолжение доказательства. Построим биекцию между $(\mathbf{Z}^\top\boldsymbol{X}, \|\boldsymbol{X}\|^2)$ и $(\hat{\boldsymbol{\theta}}(\boldsymbol{X}), \|\boldsymbol{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2)$. Заметим, что по теореме Пифагора

$$\|\boldsymbol{X}\|^2 = \|\boldsymbol{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2 + \|\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2,$$

так как $\boldsymbol{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ лежит в $L^\perp$, а $\|\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2$ — в L. Далее, согласно формуле для оценки методом наименьших квадратов

$$\mathbf{Z}^\top\boldsymbol{X} = (\mathbf{Z}^\top\mathbf{Z})\hat{\boldsymbol{\theta}}(\boldsymbol{X}).$$

Тем самым, если мы знаем значения пары $(\hat{\boldsymbol{\theta}}(\boldsymbol{X}), \|\boldsymbol{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2)$, то мы знаем значения пары $(\mathbf{Z}^\top\boldsymbol{X}, \|\boldsymbol{X}\|^2)$ (так как $\mathbf{Z}$ тоже известна). В обратную сторону всё аналогично:

$$\begin{aligned} \hat{\boldsymbol{\theta}}(\boldsymbol{X}) &= (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\boldsymbol{X}, \\ \|\boldsymbol{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2 &= \|\boldsymbol{X}\|^2 - \|\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2. \end{aligned}$$

Тем самым получаем полную взаимообратную связь между парами. Следовательно, $(\hat{\boldsymbol{\theta}}(\boldsymbol{X}), \|\boldsymbol{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2)$ есть полная достаточная статистика. □

Теперь рассмотрим некоторые следствия данной теоремы.

Следствие. $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ есть оптимальная оценка для $\boldsymbol{\theta}$, $\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ — для $\boldsymbol{\ell}$, а $\|\boldsymbol{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2/(n-k)$ — для σ^2.

Доказательство. Ранее было показано, что эти функции являются несмещёнными оценками для соответствующих параметров. Но они все являются функциями от полной и достаточной статистики. Следовательно, они оптимальны. □

10.1 Доверительные интервалы и области

Теперь мы бы хотели построить доверительные интервалы для параметров гауссовской линейной регрессии. Но для этого нужна теорема об ортогональных разложениях. Докажем её, но перед этим вспомним, что такое распределение хи-квадрат.

Определение 30. Будем говорить, что случайная величина ξ имеет *распределение хи-квадрат с n степенями свободы*, если её плотность равна

$$p_\xi(x) = \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}\mathbf{1}_{x\geqslant 0},$$

то есть $\xi\sim\Gamma(n/2, 1/2)$. Обозначение: $\xi\sim\chi^2_n$.

Далее вспомним ещё один простой факт: если $\xi_1, \dots, \xi_n$ есть независимые и одинаково распределённые случайные величины с стандартным нормальным распределением, то $\xi_1^2 + \ldots + \xi_n^2 \sim \chi_n^2$. Теперь можно приступать к теореме.

Теорема 25 (об ортогональных разложениях)**.** *Пусть $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\ell}, \sigma^2 \mathbf{I}_n)$, а $L_1 \oplus L_2 \oplus \ldots \oplus L_r$ есть разложение $\mathbb{R}^n$ в прямую сумму линейных ортогональных подпространств. Далее, пусть $\boldsymbol{Y}_i$ есть проекция $\boldsymbol{X}$ на L_i. Тогда $Y_1, \dots, Y_r$ независимы в совокупности и*

$$\frac{1}{\sigma^2}\|\boldsymbol{Y}_i - \mathsf{E}[\boldsymbol{Y}_i]\|^2 \sim \chi^2_{\dim L_i}.$$

Доказательство. Идея состоит в том, чтобы свести эту теорему к факту, сформулированному выше. Для этого возьмём в $\mathbb{R}^n$ ортонормированный базис $\boldsymbol{f}_1, \dots, \boldsymbol{f}_n$ такой, что

$$\begin{gathered}\boldsymbol{f}_1, \dots, \boldsymbol{f}_{k_1} \text{ — базис } L_1,\\ \boldsymbol{f}_{k_1+1}, \dots, \boldsymbol{f}_{k_1+k_2} \text{ — базис } L_2,\\ \vdots\\ \boldsymbol{f}_{k_1+\ldots+k_{r-1}+1}, \dots, \boldsymbol{f}_n \text{ — базис } L_r.\end{gathered}$$

В таком случае размерность L_i равна k_i для всех $i = 1, \dots, r$. Далее, введём коэффициенты $W_i = \langle \boldsymbol{X}, \boldsymbol{f}_i \rangle$ для всех $i = 1, \dots, n$. Скомпонуем эти коэффициенты в вектор и представим его в виде линейного преобразования $\boldsymbol{X}$:

$$\boldsymbol{W} = \begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix} = \mathbf{C}\boldsymbol{X}, \text{ где } \mathbf{C} = \begin{pmatrix} \boldsymbol{f}_1^\top \\ \vdots \\ \boldsymbol{f}_n^\top \end{pmatrix}.$$

Заметим, что матрица $\mathbf{C}$ ортогональна, так как её строки задают ортонормированный базис $\mathbb{R}^n$. Так как $\boldsymbol{X}$ гауссовский, то и $\boldsymbol{W}$ тоже будет гауссовским, так как он получается линейным преобразованием. Осталось заметить, что $\boldsymbol{W} \sim \mathcal{N}(\mathbf{C}\boldsymbol{\ell}, \sigma^2 \mathbf{I}_n)$, так как $\mathbf{C}\mathbf{C}^\top = \mathbf{I}_n$. Тем самым компоненты $\boldsymbol{W}$ независимы в совокупности. Тогда $\boldsymbol{Y}_i$, которые считаются следующим образом:

$$\boldsymbol{Y}_j = W_{k_1+\ldots+k_{j-1}+1}\boldsymbol{f}_{k_1+\ldots+k_{j-1}+1} + \ldots + W_{k_1+\ldots+k_j}\boldsymbol{f}_{k_1+\ldots+k_j},$$

тоже будут независимы в совокупности. Теперь докажем свойство про их распределение. Для этого заметим, что для всех $i = 1, \dots, n$

$$\frac{W_i - \mathsf{E}[W_i]}{\sigma} \sim \mathcal{N}(0, 1).$$

В таком случае, так как векторы ортонормированные, то

$$\begin{aligned}\frac{1}{\sigma^2}\|\boldsymbol{Y}_j - \mathsf{E}[\boldsymbol{Y}_j]\|^2 &= \left\| \sum_{i=1}^{k_j} \frac{W_{k_1+\ldots+k_{j-1}+i} - \mathsf{E}[W_{k_1+\ldots+k_{j-1}+i}]}{\sigma} \boldsymbol{f}_{k_1+\ldots+k_{j-1}+i} \right\|^2 \\ &= \sum_{i=1}^{k_j} \left(\frac{W_{k_1+\ldots+k_{j-1}+i} - \mathsf{E}[W_{k_1+\ldots+k_{j-1}+i}]}{\sigma} \right)^2 \sim \chi^2_{k_j}.\end{aligned}$$

□

Теперь выпишем из этого достаточно важное следствие:

Следствие. В гауссовской линейной модели $\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ (и $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$) независимо с $\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ и

$$\begin{aligned}\frac{1}{\sigma^2}\|\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X}) - \mathbf{Z}\boldsymbol{\theta}\|^2 &\sim \chi^2_k,\\ \frac{1}{\sigma^2}\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2 &\sim \chi^2_{n-k}.\end{aligned}$$

Доказательство. Как известно, $\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ есть проекция $\boldsymbol{X}$ на $L^\perp$, а $\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ проецирует $\boldsymbol{X}$ на L. Так как $\mathbb{R}^n = L \oplus L^\perp$, то применима теорема об ортогональных разложениях. Следовательно, $\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ и $\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ независимы и верны выражения из условия. □

Теперь приступим к доверительным интервалам. Перед этим надо определить, что это вообще такое.

Определение 31. Пусть $\{\mathsf{P}_\theta \mid \theta \in \Theta\}$ — параметрическое семейство, $\boldsymbol{X}$ — наблюдение из параметрического распределения $\mathsf{P} \in \{\mathsf{P}_\theta \mid \theta \in \Theta\}$, и параметр θ одномерный: $\Theta \subseteq \mathbb{R}$. Будем называть пару статистик $(T_1(\boldsymbol{X}), T_2(\boldsymbol{X})$ *доверительным интервалом* для θ с уровнем доверия γ, если для всех $\theta \in \Theta$

$$\mathsf{P}_\theta(T_1(\boldsymbol{X}) < \theta < T_2(\boldsymbol{X})) \geqslant \gamma.$$

Если в данном неравенстве всегда достигается равенство, то доверительный интервал называют *точным*.

Данное определение почти без изменений переносится на многомерный параметр.

Определение 32. Подмножество $S(\boldsymbol{X}) \subseteq \Theta$ называется доверительной областью для $\boldsymbol{\theta}$ с уровнем доверия γ, если для всех $\boldsymbol{\theta} \in \Theta$

$$\mathsf{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in S(\boldsymbol{X})) \geqslant \gamma.$$

Ну что же, сперва построим доверительный интервал для σ^2. Для этого нужно воспользоваться так называемой центральной статистикой. Дадим определение:

Определение 33. Функция от выборки $S_\theta(\boldsymbol{X})$, которая может зависеть от параметра θ, называется *центральной статистикой*, если её распределение не зависит от θ.

Заметим, что в данном случае можно использовать $\sigma^{-2}\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2$ в качестве центральной статистики, так как ранее было доказано, что она имеет распределение хи-квадрат с $n-k$ степенями свободы. Пусть u_α равно α-квантилю χ^2_{n-k}. Тогда

$$\mathsf{P}\left(u_{(1-\gamma)/2} < \frac{1}{\sigma^2}\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2 < u_{(1+\gamma)/2}\right) = \gamma.$$

Следовательно, точный доверительный интервал для σ^2 уровня доверия γ имеет вид

$$\left[\frac{1}{u_{(1+\gamma)/2}}\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2, \frac{1}{u_{(1-\gamma)/2}}\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2\right]$$

Теперь попытаемся определить доверительный интервал для θ_i. Для этого заметим, что $\hat{\boldsymbol{\theta}}(\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{A})$, где $\mathbf{A} = (\mathbf{Z}^\top\mathbf{Z})^{-1}$. В таком случае

$$\frac{\hat{\theta}_i - \theta_i}{\sqrt{\sigma^2 a_{ii}}} \sim \mathcal{N}(0,1).$$

Но возникает проблема: всё ещё остаётся σ^2. Что делать? Заменить на оценку. Оказывается, что в гауссовском случае всё остаётся весьма хорошо. Полученная случайная величина будет иметь распределение Стьюдента с $n-k$ степенями свободы:

$$\sqrt{\frac{n-k}{a_{ii}}}\frac{\hat{\theta}_i - \theta_i}{\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|} \sim \mathrm{T}_{n-k}.$$

Почему это так? Это следует из того, что $\hat{\theta}_i$ не зависит от $\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|$, и дробь имеет вид отношения стандартной нормальной случайной величины к корню из случайной величины с распределением хи-квадрат, что и есть распределение Стьюдента по определению.

Пусть t_α есть α-квантиль распределения Стьюдента T_{n-k}. Так как оно симметрично относительно нуля, то можно сразу записать точный доверительный интервал с уровнем доверия γ в следующем виде:

$$\hat{\theta}_i - t_{(1+\gamma)/2}\sqrt{\frac{a_{ii}}{n-k}}\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\| < \theta_i < \hat{\theta}_i + t_{(1+\gamma)/2}\sqrt{\frac{a_{ii}}{n-k}}\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|.$$

Осталось построить доверительную область для $\boldsymbol{\theta}$. Здесь возникает третье стандартное распределение, связанное с гауссовским — распределение Фишера.

Определение 34. Пусть ξ, η — независимые случайные величины, причём $\xi \sim \chi^2_k$, $\eta \sim \chi^2_n$. Тогда $\delta = \frac{\xi/k}{\eta/n}$ имеет распределение Фишера с (k,n) степенями свободы. Обозначение: $\delta \sim \mathrm{F}_{k,n}$.

Согласно следствию из теоремы об ортогональных разложениях

$$\frac{\|\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X}) - \mathbf{Z}\boldsymbol{\theta}\|^2}{\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2}\frac{n-k}{k} \sim \mathrm{F}_{k,n-k}.$$

Далее, пусть f_α есть α-квантиль распределения $\mathrm{F}_{k,n-k}$. Так как оно неотрицательно, то доверительной областью будет

$$S(\boldsymbol{X}) = \left\{\boldsymbol{\theta} \in \mathbb{R}^k \colon \frac{\|\mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X}) - \mathbf{Z}\boldsymbol{\theta}\|^2}{\|\boldsymbol{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|^2}\frac{n-k}{k} < f_\gamma\right\}.$$

Стоит заметить, что геометрически доверительная область будет иметь форму эллипсоида.

10.2 Проверка гипотез

На этом мы пока что простимся с линейной регрессией и вернёмся к ней несколько позже. Пока что разберёмся с такой вещью, как проверка гипотез. Вообще, проверка гипотез — это больше методология, тут математики меньше и важнее понять сам метод. Что вообще предполагается в статистике? Предполагается следующее: если вероятность события крайне мала (меньше, чем 0.01, например), то в единичном испытании данное событие не происходит никогда. Если мы так не считаем, то могут происходить события крайне малой вероятности и наша статистика нужно выбросить и думать дальше. Примерно на этом построена методология проверки гипотез: что мы должны с чем-то согласиться и действовать дальше. Обычно согласуют некоторый уровень ошибки.

Теперь формализуем вышесказанное. Пусть $\boldsymbol{X}$ — наблюдение с неизвестным распределением $\mathsf{P} \in \mathcal{P}$, где $\mathcal{P}$ есть некоторое семейство распределений. Далее, возьмём подсемейство $\mathcal{P}_0 \subset \mathcal{P}$. Поставим следующий вопрос: верно ли то, что $\mathsf{P} \in \mathcal{P}_0$? Ответ на этот вопрос позволяет сузить семейство распределений, что упростит анализ. Если ответ положительный, то $\mathcal{P}$ можно сузить до $\mathcal{P}_0$. Если же ответ отрицательный, то $\mathcal{P}$ сужается до $\mathcal{P} \setminus \mathcal{P}_0$. Компромисс в каком-то смысле невозможен: для получения компромисса нужно набрать ещё данных. Рассуждение за этим следующее: по текущим данным оказалось, что верна одна гипотеза, хотя кажется, что должна быть верна другая. Кажется, что выборка оказалась плохой и нужно набрать ещё данных для повторной проверки.

Обычно статистическая гипотеза (или предположение) обозначается следующим образом:

$$\mathrm{H}_0 \colon \mathsf{P} \in \mathcal{P}.$$

Что бывает ещё? В принципе, мы можем иметь не только одну гипотезу. Тогда параллельно с ней вводят альтернативу $\mathrm{H}_1 \colon \mathsf{P} \in \mathcal{P}_1 \subseteq \mathcal{P} \setminus \mathcal{P}_0$. Альтернатив может быть несколько:

$$\begin{gathered}\mathrm{H}_0 \colon \mathsf{P} \in \mathcal{P}_0,\\ \mathrm{H}_1 \colon \mathsf{P} \in \mathcal{P}_1,\\ \vdots\\ \mathrm{H}_k \colon \mathsf{P} \in \mathcal{P}_k.\end{gathered}$$

Методология метода проверки гипотез следующая. Допустим, что сравнивается пара гипотез H_0 против H_1. Если гипотеза H_0 принимается, то есть на вопрос «Верно ли то, что $\mathsf{P} \in \mathcal{P}_0$?» даётся положительный ответ, то семейство распределений сужается до $\mathcal{P}_0$. Однако, если гипотезу H_0 отвергают и задана альтернатива, то семейство сужают до $\mathcal{P}_1$. Если же альтернатив много, то вариантов постановки вопроса несколько. Первый вариант состоит в том, что все гипотезы предполагаются равноправными и потом каким-то образом выбирается одна из них. Обычно же принято действовать иначе: гипотезы перебираются по очереди, и, если гипотеза H_i принимается, то семейство сужается до $\mathcal{P}_i$, иначе же рассматривается гипотеза H_{i+1}.

Далее возникает вопрос: какие могут быть ошибки? Например, мы можем промахнуться: например, гипотезу H_0 приняли, хотя на самом деле она не верна. Такие ошибки разделяют на ошибки первого и второго рода.

Определение 35. Пусть сравниваются гипотезы H_0 против H_1. *Ошибкой первого рода* называется ситуация, когда H_0 отвергли, но она была верна. Аналогично, *ошибкой второго рода* называют ситуацию, когда H_0 приняли, но она была не верна.

И вопрос: какая из этих ошибок опаснее? Методологически считается, что ошибка первого рода опаснее. Предположим, что мы неверно отвергли H_0. Тогда мы вынуждены рассматривать H_1, которая заведомо неверна. В зависимости от результата проверки H_1 будет либо ошибка, либо дополнительная работа в виде проверки H_2 и оба варианта не вдохновляют. Тем самым будет проводиться много бесполезной работы. В случае ошибки второго рода мы принимаем H_0, но она неверна. В таком случае вердикт может измениться, если добавить данных, и мы сможем двинуться в сторону чего-то более правильного. Другими словами, здесь шансы совершить много лишней работы сильно меньше.

Что это означает? То, что ошибку первого рода нужно как-то заведомо избежать, то есть вероятность того, что она произойдёт, должна быть достаточно мала (не больше 0.01, например). Но для оценки вероятности ошибки нужно понимать, как мы будем проверять гипотезы.

Определение 36. Пусть $\mathcal{X}$ — выборочное пространство (множество всех возможных значений наблюдений). Тогда подмножество $S \subset \mathcal{X}$ называется *критерием* или *критическим множеством* для проверки гипотезы H_0, если правило принятия H_0 выглядит так: H_0 отвергается тогда и только тогда, когда $\boldsymbol{X} \in S$.

По сути, S есть «плохое» множество, и, если $\boldsymbol{X}$ попал туда, то H_0 считается неправильной и выбирается альтернатива. Как только появился критерий S, можно ввести вероятности ошибок первого и второго рода. Но для этого нужна функция мощности.

Определение 37. *Функцией мощности* критерия S называется

$$\beta(\mathsf{Q}, S) = \mathsf{Q}(\boldsymbol{X} \in S), \mathsf{Q} \in \mathcal{P}.$$

Далее возникает один момент: вероятность ошибки первого рода не является вероятностью, а является пачкой вероятностей. Аналогично со вторым родом.

Определение 38. Вероятностями ошибок первого рода называется набор $\{\beta(\mathsf{Q}, S) \colon \mathsf{Q} \in \mathcal{P}_0\}$.

Определение 39. Вероятностями ошибок второго рода называется набор $\{1 - \beta(\mathsf{Q}, S) \colon \mathsf{Q} \notin \mathcal{P}_0\}$.

Последний вопрос: какой критерий S будем подходить? Мы хотели построить очень хороший критерий в том смысле, что у него маленькие вероятности ошибок первого и второго рода. Это означает, что внутри критического множества функция мощности должна быть мала, а вне него должна быть велика. В идеале хочется получить нули, но это задача безнадёжна, и нужно на что-то соглашаться. Но на что? И как сказать, что один критерий лучше другого?

Для этого будем полагать, что вероятности ошибок первого рода ограничены сверху каким-то ε, который называют *уровнем значимости*. Далее, если мы зафиксировали уровень значимости, то среди всех подходящих критериев ищется тот, у которого минимальные вероятности ошибок второго рода. То, насколько получится минимизировать ошибку второго рода, сильно зависит от задачи. Хорошо себя показывают в этой истории так называемые *асимптотические критерии*: в них полагается, что у выборки растущий размер. С ними получается добиться следующего: вероятность ошибки первого рода не больше, чем ε, а вероятность второго рода стремится к нулю.

11 Лекция 11

Вернёмся к нашим гипотезам. Пусть проверяется гипотеза $\mathrm{H}_0 \colon \mathsf{P} \in \mathcal{P}_0$ против гипотезы $\mathrm{H}_1 \colon \mathsf{P} \in \mathcal{P}_1$, и есть какой-то критерий $S \subset \mathcal{X}$. Ранее мы вводили понятие уровня значимости. Формализуем его:

Определение 40. Критерий S имеет уровень значимости γ, если для любого $\mathsf{Q} \in \mathcal{P}_0$ $\beta(\mathsf{Q}, S) \leqslant \gamma$.

По сути, если γ является подходящим уровнем значимости, то и любое число, большее γ, тоже подходит. Поэтому наряду с уровнем значимости вводится минимальный уровень значимости:

Определение 41. Минимальным уровнем значимости или же размером критерия называют $\alpha(S) = \sup_{\mathsf{Q} \in \mathcal{P}_0} \beta(\mathsf{Q}, S)$.

Теперь можно ввести свойства критериев. Они называются так же, как и у оценок: несмещённость и состоятельность, но смысл у них несколько отличается. Введём их.

Определение 42. Критерий S для проверки гипотезы $\mathrm{H}_0 \colon \mathsf{P} \in \mathcal{P}_0$ против гипотезы $\mathrm{H}_1 \colon \mathsf{P} \in \mathcal{P}_1$ называется *несмещённым*,[14] если

$$\sup_{\mathsf{Q} \in \mathcal{P}_0} \beta(\mathsf{Q}, S) \leqslant \inf_{\mathsf{Q} \in \mathcal{P}_1} \beta(\mathsf{Q}, S). \tag{1}$$

Данное свойство достаточно естественно: мы хотим, чтобы на $\mathcal{P}_0$ функция мощности была маленькой, а на $\mathcal{P}_1$ — большой. Поэтому разумно попросить, чтобы значения были разделены.

Определение 43. Пусть $\{S_n\}_{n=1}^{\infty}$ — последовательность[15] критериев для проверки гипотезы $\mathrm{H}_0 \colon \mathsf{P} \in \mathcal{P}_0$ против гипотезы $\mathrm{H}_1 \colon \mathsf{P} \in \mathcal{P}_1$. Она называется *состоятельной* (или же говорят, что S_n — состоятельный критерий), если для любого $\mathsf{Q} \in \mathcal{P}_1$ $\beta(\mathsf{Q}, S_n) \to 1$ при $n \to \infty$.

По сути, это означает, что вероятности ошибки второго рода стремятся к нулю, что есть хорошо. Стоит сказать, что если асимптотический критерий не состоятелен, то он обычно бессмысленен. Но за состоятельность нужно платить, о чём расскажем позднее.

11.1 Сравнение критериев

Пусть есть два критерия S_1 и S_2. Как их можно сравнивать? Сразу же стоит сказать, что их имеет смысл сравнивать только в том случае, когда их уровни значимости[16] равны. Если это не так, то лучше взять критерий с меньшим уровнем значимости — ибо меньшая вероятность ошибки первого рода важнее, чем меньшая вероятность ошибки второго рода. Если же они равны, то сравниваем их по функции мощности.

[14] Не самое удачное название свойства, на самом деле — не понятно, что подразумевать под смещённостью.

[15] Стоит заметить, что S_n могут меняться и они живут в разных пространствах: $S_n = S_n(X_1, \ldots, X_n) \subset \mathbb{R}^n$.

[16] Стоит сказать, что размеры критерия брать не стоит, так как чем меньше уровень значимости, тем выше вероятность ошибки второго рода, и нужно найти некоторый баланс. И, что самое интересное, его можно найти.

Определение 44. Пусть S_1 и S_2 — критерии уровня значимости ε для проверки гипотезы $\mathrm{H}_0\colon \mathsf{P} \in \mathcal{P}_0$ против гипотезы $\mathrm{H}_1\colon \mathsf{P} \in \mathcal{P}_1$. Будем говорить, что S_1 (равномерное) мощнее, чем S_2, если для любого $\mathsf{Q} \in \mathcal{P}_1$ $\beta(\mathsf{Q}, S_1) \geqslant \beta(\mathsf{Q}, S_2)$, то есть вероятность ошибка второго рода для критерия S_1 равномерно меньше, чем для S_2.

Если же критерий оказывается мощнее остальных, то мы его будем называть наиболее мощным.

Определение 45. Критерий S для проверки гипотезы $\mathrm{H}_0\colon \mathsf{P} \in \mathcal{P}_0$ против гипотезы $\mathrm{H}_1\colon \mathsf{P} \in \mathcal{P}_1$ называется *равномерно наиболее мощным* критерием уровня значимости ε, если для любого другого критерия R уровня значимости ε S мощнее, чем R.

Теперь попытаемся понять, как искать равномерно наиболее мощные критерии уровня значимости ε. Самый простой вариант состоит в том, что рассматриваются только *простые* гипотезы, то есть гипотезы, в которых предлагаемые семейства содержат только одно распределение.

Рассмотрим проверку гипотезы $\mathrm{H}_0\colon \mathsf{P} = \mathsf{P}_0$ против гипотезы $\mathrm{H}_1\colon \mathsf{P} = \mathsf{P}_1$. Далее, предположим, что P_0 и P_1 имеют плотности $p_0(x)$ и $p_1(x)$ по одной и той же мере μ. По сути, оба распределения приходят из доминируемого семейства: например, различить пуассоновское и экспоненциальное распределение крайне просто — считаем, что экспоненциальное, если получили нецелое число, и пуассовское иначе. Такой критерий никогда не ошибается.

Пусть $\lambda > 0$. Введём следующее множество $S_\lambda = \{x\colon p_1(x) - \lambda p_0(x) \geqslant 0\}$ и скажем, что это критерий. Логика за этим критерием понятна: мы отвергаем H_0, если $p_1(x) > \lambda p_0(x)$. Это неравенство можно понимать в том смысле, что p_1 более правдоподобна. Теперь докажем одно интересное свойство, связанное с этим критерием.

Лемма (Нейман, Пирсон)**.** *Пусть критерий R таков, что $\mathsf{P}_0(\boldsymbol{X} \in R) \leqslant \mathsf{P}_0(\boldsymbol{X} \in S_\lambda)$. Тогда*

a) $\mathsf{P}_1(\boldsymbol{X} \in R) \leqslant \mathsf{P}_1(\boldsymbol{X} \in S_\lambda)$,

b) $\mathsf{P}_0(\boldsymbol{X} \in S_\lambda) \leqslant \mathsf{P}_1(\boldsymbol{X} \in S_\lambda)$.

Доказательство. Для начала заметим, что для любого x[17]

$$\mathbf{1}_{x\in R}(p_1(x) - \lambda p_0(x)) \leqslant \mathbf{1}_{x\in S_\lambda}(p_1(x) - \lambda p_0(x)). \tag{2}$$

Проинтегрируем неравенства по мере μ по всему выборочному пространству:

$$\int_R (p_1(x) - \lambda p_0(x))\mu(\mathrm{d}x) \leqslant \int_{S_\lambda} (p_1(x) - \lambda p_0(x))\mu(\mathrm{d}x), \tag{3}$$

$$\mathsf{P}_1(\boldsymbol{X} \in R) - \lambda \mathsf{P}_0(\boldsymbol{X} \in R) \leqslant \mathsf{P}_1(\boldsymbol{X} \in S_\lambda) - \lambda \mathsf{P}_0(\boldsymbol{X} \in S_\lambda), \tag{4}$$

$$\mathsf{P}_1(\boldsymbol{X} \in R) - \mathsf{P}_1(\boldsymbol{X} \in S_\lambda) \leqslant \lambda(\mathsf{P}_0(\boldsymbol{X} \in R) - \mathsf{P}_0(\boldsymbol{X} \in S_\lambda)) \leqslant 0. \tag{5}$$

Для второго пункта нужно рассмотреть два случая:

- Пусть $\lambda \geqslant 1$. Тогда для $x \in S_\lambda$ выполнено следующее неравенство: $p_1(x) \geqslant \lambda p_0(x) \geqslant p_0(x)$. Следовательно, $\mathsf{P}_0(\boldsymbol{X} \in S_\lambda) \leqslant \mathsf{P}_1(\boldsymbol{X} \in S_\lambda)$.
- Пусть $\lambda \in (0, 1)$. Тогда для $x \notin S_\lambda$ $p_1(x) < \lambda p_0(x) < p_0(x)$. Следовательно, $\mathsf{P}_1(\boldsymbol{X} \in \overline{S_\lambda}) \leqslant \mathsf{P}_0(\boldsymbol{X} \in \overline{S_\lambda})$ и $\mathsf{P}_0(\boldsymbol{X} \in S_\lambda) \leqslant \mathsf{P}_1(\boldsymbol{X} \in S_\lambda)$.

Тем самым для любого $\lambda > 0$ $\mathsf{P}_0(\boldsymbol{X} \in S_\lambda) \leqslant \mathsf{P}_1(\boldsymbol{X} \in S_\lambda)$. □

У данной леммы есть два приятных следствия:

Следствие. Если $\lambda > 0$ удовлетворяет уравнению $\mathsf{P}_0(\boldsymbol{X} \in S_\lambda) = \varepsilon$, то S_λ является равномерно наиболее мощным критерием уровня значимости ε для проверки гипотезы $\mathrm{H}_0\colon \mathsf{P} = \mathsf{P}_0$ против гипотезы $\mathrm{H}_1\colon \mathsf{P} = \mathsf{P}_1$.

Следствие. S_λ — несмещённый критерий.

Определение 46. Пусть $\{\mathsf{P}_\theta \mid \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}$ — это доминируемое семейство распределений с одномерным параметром и плотностью $p_\theta(x)$ по мере μ. Говорят, что это семейство имеет *монотонное отношение правдоподобия по статистике* $T(\boldsymbol{X})$, если для любых $\theta_1 < \theta_2$, $\theta_1, \theta_2 \in \Theta$

$$\frac{p_{\theta_2}(\boldsymbol{X})}{p_{\theta_1}(\boldsymbol{X})} = \psi_{\theta_1,\theta_2}(T(\boldsymbol{X})), \tag{6}$$

где ψ_{θ_1,θ_2} — это монотонная функция и её вид монотонности одинаков для всех $\theta_1 < \theta_2$.

[17] Данное рассуждение уже было в доказательстве теоремы Бахадура.

Лемма. *Пусть ψ_{θ_1,θ_2} всегда неубывает по $T(\boldsymbol{X})$. Тогда для всех $c \in \mathbb{R}$ и для всех $\theta_1 < \theta_2$, $\theta_1, \theta_2 \in \Theta$ выполнено, что*

$$\mathsf{P}_{\theta_1}(T(\boldsymbol{X}) \geqslant c) \leqslant \mathsf{P}_{\theta_2}(T(\boldsymbol{X}) \geqslant c). \tag{7}$$

Доказательство. Считаем, что $\psi_{\theta_1,\theta_2} \geqslant 0$ и не убывает на $\mathbb{R}$. Введём множество $D = \{\boldsymbol{x} \colon T(\boldsymbol{x}) \geqslant c\}$, где c такое, что $\psi_{\theta_1,\theta_2}(c) \geqslant 1$. Тогда для любого $\boldsymbol{x} \in D$

$$p_{\theta_2}(\boldsymbol{x}) = \psi_{\theta_1,\theta_2}(T(\boldsymbol{x}))p_{\theta_1}(\boldsymbol{x}) \geqslant \psi_{\theta_1,\theta_2}(c)p_{\theta_1}(\boldsymbol{x}) \geqslant p_{\theta_1}(\boldsymbol{x}). \tag{8}$$

Интегрируем это по мере μ по множеству D:

$$\int_D p_{\theta_2}(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) \geqslant \int_D p_{\theta_1}(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) \tag{9}$$

$$\mathsf{P}_{\theta_2}(T(\boldsymbol{X}) \geqslant c) \geqslant \mathsf{P}_{\theta_1}(T(\boldsymbol{X}) \geqslant c). \tag{10}$$

Если же брать c такое, что $\psi_{\theta_1,\theta_2}(c) \in [0,1]$, то рассматриваем множество $\overline{D}$ и получаем то же самое. □

Теорема 26 (о монотонном отношении правдоподобий)**.** *Допустим, что проверяется гипотеза* $\mathrm{H}_0 \colon \theta \leqslant \theta_0$ *против гипотезы* $\mathrm{H}_1 \colon \theta > \theta_0$. *Если $c \in \mathbb{R}$ удовлетворяет соотношению $\mathsf{P}_{\theta_0}(T(\boldsymbol{X}) \geqslant c) = \gamma$, то критерий $S = \{\boldsymbol{x} \colon T(\boldsymbol{x}) \geqslant c\}$ является равномерно наиболее мощным критерием уровня доверия γ для проверки гипотезы* H_0 *против гипотезы* H_1.

Доказательство. Для начала возьмём какое-нибудь $\theta < \theta_0$. Тогда по лемме

$$\mathsf{P}_{\theta}(T(\boldsymbol{X}) \geqslant c) \leqslant \mathsf{P}_{\theta_0}(T(\boldsymbol{X}) \geqslant c) = \gamma. \tag{11}$$

Из этого следует, что критерий S имеет уровень значимости γ. Проверим то, что он равномерно наиболее мощный. Пусть R — любой другой критерий с уровнем мощности γ. Возьмём какое-нибудь $\theta_1 > \theta_0$. Нужно доказать, что

$$\mathsf{P}_{\theta_1}(\boldsymbol{X} \in R) \leqslant \mathsf{P}_{\theta_1}(X \in S). \tag{12}$$

Согласно лемме Неймана-Пирсона равномерно наиболее мощный критерий уровня значимости γ для проверки гипотезы $\mathrm{H}_0 \colon \theta = \theta_0$ против гипотезы $\mathrm{H}_1 \colon \theta = \theta_1$ имеет вид

$$S_\lambda = \{\boldsymbol{x} \colon p_{\theta_1}(x) - \lambda p_{\theta_0}(x) \geqslant 0\}, \text{ где } \mathsf{P}_{\theta_0}(\boldsymbol{X} \in S_\lambda) = \gamma. \tag{13}$$

Однако

$$S_\lambda = \left\{\boldsymbol{x} \colon \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \geqslant \lambda\right\} = \{\boldsymbol{x} \colon \psi_{\theta_1,\theta_2}(T(\boldsymbol{X})) \geqslant \lambda\} = \{\boldsymbol{x} \colon T(\boldsymbol{X}) \geqslant \tilde{\lambda}\}. \tag{14}$$

Заметим, что $\tilde{\lambda} = c$ подходит под условие. Следовательно, для любого $\theta_1 > \theta_0$

$$\mathsf{P}_{\theta_1}(\boldsymbol{X} \in R) \leqslant \mathsf{P}_{\theta_1}(\boldsymbol{X} \in S_\lambda) = \mathsf{P}_{\theta_1}(T(\boldsymbol{X}) \geqslant c) = \mathsf{P}_{\theta_1}(\boldsymbol{X} \in S). \tag{15}$$

Следовательно, S является равномерно наиболее мощным критерием. □

Следствие. S является равномерно наиболее мощным критерием уровня доверия γ для проверки гипотезы $\mathrm{H}_0 \colon \theta = \theta_0$ против гипотезы $\mathrm{H}_1 \colon \theta > \theta_0$.

Теперь рассмотрим пример задачи на эту теорему.

Задача 25. *Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка из распределения Бернулли* $\mathrm{Bern}(\theta)$, $\theta \in (0,1)$. *Найти равномерно наиболее мощный критерий уровня значимости γ для проверки*

a) гипотеза $\mathrm{H}_0 \colon \theta \leqslant \theta_0$ *против гипотезы* $\mathrm{H}_1 \colon \theta > \theta_0$*;*

b) гипотеза $\mathrm{H}_0 \colon \theta \geqslant \theta_0$ *против гипотезы* $\mathrm{H}_1 \colon \theta < \theta_0$.

Доказательство. Для начала запишем функцию правдоподобия для выборки:

$$p_\theta(\boldsymbol{X}) = \prod_{i=1}^n p_\theta(X_i) = \prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i} = \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}. \tag{16}$$

Вычислим отношение правдоподобия для $\theta_1 \neq \theta_0$:

$$\frac{p_{\theta_1}(\boldsymbol{X})}{p_{\theta_0}(\boldsymbol{X})} = \left(\frac{\theta_1}{\theta_0}\right)^{\sum_{i=1}^n X_i} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-\sum_{i=1}^n X_i} \tag{17}$$

$$= \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n} \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1}\right)^{\sum_{i=1}^n X_i}. \tag{18}$$

Теперь будем искать критерии.

a) Если θ_1 из альтернативы, то $\theta_1 > \theta_0$ и $\theta_1(1-\theta_0) > \theta_1(1-\theta_0)$. Следовательно, при росте $\sum_{i=1}^n X_i$ отношение правдоподобия возрастает. Тогда искомая статистика $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$ и критерий имеет вид $S = \{\boldsymbol{x}\colon \sum_{i=1}^n x_i \geqslant c\}$. Теперь нужно найти c:

$$\mathsf{P}_{\theta_0}(\sum_{i=1}^n x_i \geqslant c) = \gamma \implies c = u_{1-\gamma}, \tag{19}$$

где $u_{1-\gamma}$ есть $(1-\gamma)$-квантиль биномиального распределения $\mathrm{Bin}(n, \theta_0)$.

b) В таком случае $\theta_1 < \theta_0$ и отношение правдоподобия уменьшается с ростом $\sum_{i=1}^n X_i$. Здесь ситуация ровно такая же, только нужно инвертировать знаки, то есть критерий имеет вид $S = \{\boldsymbol{x}\colon \sum_{i=1}^n x_i \leqslant c\}$ и $c = u_\gamma$.

□

12 Лекция 12

12.1 Критерий согласия

Рассмотрим так называемые критерии согласия. Такие критерии предлагают согласиться или же не согласиться с чем-то, что имеет простую природу.

Определение 47. Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — функция распределения с неизвестной функцией распределения $F(x)$. Далее, введём гипотезы $\mathrm{H}_0\colon F = F_0$ и $\mathrm{H}_1\colon F \neq F_0$. Критерий для проверки гипотезы H_0 против альтернативы H_1 называется *критерием согласия*.

Что мы хотели бы видеть от такого критерия? В идеале увидеть что-то наподобие равномерно наиболее мощного критерия. Но эта задача почти всегда полностью безнадёжна. Даже для хороших распределений, для которых выполнено свойство монотонного отношения правдоподобий, такого критерия может не быть. Например, для распределений $\mathrm{Bin}(1,\theta)$, $\mathrm{Exp}(\theta)$, $\mathcal{N}(\theta, 1)$ найти равномерно наиболее мощный критерий для проверки гипотезы $\mathrm{H}_0\colon \theta = \theta_0$ против альтернативы $\mathrm{H}_1\colon \theta \neq \theta_0$ не получится. С другой стороны, для некоторых распределений, для которых свойство монотонного отношения правдоподобий не выполнено, можно найти равномерно наиболее мощный критерий (например, $\mathrm{U}[0,\theta]$). Тем самым такая задача в некоторой степени осмысленна и для некоторых, казалось бы, сложных распределений её можно решить.

Для того, чтобы её решить, пока что приходится всё делать по-честному. Поэтому перейдём задаче поиска *асимптотических критериев*. Такая постановка осмысленна: если не можем решить точно, то хотя бы решим приближенно. Пусть $\boldsymbol{X} = (X_1, \ldots, X_n)$ — выборка растущего размера. Мы хотели бы получить последовательность критериев $S_n = S_n(X_1, \ldots, X_n)$ для проверки гипотезы $\mathrm{H}_0\colon \theta = \theta_0$ против альтернативы $\mathrm{H}_1\colon \theta \neq \theta_0$ такую, что

- Предел ошибки вероятности ошибки первого рода стремится к $\gamma \in (0,1)$: $\mathsf{P}_{\theta_0}(\boldsymbol{X} \in S_n) \to \gamma$.
- Последовательность критериев будет состоятельна: для любого $\theta \neq \theta_0$ $\mathsf{P}_\theta(\boldsymbol{X} \in S_n) \to 1$.

Не смотря на то, что условия кажутся сложными для выполнения, оказывается, что решения есть. Рассмотрим один из таких критериев — критерий хи-квадрат Пирсона. Пусть есть выборка $\boldsymbol{X} = (X_1, \ldots, X_n)$ из распределения P с конечным числом значений $\{a_1, \ldots, a_m\}$:

$$p_j = \mathsf{P}(X_1 = a_j) > 0, \quad j = 1, \ldots, m, \quad \sum_{i=1}^m p_i = 1. \tag{20}$$

Введём вектор вероятностей $\boldsymbol{p} = (p_1, \ldots, p_m)$. Гипотеза $\mathrm{H}_0\colon F = F_0$ бужет сводиться к равенству векторов вероятностей, то есть задача согласия сводится к проверке гипотезы $\mathrm{H}_0\colon \boldsymbol{p} = \boldsymbol{p}^0$ против альтернативы $\mathrm{H}_1\colon \boldsymbol{p} \neq \boldsymbol{p}^0$. Для проверки данной гипотезы вводится *статистика хи-квадрат*:

$$\hat{\chi}^2_n(\boldsymbol{X}) = \sum_{j=1}^m \frac{(\mu_j - np_j^0)^2}{np_j^0}, \text{ где } \mu_j = \sum_{i=1}^n \mathbf{1}_{X_i = a_j}. \tag{21}$$

Критерий же устроен следующим образом: пусть $u_{1-\gamma}$ — $(1-\gamma)$-квантиль распределения χ^2_{m-1}. Тогда гипотеза H_0 отвергается тогда и только тогда, когда $\hat{\chi}^2_n(\boldsymbol{X}) > u_{1-\gamma}$. Теперь докажем, что это хороший асимптотический критерий.

Утверждение. Критерий хи-квадрат состоятелен.

Доказательство. Для начала немного перепишем статистику хи-квадрат:

$$\hat{\chi}^2_n(\boldsymbol{X}) = n\sum_{j=1}^{m}\frac{1}{p_j^0}\left(\frac{\mu_j}{n} - p_j^0\right)^2. \tag{22}$$

Далее, если $\boldsymbol{p} \neq \boldsymbol{p}^0$, то найдётся индекс $i \in \{1, 2, \ldots, m\}$ такой, что $p_j \neq p_j^0$. Однако по усиленному закону больших чисел $\mu_j/n \to p_j$ почти наверное. Отсюда следует, что $\hat{\chi}^2_n(\boldsymbol{X})$ линейно стремится к бесконечности, так как

$$\frac{\hat{\chi}^2_n(\boldsymbol{X})}{n} \xrightarrow{\text{п.н.}} \sum_{j=1}^{m}\frac{(p_j - p_j^0)^2}{p_j^0} > 0. \tag{23}$$

Тем самым критерий состоятелен: $\mathsf{P}(\hat{\chi}^2_n(\boldsymbol{X}) > u_{1-\gamma}) \to 1$. □

Первое свойство будет получаться из следующей теоремы:

Теорема 27 (Пирсона)**.** *Если выполнена гипотеза* H_0*, то имеет место следующая сходимость по распределению:*

$$\hat{\chi}^2_n(\boldsymbol{X}) \xrightarrow[n\to\infty]{d} \chi^2_{m-1}. \tag{24}$$

Доказательство. Пусть $\boldsymbol{p} = \boldsymbol{p}^0$. Для всех $i = 1, \ldots, n$ введём случайный вектор $\boldsymbol{Y}_i = (\mathbf{1}_{X_i=a_1}, \ldots, \mathbf{1}_{X_i=a_m})$. Понятно, что такие векторы независимы и одинаково распределены, причём $\mathsf{E}[\boldsymbol{Y}_i] = \boldsymbol{p}^0$. Далее заметим, что

$$\frac{\boldsymbol{Y}_1 + \ldots + \boldsymbol{Y}_n}{n} = \frac{1}{n}\begin{pmatrix}\mu_1\\ \ldots \\ \mu_m\end{pmatrix}. \tag{25}$$

Следовательно, по многомерной центральной предельной теореме

$$\sqrt{n}\left(\frac{\boldsymbol{Y}_1 + \ldots + \boldsymbol{Y}_n}{n} - \boldsymbol{p}^0\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{26}$$

где $\boldsymbol{\Sigma} = \mathsf{D}[\boldsymbol{Y}_i]$. Заметим, что

$$\mathsf{cov}(\mathbf{1}_{X_1=a_i}, \mathbf{1}_{X_1=a_j}) = \mathsf{P}(X_1 = a_i, X_1 = a_j) - p_i^0 p_j^0 = \begin{cases} p_i^0 - p_i^0 p_j^0, & i = j \\ -p_i^0 p_j^0, & i \neq j \end{cases} \tag{27}$$

Тогда запишем матрицу ковариаций следующим образом: $\boldsymbol{\Sigma} = \mathbf{B} - \boldsymbol{p}^0(\boldsymbol{p}^0)^\top$, где $\mathbf{B} = \mathrm{diag}(\boldsymbol{p}^0)$. Теперь наша цель — поворачать вектор в ЦПТ так, чтобы матрица ковариаций стала разумной: нам нужно получить распределение хи-квадрат, поэтому желательно получить матрицу, чем-то похожую на матрицу стандартного нормального распределения). Пусть

$$\boldsymbol{\xi}_n^{'} = \sqrt{n}\left(\frac{\boldsymbol{Y}_1 + \ldots + \boldsymbol{Y}_n}{n} - \boldsymbol{p}^0\right). \tag{28}$$

Тогда по центральной предельной теореме

$$(\sqrt{\mathbf{B}})^{-1}\boldsymbol{\xi}_n^{'} \xrightarrow[n\to\infty]{d} (\sqrt{\mathbf{B}})^{-1}\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{0}, (\sqrt{\mathbf{B}})^{-1}\boldsymbol{\Sigma}(\sqrt{\mathbf{B}})^{-1}). \tag{29}$$

Новая матрица ковариаций равна $\mathbf{I}_m - \boldsymbol{z}\boldsymbol{z}^\top$, где $\boldsymbol{z} = (\sqrt{\mathbf{B}})^{-1}\boldsymbol{p}^0 = (\sqrt{p_1^0}, \ldots, \sqrt{p_m^0})$. Заметим, что вектор $\boldsymbol{z}$ единичный, а это означает, что его можно перевести поворотами в базисный вектор. Рассмотрим ортогональную матрицу $\mathbf{C}$, у которой первая строка равна $\boldsymbol{z}$, а остальные равны чему угодно. Тогда по теореме о наследовании сходимости

$$\boldsymbol{\xi}_n^{''} = \mathbf{C}(\sqrt{\mathbf{B}})^{-1}\xi_n^{'} \xrightarrow[n\to\infty]{d} \mathcal{N}(\mathbf{0}, \mathbf{C}(\mathbf{I}_m - \boldsymbol{z}\boldsymbol{z}^\top)\mathbf{C}^\top). \tag{30}$$

Однако $\mathbf{C}\boldsymbol{z} = (1, 0, \ldots, 0)^\top$, а $\mathbf{C}\mathbf{C}^\top = \mathbf{I}_m$. Поэтому $\mathbf{C}(\mathbf{I}_m - \boldsymbol{z}\boldsymbol{z}^\top)\mathbf{C}^\top = \mathrm{diag}(0, 1, \ldots, 1) = \mathbf{I}_m^{'}$. Снова по теореме о наследовании сходимости

$$\|\boldsymbol{\xi}_n^{''}\|^2 \xrightarrow[n\to\infty]{d} \|\mathcal{N}(\mathbf{0}, \mathbf{I}_m^{'})\|^2 \sim \chi^2_{m-1}. \tag{31}$$

Но, как известно, помножение вектора на ортогональную матрицу не изменяет его норму. Тогда

$$\|(\sqrt{\mathbf{B}})^{-1}\boldsymbol{\xi}_n^{'}\|^2 \xrightarrow[n\to\infty]{d} \chi^2_{m-1}. \tag{32}$$

Осталось заметить, что

$$(\sqrt{\mathbf{B}})^{-1}\boldsymbol{\xi}_n^{'} = (\sqrt{\mathbf{B}})^{-1}\sqrt{n}\begin{pmatrix}(\mu_1 - np_1^0)/n\\ \dots \\ (\mu_m - np_m^0)/n\end{pmatrix} = \begin{pmatrix}(\mu_1 - np_1^0)/\sqrt{np_1^0}\\ \dots \\ (\mu_m - np_m^0)/\sqrt{np_m^0}\end{pmatrix} \tag{33}$$

Тем самым $\hat{\chi}^2_n(\boldsymbol{X}) = \|(\sqrt{\mathbf{B}})^{-1}\boldsymbol{\xi}_n^{'}\|^2$. □

Насколько этот метод применим? На практике считается, что должно быть следующее ограничение: $np_j^0 \geqslant 5$ для всех $j = 1, \dots, m$.

12.2 Параметрический хи-квадрат

Метод достаточно мощный, хоть мы и пожертвовали тем, что вероятность ошибки первого рода не равна γ, а только стремится к ней. Что самое ироничное: отказ от точности дал некоторые хорошие свойства. Например, его можно применять, когда модель параметрическая. В таком случае критерий называется *параметрическим критерием хи-квадрат*.

Пусть $\boldsymbol{X} = (X_1, \dots, X_n)$ — выборка из неизвестного распределения со значениями $\{a_1, \dots, a_m\}$:

$$p_j(\theta) = \mathsf{P}_\theta(X_1 = a_j), \quad \theta \in \Theta \tag{34}$$

В данном случае гипотеза и альтернатива формулируются следующим образом: $\mathrm{H}_0\colon \boldsymbol{p} \in \{\mathsf{P}_\theta, \theta \in \Theta\}$ против $\mathrm{H}_1\colon \boldsymbol{p} \notin \{\mathsf{P}_\theta, \theta \in \Theta\}$. Теперь хотелось бы составить статистику хи-квадрат, но особо не получится, так как $\boldsymbol{p}^0$ больше нет. Можно ли его чем-то заменить? Заменим на оценку:

$$\hat{\chi}^2_n(\boldsymbol{X}) = \sum_{i=1}^m \frac{(\mu_j - n\hat{p}_j(\boldsymbol{X}))^2}{n\hat{p}_j(\boldsymbol{X})}. \tag{35}$$

Какую оценку взять? Самым разумным выбором является оценка максимального правдоподобия: $\hat{p}_j(\boldsymbol{X}) = p_j(\hat{\theta}(\boldsymbol{X}))$, где $\hat{\theta}(\boldsymbol{X})$ — оценка максимального правдоподобия для θ. Оказывается, что в таком случае можно доказать результат, похожий на теорему Пирсона: если $\boldsymbol{\theta} \in \mathbb{R}^k$, $k < m-1$, то при выполнении определённых условий

$$\hat{\chi}^2_n(\boldsymbol{X}) \xrightarrow[n\to\infty]{d} \chi^2_{m-1-k}. \tag{36}$$

Данную теорему мы докажем на следующей лекции.

13 Лекция 13

Вернёмся к параметрическому критерию хи-квадрат. Ранее мы сказали, что при выполнении определённых условий будет выполнен аналог теоремы Пирсона. Сформулируем их:

a) $\Theta \subseteq \mathbb{R}^s$, $s < m$ — открытое множество.

b) Для любого значения параметра все вероятности отделены от нуля: $p_i(\boldsymbol{\theta}) \geqslant c^2 > 0$.

c) Будем считать, что $\frac{\partial p_i(\boldsymbol{\theta})}{\partial \theta_j}$ и $\frac{\partial^2 p_i(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}$ непрерывны на всём Θ.

d) Матрица $\mathbf{D} = \|\frac{\partial p_i(\boldsymbol{\theta})}{\partial \theta_j}\|_{i,j=1}^{m,s}$ имеет ранг s для любого $\theta \in \Theta$.

Теперь можно сформулировать теорему про параметрический критерий хи-квадрат.

Теорема 28. *Пусть выполнены условия модели. Введём следующую систему уравнений:*

$$\sum_{i=1}^m \frac{\mu_i}{p_i(\boldsymbol{\theta})}\frac{\partial p_i(\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1, \dots, s, \quad \mu_j = \sum_{i=1}^n \mathbf{1}_{X_i = a_j}. \tag{37}$$

Если верна гипотеза H_0*, то с вероятностью, стремящейся к 1, данная система имеет единственное решение* $\hat{\theta}(\boldsymbol{X})$ *такое, что* $\hat{\theta}(\boldsymbol{X})$ *сходится по вероятности к истинному значению параметра* $\boldsymbol{\theta}$ *и*

$$\hat{\chi}^2_n(\boldsymbol{X}) = \sum_{j=1}^m \frac{(\mu_j - np_j(\hat{\theta}(\boldsymbol{X})))^2}{np_j(\hat{\theta}(\boldsymbol{X}))} \xrightarrow[n\to\infty]{d} \chi^2_{m-1-s}. \tag{38}$$

Доказательство. Пусть $\boldsymbol{\theta}_0$ — истинное значение параметра $\boldsymbol{\theta}$. Далее, для удобства скажем, что

$$p_i(\boldsymbol{\theta}_0) = p_i^0, \quad \left.\frac{\partial p_i(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \left(\frac{\partial p_i}{\partial \theta_j}\right)_0.$$

Теперь введём матрицы

$$\mathbf{P}_0 = \operatorname{diag}\left(\frac{1}{\sqrt{p_1^0}}, \ldots, \frac{1}{\sqrt{p_m^0}}\right) \in \mathbb{R}^{m\times m}, \quad \mathbf{D}_0 = \mathbf{D}(\boldsymbol{\theta}_0) = \left\|\left(\frac{\partial p_i}{\partial \theta_j}\right)_0\right\|_{i,j=1}^{m,s} \in \mathbb{R}^{m\times s}, \quad \mathbf{B}_0 = \mathbf{P}_0\mathbf{D}_0.$$

Теперь выразим систему из условия через данные матрицы. Введём ещё одно обозначение:

$$\begin{aligned}\omega_j(\boldsymbol{\theta}) = \sum_{i=1}^m \frac{\mu_i - np_i^0}{n}\left(\frac{1}{p_i}\frac{\partial p_i(\boldsymbol{\theta})}{\partial \theta_j} - \frac{1}{p_i^0}\left(\frac{\partial p_i}{\partial \theta_j}\right)_0\right) - \sum_{i=1}^m (p_i - p_i^0)\left(\frac{1}{p_i}\frac{\partial p_i(\boldsymbol{\theta})}{\partial \theta_j} - \frac{1}{p_i^0}\left(\frac{\partial p_i}{\partial \theta_j}\right)_0\right)\\ - \sum_{i=1}^m \frac{1}{p_i^0}\left(\frac{\partial p_i}{\partial \theta_j}\right)_0 \left(p_i - p_i^0 - \sum_{k=1}^s \left(\frac{\partial p_i}{\partial \theta_k}\right)_0 (\theta_k - \theta_k^0)\right).\end{aligned}$$

Зачем вводить это безобразие? Для того, чтобы доказать, что оно мало. Воспользуемся системой из условия и тем, что сумма p_i равна единице:

$$\omega_j(\boldsymbol{\theta}) = -\sum_{i=1}^m \frac{\mu_i - np_i^0}{n}\left(\frac{\partial p_i}{\partial \theta_j}\right)_0 + \sum_{i=1}^m\sum_{k=1}^s \frac{1}{p_i^0}\left(\frac{\partial p_i}{\partial \theta_j}\right)_0\left(\frac{\partial p_i}{\partial \theta_k}\right)_0 (\theta_k - \theta_k^0).$$

Несложно показать, что если выполнено такое равенство, то верна система из условия. Теперь перепишем это равенство в матричном виде. Введём вектор $\boldsymbol{Y}_j$ по следущему правилу: $Y_j = (\mu_j - np_j^0)/\sqrt{np_j^0}$. Тогда (проверьте!)

$$\boldsymbol{\omega}(\boldsymbol{\theta}) = -\frac{1}{\sqrt{n}}\mathbf{B}_0^\top \boldsymbol{Y} + \mathbf{B}_0^\top\mathbf{B}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

Далее, заметим, что матрица $\mathbf{B}_0^\top\mathbf{B}_0$ обратима, так как $\mathbf{D}_0$ имеет ранг s. В таком случае

$$\boldsymbol{\theta} - \boldsymbol{\theta}_0 = \frac{1}{\sqrt{n}}(\mathbf{B}_0^\top\mathbf{B}_0)^{-1}\mathbf{B}_0^\top\boldsymbol{Y} + (\mathbf{B}_0^\top\mathbf{B}_0)^{-1}\boldsymbol{\omega}(\boldsymbol{\theta}).$$

Теперь нужно доказать несколько фактов.

Лемма. *Пусть $\lambda = \lambda(n) \to \infty$, $\lambda^2 = o(n)$. Тогда с вероятностью $1 - \lambda^{-2}$ для всех $i \in \{1, \ldots, m\}$*

$$|Y_i| \leqslant \frac{\lambda}{c}.$$

Доказательство. Заметим, что $\mu_i \sim \operatorname{Bin}(n, p_i^0)$. Тогда по неравенству Чебышева

$$\mathsf{P}_{\boldsymbol{\theta}_0}\left(\left|\mu_i - np_i^0\right| \geqslant \lambda\sqrt{n}\right) \leqslant \frac{\mathsf{D}_{\boldsymbol{\theta}_0}[\mu_i]}{n\lambda^2} \leqslant \frac{p_i^0}{\lambda^2}.$$

Следовательно,

$$\mathsf{P}_{\boldsymbol{\theta}_0}\left(\exists i\colon \left|\mu_i - np_i^0\right| \geqslant \lambda\sqrt{n}\right) \leqslant \sum_{i=1}^m \mathsf{P}_{\boldsymbol{\theta}_0}\left(\left|\mu_i - np_i^0\right| \geqslant \lambda\sqrt{n}\right) \leqslant \sum_{i=1}^m \frac{p_i^0}{\lambda^2} = \frac{1}{\lambda^2}.$$

Тогда, с вероятностью не меньше $1 - \lambda^{-2}$ для всех i

$$|Y_i| = \frac{|\mu_i - np_i^0|}{\sqrt{np_i^0}} \leqslant \frac{\lambda\sqrt{n}}{\sqrt{np_i^0}} \leqslant \frac{\lambda}{c}.$$ □

Лемма.

$$|\omega_j(\boldsymbol{\theta}') - \omega_j(\boldsymbol{\theta}'')| \leqslant K_1\|\boldsymbol{\theta}' - \boldsymbol{\theta}''\|\left(\|\boldsymbol{\theta}' - \boldsymbol{\theta}_0\| + \|\boldsymbol{\theta}'' - \boldsymbol{\theta}_0\| + \frac{\lambda}{\sqrt{n}}\right),$$

где K_1 — некоторая абсолютная константа.

Доказательство. Оставляется читателю в качестве упражнения.[18] *Указание:* Рассмотрим разность (сразу учтём, что p_i суммируются в 1):

$$\begin{aligned} w_j(\boldsymbol{\theta}') - w_j(\boldsymbol{\theta}'') = &\sum_{i=1}^{m} \frac{\mu_i - np_i^0}{n} \left(\frac{1}{p_i(\boldsymbol{\theta}')} \frac{\partial p_i(\boldsymbol{\theta}')}{\partial \theta_j'} - \frac{1}{p_i(\boldsymbol{\theta}'')} \frac{\partial p_i(\boldsymbol{\theta}'')}{\partial \theta_j''} \right) \\ &+ \sum_{i=1}^{m} \frac{p_i(\boldsymbol{\theta}') - p_i(\boldsymbol{\theta}'')}{p_i^0} \left(\frac{\partial p_i}{\partial \theta_j} \right)_0 + \sum_{i=1}^{m} p_i^0 \left(\frac{1}{p_i(\boldsymbol{\theta}')} \frac{\partial p_i(\boldsymbol{\theta}')}{\partial \theta_j'} - \frac{1}{p_i(\boldsymbol{\theta}'')} \frac{\partial p_i(\boldsymbol{\theta}'')}{\partial \theta_j''} \right) \\ &- \sum_{i=1}^{m} \frac{1}{p_i^0} \left(\frac{\partial p_i}{\partial \theta_j} \right)_0 \left(p_i(\boldsymbol{\theta}') - p_i(\boldsymbol{\theta}'') - \sum_{k=1}^{s} \left(\frac{\partial p_i}{\partial \theta_k} \right)_0 (\theta_k' - \theta_k'') \right). \end{aligned}$$

Далее, воспользуйтесь разложением $p_i(\boldsymbol{\theta}')$ и $p_i(\boldsymbol{\theta}'')$ в ряд Тейлора до второго порядка в точке $\boldsymbol{\theta}_0$. □

Теперь осталось построить саму оценку. Рассмотрим уравнение

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}} (\mathbf{B}_0^\top \mathbf{B}_0)^{-1} \mathbf{B}_0^\top \boldsymbol{Y} + (\mathbf{B}_0^\top \mathbf{B}_0)^{-1} \boldsymbol{\omega}(\boldsymbol{\theta}).$$

Заметим, что $\boldsymbol{\omega}(\boldsymbol{\theta}_0) = 0$. Далее, определим последовательность приближений $(\hat{\boldsymbol{\theta}}_l, l \in \mathbb{N})$ по следущему правилу:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_1 &= \boldsymbol{\theta}_0 + \frac{1}{\sqrt{n}} (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B} \boldsymbol{Y}, \\ \hat{\boldsymbol{\theta}}_l &= \hat{\boldsymbol{\theta}}_1 + (\mathbf{B}^\top \mathbf{B})^{-1} \boldsymbol{\omega}(\hat{\boldsymbol{\theta}}_{l-1}). \end{aligned}$$

Заметим, что $\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0\| \leqslant K_2 \lambda / \sqrt{n}$ для какой-то константы K_2. □

14 Лекция 14

15 Лекция 15

15.1 Слабая сходимость

Определение 48. *Метрическим пространством* будем называть (X, d), где X есть некоторое множество, а $d\colon X \times X \mapsto \mathbb{R}_+$ — функция, называемая *метрикой*, и обладающая следующими свойствами:

a) Аксиома тождества: $d(x, x) = 0$ тогда и только тогда, когда $x = 0$,

b) Аксиома симметрии: $d(x, y) = d(y, x)$,

c) Неравенство треугольника: $d(x, z) \leqslant d(x, y) + d(y, z)$.

Определение 49. Пусть (S, ρ) — метрическое пространство. Минимальную сигма-алгебру, порождённую открытыми шарами в S, будем называть *борелевской сигма-алгеброй* $\mathcal{B}(S)$.

Определение 50. Пусть задано метрическое пространство (S, ρ) и последовательность $\{\mathsf{Q}_n\}_{n=1}^{\infty}$ вероятностных мер на S. Будем говорить, что Q_n *слабо сходятся* к вероятностной мере Q на S, если для любой ограниченной непрерывной функции $f\colon S \mapsto \mathbb{R}$

$$\lim_{n \to \infty} \int_S f(x) \mathsf{Q}_n(\mathrm{d}x) = \int_S f(x) \mathsf{Q}(\mathrm{d}x). \tag{39}$$

Обозначение: $\mathsf{Q}_n \xrightarrow{W} \mathsf{Q}$.

Теорема 29 (Александров)**.** *Пусть* $\{\mathsf{Q}_n\}_{n=1}^{\infty}$ *и* Q *— вероятностные меры на метрическом пространстве* (S, ρ)*. Тогда следующие утверждения эквивалентны:*

a) $\mathsf{Q}_n \xrightarrow{W} \mathsf{Q}$,

b) Для любого замкнутого множества $F \subset S$ $\overline{\lim}_{n \to \infty} \mathsf{Q}_n(F) \leqslant \mathsf{Q}(F)$,

c) Для любого открытого множества $G \subset S$ $\underline{\lim}_{n \to \infty} \mathsf{Q}_n(G) \geqslant \mathsf{Q}(G)$,

[18] Если кто сможет довести эту бандуру до конца — пинганите в Telegram.

d) Для любого борелевского множества $B \in \mathcal{B}(S)$ такого, что $\mathsf{Q}(\partial B) = 0$, $\mathsf{Q}_n(B) \to \mathsf{Q}(B)$ при $n \to \infty$.

Доказательство. Для начала докажем, что если выполнена слабая сходимость, то для произвольного замкнутого множества F выполнено ограничение снизу. Для произвольного $\varepsilon > 0$ введём функцию

$$f_\varepsilon(x) = \left(1 - \frac{\rho(x, F)}{\varepsilon}\right)^+, \text{ где } \rho(x, F) = \inf_{y \in F} d(x, y). \tag{40}$$

Заметим, что она непрерывна и ограничена. Тогда верна следующая цепочка:

$$\overline{\lim_{n\to\infty}}\, \mathsf{Q}_n(F) = \overline{\lim_{n\to\infty}} \int_S \mathbf{1}_{x\in F} \mathsf{Q}_n(\mathrm{d}x) \leqslant \overline{\lim_{n\to\infty}} \int_S f_\varepsilon(x) \mathsf{Q}_n(\mathrm{d}x) = \int_S f_\varepsilon(x) \mathsf{Q}(\mathrm{d}x) \leqslant \mathsf{Q}(F^\varepsilon), \tag{41}$$

где $F^\varepsilon = \{x \in S \colon \rho(x, F) \leqslant \varepsilon\}$. Осталось заметить, что F^ε монотонно сжимается к F при $\varepsilon \to 0$. Тогда по непрерывности вероятностной меры $\mathsf{Q}(F^\varepsilon) \to \mathsf{Q}(F)$. Следовательно, при устремлении ε к нулю получаем второй пункт.

Далее, эквивалентность второго и третьего пунктов очевидна, так как от одного к другому можно переходить, рассматривая дополнения множеств. Теперь докажем, что из второго и третьих пунктов следует четвёртый пункт. Для этого возьмём произвольное борелевское множество $B \in \mathcal{B}(S)$ такое, что $\mathsf{Q}(\partial B) = 0$. Теперь введём два множества: $F = [B]$ и $G = B \setminus \partial B$. Заметим, что F замкнуто, а G открыто, причём их меры равны мере B (так как мера границы есть ноль). Тогда несложно показать, что искомый предел существует и равен $Q(B)$:

$$\overline{\lim_{n\to\infty}}\, \mathsf{Q}_n(B) \leqslant \overline{\lim_{n\to\infty}}\, \mathsf{Q}_n(F) \leqslant Q(F) = Q(B), \tag{42}$$

$$\underline{\lim_{n\to\infty}}\, \mathsf{Q}_n(B) \geqslant \underline{\lim_{n\to\infty}}\, \mathsf{Q}_n(G) \geqslant Q(G) = Q(B). \tag{43}$$

Осталось доказать, что из последнего пункта следует первый. Возьмём некоторую ограниченную непрерывную функцию $f \colon S \mapsto \mathbb{R}$ такую, что $|f(x)| < M$ для всех $x \in S$. Далее, построим следующее множество:

$$D = \{t \in [-M, M] \colon \mathsf{Q}(\{x \in S \colon f(x) = t\}) > 0\}. \tag{44}$$

Данное множество не более, чем счётно. Теперь зафиксируем произвольное натуральное k и возьмём разбиение $-M = t_0 < t_1 < \ldots < t_k = M$ отрезка $[-M, M]$ такое, что ни одно из t_i не содержится в D. Далее, построим набор множеств $\{B_i\}_{i=1}^k$ по следующему правилу: $B_i = \{x \in S \colon t_{i-1} \leqslant f(x) < t_i\}$. Заметим, что $\partial B_i \subseteq f^{-1}(\{t_{i-1}\}) \cup f^{-1}(\{t_i\})$. Но оба прообраза имеют нулевую меру, поэтому граница B_i тоже имеет нулевую меру. Следовательно, $\mathsf{Q}_n(B_i) \to \mathsf{Q}(B_i)$ при $n \to \infty$ для всех $i = 1, \ldots, k$. Теперь рассмотрим следующий верхний предел:

$$\Delta = \overline{\lim_{n\to\infty}} \left| \int_S f(x) \mathsf{Q}_n(\mathrm{d}x) - \int_S f(x) \mathsf{Q}(\mathrm{d}x) \right|. \tag{45}$$

Ограничим её сверху суммой трёх верхних пределов:

$$\Delta \leqslant \overline{\lim_{n\to\infty}} \left| \int_S f(x) \mathsf{Q}_n(\mathrm{d}x) - \sum_{i=1}^k t_{i-1} \mathsf{Q}_n(B_i) \right| \tag{46}$$

$$+ \overline{\lim_{n\to\infty}} \left| \sum_{i=1}^k t_{i-1} \mathsf{Q}_n(B_i) - \sum_{i=1}^k t_{i-1} \mathsf{Q}(B_i) \right| \tag{47}$$

$$+ \overline{\lim_{n\to\infty}} \left| \int_S f(x) \mathsf{Q}(\mathrm{d}x) - \sum_{i=1}^k t_{i-1} \mathsf{Q}(B_i) \right| \tag{48}$$

Второй предел сразу равен нулю. Теперь покажем, что первый (а так же и третий) предел не превосходит $\max_{i=1,\ldots,k} |t_i - t_{i-1}|$. Для этого заметим, что

$$\int_S f(x) \mathsf{Q}_n(\mathrm{d}x) = \sum_{i=1}^k \int_{B_i} f(x) \mathsf{Q}_n(\mathrm{d}x) \leqslant \sum_{i=1}^k t_i \mathsf{Q}_n(B_i). \tag{49}$$

Следовательно,

$$\left| \int_S f(x) \mathsf{Q}_n(\mathrm{d}x) - \sum_{i=1}^k t_{i-1} \mathsf{Q}_n(B_i) \right| \leqslant \left| \sum_{i=1}^k (t_i - t_{i-1}) \mathsf{Q}_n(B_i) \right| \leqslant \max_{i=1,\ldots,k} |t_i - t_{i-1}|. \tag{50}$$

Отсюда получаем, что

$$\Delta \leqslant 2 \max_{i=1,\ldots,k} |t_i - t_{i-1}|. \tag{51}$$

Но этот максимум стремится к нулю при устремлении диаметра разбиения к нулю. Следовательно, имеет место слабая сходимость. □

15.2 Случайные процессы

Пусть $X = (X_t, t \in [0,1])$ — случайный процесс. Тогда X — случайный элемент со значением в пространстве функций на $[0,1]$ с цилиндрической сигма-алгеброй.

Определение 51. Пространство функций на $[0,1]$ — это множество $\mathbb{R}^{[0,1]} = \{y = (y(t), t \in [0,1]), y(t) \in \mathbb{R}\}$.

Определение 52. Пусть $B \in \mathcal{B}(\mathbb{R})$ и $t \in [0,1]$. *Элементарным цилиндром* называется множество $C(t,B) = \{y \in \mathbb{R}^{[0,1]} \colon y(t) \in B\}$.

Определение 53. *Цилиндрической сигма-алгеброй* называется минимальная сигма-алгебра, порождённая элементарными цилиндрами:

$$\mathrm{Cyl}(\mathbb{R}^{[0,1]}) = \sigma(\{C(t,B) \colon t \in [0,1], B \in \mathcal{B}(\mathbb{R})\}) \tag{52}$$

Утверждение. Отображение $X \colon \Omega \mapsto (\mathbb{R}^{[0,1]}, \mathrm{Cyl}(\mathbb{R}^{[0,1]}))$ является измеримым.

Из этого следует, что имеет смысл вводить распределение случайного процесса $\mathsf{P}_X(C) = \mathsf{P}(X \in C)$, $C \in \mathrm{Cyl}(\mathbb{R}^{[0,1]})$. Но пространство $\mathbb{R}^{[0,1]}$ плохое. Поэтому обычно его сужают до пространства непрерывных функций из $[0,1]$ в $\mathbb{R}$, которое обозначают через $\mathcal{C}[0,1]$. На нём есть норма $\|x\| = \max_{t\in[0,1]} |x(t)|$. Следовательно, можно построить борелевскую сигма-алгебру $\mathcal{B}(\mathcal{C}[0,1])$.

Лемма. *Для пространства $\mathcal{C}[0,1]$ борелевская сигма-алгебра совпадает с цилиндрической сигма-алгеброй:* $\mathcal{B}(\mathcal{C}[0,1]) = \mathrm{Cyl}(\mathcal{C}[0,1])$.

Доказательство. Для начала докажем, что выполнено следующее вложение: $\mathcal{B}(\mathcal{C}[0,1]) \subseteq \mathrm{Cyl}(\mathcal{C}[0,1])$. Для этого возьмём замкнутый шар $B_r[x] = \{y \in \mathcal{C}[0,1] \colon \|y - x\| \leqslant r\}$. Тогда по непрерывности можно сказать, что

$$B_r[x] = \bigcap_{t\in\mathbb{Q}} \{y \in \mathcal{C}[0,1] \colon |y(t) - x(t)| \leqslant r\}. \tag{53}$$

Но это есть ни что иное, как счётное пересечение элементарных цилиндров. Следовательно, $B_r[x] \in \mathrm{Cyl}(\mathcal{C}[0,1])$.

Теперь докажем обратное вложение. Для этого рассмотрим следующий цилиндр:

$$C(t,[a,b]) = \{y \in \mathcal{C}[0,1] \colon y(t) \in [a,b]\}. \tag{54}$$

Теперь введём функцию $h_t \colon \mathcal{C}[0,1] \mapsto \mathbb{R}$, действующую по правилу $h_t(x) = x(t)$. Несложно понять, что это непрерывная функция. Теперь заметим, что $C(t,[a,b]) = h_t^{-1}([a,b])$. Но это будет замкнутое множество, так как для непрерывной функции прообраз замкнутого множества замкнут. Следовательно, $C(t,[a,b]) \in \mathcal{B}(\mathcal{C}[0,1])$. □

Определение 54. Пусть $X^{(n)} = (X_t^{(n)}, t \in [0,1])$ — последовательность случайных процессов. Будем говорить, что она сходится по распределению к случайному процессу $X = (X_t, t \in [0,1])$, если имеет место слабая сходимость распределений: $\mathsf{P}_{X^{(n)}} \xrightarrow{W} \mathsf{P}_X$ при $n \to \infty$. Обозначение: $X^{(n)} \xrightarrow{\mathcal{D}} X$.

Теорема 30 (о наследовании сходимости)**.** *Пусть $h \colon \mathcal{C}[0,1] \mapsto \mathbb{R}$ — непрерывная функция и $X^{(n)} \xrightarrow{\mathcal{D}} X$ при $n \to \infty$. Тогда $h(X^{(n)}) \xrightarrow{d} h(X)$.*

Доказательство. Пусть $y \colon \mathbb{R} \mapsto \mathbb{R}$ — ограниченная непрерывная функция. Тогда композиция $g = y \odot h \mapsto \mathcal{C}[0,1] \mapsto \mathbb{R}$ будет ограниченной непрерывной функцией. Рассмотрим $\mathsf{E}[y(h(X^{(n)})]$. Оно равно

$$\mathsf{E}[y(h(X^{(n)}))] = \mathsf{E}[g(X^{(n)})] = \int_{\mathcal{C}[0,1]} g(x) \mathsf{P}_{X^{(n)}}(\mathrm{d}x). \tag{55}$$

Воспользуемся слабой сходимостью распределений:

$$\lim_{n\to\infty} \int_{\mathcal{C}[0,1]} g(x) \mathsf{P}_{X^{(n)}}(\mathrm{d}x) = \int_{\mathcal{C}[0,1]} g(x) \mathsf{P}_X(\mathrm{d}x) = \mathsf{E}[g(X)]. \tag{56}$$

Следовательно, для любой ограниченной непрерывной функции $y \colon \mathbb{R} \mapsto \mathbb{R}$ выполнена следующая сходимость: $\mathsf{E}[y(h(X^{(n)})] \to \mathsf{E}[y(h(X))]$ при $n \to \infty$. Но это означает, что $h(X^{(n)}) \xrightarrow{d} h(X)$. □

15.3 Назад к критерию Колмогорова

Теорема 31 (принцип инвариантности Донскера-Прохорова)**.** *Пусть $\{\xi_n\}_{n=1}^{\infty}$ — последовательность независимых и одинаково распределённых случайных величин таких, что $\mathsf{E}[\xi_1] = 0$ и $\mathsf{D}[\xi_1] = 1$. Введём последовательность случайных величин $\{S_n\}_{n=0}^{\infty}$ по следующему правилу: $S_0 = 0$, $S_n = \xi_1 + \ldots + \xi_n$. Далее, построим случайный процесс $X^{(n)} = (X_t^{(n)}, t \in [0,1])$, как линейную интерполяцию $S_0, \ldots, S_n$:*

$$X_t^{(n)} = \frac{S_k}{\sqrt{n}}(k+1-nt) + \frac{S_{k+1}}{\sqrt{n}}(nt-k) \text{ при } t \in \left[\frac{k}{n}, \frac{k+1}{n}\right], \ k = 0, \ldots, n-1. \tag{57}$$

Тогда $X^{(n)} \xrightarrow{\mathcal{D}} W$, где $W = (W_t, t \in [0,1])$ — винеровский процесс.

Применим её к доказательству критерия Колмогорова.

Теорема 32. *Пусть $X_1, \ldots, X_n$ — независимые и одинаково распределённые случайные величины с распределением $\mathrm{U}[0,1]$, а $x \in [0,1]$. Далее, вводится статистика*

$$D_n = \sup_{x \in [0,1]} |\hat{F}_n(x) - x|. \tag{58}$$

Тогда имеет место следующая сходимость:

$$\sqrt{n} D_n \xrightarrow{d} \sup_{t \in [0,1]} |W_t - tW_1|. \tag{59}$$

Доказательство. Для начала видоизменим определение D_n. Для этого поймём, где может достигаться супремум. Так как $\hat{F}_n$ является кусочно-постоянной функцией, то супремум может достигаться только в точках разрыва, ибо иначе значение можно увеличить, немного уменьшив x — от этого значение $\hat{F}_n(x)$ не изменится. Следовательно,

$$D_n = \max_{k=1,\ldots,n} \left| X_{(k)} - \frac{k}{n} \right|. \tag{60}$$

Ранее доказывалось, что если $\xi_1, \ldots, \xi_{n+1}$ — независимые и одинаково распределёные случайные величины с стандартным экспоненциальным распределением $\mathrm{Exp}(1)$, а $S_k = \xi_1 + \ldots + \xi_k$, то

$$(X_{(1)}, \ldots, X_{(n)}) \stackrel{d}{=} \left(\frac{S_1}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}} \right). \tag{61}$$

Тогда распределение $\sqrt{n} D_n$ равно распределению

$$\sqrt{n} \max_{k=1,\ldots,n} \left| \frac{S_k}{S_{n+1}} - \frac{k}{n} \right|. \tag{62}$$

Теперь заметим, что в пределе распределение выражения выше будет распределению

$$\Delta = \sqrt{n+1} \max_{k=1,\ldots,n} \left| \frac{S_k}{S_{n+1}} - \frac{k}{n+1} \right|. \tag{63}$$

Преобразуем выражение:

$$\Delta = \frac{n+1}{S_{n+1}} \max_{k=1,\ldots,n} \left| \frac{S_k}{\sqrt{n+1}} - \frac{kS_{n+1}}{(n+1)\sqrt{n+1}} \right| \tag{64}$$

$$= \frac{n+1}{S_{n+1}} \max_{k=1,\ldots,n} \left| \frac{S_k - k}{\sqrt{n+1}} - \frac{k}{n+1} \frac{S_{n+1} - (n+1)}{\sqrt{n+1}} \right| = T_n. \tag{65}$$

Теперь введём случайный процесс $X^{(n)} = (X_t^{(n)}, t \in [0,1])$ по следующему правилу:

$$X_t^{(n)} = \frac{S_k}{\sqrt{n+1}}(k+1-(n+1)t) + \frac{S_{k+1}}{\sqrt{n+1}}((n+1)t-k) \text{ при } t \in \left[\frac{k}{n+1}, \frac{k+1}{n+1}\right], \ k = 0, \ldots, n. \tag{66}$$

Далее заметим, что траектории случайного процесса $X^{(n)}$ — это ломаные. Поэтому

$$T_n \stackrel{d}{=} \frac{n+1}{S_{n+1}} \sup_{t \in [0,1]} |X_t^{(n)} - tX_1^{(n)}|. \tag{67}$$

Согласно лемме Слуцкого, усиленному закону больших чисел, принципу инвариантности и теореме о наследовании сходимости получаем, что

$$\Delta \xrightarrow{d} \sup_{t \in [0,1]} |W_t - tW_1|. \tag{68}$$

Тем самым получаем желаемое. □

Определение 55. Случайныц процесс $W^0 = (W^0_t, t \in [0,1])$, где $W^0_t = W_t - tW_1$, называется *броуновским мостом.*

Остался последний вопрос: как показать, что распределение супремума броуновского моста будет равно распределению Колмогорова? Это уже не самая тривиальная задача.

16 Лекция 16

17 Лекция 17

17.1 Ранговые методы

Вернёмся к критериям однородности, но несколько изменим формулировку. Пусть $\boldsymbol{X} = (X_1, \ldots, X_m)$ и $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ — независимые выборки с некоторыми неизвестными непрерывными распределениями. Можно ли сказать, что X_i равно Y_j по распределению для всех $i = 1, \ldots, m$, $j = 1, \ldots, n$?

Определение 56. Соберём выборки $\boldsymbol{X}$ и $\boldsymbol{Y}$ в вектор $\boldsymbol{Z}$. *Рангом* случайной величины Y_j называется

$$R(Y_j) = \sum_{i=1}^{n+m} \mathbf{1}_{Z_i \leqslant Y_j}. \tag{69}$$

Определение 57. *Статистикой ранговых сумм Вилкоксона* называется

$$W_{m,n}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{j=1}^{n} R(Y_j). \tag{70}$$

Утверждение. Если верна гипотеза однородности, то статистика ранговых сумм Вилкоксона не зависит от распределения выборок.

Доказательство. Пусть $F(x)$ — функция распределения элементов выборок. Так как она непрерывна, то $\mathbf{1}_{X_i \leqslant Y_j} = \mathbf{1}_{F(X_i) \leqslant F(Y_j)}$. Но $F(X_1) \sim \mathrm{U}[0,1]$. □