♂ФИВТ♂ МФТИ

Семинары по мат. статистике

Семинарист: Алексей Волостнов

Составитель: Андрей Куссев

Осень, 2022 год

Содержание

Примечание. Эта методичка написана ещё тем оболтусом и профаном, который даже на лекции не ходит, да ещё и настолько ЧСВ, что говорит о себе в третьем лице (мерзкий тип, советую его избегать). Доверять всему тому, что здесь написано, нельзя от слова совсем. А ещё тут куча опечаток, и как минимум формулировки определений и теорем лучше перепроверять у нормальных авторов. Если Вы видите здесь лажу или непонятный момент – не стесняйтесь и пинайте меня, чтобы я это исправил.

1 Оценки и их свойства

В теории вероятности мы в основном работали с заранее известными распределениями: изучали их свойства, вводили их характеристики и занимались прочими вещами. Математическая статистика делает всё с точностью до наоборот: по свойствам выборки надо определить, из какого распределения она пришла. Часто набор распределений, которые являются кандидатами на роль истинного распределения, можно описать набором параметров, поэтому в основном нашей задачей будет определить с некоторой точностью значение параметра по реализации распределения.

Напомним, что мы работаем в вероятностно-статистической модели $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathcal{P})$, где $\mathcal{X}$ – множество результатов наблюдения (обычно под ним мы подразумеваем $\mathbb{R}$ или $\mathbb{R}^n$, но также это может быть и $\mathbb{R}^\infty$, когда мы хотим работать с бесконечными выборками), $\mathcal{B}(\mathcal{X})$ – борелевская σ-алгебра на $\mathcal{X}$, $\mathcal{P}$ – семейство вероятностных мер на $\mathcal{B}(\mathcal{X})$ (чаще всего оно будет иметь вид $\{\mathsf{P}_\theta \colon \theta \in \Theta\}$, где θ – неизвестный параметр). Так как распределение теперь не фиксировано, то часто в обозначения матожидания, дисперсии, сходимости по вероятности и прочих вещах, которые используют конкретное распределение, мы будем писать индекс θ, чтобы подчеркнуть, какое распределение используется в данный момент.

Определение. Пусть $(\Omega, \mathcal{F})$ – измеримое пространство. Произвольная $(\mathcal{B}(\mathcal{X})|\mathcal{F})$-измеримая функция $S \colon \mathcal{X} \to \Omega$ называется *статистикой*.

В случае, когда $\mathcal{P}$ описывается параметром $\theta \in \Theta$, а $\Omega = \Theta$, такая статистика называется *оценкой* параметра θ.

Замечание. 1. Статистика по умолчанию не зависит от неизвестного параметра, иначе получается странно: хотим найти неизвестный параметр с помощью статистики, при этом она почему-то включает в себя этот параметр. Иногда всё же мы будем рассматривать функции, использующие фиксированное значение параметра (например, в построении доверительных интервалов).

2. Обратите внимание, что оценкой мы априори считаем что-то, что имеет значения лишь в допустимом множестве Θ. Какой бы хорошей статистикой $S(X)$ не была, если она периодически выдаёт недопустимое значение параметра, то она неадекватна. Иногда этим свойством мы будем пользоваться, так что имейте это в виду.

Чтобы понимать, какие оценки хорошие, а какие – не очень, нужно выделить некоторые свойства оценок, которые было бы крайне желательно иметь.

Первое из них говорит, что если нам будут поступать раз за разом выборки, то оценки для них будут в среднем похожи на истинный параметр:

Определение. Оценка $\widehat{\theta}(X)$ называется *несмещённой* оценкой параметра $\tau(\theta)$, если для любого $\theta \in \Theta$ $\mathsf{E}_\theta \widehat{\theta} = \tau(\theta)$.

Это свойство весьма логичное, но его очевидно недостаточно, чтобы утверждать, что оценка хоть сколь-нибудь пригодна. Например, если $\mathsf{E}_\theta X_1 = \theta$, то X_1 – несмещённая оценка параметра θ, хотя она не использует всю мощь выборки. Это подводит нас к асимптотическим свойствам оценок: нам бы хотелось, чтобы при увеличении размера выборки увеличилась бы и точность в предсказании параметра.

Определение. Пусть $(X_1, \ldots, X_n, \ldots)$ – выборка. Оценка θ_n^* называется *состоятельной* оценкой параметра $\tau(\theta)$, если для любого $\theta \in \Theta$ выполнено $\theta_n^* \xrightarrow{\mathsf{P}_\theta} \theta$. Оценка θ_n^* называется *сильно состоятельной* оценкой параметра $\tau(\theta)$, если для любого $\theta \in \Theta$ выполнено $\theta_n^* \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \theta$.

Вообще под (сильно) состоятельной оценкой подразумевают последовательность оценок, но обычно все и так понимают, о чём речь. Если из контекста понятно, как именно зависит оценка от параметра n, то нижний индекс оценки убирают.

Не менее интересной для рассмотрения оказывается скорость сходимости оценки к истинному

значению параметра.

Определение. Оценка θ_n^* называется *асимптотически нормальной* оценкой параметра $\tau(\theta)$, если для любого $\theta \in \Theta$

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

Величина $\sigma^2(\theta)$ называется *асимптотической дисперсией.*

Так как нормально распределённая случайная величина по "правилу трёх сигм" принимает своё значение на интервале $(-3\sigma(\theta); 3\sigma(\theta))$ с очень высокой вероятностью, то можно считать, что оценка стремится к истинному значению параметра со скоростью порядка $3\sigma(\theta)/\sqrt{n}$.

Заметим, что из сильной состоятельности или асимптотической нормальности оценки следует её состоятельность (по поводу последнего смотрите задачу 1.2).

Методы доказательства асимптотических свойств:

1. ЗБЧ, УЗБЧ, ЦПТ и их многомерные аналоги
2. Теорема о наследовании сходимостей
3.

Теорема (о наследовании асимптотической нормальности)**.** *Пусть $\widehat{\theta}(X)$ – асимптотически нормальная оценка с асимптотической дисперсией $\sigma^2(\theta)$. $\tau(\theta)$ – дифференцируемая на Θ функция. Тогда*

$$\sqrt{n}(\tau(\widehat{\theta}) - \tau(\theta)) \xrightarrow{d_\theta} \mathcal{N}(0, \sigma^2(\theta) \cdot (\tau'(\theta))^2).$$

Замечание. Можно (хотя не рекомендуется) понимать это так: к $\widehat{\theta}$, которая является "примерно" нормально распределённой при больших n, применяется преобразование, которое в малой окрестности θ "почти что" линейное с коэффициентом $\tau'(\theta)$, отчего и дисперсия увеличилась в $(\tau'(\theta))^2$ раз.

Задача 1.1. Пусть $X_1, \ldots, X_n$ – выборка из $U(0, \theta)$. Проверьте на несмещённость, состоятельность, сильную состоятельность и асимптотическую нормальность следующие оценки параметра θ: **(а)** $\overline{X} + X_{(n)}/2$, **(б)** $(n+1)X_{(1)}$, **(в)** $X_{(1)} + X_{(n)}$. **(г)** Найдите число $\delta > 0$ и невырожденное распределение F_θ такие, что $n(\theta - X_{(n)}) \xrightarrow{d_\theta} \xi \sim F_\theta$

Решение. Для начала поймём, как распределены первая и последняя порядковые статистики. Для $t \in (0; 1)$:

$$\mathsf{P}_\theta(X_{(n)} \leqslant t) = \mathsf{P}_\theta(X_1, \ldots, X_n \leqslant t) = \prod \mathsf{P}_\theta(X_i \leqslant t) = \mathsf{P}_\theta(X_1 \leqslant t)^n = \frac{t^n}{\theta^n};$$

$$\rho_{X_{(n)}}(t) = \frac{nt^{n-1}}{\theta^n} I(0 < t < \theta).$$

$$\mathsf{P}_\theta(X_{(1)} \leqslant t) = 1 - \mathsf{P}_\theta(X_{(1)} > t) = 1 - \mathsf{P}_\theta(X_1, \ldots, X_n > t) = 1 - \prod \mathsf{P}_\theta(X_i > t) =$$

$$= 1 - (1 - \mathsf{P}_\theta(X_1 \leqslant t))^n = 1 - \left(1 - \frac{t}{\theta}\right)^n; \quad \rho_{X_{(1)}}(t) = \frac{n}{\theta}\left(1 - \frac{t}{\theta}\right)^{n-1} I(0 < t < \theta).$$

Также полезным для проверки на несмещённость будут их матожидания:

$$\mathsf{E}_\theta X_{(n)} = \int_0^\theta t \cdot \frac{nt^{n-1}}{\theta^n}\, dt = \frac{n}{n+1}\theta,$$

$$\mathsf{E}_\theta X_{(1)} = \int_0^\theta t \cdot \frac{n}{\theta}\left(1 - \frac{t}{\theta}\right)^{n-1} dt = n\theta \int_0^1 s(1-s)^{n-1}\, ds = n\theta B(2, n) = n\theta \frac{\Gamma(2)\Gamma(n)}{\Gamma(n+2)} = \frac{\theta}{n+1}.$$

Теперь мы готовы к решению задачи.

Несмещённость. Из линейности матожидания легко видеть, что несмещёнными будут $(n+1)X_{(1)}$ и $X_{(1)} + X_{(n)}$.

Состоятельность. $\overline{X} \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \theta/2$ из УЗБЧ. Для произвольного $0 < \varepsilon < \theta$ (для $\varepsilon > \theta$ всё ясно):

$$\mathsf{P}_\theta(|X_{(n)} - \theta| > \varepsilon) = \underbrace{\mathsf{P}_\theta(X_{(n)} > \theta + \varepsilon)}_{=0} + \mathsf{P}_\theta(X_{(n)} < \theta - \varepsilon) = \frac{(\theta - \varepsilon)^n}{\theta^n} \to 0.$$

Что же насчёт первой порядковой статистики, то

$$\mathsf{P}_\theta(|(n+1)X_{(1)} - \theta| > \varepsilon) \geqslant \mathsf{P}_\theta((n+1)X_{(1)} > \theta + \varepsilon) = \left(1 - \frac{\theta + \varepsilon}{\theta(n+1)}\right)^n \to \exp\left(-\frac{\theta + \varepsilon}{\theta}\right) \neq 0$$

С другой стороны, для

$$\mathsf{P}_\theta(|X_{(1)}| > \varepsilon) = \mathsf{P}_\theta(X_{(1)} > \varepsilon) = \left(1 - \frac{\varepsilon}{\theta}\right)^n \to 0.$$

Таким образом, $X_{(n)} \xrightarrow{\mathsf{P}_\theta} \theta$, $X_{(1)} \xrightarrow{\mathsf{P}_\theta} 0$, но $(n+1)X_{(1)} \xrightarrow{\mathsf{P}_\theta} \theta$. Из всего этого получаем, что $\overline{X} + X_{(n)}/2$ в силу того факта, что сходимости по вероятности можно складывать, будет состоятельной, $(n+1)X_{(1)}$ не является состоятельной оценкой θ, а вот $X_{(1)} + X_{(n)}$ уже будет являться как сумма $X_{(n)}$, стремящейся по вероятности к θ, и $X_{(1)}$, стремящейся по вероятности к нулю.

Сильная состоятельность. Тут всё куда проще, ведь при фиксированной выборке что $X_{(n)}$, что $X_{(1)}$ – монотонны при увеличении n, а это значит, что из их сходимости по вероятности будет следовать сходимость P_θ-п.н. Действительно, как известно из курса теории вероятностей, у последовательности, сходящейся по вероятности, есть подпоследовательность, сходящаяся почти наверное. Тогда из монотонности следует, что и вся последовательность такая. Отсюда, $X_{(n)} \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \theta$, $X_{(1)} \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} 0$, поэтому оценки $\overline{X} + X_{(n)}/2$ и $X_{(1)} + X_{(n)}$ будут сильно состоятельными. $(n+1)X_{(1)}$ же таковой не является, так как она даже не состоятельна.

Асимптотическая нормальность. Аналогично предыдущему пункту и по задаче 1.2 $(n+1)X_{(1)}$ не асимптотически нормальна. Проверим, как себя ведут порядковые статистики:

$$\mathsf{P}_\theta(\sqrt{n}(X_{(n)} - \theta) \leqslant t) = \mathsf{P}_\theta\left(X_{(n)} \leqslant \theta + \frac{t}{\sqrt{n}}\right) = \begin{cases} 1, & t \geqslant 0; \\ \left(1 + \dfrac{t}{\theta\sqrt{n}}\right)^n \to 0, & t < 0. \end{cases}$$

$$\mathsf{P}_\theta(\sqrt{n}X_{(1)} \leqslant t) = \mathsf{P}_\theta\left(X_{(1)} \leqslant \frac{t}{\sqrt{n}}\right) = \begin{cases} 0, & t < 0; \\ 1 - \left(1 - \dfrac{t}{\theta\sqrt{n}}\right)^n \to 1, & t > 0. \end{cases}$$

Стало быть, $\sqrt{n}(X_{(n)} - \theta), \sqrt{n}X_{(1)} \xrightarrow{d_\theta} 0 \sim \mathcal{N}(0,0)$ (мы натуралы, поэтому считаем нуль нормально распределённым), и по лемме Слуцкого $\sqrt{n}(\overline{X} + X_{(n)}/2 - \theta) \xrightarrow{d_\theta} \mathcal{N}(0, \mathsf{D}_\theta\overline{X})$, $\sqrt{n}(X_{(1)} + X_{(n)} - \theta) \xrightarrow{d_\theta} 0$. Таким образом, эти оценки будут ещё и асимптотически нормальными.

(г) Подберём δ так, чтобы распределение $n^\delta(\theta - X_{(n)})$ было чем-то нетривиальным.

$$\mathsf{P}_\theta(n^\delta(\theta - X_{(n)}) \leqslant t) = \mathsf{P}_\theta(X_{(n)} \geqslant \theta - tn^{-\delta}) = \begin{cases} 0, & t \leqslant 0; \\ 1 - \left(1 - \dfrac{t}{\theta n^\delta}\right)^n, & t > 0. \end{cases} \circledast$$

Как мы видим, при $\delta < 1$ распределение будет тривиальным, а при $\delta > 1$ и вовсе получается что-то неадекватное. При $\delta = 1$ же:

$$\circledast \begin{cases} 0, & t \leqslant 0; \\ 1 - e^{-\frac{t}{\theta}}, & t > 0. \end{cases},$$

что есть функция распределения для $Exp\left(\frac{1}{\theta}\right)$. ■

Задача 1.2. Пусть $\theta^*(X)$ – асимптотически нормальная оценка параметра θ. Докажите, что тогда $\theta^*(X)$ является состоятельной оценкой θ.

Решение. С одной стороны, по условию $\sqrt{n}(\theta^* - \theta) \xrightarrow{d_\theta} \mathcal{N}(0, \sigma^2(\theta))$. С другой, очевидно выполняется $1/\sqrt{n} \xrightarrow{d_\theta} 0$. Тогда по лемме Слуцкого $\theta^* - \theta \xrightarrow{d_\theta} 0$. Из сходимости по распределению к константе следует сходимость к ней по вероятности, а значит, $\theta^* \xrightarrow{\mathsf{P}_\theta} \theta$. ■

Задача 1.3. Пусть $X_1, \ldots, X_n$ – выборка из некоторого распределения с параметром σ^2. Пусть, кроме того, $\mathsf{D}_\theta X_1 = \sigma^2$. Назовём *выборочной дисперсией* статистику $s^2 = \sum(X_i - \overline{X})^2/n$ Докажите, что:

(а) $s^2 = \overline{X^2} - \overline{X}^2$;

(б) s^2 является сильно состоятельной оценкой для σ^2;

(в) если $\mathsf{E}_\theta X_1^4 < \infty$, то s^2 является асимптотически нормальной оценкой для σ^2.

(г) Является ли она несмещённой оценкой для σ^2?

Решение. **(а)**

$$s^2 = \sum\left(\frac{X_i^2}{n} - \frac{2X_i\overline{X}}{n} + \frac{\overline{X}^2}{n}\right) = \frac{\sum X_i^2}{n} - 2\overline{X}\frac{\sum X_i}{n} + \overline{X}^2 = \overline{X^2} - 2\overline{X}^2 + \overline{X}^2 = \overline{X^2} - \overline{X}^2.$$

(б) По УЗБЧ $\overline{X} \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \mathsf{E}_\theta X_1$, а $\overline{X^2} \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \mathsf{E}_\theta X_1^2$. Значит, по теореме о наследовании сходимости почти наверное и предыдущему пункту: $s^2 \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \mathsf{E}_\theta X_1^2 - (\mathsf{E}_\theta X_1)^2 = \mathsf{D}_\theta X_1 = \sigma^2$.

(в) По многомерному ЦПТ (её можно применять, так как из конечности $\mathsf{E}_\theta X_1^4$ следует конечность вторых моментов у координат вектора):

$$\sqrt{n}\left(\begin{pmatrix}\overline{X}\\ \overline{X^2}\end{pmatrix} - \begin{pmatrix}\mathsf{E}_\theta X_1\\ \mathsf{E}_\theta X_1^2\end{pmatrix}\right) \xrightarrow{d_\theta} \mathcal{N}(0, \Sigma),$$

где Σ – некоторая ковариационная матрица. Применяя теорему о наследовании асимптотической нормальности для $\tau(x, y) = y - x^2$:

$$\sqrt{n}\left(s^2 - \sigma^2\right) = \sqrt{n}\left(\tau(\overline{X}, \overline{X^2}) - \tau(\mathsf{E}_\theta X_1, \mathsf{E}_\theta X_1^2)\right) \xrightarrow{d_\theta} \mathcal{N}(0, \nabla\tau^T\Sigma\nabla\tau),$$

что и требовалось.

(г) Несмотря на то что оценка s^2 обладает такими потрясающими свойствами, она имеет смещение:

$$\mathsf{E}_\theta S^2 = \mathsf{E}_\theta\left(\frac{1}{n}\sum X_i^2 - \frac{1}{n^2}\sum_{i,j} X_iX_j\right) = \frac{1}{n}\sum \mathsf{E}_\theta X_i^2 - \frac{1}{n^2}\sum \mathsf{E}_\theta X_i^2 - \frac{1}{n^2}\sum_{i\neq j}\mathsf{E}_\theta(X_iX_j) =$$

$$= \mathsf{E}_\theta X_1^2 - \frac{1}{n}\mathsf{E}_\theta X_1^2 - \frac{1}{n^2}\underbrace{\sum_{i\neq j}\mathsf{E}_\theta X_i\mathsf{E}_\theta X_j}_{n^2-n \text{ слагаемых}} = \frac{n-1}{n}\mathsf{E}_\theta X_1^2 - \frac{n-1}{n}(\mathsf{E}_\theta X_1)^2 = \frac{n-1}{n}\mathsf{D}_\theta X_1 = \frac{n-1}{n}\sigma^2.$$

■

Задача 1.4. Пусть $X_1, \ldots, X_n$ – выборка из экспоненциального распределения с параметром θ, т.е. $p_\theta(t) = \theta e^{-\theta t} I(t > 0)$. Покажите, что для любого $k \in \mathbb{N}$ статистика $\left(k!/\overline{X^k}\right)^{1/k}$ является асимптотически нормальной оценкой параметра θ. Найдите её асимптотическую дисперсию.

Решение. Для начала вспомним, что $\mathsf{E}_\theta X_1^k = k!/\theta^k$ (это можно получить, честно найдя интеграл или рассмотрев хар. функцию). По ЦПТ:

$$\sqrt{n}\left(\overline{X^k} - \frac{k!}{\theta^k}\right) \xrightarrow{d_\theta} \mathcal{N}(0, \mathsf{D}_\theta X_1^k) = \mathcal{N}(0, \mathsf{E}_\theta X_1^{2k} - (\mathsf{E}_\theta X_1^k)^2) = \mathcal{N}\left(0, \frac{(2k)! - k!^2}{\theta^{2k}}\right).$$

Применим теорему о наследовании асимптотической нормальности для $\tau(x) = \left(\frac{k!}{x}\right)^{1/k}$, но сначала посчитаем её производную

$$\tau'(x) = -\frac{k!^{1/k}}{kx^{1+1/k}}; \quad \tau'\left(\frac{k!}{\theta^k}\right) = \frac{\theta^{k+1}}{k! \cdot k}.$$

Таким образом,

$$\sqrt{n}\left(\left(k!/\overline{X^k}\right)^{1/k} - \theta\right) \xrightarrow{d_\theta} \mathcal{N}\left(0, \frac{(2k)! - k!^2}{\theta^{2k}} \cdot \frac{\theta^{2k+2}}{k!^2 \cdot k^2}\right) = \mathcal{N}\left(0, \frac{\theta^2((2k)! - k!^2)}{k!^2 \cdot k^2}\right).$$

■

Задача 1.5. Пусть $X_1, \ldots, X_n$ – выборка из распределения $Bern(\theta)$. Предположим, что функция τ такова, что существует несмещённая оценка для $\tau(\theta)$. Докажите, что τ является многочленом степени не выше n. Любой ли такой многочлен подойдёт?

Решение. Пусть $\widehat{\theta}(X)$ – несмещённая оценка для $\tau(\theta)$. Тогда можно честно посчитать её матожидание:

$$\tau(\theta) = \mathsf{E}_\theta \widehat{\theta}(X) = \sum_{\mathbf{x} \in \{0,1\}^n} \widehat{\theta}(x_1, \ldots, x_n) \cdot \theta^{\sum x_i}(1-\theta)^{n - \sum x_i}$$

Каждое слагаемое есть произведение многочлена $\theta^{\sum x_i}$ степени $\sum x_i$, многочлена $(1-\theta)^{n-\sum x_i}$ степени $n - \sum x_i$ и какой-то константы $\widehat{\theta}(x_1, \ldots, x_n)$, что есть многочлен степени n. Стало быть, $\tau(\theta)$ как сумма таких слагаемых есть многочлен степени не выше n.

Второй вопрос кажется несложным: произвольный многочлен можно однозначно разложить по $\theta^k(1-\theta)^{n-k}$, откуда можно явно построить нужную $\widehat{\theta}$. Но на самом деле ответ отрицательный – для некоторых многочленов полученные «оценки» не будут вообще *оценками*, так как их значения могут не лежать в $\tau([0;1])$ (помните второй пункт замечания к определению оценки?).

Подтвердим вышесказанное контрпримером. Рассмотрим $\tau(x) = x(1-x)$ и двухэлементную выборку. Многочлен очевидным образом раскладывается по $x^k(1-x)^{2-k}$, поэтому если и имеется несмещённая оценка $\widehat{\theta}(X)$, то выполнено

$$\begin{cases} \widehat{\theta}(0,0) = 0 \\ \widehat{\theta}(1,1) = 0 \\ \widehat{\theta}(1,0) + \widehat{\theta}(0,1) = 1. \end{cases}$$

Последнее условие значит, что либо $\widehat{\theta}(1,0) \geqslant 1/2$, либо $\widehat{\theta}(0,1) \geqslant 1/2$. Но $\tau([0;1]) = [0;1/4]$, то есть значения $\widehat{\theta}$ априори не будут лежать в заданном множестве. ■

2 Основные методы нахождения оценок

2.1 Метод моментов

Допустим, что распределение элементов выборки зависит от k неизвестных параметров $\theta_1, \dots, \theta_k$, где вектор $\theta = (\theta_1, \dots, \theta_k)$ принадлежит некоторой области Θ в $\mathbb{R}^k$. Для построения оценки по методу моментов возьмём такие борелевские $g_1, \dots, g_k \colon \mathbb{R} \to \mathbb{R}$, что $\forall i \in \{1, \dots, k\}$ определено $\mathsf{E} g_i(X_1) = m_i(\theta)$. Предположим, что у уравнения $m(\theta) = \overline{g(X)}$ имеется единственное решение, где $m = (m_1, \dots, m_k)$, $g = (g_1, \dots, g_k)$. Так как из закона больших чисел мы знаем, что $\overline{g_i(X)}$ примерно равно $\mathsf{E} g_i(X_1)$, то логично положить решение уравнения выше за оценку параметра. А именно:

Теорема (сильная состоятельность оценки по методу моментов)**.** *Пусть $m \colon \Theta \to m(\Theta)$ - биекция, и функцию m^{-1} можно доопределить до функции, заданной на всем $\mathbb{R}^k$, и непрерывной в каждой точке множества $m(\Theta)$. Также, $\mathsf{E} g_i(X_1) < \infty$ $\forall i \in \{1, \dots, k\}$, $\forall \theta \in \Theta$. Тогда оценка по методу моментов является сильно состоятельной оценкой параметра θ.*

Часто в задачах не получается доопределить m^{-1} на всё $\mathbb{R}^k$, но это и не надо, если $g(X) \in m(\Theta)$.

Для упрощения вычислений часто в качестве $g_i(t)$ берут t^i (такие фукнции называют *пробными*), и тогда соответствующая $m_i(\theta)$ называется *моментом i-ого порядка*, откуда собственно и пошло название метода.

Схожее утверждение можно дать про асимптотическую нормальность:

Теорема (асимптотическая нормальность оценки по методу моментов)**.** *Если в условиях предыдущей теоремы функция m_{-1}, доопределенная на $\mathbb{R}^k$, дифференцируема на $m(\Theta)$, и $\mathsf{E} g_i(X_1)^2 < \infty$ $\forall i \in \{1, \dots, k\}$, $\forall \theta \in \Theta$, то оценка, полученная по методу моментов, асимптотически нормальна.*

Как можно видеть, оценки по методу моментов интуитивно понятны и легки в построении. Однако обычно асимптотическая дисперсия оценок, полученных по методу моментов, довольно велика, в то время как оценки, построенные другими методами, оказываются более выигрышными (например, как в задаче 2.5). К тому же не факт, что они будут несмещёнными.
Задача 2.1. Найдите оценку по методу моментов для параметров распределений **(а)** $Pois(\lambda)$, **(б)** $Geom(p)$, **(в)** $Beta(\alpha, \beta)$, **(г)** $U(a, b)$.

Решение. **(а)** Так как $m(\lambda) = \mathsf{E}_\lambda X = \lambda$ – тождественная, то в качестве оценки параметра можно взять $\overline{X}$.

(б) В данном случае $m(p) = \mathsf{E}_p X = \frac{1-p}{p} = \frac{1}{p} - 1$, тогда $m^{-1}(t) = \frac{1}{t+1}$. Таким образом, оценка по методу моментов $\widehat{p} = \frac{1}{\overline{X}+1}$.

(в) Тут уже надо оценивать двумерный параметр $\theta = (\alpha, \beta)$, поэтому найдём первый и второй моменты: $m_1(\alpha, \beta) = \mathsf{E}_\theta X_1 = \frac{\alpha}{\alpha+\beta} = x$, $m_2(\alpha, \beta) = \mathsf{E}_\theta X_1^2 = \mathsf{D}_\theta X_1 + (\mathsf{E}_\theta X_1)^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha^3+\alpha^2\beta+\alpha^2+\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} = y$. Тогда

$$\left.\begin{aligned} \frac{x}{y} = \frac{\alpha+\beta+1}{\alpha+1} = 1 + \frac{\beta}{\alpha+1}, \quad \frac{\beta}{\alpha+1} = \frac{x}{y} - 1, \quad \frac{\alpha}{\beta} + \frac{1}{\beta} = \frac{y}{x-y} \\ \frac{1}{x} = \frac{\alpha+\beta}{\alpha} = 1 + \frac{\beta}{\alpha}, \quad \frac{\alpha}{\beta} = \frac{x}{1-x} \end{aligned}\right\} \Longrightarrow \frac{1}{\beta} = \frac{y}{x-y} - \frac{x}{1-x} \Longrightarrow$$

$$\beta = \frac{(x-y)(1-x)}{y-x^2}, \quad \alpha = \frac{x(x-y)}{y-x^2}.$$

С учётом того, что $\overline{X^2} - \overline{X}^2 = s^2$, оценку по методу моментов можно записать как

$$\widehat{\alpha} = \frac{\overline{X}(\overline{X} - \overline{X^2})}{s^2}, \quad \widehat{\beta} = \frac{(1-\overline{X})(\overline{X} - \overline{X^2})}{s^2}.$$

(г) $m_1(a,b) = \mathsf{E}_\theta X_1 = \frac{a+b}{2} = x$, $m_2(a,b) = \mathsf{E}_\theta X_1^2 = \mathsf{D}_\theta X_1 + (\mathsf{E}_\theta X_1)^2 = \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4} = \frac{a^2+b^2+ab}{3} = y$. Откуда $a + b = 2x$, $ab = a^2 + 2ab + b^2 - 3y = 4x^2 - 3y$, что есть коэффициенты квадратного уравения с корнями a и b. С учётом того, что $a \leqslant b$, получаем, что $a = x - \sqrt{3y - 3x^2}$, $b = x + \sqrt{3y - 3x^2}$. Итого имеется следующая оценка по методу моментов:

$$\widehat{a} = \overline{X} - \sqrt{3s^2}, \quad \widehat{b} = \overline{X} + \sqrt{3s^2}.$$

■

2.2 Метод максимального правдоподобия

Определение. Пусть $\mathcal{P}_\theta$ – доминируемое семейство распределений с плотностью $\rho_\theta(x)$. *Функцией правдоподобия* выборки $X_1, \ldots, X_n$ называется плотность их совместного распределения

$$f_\theta(X_1, \ldots, X_n) = \rho_\theta(X_1) \ldots \rho_\theta(X_n).$$

Величина

$$L_\theta(X_1, \ldots, X_n) = \ln f_\theta(X_1, \ldots, X_n)$$

называется *логарифмической функцией правдоподобия.*

Оценкой максимального правдоподобия параметра θ называется статистика

$$\widehat{\theta}(X) = \arg\max_{\theta \in \Theta} f_\theta(X_1, \ldots, X_n)$$

Заметим, что точки максимума функции правдоподобия и её логарифмического братаблизнеца совпадают в силу монотонности логарифма, поэтому максимизировать можно любую из этих функций. Практически всегда мы будем просто находить нули производных и максимально забивать на доказательство того, что они будут являться точками максимума. Ну, или считайте, что доказательство даётся читателю в качестве упражнения.

Задача 2.2. Найдите оценки по методу максимального правдоподобия для параметров следующих распределений: **(а)** $Geom(p)$, **(б)** $U(0,a)$, **(в)** $\mathcal{N}(a,\sigma^2)$, **(г)** оценку параметра λ распределения $\Gamma(\alpha,\lambda)$, считая α известным, **(д)** $Pareto(k,a)$.

Решение. **(а)**

$$f_p(X_1, \ldots, X_n) = \prod (1-p)^{X_i} p, \quad L_p(X_1, \ldots, X_n) = n \ln p + \sum X_i \ln(1-p),$$

$$\frac{\partial}{\partial p} L_p(X_1, \ldots, X_n) = \frac{n}{p} - \sum \frac{X_i}{1-p} = 0, \quad \frac{n(1-p) - p\sum X_i}{p(1-p)} = 0,$$

$$p\left(n + \sum X_i\right) = n \Longrightarrow \widehat{p} = \frac{n}{n + \sum X_i} = \frac{1}{1 + \overline{X}}.$$

(б) Один из немногих случаев, когда тупо взять и продифференцировать не выйдет. Функция правдоподобия выглядит так

$$f_a(X_1, \ldots, X_n) = \frac{1}{a^n} I(0 \leqslant X_1, \ldots, X_n \leqslant a).$$

Там, где f_a не равна нулю, она равна некоторой константе $\frac{1}{a^n}$, которую надо максимизировать, то есть надо минимизировать a. Но сделать её меньше $X_{(n)}$ не получится, так как иначе не выполнится условие под индикатором. Следовательно, $\widehat{a} = X_{(n)}$ будет искомой ОМП.

(в)

$$f_\theta(X_1,\dots,X_n)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\sum\frac{(X_i-a)^2}{2\sigma^2}\right),\quad L_\theta(X_1,\dots,X_n)=-\frac{n}{2}\ln 2\pi\sigma^2-\sum\frac{(X_i-a)^2}{2\sigma^2}$$

$$\begin{cases}\dfrac{\partial}{\partial a}=\sum\dfrac{X_i-a}{\sigma^2}=0,\\ \dfrac{\partial}{\partial\sigma^2}=-\dfrac{n}{2\sigma^2}+\sum\dfrac{(X_i-a)^2}{2\sigma^4}=0.\end{cases}\quad\begin{cases}\widehat{a}=\overline{X},\\ \dfrac{1}{\sigma^2}\underbrace{\sum(X_i-\overline{X})^2}_{=ns^2}-n=0.\end{cases}$$

$$\begin{cases}\widehat{a}=\overline{X},\\ \widehat{\sigma^2}=s^2.\end{cases}$$

(г)

$$f_\lambda(X_1,\dots,X_n)=\frac{\lambda^{n\alpha}\left(\prod X_i\right)^{\alpha-1}}{\Gamma(\alpha)^n}e^{-\lambda\sum X_i}I(X_1,\dots,X_n>0).$$

Для $X_1,\dots,X_n>0$ имеем

$$L_\lambda(X_1,\dots,X_n)=n\alpha\ln\lambda-\lambda\sum X_i+\ln\left(\left(\prod X_i\right)^{\alpha-1}\right)-n\ln\Gamma(\alpha),$$

$$\frac{\partial}{\partial\lambda}L_\lambda(X_1,\dots,X_n)=\frac{n\alpha}{\lambda}-\sum X_i=0\Longrightarrow\widehat{\lambda}=\frac{\alpha}{\overline{X}}.$$

(д)

$$f_\lambda(X_1,\dots,X_n)=\frac{k^n a^{nk}}{\left(\prod X_i\right)^{k+1}}I(X_1,\dots,X_n\geqslant a).$$

Для фиксированного k максимум f_θ достигает при $\widehat{a}=X_{(1)}$ (аналогично пункту **(б)**). Тогда если принять a равным первой порядковой статистике, получаем

$$L_\theta(X_1,\dots,X_n)=n\ln k+nk\ln X_{(1)}-(k+1)\sum\ln X_i$$

$$\frac{\partial}{\partial\theta}L_\theta(X_1,\dots,X_n)=\frac{n}{k}+n\ln X_{(1)}-\sum\ln X_i=0,\quad\frac{1}{k}=\overline{\ln X}-\ln X_{(1)}\Longrightarrow$$

$$\widehat{k}=\frac{1}{\overline{\ln X}-\ln X_{(1)}}.$$

■

Задача 2.3. Найдите оценку максимального правдоподобия для параметра сдвига в модели распределения Коши,

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

если выборка состоит из **(а)** одного наблюдения, **(б)** двух наблюдений (т.е. $n=1,2$).

Решение. **(а)** Тут всё весьма просто:

$$f_\theta(x)=\frac{1}{\pi(1+(x-\theta)^2)}$$

$$f'_\theta(x)=\frac{1}{\pi(1+(x-\theta)^2)^2}\cdot 2(x-\theta)=0\Longrightarrow\widehat{\theta}=x.$$

(б) И тут становится понятно, почему распределения Коши не было в предыдущей задаче...

$$f_\theta(x_1,x_2)=\frac{1}{\pi^2(1+(x_1-\theta)^2)(1+(x_2-\theta)^2)}$$
$$L_\theta(x_1,x_2)=-\ln\left(1+(x_1-\theta)^2\right)-\ln\left(1+(x_2-\theta)^2\right)-2\ln\pi$$
$$\frac{\partial}{\partial\theta}L_\theta(x_1,x_2)=\frac{2(x_1-\theta)}{1+(x_1-\theta)^2}+\frac{2(x_2-\theta)}{1+(x_2-\theta)^2}=0$$
$$x_1-\theta+(x_1-\theta)(x_2-\theta)^2+x_2-\theta+(x_2-\theta)(x_1-\theta)^2=0$$
$$\left((x_1-\theta)(x_2-\theta)+1\right)(x_1-\theta+x_2-\theta)=0\Longrightarrow$$
$$\begin{bmatrix}\widehat{\theta}=\overline{X}\\ \widehat{\theta}^2-\widehat{\theta}(X_1+X_2)+X_1X_2+1=0\end{bmatrix};\quad\begin{bmatrix}\widehat{\theta}=\overline{X}\\ \widehat{\theta}=\overline{X}\pm\dfrac{\sqrt{(X_1-X_2)^2-4}}{2}\end{bmatrix}$$

При $|X_1-X_2|\leqslant 2$ второго решения нет, а значит, имеется ОМП $\widehat{\theta}=\overline{X}$. Иначе определено ещё два решения, причём несложно проверить, что именно они будут доставлять максимум, а $\overline{X}$ – локальный минимум. То есть в случае $|X_1-X_2|>2$ оценками максимального правдоподобия будут $\widehat{\theta}=\overline{X}\pm\frac{\sqrt{(X_1-X_2)^2-4}}{2}$. ■

Задача 2.4. Напомним, что плотность гауссовского вектора размерности k равна

$$\rho(x_1,\ldots,x_k)=(2\pi)^{-k/2}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-a)^T\Sigma^{-1}(\mathbf{x}-a)\right).$$

Пусть $X_1,\ldots,X_n\sim\mathcal{N}(a,\Sigma)$ – независимые гауссовские векторы. Найдите оценку максимального правдоподобия для вектора средних a и ковариационной матрицы Σ, где $a\in\mathbb{R}^k$, $\Sigma\in\mathbb{S}^k_{++}$.

Решение. Найдём логарифмическую функцию правдоподобия:

$$f_\theta(X_1,\ldots,X_n)=(2\pi)^{-nk/2}(\det\Sigma)^{-n/2}\exp\left(\sum_{i=1}^n-\frac{1}{2}(X_i-a)^T\Sigma^{-1}(X_i-a)\right),$$
$$\ln f_\theta(X_1,\ldots,X_n)=-\frac{nk}{2}\ln 2\pi-\frac{n}{2}\ln\det\Sigma-\frac{1}{2}\sum_{i=1}^n(X_i-a)^T\Sigma^{-1}(X_i-a).$$

С производной по a всё плюс-минус ясно, хотя для дальнейшего понимания выпишем её через дифференциал:

$$d_a\ln f_\theta(X_1,\ldots,X_n)=-\frac{1}{2}d_a\left(\sum_{i=1}^n(X_i-a)^T\Sigma^{-1}(X_i-a)\right)=$$
$$=-\frac{1}{2}\sum_{i=1}^n\left(d_a(X_i-a)^T\Sigma^{-1}(X_i-a)+(X_i-a)^T\Sigma^{-1}d_a(X_i-a)\right)=$$
$$=\frac{1}{2}\sum_{i=1}^n\left(d_aa^T\Sigma^{-1}(X_i-a)+(X_i-a)^T\Sigma^{-1}d_aa\right)=\frac{1}{2}\sum_{i=1}^n\left(\langle d_aa,\Sigma^{-1}(X_i-a)\rangle+\langle\Sigma^{-T}(X_i-a),d_aa\rangle\right)=$$
$$=\left\langle\sum_{i=1}^n\Sigma^{-1}(X_i-a),d_aa\right\rangle=0\Longrightarrow\sum_{i=1}^n\Sigma^{-1}(X_i-a)=0\Longrightarrow\widehat{a}=\overline{X}.$$

■

Тут даже представляется возможным найти второй дифференциал, тем самым можно показать, что при фиксированной Σ полученная оценка $\widehat{a}$ доставляет максимум функции правдоподобия, но мы на это как обычно забьём.

Куда интереснее найти дифференциал по Σ. Для этого сделаем следующий трюк: сумму в логарифмической функции правдоподобия представим как след от одноэлементной матрицы:

$$\ln f_\theta(X_1,\dots,X_n) = -\frac{nk}{2}\ln 2\pi - \frac{n}{2}\ln\det\Sigma - \frac{1}{2}\operatorname{tr}\sum_{i=1}^{n}(X_i-a)^T\Sigma^{-1}(X_i-a).$$

Это окажется весьма удобным, так как по свойству следа функцию теперь можно записать так:

$$\ln f_\theta(X_1,\dots,X_n) = -\frac{nk}{2}\ln 2\pi - \frac{n}{2}\ln\det\Sigma - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}(X_i-a)(X_i-a)^T\Sigma^{-1} =$$

$$= -\frac{nk}{2}\ln 2\pi - \frac{n}{2}\ln\det\Sigma - \frac{1}{2}\sum_{i=1}^{n}\langle (X_i-a)(X_i-a)^T, \Sigma^{-1}\rangle.$$

Осталось также вспомнить (или загуглить) формулу для дифференциала определителя:

$$d(\det\Sigma) = \det\Sigma\cdot\langle\Sigma^{-T}, d\Sigma\rangle,$$

где $\Sigma^{-T} = (\Sigma^T)^{-1}$ (что в нашем случае просто Σ^{-1}).

Эту формулу на самом деле несложно вывести, если вспомнить, что частная производная определителя по элементу матрицы – это соответствующее алгебраическое дополнение, а у нас как раз есть формула из алгема, связывающее матрицы из алгебраических дополнений и обратную. Но вернёмся к нашим баранам.

Чтобы не вспоминать дифференциал для обратной матрицы, введём замену $\Xi = \Sigma^{-1}$ и будем дифференцировать по ней:

$$d_\Xi \ln f_\theta(X_1,\dots,X_n) = \frac{n}{2}d_\Xi(\ln\det\Xi) - \frac{1}{2}d_\Xi\left(\sum_{i=1}^{n}\langle (X_i-a)(X_i-a)^T,\Xi\rangle\right) =$$

$$= \frac{n}{2\det\Xi}d_\Xi(\det\Xi) - \frac{1}{2}\sum_{i=1}^{n}\left\langle (X_i-a)(X_i-a)^T, d_\Xi\Xi\right\rangle =$$

$$= \left\langle \frac{n}{2}\Xi^{-1} - \frac{1}{2}\sum_{i=1}^{n}(X_i-a)(X_i-a)^T, d_\Xi\Xi\right\rangle = 0 \Longrightarrow \Xi^{-1} = \frac{1}{n}\sum_{i=1}^{n}(X_i-a)(X_i-a)^T.$$

С учётом того, что оценку для a мы нашли ранее, получаем итоговый ответ:

$$\widehat{a} = \overline{X},\quad \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})^T.$$

2.3 Метод выборочных квантилей

Один из способов нахождения оценок является рассмотрение порядковых статистик из выборки. Идея в следующем: для непрерывных функций распределения $F_\theta(X_1)\sim U[0,1]$ (строго это доказано в задаче 5.5 из листка по доверительным интервалам). Поэтому если мы возьмём p-ую часть от выборки в порядке возрастания, то последний её элемент при действии F_θ будет *примерно* равен p, а значит, сам он примерно равен $F_\theta^{-1}(p)$. Формализуем вышесказанное.

Определение. *p-квантилью* распределения P называется $z_p = \inf\{x\colon F_{\mathsf{P}}(X)\geqslant p\}$, где $p\in(0;1)$.

Определение. Пусть $X_1,\dots X_n$ – выборка. Статистика

$$z_{n,p} = \begin{cases} X_{(np)+1}, & np\notin\mathbb{Z},\\ X_{(np)}, & np\in\mathbb{Z}\end{cases}$$

называется *выборочным квантилем.*

Теорема (о выборочной квантиле)**.** *Пусть $X_1, \dots X_n$ – выборка из распределения P с плотностью $\rho(x)$. Пусть z_p – p-квантиль распределения P, причем $\rho(x)$ непрерывно дифференцируема в окрестности z_p и $\rho(z_p) > 0$. Тогда*

$$\sqrt{n}(z_{n,p} - z_p) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{\rho^2(z_p)}\right).$$

Особенно часто выделяют случай, когда $p = 1/2$. Этот случай настолько исключительный, что для него придумали другое определение выборочного квантиля:
Определение. *Выборочной медианой* для выборки $X_1, \dots X_n$ называется

$$\widehat{\mu} = \begin{cases} X_{(k+1)}, & n = 2k+1, \\ \dfrac{X_{(k)} + X_{(k+1)}}{2}, & n = 2k \end{cases}$$

Для неё также справедлива теорема выше, только обычно $z_{n,1/2}$ заменяют на $\widehat{\mu}$.
Задача 2.5. Постройте асимптотически нормальную оценку для параметра масштаба в модели распределения Коши:

$$\rho_\theta(x) = \frac{\theta}{\pi(\theta^2 + x^2)}.$$

Решение. Функция распределения будет равняться

$$F_\theta(x) = \int_{-\infty}^{x} \frac{\theta}{\pi(\theta^2+t^2)}\,dt = \left[s = \frac{t}{\theta}\right] = \int_{-\infty}^{\frac{x}{\theta}} \frac{1}{\pi(1+s^2)}\,ds = \frac{1}{\pi}\operatorname{arctg} s\Big|_{-\infty}^{\frac{x}{\theta}} = \frac{1}{\pi}\operatorname{arctg}\frac{x}{\theta} + \frac{1}{2}.$$

Следовательно, $\frac{3}{4}$-квантилью для данного семейства распределений будет $z_{3/4} = \theta$. Тогда по теореме о выборочном квантиле $z_{n,3/4}$ будет асимптотически нормальной оценкой параметра θ:

$$\sqrt{n}(z_{n,3/4} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{3/4(1-3/4)}{\rho^2(\theta)}\right) = \mathcal{N}\left(0, \frac{3\pi^2\theta^2}{4}\right).$$

В качестве поучительного примера решим задачу методом моментов. С пробными функциями он работать не будет, так как у распределения Коши нет матожидания. Поэтому рассмотрим $g(t) = \frac{1}{1+t^2}$. Для неё

$$m(\theta) = \mathsf{E}_\theta g(X_1) = \int_{\mathbb{R}} \frac{\theta}{\pi(\theta^2+t^2)(1+t^2)}\,dt = \frac{\theta}{\pi(1-\theta^2)}\int_{\mathbb{R}}\left(\frac{1}{\theta^2+t^2} - \frac{1}{1+t^2}\right)dt =$$
$$= \frac{\theta}{\pi(1-\theta^2)}\left(\frac{\pi}{\theta} - \pi\right) = \frac{1}{1+\theta} \Longrightarrow \widehat{\theta} = 1\Big/\overline{\frac{1}{1+X^2}} - 1$$

Получилось весьма недурно. Но проверим, как наша оценка в плане асимптотической дисперсии:

$$\sqrt{n}\left(\overline{g(X)} - \frac{1}{1+\theta}\right) \xrightarrow{d_\theta} \mathcal{N}(0, \mathsf{D}_\theta g(X_1))$$

Дисперсию посчитаем с помощью Wolfram Alpha: $\mathsf{D}_\theta g(X_1) = \mathsf{E}_\theta g(X_1)^2 - (\mathsf{E}_\theta g(X_1))^2 = \frac{\theta+2}{2(\theta+1)^2} - \frac{1}{(\theta+1)^2} = \frac{\theta}{2(\theta+1)^2}$. Применяя дельта-метод, получаем, что

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d_\theta} \mathcal{N}\left(0, \frac{\theta}{2(\theta+1)^2}\cdot(1+\theta)^4\right) = \mathcal{N}\left(0, \frac{\theta(\theta+1)^2}{2}\right).$$

Как мы видим, асимптотическая дисперсия оценки по методу моментов получилась на порядок хуже, чем через выборочный квантиль (хотя стоит признать, для маленьких значений θ она будет всё же меньше). Можно воспринимать это как напоминание о том, что топорный метод моментов не всегда даёт лучший результат. ■

3 Сравнение оценок и эффективные оценки

3.1 Сравнение оценок в равномерном подходе

Итак, мы научились с горем пополам строить оценки с различными свойствами. Как же их сравнивать? Один из способов сравнения оценок – введение некоторой функции, которая будет показывать, насколько сильно оценка отличается от истинного значения параметра.

Определение. Борелевская неотрицательная функция двух переменных $g(x, y)$ называется *функцией потерь*.

Пример. 1. $g(x, y) = |x - y|$

2. $g(x, y) = (x - y)^2$ – *квадратичная функция потерь.*

Определение. Пусть θ^* – оценка параметра θ, $g(x, y)$ – функция потерь. Функция $R(\theta^*, \theta) = \mathsf{E}_\theta g(\theta^*, \theta)$ называется *функцией риска* оценки θ^*. Говорят, что оценка θ^* *лучше* оценки $\hat{\theta}$ *в равномерном подходе с функцией потерь* g, если для любого $\theta \in \Theta$ $R(\theta^*, \theta) \leqslant R(\hat{\theta}, \theta)$, причём существует такое θ, что неравенство является строгим. Равномерный подход с квадратичной функцией потерь называют *среднеквадратичным*.

Задача 3.1. Пусть $X_1, \ldots, X_n$ выборка из равномерного распределения на отрезке $U[0, \theta]$. Сравните следующие оценки параметра θ в среднеквадратичном подходе: $2\overline{X}$, $(n + 1)X_{(1)}$, $\frac{n+1}{n}X_{(n)}$.

Решение. Данные оценки являются несмещёнными, а значит, их функция риска будет являться дисперсией.

(а) $R(2\overline{X}, \theta) = \mathsf{D}_\theta 2\overline{X} = \frac{4}{n^2}\mathsf{D}_\theta \sum X_i = \frac{4}{n}\mathsf{D}_\theta X_i = \frac{4}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n}$.

(б) Вспомним, что для $t \in [0, \theta]$ выполнено

$$\mathsf{P}_\theta(X_{(1)} \leqslant t) = 1 - \mathsf{P}_\theta(X_{(1)} > t) = 1 - \prod \mathsf{P}_\theta(X_i > t) = 1 - (1 - \mathsf{P}_\theta(X_1 \leqslant t))^n = 1 - \left(1 - \frac{t}{\theta}\right)^n,$$

а стало быть $\rho_{X_{(1)}}(t) = \frac{n}{\theta}\left(1 - \frac{t}{\theta}\right)^{n-1} I_{[0,\theta]}(t)$. Отсюда можно в лоб посчитать дисперсию

$$R((n+1)X_{(1)}, \theta) = \mathsf{D}_\theta(n+1)X_{(1)} = \mathsf{E}_\theta(n+1)^2 X_{(1)}^2 - (\mathsf{E}_\theta(n+1)X_{(1)})^2 =$$

$$= -\theta^2 + (n+1)^2 \int_0^\theta t^2 \frac{n}{\theta}\left(1 - \frac{t}{\theta}\right)^{n-1} dt = -\theta^2 + n(n+1)^2\theta^2 \int_0^1 s^2(1-s)^{n-1} ds =$$

$$= -\theta^2 + n(n+1)^2\theta^2 B(n, 3) = -\theta^2 + n(n+1)^2\theta^2 \frac{2!(n-1)!}{(n+2)!} = \frac{\theta^2 n}{n+2}.$$

Как видим, дисперсия не стремится к нулю при увеличении размера выборки, поэтому данная оценка весьма плохая в среднеквадратичном подходе.

(в) Тут распределение ищется по-проще: для $t \in [0, \theta]$

$$\mathsf{P}_\theta(X_{(n)} \leqslant t) = \prod \mathsf{P}_\theta(X_i \leqslant t) = \frac{t^n}{\theta^n},$$

поэтому $\rho_{X_{(n)}}(t) = \frac{nt^{n-1}}{\theta^n} I_{[0,\theta]}(t)$. Значит,

$$R\left(\frac{n+1}{n}X_{(n)}, \theta\right) = \mathsf{D}_\theta \frac{n+1}{n}X_{(n)} = \mathsf{E}_\theta\left(\frac{n+1}{n}X_{(n)}\right)^2 - \left(\mathsf{E}_\theta \frac{n+1}{n}X_{(1)}\right)^2 =$$

$$= -\theta^2 + \frac{(n+1)^2}{n^2}\int_0^\theta t^2 \frac{nt^{n-1}}{\theta^n}\, dt = -\theta^2 + \frac{(n+1)^2}{n} \cdot \frac{t^{n+2}}{\theta^n(n+2)}\bigg|_0^\theta =$$

$$= -\theta^2 + \frac{(n+1)^2\theta^2}{n(n+2)} = \frac{\theta^2}{n(n+2)}.$$

Полученная дисперсия убывает даже быстрее, чем, казалось бы, самая логичная и простая оценка $2\overline{X}$, что наводит на определённые мысли. ■

3.2 Информация Фишера и эффективные оценки

— Просто верить? И всё?
— Просто верить. Этого вполне достаточно.

Дети против волшебников

Далее мы по умолчанию предполагаем в вероятностно-статистической модели выполнены *условия регулярности*:

a) носитель распределения (т.е. $\{x\colon \rho_\theta(x) > 0\}$) не зависит от θ;

b) Θ – открытое связное множество в $\mathbb{R}^k$;

c) для любого $\theta \in \Theta$ и для любой статистики $S(X)$ с условием $\mathsf{E}_\theta S(X)^2 < \infty$ выполнено

$$\frac{\partial}{\partial\theta}\mathsf{E}_\theta S(X) = \mathsf{E}_\theta\left(S(X)\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\right);$$

d) $0 < I_X(\theta) < \infty$ для любого $\theta \in \Theta$.

Третье условие может ввести в ужас, хотя на самом деле мы всего лишь хотим дифференцировать по параметру θ:

$$\frac{\partial}{\partial\theta}\mathsf{E}_\theta S(X) = \frac{\partial}{\partial\theta}\int S(\mathbf{x})\rho_\theta(\mathbf{x})\,d\mathbf{x} = \int\frac{\partial}{\partial\theta}\left(S(\mathbf{x})\rho_\theta(\mathbf{x})\right)\,d\mathbf{x} = \int S(\mathbf{x})\rho'_\theta(\mathbf{x})\,d\mathbf{x} =$$
$$= \int S(\mathbf{x})\frac{\rho'_\theta(\mathbf{x})}{\rho_\theta(\mathbf{x})}\rho_\theta(\mathbf{x})\,d\mathbf{x} = \int\left(S(\mathbf{x})\frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{x})\right)\rho_\theta(\mathbf{x})\,d\mathbf{x} = \mathsf{E}_\theta\left(S(X)\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\right),$$

а проверять, можно ли так делать, мы умеем с матана 4 семестра. Для разнообразия можете проверить, что некоторые стандартные семейства удовлетворяют условиям регулярности, но мы будем в них искренне и бесповоротно верить.

Одной из важнейших характеристик выборок является следующая величина:

Определение. *Информацией Фишера* выборки X называется величина

$$I_X(\theta) = \mathsf{D}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\right).$$

В случае многомерного параметра $I_X(\theta)$ определяется как матрица ковариаций градиента $\ln p_\theta(X)$ и называется *информационной матрицей.*

Задача 3.2. (а) Убедитесь, что $I_X(\theta) = \mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\right)^2$

(б) Докажите аддитивность информации Фишера, а именно

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta)$$

для любых двух независимых выборок X и Y. Как следствие, $I_X(\theta) = ni(\theta)$, где $i(\theta)$ – информация, содержащаяся в одном элементе выборки.

Решение. **(а)** Достаточно показать, что $\mathsf{E}_\theta\frac{\partial}{\partial\theta}\ln\rho_\theta(X) = 0$. Это можно доказать, воспользовавшись пунктом c) из условий регулярности и взяв статистику $S(X) \equiv 1$:

$$0 = \frac{\partial}{\partial\theta}(1) = \frac{\partial}{\partial\theta}\mathsf{E}_\theta 1 = \mathsf{E}_\theta\left(1\cdot\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\right).$$

(б) Воспользуемся линейностью дисперсии для независимых случайных величин:

$$I_{(X,Y)}(\theta) = \mathsf{D}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X,Y)\right) = \mathsf{D}_\theta\left(\frac{\partial}{\partial\theta}\ln\left(\rho_\theta(X)\cdot\rho_\theta(Y)\right)\right) = \mathsf{D}_\theta\left(\frac{\partial}{\partial\theta}(\ln\rho_\theta(X)+\ln\rho_\theta(Y))\right) =$$
$$= \mathsf{D}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X)+\frac{\partial}{\partial\theta}\ln\rho_\theta(Y)\right) = \mathsf{D}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\right)+\mathsf{D}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(Y)\right) = I_X(\theta)+I_Y(\theta).$$

■

Следующая оценка показывает, что есть предел мечтаний о дисперсии нашей оценки:

Теорема (неравенство Рао-Крамера)**.** *Для любой несмещённой оценки θ^* для $\tau(\theta)$ с $\mathsf{E}_\theta\theta^{*2}<\infty$ и для любого $\theta\in\Theta$ справедливо неравенство*

$$\mathsf{D}_\theta(\theta^*(X)) \geqslant \frac{(\tau'(\theta))^2}{I_X(\theta)}.$$

В многомерном случае неравенство Рао-Крамера принимает вид

$$\mathsf{cov}(\theta^*(X),\theta^*(X)) \geqslant \frac{\partial\tau}{\partial\theta}I_X^{-1}(\theta)\left(\frac{\partial\tau}{\partial\theta}\right)^T.$$

Неравенство $A\geqslant B$ двух симметричных матриц A и B понимают как неотрицательную определённость матрицы $A-B$.

Возникает вопрос: а как понять, что наша оценка имеет наименьшую дисперсию, которую позволяет нам это неравенство? На него есть весьма простой ответ:
Определение. Оценка θ^* называется *эффективной*, если для неё выполнено равенство в неравенстве Рао-Крамера.

Теорема (критерий эффективности)**.** *Оценка θ^* эффективна тогда и только тогда, когда для любого $\theta\in\Theta$*

$$\theta^*(X)-\tau(\theta) = \tau'(\theta)I_X^{-1}(\theta)(\ln\rho_\theta(X))'_\theta.$$

Но не спешите радоваться: отнюдь не любое семейство распределений позволяет иметь эффективную оценку. Впрочем, есть критерий, дающий понять, для какого класса семейств она имеется, и он весьма широк:
Определение. Семейство $\{\mathsf{P}_\theta\}$, где $\theta=(\theta_1,\ldots,\theta_k)$, принадлежит *экспоненциальному классу распределений*, если обобщённая плотность распределения P_θ имеет вид

$$\rho_\theta(x) = g(x)\exp\left(a_0(\theta)+\sum_{i=1}^k a_i(\theta)u_i(x)\right).$$

Теорема. *Пусть $\theta\subset\mathbb{R}$. Тогда эффективная оценка существует только для экспоненциальных семейств. Более того, в этом случае $\theta^*(X)=\overline{u(X)}$ является эффективной оценкой для $-a_0'(\theta)/a_1'(\theta)$, а любая другая эффективная оценка будет линейной функцией от $\theta^*(X)$.*

Задача 3.3. Убедитесь, что **(а)** $\mathcal{N}(a,\sigma^2)$, **(б)** $Exp(\lambda)$, **(в)** $\Gamma(\alpha,\lambda)$, **(г)** $Beta(\alpha,\beta)$, **(д)** $Bern(p)$, **(е)** $Pois(\lambda)$ принадлежат экспоненциальному классу распределений. Найдите соответствующие значения $u_i(x)$.

Решение. **(а)** $\rho_{(a,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) = \exp\left(\left(\frac{1}{2}\ln\frac{1}{2\pi\sigma^2}-\frac{a^2}{2\sigma^2}\right)-\frac{1}{2\sigma^2}x^2+\frac{a}{\sigma^2}x\right)$. $u_1(x)=x$, $u_2(x)=x^2$.

(б) $\rho_\lambda(x)=\lambda e^{-\lambda x}I(x>0) = I(x>0)\cdot\exp(\ln\lambda-\lambda x)$. $u_1(x)=x$.

(в) $\rho_{(\alpha,\lambda)}(x) = \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} I(x > 0) = I(x > 0) \cdot \exp\left((\alpha \ln\lambda - \ln\Gamma(\alpha)) - \lambda x + (\alpha-1)\ln x\right)$. $u_1(x) = x$, $u_2(x) = \ln x$.

(г) $\rho_{(\alpha,\beta)}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} I(0<x<1) = I(0<x<1) \cdot \exp(-\ln B(\alpha,\beta) + (\alpha-1)\ln x + +(\beta-1)\ln(1-x))$. $u_1(x) = \ln x$, $u_2(x) = \ln(1-x)$.

(д) $\rho_p(x) = p^x(1-p)^{1-x} = \exp(x\ln p + (1-x)\ln(1-p)) = \exp(\ln(1-p) + x\ln\frac{p}{1-p})$. $u_1(x) = x$.

(е) $\rho_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!}\exp(-\lambda + x\ln\lambda)$. $u_1(x) = x$. ■

Задача 3.4. Пусть $X_1, \ldots, X_n$ – выборка из $Pois(\lambda)$. Для каких функций $\tau(\lambda)$ существует эффективная оценка? Найдите $i(\lambda)$ – информацию одного наблюдения выборки.

Решение. Воспользуемся последней теоремой и предыдущей задачей. В совокупности они утверждают, что эффективной оценкой является линейная функция от $\lambda^* = \overline{u_1(X)} = \overline{X}$, причём она оценивает $\tau(\lambda) = -a_0'(\lambda)/a_1'(\lambda)$, где $a_0(\lambda) = -\lambda$, $a_1(\lambda) = \ln\lambda$, то есть $\tau(\lambda) = \lambda$. Из критерия эффективности

$$i(\lambda) = \frac{\tau'(\lambda)\cdot(\ln\rho_\lambda(x))'_\lambda}{\lambda^* - \tau(\lambda)} = \frac{1\cdot(-\lambda + x\ln\lambda - \ln x!)'_\lambda}{x-\lambda} = \frac{x/\lambda - 1}{x-\lambda} = \frac{1}{\lambda}.$$

Впрочем, её несложно найти и напрямую, найдя дисперсию вклада выборки. ■

Задача 3.5. Пусть $X_1, \ldots, X_n$ – выборка из нормального распределения с параметрами (a, σ^2). Найдите эффективную оценку **(а)** параметра a, если σ^2 известно; **(б)** параметра σ^2, если a известно; **(в)** параметра $(a, a^2+\sigma^2)$. **(г)** Существует ли эффективная оценка для параметра (a, σ^2)? Вычислите информацию Фишера одного наблюдения во всех случаях.

Решение. **(а)** Имеем $\rho_a(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) = \exp\left(-\frac{1}{2\sigma^2}x^2\right)\cdot\exp\left(\left(\frac{1}{2}\ln\frac{1}{2\pi\sigma^2} - \frac{a^2}{2\sigma^2}\right) + \frac{a}{\sigma^2}x\right)$. Стало быть, наше модель принадлежит экспоненциальному семейству, а значит, по теореме $\overline{u_1(x)} = \overline{X}$ является эффективной оценкой для $-a_0'(a)/a_1'(a) = -\left(\frac{1}{2}\ln\frac{1}{2\pi\sigma^2} - \frac{a^2}{2\sigma^2}\right)'_a \Big/ \left(\frac{a}{\sigma^2}\right)'_a = a$. Для разнообразия посчитаем информацию одного наблюдения по определению:

$$i(a) = \mathsf{D}_a\left(\frac{\partial}{\partial a}\ln\rho_a(x)\right) = \mathsf{D}_a\left(\frac{x-a}{\sigma^2}\right) = \mathsf{D}_\theta\frac{x}{\sigma^2} = \frac{1}{\sigma^4}\mathsf{D}_a x = \frac{1}{\sigma^2}.$$

(б) Модель всё ещё лежит в экспоненциальном семействе:

$$\rho_{\sigma^2}(x) = \exp\left(-\frac{1}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x-a)^2\right).$$

По теореме имеется эффективная оценка $\widehat{\sigma^2} = \overline{(X-a)^2}$ для $\tau(\sigma^2) = \left(\frac{1}{2}\ln 2\pi\sigma^2\right)'_{\sigma^2} \Big/ \left(\frac{-1}{2\sigma^2}\right)'_{\sigma^2} = \frac{2\pi}{4\pi\sigma^2}\cdot 2\sigma^4 = \sigma^2$. Из критерия эффективности для одноэлементной выборки

$$i(\sigma^2) = \frac{\tau'(\sigma^2)\cdot(\ln\rho_{\sigma^2}(x))'_{\sigma^2}}{\widehat{\sigma^2} - \tau(\sigma^2)} = \frac{1\cdot\left(\frac{-1}{2\sigma^2} + \frac{(x-a)^2}{2\sigma^4}\right)}{(x-a)^2 - \sigma^2} = \frac{1}{2\sigma^4}.$$

(в) Будем пользоваться критерием эффективности. Тут-то и начинается многомерное веселье:

$$\frac{\partial}{\partial a}\ln\rho(X) = \sum\frac{X_i - a}{\sigma^2} \quad \frac{\partial}{\partial\sigma^2}\ln\rho(X) = -\frac{n}{2\sigma^2} + \sum\frac{(X_i-a)^2}{2\sigma^4}.$$

Нахождение информационной матрицы упрощается тем, что на её диагонали стоят $\mathsf{cov}\left(\frac{\partial}{\partial a}\ln\rho(X), \frac{\partial}{\partial a}\ln\rho(X)\right) = I_X(a) = \frac{n}{\sigma^2}$ и $\mathsf{cov}\left(\frac{\partial}{\partial\sigma^2}\ln\rho(X), \frac{\partial}{\partial\sigma^2}\ln\rho(X)\right) = I_X(\sigma^2) = \frac{n}{2\sigma^4}$, которые мы посчитали ранее. С остальными элементами матрицы нам не очень повезло:

$$\mathsf{cov}\left(\frac{\partial}{\partial a}\ln\rho(X), \frac{\partial}{\partial\sigma^2}\ln\rho(X)\right) = \mathsf{cov}\left(\sum\frac{X_i-a}{\sigma^2}, -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\sum X_i^2 - 2a\sum X_i + na^2\right)\right) \ominus$$

Заметим, что ковариация с константой равна 0, а значит, вынося множители за знак ковариации:

$$\circledcirc\frac{1}{2\sigma^6}\mathsf{cov}\left(\sum X_i,\sum X_i^2-2a\sum X_i\right)=\frac{1}{2\sigma^6}\mathsf{cov}\left(\sum X_i,\sum X_i^2\right)-\frac{a}{\sigma^6}\mathsf{cov}\left(\sum X_i,\sum X_i\right)\circledcirc$$

Вспоминаем, что элементы выборки независимы, а в сумме выше выживают лишь ковариации по одинаковым индексам:

$$\circledcirc\frac{1}{2\sigma^6}\sum\mathsf{cov}\left(X_i,X_i^2\right)-\sum\frac{a}{\sigma^6}\mathsf{cov}\left(X_i,X_i\right)=\frac{n}{2\sigma^2}(\mathsf{E}_\theta X_1^3-\mathsf{E}_\theta X_1\cdot\mathsf{E}_\theta X_1^2-2a\mathsf{D}_\theta X_1)\circledcirc$$

3-ий момент либо считаем ручками, либо смотрим в Википедию:

$$\circledcirc\frac{n}{2\sigma^2}(a^3+3a\sigma^2-a(\sigma^2+a^2)-2a\sigma^2)=0.$$

Находим матрицу Якоби для $\tau(a,\sigma^2)=(a,a^2+\sigma^2)$, обращаем информационную матрицу $I_X(\theta)$ (благо это нетрудно) и считаем ответ:

$$\theta^*=\tau'(\theta)I_X^{-1}(\theta)(\ln\rho_\theta(X))'_\theta=\begin{pmatrix}a\\a^2+\sigma^2\end{pmatrix}+\begin{pmatrix}1&0\\2a&1\end{pmatrix}\begin{pmatrix}\frac{\sigma^2}{n}&0\\0&\frac{2\sigma^4}{n}\end{pmatrix}\begin{pmatrix}\sum\frac{X_i-a}{\sigma^2}\\-\frac{n}{2\sigma^2}+\sum\frac{(X_i-a)^2}{2\sigma^4}\end{pmatrix}=$$

$$=\begin{pmatrix}a\\a^2+\sigma^2\end{pmatrix}+\begin{pmatrix}\frac{\sigma^2}{n}&0\\\frac{2a\sigma^2}{n}&\frac{2\sigma^4}{n}\end{pmatrix}\begin{pmatrix}\frac{n}{\sigma^2}(\overline{X}-a)\\-\frac{n}{2\sigma^2}+\frac{n}{2\sigma^4}\left(\overline{X^2}-2a\overline{X}+a^2\right)\end{pmatrix}=$$

$$=\begin{pmatrix}a\\a^2+\sigma^2\end{pmatrix}+\begin{pmatrix}\overline{X}-a\\\overline{X^2}-\sigma^2-a^2\end{pmatrix}=\begin{pmatrix}\overline{X}\\\overline{X^2}.\end{pmatrix}$$

(г) Предположим, что такая оценка θ^* существует. Тогда по критерию эффективности

$$\theta^*=\begin{pmatrix}a\\\sigma^2\end{pmatrix}+\begin{pmatrix}\frac{\sigma^2}{n}&0\\0&\frac{2\sigma^4}{n}\end{pmatrix}\begin{pmatrix}\frac{n}{\sigma^2}(\overline{X}-a)\\-\frac{n}{2\sigma^2}+\frac{n}{2\sigma^4}\left(\overline{X^2}-2a\overline{X}+a^2\right)\end{pmatrix}=\begin{pmatrix}\overline{X}\\\overline{X^2}-2a\overline{X}+a^2\end{pmatrix}.$$

Полученная статистика не является статистикой, так как имеется явная зависимость от параметра. Значит, эффективной оценки для данной модели не существует. ■

Следующая задача носит довольно технический характер, но она проясняет понятие информации Фишера, а также подводит нас к новой теме.

Задача 3.6. Пусть $S(X)$ – статистика, обобщённая плотность которой равна $g_\theta(s)$. Определим информацию Фишера $I_S(\theta)=\mathsf{D}_\theta\left(\frac{\partial}{\partial\theta}\ln g_\theta(S(X))\right)$ и потребуем выполнения условия регулярности $\frac{\partial}{\partial\theta}\mathsf{E}_\theta T(X)=\mathsf{E}_\theta\left(T(X)\frac{\partial}{\partial\theta}\ln g_\theta(S(X))\right)$ для любой $S(X)$-измеримой статистики $T(X)$. Докажите, что

(а) $\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X)|S(X)\right)=\frac{\partial}{\partial\theta}\ln g_\theta(S(X))$;
(б) $I_S(\theta)=\mathsf{cov}(\frac{\partial}{\partial\theta}\ln\rho_\theta(X),\frac{\partial}{\partial\theta}\ln g_\theta(S(X)))$;
(в) $I_S(\theta)\leqslant I_X(\theta)$.

Решение. **(а)** Проверяется по определению. $\frac{\partial}{\partial\theta}\ln g_\theta(S(X))$ будет $S(X)$-измеримой как функция от $S(X)$. По условиям регулярности для любого $C\in\sigma(S)$:

$$\mathsf{E}_\theta\left(I_C(X)\frac{\partial}{\partial\theta}\ln g_\theta(S(X))\right)=\frac{\partial}{\partial\theta}\mathsf{E}_\theta I_C(X)=\mathsf{E}_\theta\left(I_C(X)\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\right).$$

(б) По определению

$$\mathsf{cov}\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X),\frac{\partial}{\partial\theta}\ln g_\theta(S(X))\right)=\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\cdot\frac{\partial}{\partial\theta}\ln g_\theta(S(X))\right)-$$

$$-\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\right)\cdot\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln g_\theta(S(X))\right)=\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\cdot\frac{\partial}{\partial\theta}\ln g_\theta(S(X))\right)\circledcirc$$

так как по задаче 3.2 $\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\right)=0$. Тогда по пункту **(а)** и формуле полного матожидания

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

То, что дисперсия и второй момент у $\frac{\partial}{\partial\theta}\ln g_\theta(S(X))$ совпадают, доказывается аналогично задаче 3.2.

(в) Применяем предыдущий пункт и неравенство Коши-Буняковского:

$$I_S(X)=\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\cdot\frac{\partial}{\partial\theta}\ln g_\theta(S(X))\right)\leqslant\sqrt{\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\right)^2\cdot\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln g_\theta(S(X))\right)^2}=$$
$$=\sqrt{I_X(\theta)\cdot I_S(\theta)}\Longrightarrow I_S(\theta)\leqslant I_X(\theta).$$

■

Мы получили очень любопытный результат: если мы *редуцируем* данные и рассматриваем не всю выборку X, а лишь статистику от неё $S(X)$, то информация Фишера либо остаётся той же, либо уменьшается, что соответствует нашим ожиданиям. Это в очередной раз подтверждает, что $I_X(\theta)$ является показательной мерой того, насколько много данных содержится в выборке.

Но возникает вопрос: а когда достигается равенство в пункте **(в)**? Так как при доказательстве мы использовали неравенство Коши-Буняковского, то равенство будет достигаться, если $\frac{\partial}{\partial\theta}\ln\rho_\theta(X)$ и $\frac{\partial}{\partial\theta}\ln g_\theta(S(X))$ будут линейно зависимыми почти наверное. Но из пункта **(а)** мы знаем, что одно есть УМО от другого, поэтому они обязаны почти наверное совпадать, то есть для каждого $\theta\in\Theta$

$$\frac{\partial}{\partial\theta}\ln\rho_\theta(X)\overset{\mathsf{P}_\theta\text{-п.н.}}{=}\frac{\partial}{\partial\theta}\ln g_\theta(S(X)).$$

Стало быть, выражения под знаками $\frac{\partial}{\partial\theta}$ отличаются на константу, не зависящую от θ (но может быть от X), то есть

$$\rho_\theta(X)=g_\theta(S(X))\cdot h(X).\tag{1}$$

Итог: только статистики $S(X)$, которые удовлетворяют равенству (1), сохраняют информацию при редуцировании данных. Из-за подобного свойства при "сжатии" эти статистики представляют особый интерес с практической точки зрения. Рассмотрим их поподробнее.

4 Достаточные статистики

Определение. Статистика $S(X)$ называется *достаточной* для семейства распределений $\mathcal{P}_\theta$, если условное распределение $\mathsf{P}_\theta(X \in B|S(X) = t)$ одинаково для всех $\theta \in \Theta$.

Интуиция. Неформально "розыгрыш" значения вектора выборки можно разделить на два этапа: сначала выбираем значение для достаточной статистики $S(X)$, а после этого – само значение X в соответствии с условным распределением $\mathsf{P}_\theta(X \in B|S(X) = t)$. Так как оно не зависит от параметра θ, то вся информация о нём хранится в первом этапе, то есть то, какое именно значение X доставляет равенство $S(X) = t$, нас не интересует.

Теорема (критерий факторизации)**.** *Пусть $\mathcal{P}_\theta$ – доминируемое семейство распределений с обобщённой плотностью ρ_θ. Тогда $S(X)$ – достаточная статистика тогда и только тогда, когда обобщённая плотность допускает представление*

$$\rho_\theta(\mathbf{x}) = g_\theta(S(\mathbf{x}))h(\mathbf{x}),$$

где для всех θ функции g_θ и h – борелевские и неотрицательные (сравните с выводом задачи 3.6).

Заметим, что распределения из экспоненциального семейства $\mathcal{P}_\theta$ автоматически имеют достаточные статистики $T(X)$, ведь

$$\rho_\theta(\mathbf{x}) = h(\mathbf{x}) \exp\left(a_0(\theta) + \sum_{i=1}^{k} a_i(\theta)T_i(\mathbf{x})\right), \tag{2}$$

и из критерия факторизации получаем требуемое.

Пример. Рассмотрим пример достаточной статистики не для экспоненциального семейства. Введём модель с равномерным распределением $U(0, \theta)$, где θ – неизвестный параметр. Это семейство распределений имеет плотность $\rho_\theta(x) = \frac{1}{\theta} I(0 < x < \theta)$, а значит, совместная плотность имеет вид

$$\rho_\theta(X) = \frac{1}{\theta^n} I(0 < X_1, \dots X_n < \theta) = I(0 < X_{(1)}) \cdot \frac{I(X_{(n)} < \theta)}{\theta^n} = h(X) \cdot g(T(X), \theta),$$

где $h(X) = I(0 < X_{(1)})$, $g(t, \theta) = \frac{I(t<\theta)}{\theta^n}$, $T(X) = X_{(n)}$. Таким образом, $X_{(n)}$ является достаточной статистикой по критерию факторизации.

Определение. Статистика $S(X)$ называется *полной* для семейства распределений $\mathcal{P}_\theta$, если для любой борелевской функции $f(x)$ выполнено

$$\forall \theta \in \Theta\colon \mathsf{E}_\theta f(S(X)) = 0 \Longrightarrow$$
$$\forall \theta \in \Theta\colon f(S(X)) = 0 \ (\mathsf{P}_\theta\text{-п. н.})$$

Определение по сути говорит, что статистика $S(X)$ выражает параметр единственным образом, то есть вы не можете двумя разными способами несмещённо оценить функцию от параметра, так как иначе матожидание их разности даёт нуль (пример применения подобного рассуждения есть в задаче 4.9). Весьма полезной для нахождения полных достаточных статистик оказывается следующая

Теорема. *Пусть $\theta \in \Theta \subset \mathbb{R}^k$ и для семейства $\mathcal{P}_\theta$ выполняется (2). Пусть кроме того множество значений $(a_1(\theta), \dots, a_k(\theta))$ для $\theta \in \Theta$ содержит внутреннюю точку. Тогда $T(X)$ является полной достаточной статистикой.*

Ключевой особенностью полных достаточных статистик является тот факт, что они из оценки любой степени паршивости могут сделать конфетку:

Определение. Оценка $\widehat{\theta}$ называется *оптимальной*, если она является наилучшей в классе несмещённых оценок в среднеквадратичном подходе.

Теорема (Леман-Шеффе)**.** *Если $S(X)$ – полная достаточная статистика для $\mathcal{P}_\theta$ и $\mathsf{E}_\theta\widehat{\theta}(X) = \tau(\theta)$, то $\theta^*(X) = \mathsf{E}_\theta(\widehat{\theta}(X)|S(X))$ – оптимальная оценка для $\tau(\theta)$.*

Как следствие, функция от полной достаточной статистики заведомо является оптимальной оценкой своего матожидания, так как в силу её S-измеримости её УМО – она же сама.

Задача 4.1. Найдите оптимальные оценки для параметров распределений **(а)** $Bern(p)$; **(б)** $Pois(\lambda)$.

Решение. В обоих пунктах из представления распределений в виде экспоненциального семейства получаем, что $\overline{X}$ будет являться достаточной статистикой. Полнота следует из достаточного условия полноты. Значит, по теореме Лемана-Шеффе данная оценка будет оптимальной. ■

Задача 4.2. По выборке размера $n \geqslant 2$ из распределения $Exp(\lambda)$ найдите оптимальные оценки для **(а)** λ; **(б)** $\tau(\lambda) = \lambda^{1/2}$.

Решение. **(а)** Так как $\mathsf{E}_\theta\overline{X} = \frac{1}{\lambda}$, то логично предположить, что $1/\overline{X}$ даст нам что-то подходящее. Проверим эту догадку. Так как $X_i \sim Exp(\lambda) \sim \Gamma(1,\lambda)$, то $\sum X_i \sim \Gamma(n,\lambda)$ (в силу независимости X_i). Это в свою очередь означает, что

$$\mathsf{E}_\theta\frac{1}{\sum X_i} = \int_0^{+\infty}\frac{1}{x}\frac{\lambda^n x^{n-1}}{\Gamma(n)}e^{-\lambda x}\,dx = \frac{\Gamma(n-1)\lambda}{\Gamma(n)}\underbrace{\int_0^{+\infty}\frac{\lambda^{n-1}x^{n-2}}{\Gamma(n-1)}e^{-\lambda x}\,dx}_{\text{интеграл плотности }\Gamma(n-1,\lambda)} = \frac{\Gamma(n-1)\lambda}{\Gamma(n)} = \frac{\lambda}{n-1}.$$

Таким образом, из теоремы Лемана-Шеффе получаем, что $\frac{n-1}{\sum X_i}$ является требуемой оптимальной оценкой.

(б) Решение аналогично:

$$\mathsf{E}_\theta\frac{1}{\sqrt{\sum X_i}} = \int_0^{+\infty}\frac{1}{\sqrt{x}}\frac{\lambda^n x^{n-1}}{\Gamma(n)}e^{-\lambda x}\,dx = \frac{\Gamma(n-1/2)\sqrt{\lambda}}{\Gamma(n)}\underbrace{\int_0^{+\infty}\frac{\lambda^{n-1/2}x^{n-3/2}}{\Gamma(n-1/2)}e^{-\lambda x}\,dx}_{\text{интеграл плотности }\Gamma(n-1/2,\lambda)} = \frac{\Gamma(n-1/2)\sqrt{\lambda}}{\Gamma(n)}.$$

В данном случае оптимальной оценкой уже будет являться нечто более сложное: $\hat{\theta} = \frac{\Gamma(n)}{\Gamma(n-1/2)\sqrt{\sum X_i}}$

Возникает логичный вопрос: а зачем условие $n \geqslant 2$? Что будет, если рассмотреть случай $n = 1$? Есть два объяснения разной степени обоснованности. Первое: наша оценка в пункте **(а)** тупо перестаёт работать. Такую аргументацию даже могут принять, но куда более удачным является второе объяснение – оптимальной оценки в данном случае просто нет. Действительно, пусть существует $T(X_1)$ такая, что $\mathsf{E}_\theta T(X_1) = \lambda$. В данной модели выполняются условия регулярности (можете проверить), причём так как $T(X_1)$ – оптимальна, то у неё существует второй момент, поэтому можно применить свойство c) регулярности:

$$1 = \frac{\partial}{\partial\lambda}\lambda = \frac{\partial}{\partial\lambda}\mathsf{E}_\theta T(X_1) = \mathsf{E}_\theta\left(T(X_1)\frac{\partial}{\partial\lambda}\ln\rho_\lambda(X_1)\right) = \mathsf{E}_\theta\left(T(X_1)\left(\frac{1}{\lambda} - X_1\right)\right) =$$

$$= \frac{1}{\lambda}\mathsf{E}_\theta T(X_1) - \mathsf{E}_\theta(X_1\cdot T(X_1)) = 1 - \mathsf{E}_\theta(X_1\cdot T(X_1)).$$

Таким образом, $\mathsf{E}_\theta(X_1\cdot T(X_1)) = 0$. Но так как $T(X_1)$ – оценка для λ, то её значения неотрицательны. Если интеграл неотрицательной функции равен нулю, то она почти наверное равна нулю, чего быть не может – противоречие. ■

Задача 4.3. По выборке из распределения $\mathcal{N}(a, \sigma^2)$ постройте оптимальную оценку для вектора параметров $\theta = (a, \sigma^2)$.

Решение. Распишем более подробно функцию правдоподобия для нормального распределения:

$$\rho_\theta(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\sum \frac{(x_i - a)^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{a}{\sigma^2}\sum x_i - \frac{na^2}{2\sigma^2}\right),$$

что является функцией от $(\sum x_i^2, \sum x_i)$ и вектора параметров $\theta = (a, \sigma^2)$. Значит, по критерию факторизации статистика $T(X) = (\sum X_i^2, \sum X_i)$ является достаточной, а следовательно

$$\widehat{a} = \overline{X}, \;\; \widehat{\sigma^2} = \frac{n}{n-1}s^2 = \frac{n}{n-1}(\overline{X^2} - \overline{X}^2)$$

будут являться оптимальными оценками. ■

Задача 4.4. Докажите, что если $\widehat{\theta}(X)$ и $\theta^*(X)$ – две оптимальные оценки, то они равны P_θ-п.н. для любого θ.

Решение. По определению эти оценки несмещённые, а значит, по линейности матожидания оценка $(\widehat{\theta} + \theta^*)/2$ также не смещена. В силу оптимальности $4\mathsf{E}_\theta\widehat{\theta}^2 \leqslant \mathsf{E}_\theta(\widehat{\theta} + \theta^*)^2$ и $4\mathsf{E}_\theta\left(\theta^*\right)^2 \leqslant \mathsf{E}_\theta(\widehat{\theta} + \theta^*)^2$, что при сложении даёт $2\mathsf{E}_\theta\widehat{\theta}^2 + 2\mathsf{E}_\theta\left(\theta^*\right)^2 \leqslant 4\mathsf{E}_\theta\widehat{\theta}\theta^*$. При выделении полного квадрата получаем $\mathsf{E}_\theta(\widehat{\theta} - \theta^*)^2 \leqslant 0$, что, конечно, означает, что для каждого θ выражение под знаком матожидания почти наверное равно 0. ■

Задача 4.5. Пусть $X_1, \ldots, X_n$ – н.о.р.с.в. из распределения $Pois(\lambda)$. Найдите $\mathsf{E}_\lambda(X_1^2 | X_1 + \ldots + X_n)$.

Решение. Из задачи 4.1 мы знаем, что правая часть УМО – полная достаточная статистика, а значит, по теореме Лемана-Шеффе УМО будет являться оптимальной оценкой для матожидания X_1^2, то есть для $\mathsf{E}_\lambda X_1^2 = \mathsf{D}_\lambda X_1 + (\mathsf{E}_\lambda X_1)^2 = \lambda^2 + \lambda$. Таким образом, нам надо каким-то образом найти такую φ, что $\mathsf{E}_\lambda\varphi\left(\sum X_i\right) = \lambda^2 + \lambda$. Что ж, логично начать с

$$\mathsf{E}_\lambda\left(\sum X_i\right)^2 = \mathsf{D}_\lambda\sum X_i + \left(\mathsf{E}_\lambda\sum X_i\right)^2 = n\mathsf{D}_\lambda X_1 + (n\mathsf{E}_\lambda X_1)^2 = n\lambda + n^2\lambda^2.$$

Мы почти у цели. Осталось слегка поменять коэффициент у λ:

$$\mathsf{E}_\lambda\left[(n-1)\sum X_i + \left(\sum X_i\right)^2\right] = (n-1)n\lambda + n\lambda + n^2\lambda^2 = n^2(\lambda + \lambda^2) \Longrightarrow$$

$$\mathsf{E}_\lambda(X_1^2 | X_1 + \ldots + X_n) = \frac{n-1}{n^2}\sum X_i + \frac{1}{n^2}\left(\sum X_i\right)^2.$$

■

Задача 4.6. (теорема Басу) Пусть $S(X)$ – полная достаточная статистика, $A(X)$ – статистика, распределение которой одинаково при всех $\theta \in \Theta$ (*англ.* `ancillary statistic`). Докажите, что $A(X)$ и $S(X)$ независимы.

Решение. Ничего более умного, чем по определению показать независимость событий из σ-алгебр, порождённых $S(X)$ и $A(X)$, тут придумать нельзя, поэтому так и поступим. Пусть $T \in \sigma(A)$, то есть $\exists B \in \mathcal{B}(\mathbb{R}) \quad A^{-1}(B) = T$. Рассмотрим $I_B \circ A(X) = I_T(X)$. Её распределение также независит от θ, так как определяется распределением $A(X)$. Это значит, что $\mathsf{E}_\theta I_T(X)$ является некоторой константой, независящей от θ. То есть $I_T(X)$ является несмещённой оценкой константы $\mathsf{E}_\theta I_T(X)$. Возникает вопрос: а какая есть у этой константы оптимальная оценка, то есть с минимальной дисперсией? Так она же сама и является! Её дисперсия попросту равна нулю, куда уж меньше? Следовательно, по теореме Лемана-Шеффе

$$\mathsf{E}_\theta(I_T(X) | S(X)) = \mathsf{E}_\theta I_T(X).$$

По определению УМО это означает, что для любого $C \in \sigma(S)$:

$$\int_C I_T(X)\, d\mathsf{P}_\theta = \int_C \mathsf{E}_\theta I_T(X)\, d\mathsf{P}_\theta.$$

Но первый интеграл равен $\int I_{T\cap C}(X)\, d\mathsf{P}_\theta = \mathsf{P}_\theta(T \cap C)$, а второй – $\mathsf{E}_\theta I_T(X) \int I_C(X)\, d\mathsf{P}_\theta = \mathsf{E}_\theta I_T(X) \cdot \mathsf{P}_\theta(C) = \mathsf{P}_\theta(T) \cdot \mathsf{P}_\theta(C)$, что и требовалось. ■

Задача 4.7. Докажите, что статистики $\overline{X}$ и s^2, построенные по выборке из нормального распределения, независимы.

Решение. Рассмотрим модель сдвига $\mathcal{N}(a, \sigma^2)$, где a – параметр, а σ^2 – *известная* величина. Распишем плотность, как в задаче 3:

$$\rho_\theta(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\sum \frac{(x_i - a)^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{a}{\sigma^2}\sum x_i - \frac{na^2}{2\sigma^2}\right) =$$

$$= \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum x_i^2\right)}_{h(\mathbf{x})} \cdot \underbrace{\exp\left(\frac{a}{\sigma^2}\sum x_i - \frac{na^2}{2\sigma^2}\right)}_{g(T(\mathbf{x}),a)},$$

где $T(X) = \sum X_i$. Применяем критерий факторизации и понимаем, что $\overline{X}$ – достаточная статистика. Она же будет полной, так как модель принадлежит экспоненциальному семейству, и функция перед $T(X)$ в экспоненте, а именно a/σ^2, подходит под достаточное условие полноты. Осталось только показать, что распределение s^2 не зависит от a, и дело в шляпе – применима теорема Басу. Но это несложно показать, избавившись от параметра в формуле выборочной дисперсии: так как a – матожидание X_i, а σ^2 – её дисперсия, то X_i можно представить в виде $a + \sigma\xi_i$, где ξ_i имеет стандартное нормальное распределение. Но тогда

$$s^2 = \sum \left(X_i - \overline{X}\right)^2 = \sum \left(a + \sigma\xi_i - \overline{a + \sigma\xi}\right)^2 = \sigma^2 \sum \left(\xi_i - \overline{\xi}\right)^2.$$

Последнее выражение – функция от выборки из независимых величин, которые распределены одинаково вне зависимости от a, поэтому её распределение также не зависит от a. ■

Задача 4.8. Пусть X_1, X_2, X_3 – н.о.р.с.в. из распределения $Exp(\lambda)$. Докажите, что $X_1+X_2+X_3 \perp\!\!\!\perp \frac{X_1}{X_1+X_2+X_3}$.

Решение. Идея не шибко отличается от предыдущей задачи: полнота и достаточность $X_1 + X_2 + X_3$ проверяется аналогично. Так как $X_i \sim Exp(\lambda)$, то $X_i = \xi_i/\lambda$, где $\xi_i \sim Exp(1)$. Значит, для любого $c \in \mathbb{R}$

$$\mathsf{P}_\lambda\left(\frac{X_1}{X_1 + X_2 + X_3} \leqslant c\right) = \mathsf{P}_\lambda\left(\frac{\xi_1}{\xi_1 + \xi_2 + \xi_3} \leqslant c\right) = \mathsf{P}_\lambda\left(\xi_1 \leqslant c(\xi_1 + \xi_2 + \xi_3)\right) = F_{(1-c)\xi_1 - c\xi_2 - c\xi_3}(0),$$

что определяется свёрткой независимых случайных величин ξ_i, а стало быть определяется целиком и полностью c, и от λ не зависит. ■

Задача 4.9. Убедитесь, что для семейства распределений $\mathcal{N}(\theta, \theta^2)$ не существует полной достаточной статистики.

Решение. *Идея.* Мы уже знаем, что $(\sum X_i^2, \sum X_i)$ является достаточной статистикой, но теперь параметр лишь один, и есть подозрения, что сейчас эта оценка несёт в себе слишком много информации, что наводит на мысли о неполноте. Надо показать, что, во-первых, она не будет полной, а во-вторых, и это самое главное, любая другая достаточная статистика будет априори *богаче* данной, и из этого мы выведем, что она тем паче не будет полной.

Неполнота $(\sum X_i^2, \sum X_i)$ получается из выражения параметра θ двумя способами:

$$\left.\begin{aligned}\mathsf{E}_\theta\left(\sum X_i\right)^2 = \mathsf{D}_\theta \sum X_i + \left(\mathsf{E}_\theta \sum X_i\right)^2 = (n+n^2)\theta^2 \\ \mathsf{E}_\theta \sum X_i^2 = n\mathsf{E}_\theta X_i^2 = n(\mathsf{D}_\theta X_i + (\mathsf{E}_\theta X_i)^2) = 2n\theta^2\end{aligned}\right\} \Longrightarrow \mathsf{E}_\theta\left[2\left(\sum X_i\right)^2 - (n+1)\sum X_i^2\right] = 0,$$

при этом выражение под знаком матожидания не равно нулю почти наверное. Обратите внимание, что мы не используем здесь признак полноты из начала параграфа, так как это лишь достаточное условие!

Пусть нашлась достаточная статистика $T(X)$, то есть по критерию факторизации $\rho_\theta(X) = g(T(X),\theta)\cdot h(X)$. Также мы знаем, что

$$\rho_\theta(X) = \frac{1}{(2\pi\theta^2)^{n/2}}\exp\left(-\frac{1}{2\theta^2}\sum X_i^2 + \frac{1}{\theta}\sum X_i - \frac{n}{2}\right),$$

Очень хочется показать, что $\sum X_i^2$ и $\sum X_i$ на самом деле выражаются через $T(X)$. Постараемся подобрать θ так, чтобы и избавиться от ненужной $h(X)$, и убрать, например, $\sum X_i$, оставив наедине $T(X)$ и $\sum X_i^2$:

$$\frac{\rho_1(X)\rho_{1/3}(X)}{\rho_{1/2}(X)\rho_{1/2}(X)} = \frac{g(T(X),1)g(T(X),1/3)}{g(T(X),1/2)g(T(X),1/2)} = C\cdot\exp\left((-1/2-9/2+2+2)\sum X_i^2\right) \Longrightarrow$$

$$\sum X_i^2 = -\ln\left(\frac{g(T(X),1)g(T(X),1/3)}{Cg(T(X),1/2)g(T(X),1/2)}\right)$$

В правой части – функция от $T(X)$, что мы и хотели. Аналогично можно показать, что $\sum X_i$ – функция от $T(X)$, то есть $(\sum X_i^2, \sum X_i) = \varphi(T(X))$, где φ – некая борелевская функция.

Это и показывает тот факт, что $T(X)$ богаче $\left(\sum X_i^2, \sum X_i\right)$, ведь если статистика есть борелевская функция от другой статистики, то σ-алгебра первой есть подмножество второй. Иными словами, $\left(\sum X_i^2, \sum X_i\right)$ является *минимальной достаточной статистикой*. Занятно, что минимальная достаточная σ-алгебра существует *всегда*. Правда нам всё же удобнее работать со статистиками, а с отысканием минимальных достаточных статистик всё не так просто, и этому можно посвятить отдельный параграф (например, [1, § 23]).

Почему же из этого следует, что $T(X)$ точно не полная? Предположим, что это не так, и $T(X)$ всё-таки полная. Но тогда несложно по определению проверить, что полной окажется $(\sum X_i^2, \sum X_i)$. Действительно, если

$$\mathsf{E}_\theta f\left(\sum X_i^2, \sum X_i\right) = 0 \Longrightarrow \mathsf{E}_\theta f(\varphi(T(X))) = 0 \Longrightarrow \text{из полноты } T(X)$$

$$f(\varphi(T(X))) = f\left(\sum X_i^2, \sum X_i\right) = 0 \ (\mathsf{P}_\theta - \text{п. н.}) \Longrightarrow \left(\sum X_i^2, \sum X_i\right) \text{ – полная},$$

откуда и получаем заветное противоречие. ■

5 Доверительные интервалы

Конечно же, точечные оценки, которые мы составляли ранее, почти наверное не совпадут с истинным значением параметра. Но нам ведь этого и не надо: достаточно того, чтобы они различались не очень сильно. Можно подойти к этой проблеме иначе: локализовать параметр в некотором интервале, куда он попадёт с некоторой высокой фиксированной вероятностью.

Определение. *Доверительным интервалом уровня доверия* γ для параметра θ называется пара статистик $(T_1(X), T_2(X))$, такая, что для любого $\theta \in \Theta$

$$\mathsf{P}_\theta(T_1(X) < \theta < T_2(X)) \geqslant \gamma.$$

Если выполнено равенство $\mathsf{P}_\theta(T_1(X) < \theta < T_2(X)) = \gamma$, то доверительный интервал называется *точным*.

Не всегда получается легко построить такие интервалы, чтобы вероятность попадания в них была априори выше нужного числа. Но можно лишь потребовать, чтобы неравенство выполнялось в пределе, что в случае большой выборки не будет нас сильно ограничивать.

Определение. Если

$$\varliminf_{n\to\infty} \mathsf{P}_\theta(T_1^{(n)}(X) < \theta < T_2^{(n)}(X)) \geqslant \gamma,$$

то $(T_1(X), T_2(X))$ называется *асимптотическим доверительным интервалом уровня доверия* γ. Аналогично предыдущему определению, интервал называется *точным*, если предел в точности равен γ.

Хотя формально нет никаких условий на длину интервала, имеет смысл выбирать T_1, T_2 такими, чтобы длина интервала была как можно меньше, в частности, $T_2^{(n)}(X) - T_1^{(n)}(X)$ должно стремиться к 0 для всех θ (если это возможно). Приведём два самых простых метода нахождения обычного и асимптотического доверительного интервала.

Метод 1. Использование центральной статистики

Определение. Функция $G(\mathbf{x}, \theta)$ называется *центральной статистикой*, если

1. распределение $G(X, \theta)$ не зависит от θ для всех $\theta \in \Theta$;
2. при каждом $\mathbf{x} \in \mathbb{R}^n$ функция $g(\mathbf{x}, \theta)$ непрерывна и строго убывает (возрастает) по θ.

Обозначим p-квантиль распределения $G(X, \theta)$ через x_p. Возьмем $0 \leqslant p_1 < p_2 \leqslant 1$ такие, что $p_2 - p_1 = \gamma$. Определим $T_1(x)$ и $T_2(x)$ как решения относительно θ соответственно уравнений $G(X, \theta) = x_{p_1}$ и $G(X, \theta) = x_{p_2}$. Наличие решения гарантируется определением выше. Тогда из монотонности $G(X, \theta)$ получаем, что $\mathsf{P}_\theta(T_1(X) < \theta < T_2(X)) = \mathsf{P}_\theta(x_{p_1} < G(X, \theta) < x_{p_2}) = p_2 - p_1 = \gamma$. Удобно брать $p_2 = \frac{1+\gamma}{2}$ и $p_1 = \frac{1-\gamma}{2}$ (в таком случае интервал называется *центральным*), особенно когда распределение $G(X, \theta)$ симметрично относительно начала координат.

В последних задачах мы познакомимся с универсальным способом построения центральной статистики для непрерывных распределений.

Метод 2. Использование асимптотически нормальных оценок

Допустим на нас с неба свалилась асимптотически нормальная оценка θ_n^*, то есть $\sqrt{n}(\theta_n^* - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$ при $n \to \infty$. Потребуем, чтобы асимптотическая дисперсия $\sigma^2(\theta)$ была положительна и непрерывна при всех $\theta \in \Theta$. Рассмотрим

$$\frac{\sqrt{n}(\theta_n^* - \theta)}{\sigma(\theta_n^*)} = \frac{\sqrt{n}(\theta_n^* - \theta)}{\sigma(\theta)} \cdot \frac{\sigma(\theta)}{\sigma(\theta_n^*)}$$

Первый множитель сходится к стандартно нормально распределённой случайной величине при $n \to \infty$. Разберёмся со вторым множителем. Так как θ_n^* асимптотически нормальна, то она состоятельна, то есть $\theta_n^* \xrightarrow{\mathsf{P}_\theta} \theta$. Тогда из непрерывности асимптотической дисперсии $\sigma(\theta_n^*) \xrightarrow{\mathsf{P}_\theta} \sigma(\theta)$, то есть $\frac{\sigma(\theta)}{\sigma(\theta_n^*)} \xrightarrow{\mathsf{P}_\theta} 1$. Отсюда из леммы Слуцкого получаем, что всё произведение сходится к

чему-то нормальному. Если обозначить за x_p p-квантиль для $\mathcal{N}(0,1)$, то из сходимости по распределению следует

$$\mathsf{P}_\theta\left(\theta_n^* - \frac{x_{(1+\gamma)/2}\sigma(\theta_n^*)}{\sqrt{n}} < \theta < \theta_n^* + \frac{x_{(1+\gamma)/2}\sigma(\theta_n^*)}{\sqrt{n}}\right) = \mathsf{P}_\theta\left(\left|\sqrt{n}\cdot\frac{\theta_n^*-\theta}{\sigma(\theta_n^*)}\right| < x_{(1+\gamma)/2}\right) \to \gamma.$$

Обратите внимание, что тут мы используем квантили с одинаковым индексом $x_{(1+\gamma)/2}$, но при этом знаки перед ними в левой и правой части неравенства разные. Правильным будет также написать интервал как

$$\left(\theta_n^* - \frac{x_{(1+\gamma)/2}\sigma(\theta_n^*)}{\sqrt{n}}; \theta_n^* - \frac{x_{(1-\gamma)/2}\sigma(\theta_n^*)}{\sqrt{n}}\right),$$

так как в силу симметричности распределения $x_{(1-\gamma)/2} = -x_{(1+\gamma)/2}$. Чаще всего для симметричных распределений мы будем расписывать интервалы через одинаковые квантили, потому что эстетически так красивее.

Задача 5.1. По выборке из распределения $U(0,\theta)$ постройте точный доверительный интервал уровня доверия γ для параметра θ, границы которого являются функциями от $X_{(n)}$. Посчитайте асимптотику длины интервала при $n \to \infty$.

Решение. Рассмотрим функцию $G(X,\theta) = \frac{X_{(n)}}{\theta}$. Она будет центральной статистикой, поскольку для $t \in [0,1]$ $\mathsf{P}_\theta(G(X,\theta) \leqslant t) = \mathsf{P}_\theta(X_{(n)} \leqslant t\theta) = \frac{(t\theta)^n}{\theta^n} = t^n$, а значит, её распределение не зависит от θ. Из этого представления легко найти квантиль распределения: $x_p = \sqrt[n]{p}$. Таким образом,

$$\mathsf{P}_\theta\left(\frac{X_{(n)}}{\sqrt[n]{\frac{1+\gamma}{2}}} < \theta < \frac{X_{(n)}}{\sqrt[n]{\frac{1-\gamma}{2}}}\right) = \mathsf{P}_\theta\left(\sqrt[n]{\frac{1-\gamma}{2}} < \frac{X_{(n)}}{\theta} < \sqrt[n]{\frac{1+\gamma}{2}}\right) = \gamma.$$

Примем для простоты $\alpha = \frac{1-\gamma}{2}$, $\beta = \frac{1+\gamma}{2}$. Длина интервала может быть высчитана как

$$X_{(n)}(\alpha^{-1/n} - \beta^{-1/n}) \approx \theta\left(1 - \frac{\ln\alpha}{n} - 1 + \frac{\ln\beta}{n}\right) = \frac{\theta}{n}\ln\frac{\beta}{\alpha}$$

■

Задача 5.2. По выборке из распределения Коши с плотностью $\rho_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)}$ постройте точный асимптотический доверительный интервал для θ уровня доверия γ.

Решение. Как мы знаем, медиана μ для распределения Коши является асимптотически нормальной оценкой параметра θ:

$$\sqrt{n}(\mu-\theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4\rho_\theta^2(\theta)}\right) = \mathcal{N}\left(0, \frac{\pi^2}{4}\right).$$

Таким образом, по методу 2 мы получаем точный асимптотический доверительный интервал:

$$\mathsf{P}_\theta\left(\mu - \frac{x_{(1+\gamma)/2}\pi}{2\sqrt{n}} < \theta < \mu + \frac{x_{(1+\gamma)/2}\pi}{2\sqrt{n}}\right) \to \gamma,$$

где x_p – p-квантиль для стандартного нормального распределения. ■

Задача 5.3. По выборке из распределения $\Gamma(\alpha,\gamma)$ постройте доверительный интервал уровня доверия γ для параметра λ, если α известно.

Решение. Заметим, что если $X_i \sim \Gamma(\alpha,\lambda)$, то $\lambda X_i \sim \Gamma(\alpha,1)$. Это значит, что $\lambda\sum X_i$ является центральной статистикой с распределением $\Gamma(n\alpha,1)$. Поэтому

$$\mathsf{P}_\lambda\left(\frac{y_{(1-\gamma)/2}}{\sum X_i} < \lambda < \frac{y_{(1+\gamma)/2}}{\sum X_i}\right) = \gamma,$$

где y_p – p-квантиль распределения $\Gamma(n\alpha,1)$. ■

Задача 5.4. Дана выборка из распределения $\mathcal{N}(a, \sigma^2)$.

(а) Проверьте, что $\frac{ns^2}{\sigma^2} \sim \chi^2(n-1)$.

(б) Найдите точную доверительную область уровня доверия γ для вектора параметров $\theta = (a, \sigma^2)$.

Решение. **(а)** Представим $X_i = a + \sigma\xi_i$, где $\xi_i \sim \mathcal{N}(0,1)$. Распишем выборочную дисперсию:

$$\frac{ns^2}{\sigma^2} = \frac{1}{\sigma^2}\sum \left(X_i - \overline{X}\right)^2 = \frac{1}{\sigma^2}\sum \left(a + \sigma\xi_i - \overline{a+\sigma\xi}\right)^2 = \sum \left(\xi_i - \overline{\xi}\right)^2 =$$
$$= \sum \xi_i^2 - \frac{1}{n}\left(\sum \xi_i\right)^2 = \sum \xi_i^2 - \left(\sum \frac{\xi_i}{\sqrt{n}}\right)^2.$$

Способ I. Докажем сперва следующее

Утверждение. *Если $\xi_1 \perp\!\!\!\perp \xi_2$, и $\xi_1 \sim \chi^2(n)$, $\xi_1 + \xi_2 \sim \chi^2(n+m)$, то $\xi_2 \sim \chi^2(m)$.*

Доказательство. Из свойства хар. функций: $\varphi_{\xi_1}(t) \cdot \varphi_{\xi_2}(t) = \varphi_{\xi_1+\xi_2}(t)$, поэтому $\varphi_{\xi_2}(t) = \frac{\varphi_{\xi_1+\xi_2}(t)}{\varphi_{\xi_1}(t)}$. Но так как $\chi^2(n) = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$, то в равенство выше можно подставить хар. функцию от гамма-распределения (напомним, что для $\zeta \sim \Gamma(\alpha, \lambda)$ она равна $\varphi_\zeta(t) = \left(1 - \frac{it}{\lambda}\right)^{-\alpha}$):

$$\varphi_{\xi_2}(t) = \frac{(1-2it)^{-\frac{n+m}{2}}}{(1-2it)^{-\frac{n}{2}}} = (1-2it)^{-\frac{m}{2}},$$

что есть хар. функция для $\Gamma\left(\frac{m}{2}, \frac{1}{2}\right) = \chi^2(m)$. Значит, по теореме о единственности ξ_2 имеет в точности распределение $\chi^2(m)$. □

Осталось осознать, что $\eta = \left(\sum \frac{\xi_i}{\sqrt{n}}\right) \sim \mathcal{N}(0,1)$ (а значит, $\eta^2 \sim \chi^2(1)$), $ns^2/\sigma^2 + \eta^2 = \sum \xi_i^2 \sim \chi^2(n)$, а $ns^2/\sigma^2 \perp\!\!\!\perp \eta^2$ по задаче 4.7 из предыдущего листка.

Способ II. Рассмотрим такую ортогональную матрицу A (то есть $AA^T = 1$), что первая её строчка равна $(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$. Очевидно, этот вектор можно дополнить до ортонормированного базиса, а стало быть такая A существует. Тогда $\eta = A\xi$, где $\xi = (\xi_1, \ldots, \xi_n)^T$ и $\eta = (\eta_1, \ldots, \eta_n)^T$, является гауссовым вектором с нулевым матожиданием и матрицей ковариаций $AEA^T = E$, то есть η_i – независимые и стандартно нормально распределены. При этом в силу ортогональности длина вектора не меняется, что есть $\sum \xi_i^2 = \sum \eta_i^2$. Таким образом, выборочная дисперсия выше перепишется как

$$\frac{ns^2}{\sigma^2} = \sum_{i=1}^n \xi_i^2 - \left(\sum_{i=1}^n \frac{\xi_i}{\sqrt{n}}\right)^2 = \sum_{i=1}^n \eta_i^2 - \eta_1^2 = \sum_{i=2}^n \eta_i^2 \sim \chi^2(n-1).$$

(б) Теперь у нас есть две статистики, распределения которых не зависят от вектора параметров θ – это $\frac{ns^2}{\sigma^2} \sim \chi^2(n-1)$ и $\frac{\sqrt{n}(\overline{X}-a)}{\sigma} \sim \mathcal{N}(0,1)$. Из задачи 4.7 предыдущего листка следует, что эти статистики будут независимыми, а стало быть вероятность попадания в окрестность этих статистик равна произведению вероятностей. Это и даёт нам нужную область:

$$\gamma = \sqrt{\gamma}\sqrt{\gamma} = \mathsf{P}_\theta\left(x_{(1-\sqrt{\gamma})/2} < \frac{\sqrt{n}(\overline{X}-a)}{\sigma} < x_{(1+\sqrt{\gamma})/2}\right)\mathsf{P}_\theta\left(z_{(1-\sqrt{\gamma})/2} < \frac{ns^2}{\sigma^2} < z_{(1+\sqrt{\gamma})/2}\right) =$$
$$= \mathsf{P}_\theta\left(\overline{X} - \frac{\sigma x_{(1+\sqrt{\gamma})/2}}{\sqrt{n}} < a < \overline{X} + \frac{\sigma x_{(1+\sqrt{\gamma})/2}}{\sqrt{n}},\ \frac{ns^2}{z_{(1+\sqrt{\gamma})/2}} < \sigma^2 < \frac{ns^2}{z_{(1-\sqrt{\gamma})/2}}\right).$$

■

Задача 5.5. Дана выборка $X_1, \ldots, X_n$ из семейства распределений с непрерывными функциями распределения $F_\theta(x)$. Убедитесь, что

$$G(X, \theta) = -\sum_{i=1}^{n} \ln F_\theta(X_i) \sim \Gamma(n, 1),$$

и, как следствие, является центральной статистикой.

Решение. Убедимся, что $F_\theta(X_i)$ распределена как $U(0, 1)$. Для $t \notin (0, 1)$ равенство их функций распределения очевидно. Иначе

$$\mathsf{P}_\theta(F_\theta(X_i) \leqslant t) = \mathsf{P}_\theta(X_i \leqslant F_\theta^{-1}(t)) = F_\theta(F_\theta^{-1}(t)) = t,$$

где в качестве $F_\theta^{-1}(t)$ можно взять любую точку из прообраза t при действии F_θ (он не пуст в силу непрерывности F_θ). Под действием $\varphi(t) = -\ln t$ распределение становится экспоненциальным. Действительно, если раньше $\rho(t) = I(0 < t < 1)$, то теперь плотность равна $\tilde{\rho}(x) = |(\varphi^{-1}(x))'|\rho(\varphi^{-1}(x)) = e^{-x} I(0 < e^{-x} < 1) = e^{-x} I(x > 0)$, то есть $-\ln F_\theta(X_i) \sim Exp(1) = \Gamma(1, 1)$. Отсюда $G(X, \theta)$ как сумма независимых случайных величин распределена как $\Gamma(n, 1)$. ■

Задача 5.6. Дана выборка из распределения $Pareto(\theta, 1)$, $\theta > 0$. Постройте точный доверительный интервал уровня доверия γ для параметра θ.

Решение. Просто применяем предыдущую задачу. Для распределения Парето функция распределения имеет вид $F_\theta(t) = 1 - t^{-\theta}$, $t \geqslant 1$. Для упрощения заметим, что $1 - F_\theta(X_i)$ в силу симметричности распределена также равномерно на $[0, 1]$, поэтому статистика

$$G(X, \theta) = -\sum_{i=1}^{n} \ln\left(1 - F_\theta(X_i)\right) = \theta \sum_{i=1}^{n} \ln X_i$$

является центральной и распределена как $\Gamma(n, 1)$. Поэтому если принять u_p за p-квантиль такого распределения, то

$$\mathsf{P}_\theta\left(\frac{u_{(1-\gamma)/2}}{\sum \ln X_i} < \theta < \frac{u_{(1+\gamma)/2}}{\sum \ln X_i}\right) = \mathsf{P}_\theta\left(u_{(1-\gamma)/2} < G(X, \theta) < u_{(1+\gamma)/2}\right) = \gamma.$$

■

6 Контрольная работа №1: условия

Следующие задачи предлагались группе Б05-024 на контрольной работе. Задачи 3 и 4 у обоих вариантов совпадают

Левый вариант

1. Пусть $X_1, \ldots, X_n$ – выборка из распределения с плотностью

$$\rho_\theta(x) = \frac{1+\theta x}{2} I(-1 \leqslant x \leqslant 1), \quad \theta \in \Theta = [-1; 1].$$

Рассмотрим оценку $\widehat{\theta}(X) = 3\overline{X}$. Является ли она несмещённой, состоятельной, сильно состоятельной, асимптотически нормальной?

2. Дана выборка $X_1, \ldots, X_n$ из распределения Лапласа со сдвигом $\theta > 0$ (плотность равна $\rho_\theta(x) = \frac{1}{2}e^{-|x-\theta|}$). Постройте асимптотический доверительный интервал уровня доверия γ для параметра θ.

3. **(а)** Пусть обобщённая плотность экспоненциального семейства распределений P_θ имеет вид

$$\rho_\theta(x) = g(x)\exp\left(u(x)\theta - b(\theta)\right).$$

Докажите, что $\mathsf{E}_\theta u(X) = \frac{\partial}{\partial\theta} b(\theta)$ и $\mathsf{D}_\theta u(X) = \frac{\partial^2}{\partial\theta^2} b(\theta)$.

(б) По выборке $X_1, \ldots, X_n$ из распределения с плотностью

$$\rho_\theta(x) = \frac{\theta e^{-x}}{(1+e^{-x})^{\theta+1}}$$

постройте оптимальную оценку для $\tau(\theta) = \frac{1}{\theta^2}$.

4. Приведите пример состоятельной, но не асимптотически нормальной оценки

Замечание. Мы считаем константу нормально распределённой.

Правый вариант

1. По выборке $X_1, \ldots, X_n$ из распределения с плотностью

$$\rho_\theta(x) = \frac{1}{x\theta\sqrt{2\pi}} \exp\left(-\frac{\ln^2 x}{2\theta^2}\right) I(x > 0)$$

постройте сильно состоятельную оценку параметра θ.

2. Пусть $X_1, \ldots, X_n$ – выборка из отрицательного биномиального распределения $\mathrm{NB}(M, \theta)$, то есть $\mathsf{P}_\theta(X_i = k) = C_{k+m-1}^{m-1}(1-\theta)^k\theta^m$ для $k \in \{0, 1, \ldots\}$. Параметр m известен. Для каких функций $\tau(\theta)$ существует эффективная оценка? Найдите $i(\theta)$ – информацию одного наблюдения выборки.

3. **(а)** Пусть обобщённая плотность экспоненциального семейства распределений P_θ имеет вид

$$\rho_\theta(x) = g(x)\exp\left(u(x)\theta - b(\theta)\right).$$

Докажите, что $\mathsf{E}_\theta u(X) = \frac{\partial}{\partial\theta} b(\theta)$ и $\mathsf{D}_\theta u(X) = \frac{\partial^2}{\partial\theta^2} b(\theta)$.

(б) По выборке $X_1, \ldots, X_n$ из распределения с плотностью

$$\rho_\theta(x) = \frac{\theta e^{-x}}{(1+e^{-x})^{\theta+1}}$$

постройте оптимальную оценку для $\tau(\theta) = \frac{1}{\theta^2}$.

4. Приведите пример состоятельной, но не асимптотически нормальной оценки

Замечание. Мы считаем константу нормально распределённой.

7 Байесовские оценки

7.1 Мотивация и определения

Всё это время мы не воспринимали множество параметров как нечто, имеющее сложную структуру. Максимум мы пользовались какими-то свойствами $\mathbb{R}^n$, откуда чаще всего и приходят параметры, но дополнительными свойствами или информацией множество Θ мы не наделяли. Это может показаться логичным, весь если мы не знаем истинного значения параметра, то и само множество его теоретических значений скорее всего также покрыто туманом войны. Но в реальной жизни мы не только имеем дело с «приятными» множествами допустимых параметров, но также имеем некоторую информацию о них.

Рассмотрим пример с распределением Вейбулла из второй домашки по практикуму. В ней мы убедились, что состояние банковского счёта имеет вышеуказанное распределение, и даже с помощью ОМП находили оценку для параметра этого распределения. Но ведь этот параметр может быть разным для разных счетов, он зависит от человека, который этим счётом владеет. Стало быть, параметр в данной ситуации – это не что-то фиксированное, что осталось лишь оценить, а в каком-то смысле *случайная величина*, которая к тому же обладает некоторым *распределением*. И если мы знаем это распределение или догадываемся о его природе, исходя из опыта предыдущих наблюдений, то это может помочь в оценке очередного параметра в очередной такой модели.

Это подводит нас к *байесовскому подходу* в оценивании параметра, в котором известную информацию о нём мы заключаем в некотором распределении Q на множестве Θ (чаще всего под Θ будет подразумеваться множество из $\mathbb{R}^n$, поэтому эта мера определяется на борелевских подмножествах из $\mathcal{B}(\Theta)$, если не обговорено иного).

Более формально, теперь мы работаем в новом вероятностном пространстве

$$(\Theta \times \mathcal{X}, \mathcal{B}(\Theta) \otimes \mathcal{B}(\mathcal{X}), \widetilde{\mathsf{P}}),$$

где $\mathcal{B}(\Theta)\otimes\mathcal{B}(\mathcal{X})$ – прямое произведение сигма-алгебр, а мера $\widetilde{\mathsf{P}}$ задаётся обобщённой плотностью $\rho_{\theta,X}(t, \mathbf{x}) = q(t) \cdot \rho_t(\mathbf{x})$. Обобщённая плотность $\rho_t(\mathbf{x})$, как и ранее, отвечает за распределение выборки при фиксированном значении параметра, а новый персонаж – $q(t)$ – за распределение на множестве параметров.

Определение. Плотность $q(t)$, $t \in \Theta$, называется *априорной*.

Априори означает то, что эта плотность известна нам *до* момента проведения наблюдения, то есть она является чем-то типа прикидки того, каким может быть параметр.

При этом когда наблюдение уже проведено, ясно, что наше мнение о параметре изменилось – выборка подсказывает нам, в какую сторону нужно идти, чтобы оценить параметр.

Определение. Условная плотность параметра θ при условии выборки $X_1, \ldots, X_n$, которая (напоминаем) может быть высчитана по формуле

$$\rho_{\theta|X}(t|\mathbf{x}) = \frac{\rho_{\theta,X}(t, \mathbf{x})}{\rho_X(\mathbf{x})} = \frac{q(t)\rho_t(\mathbf{x})}{\int_\Theta q(s)\rho_s(\mathbf{x})\, d\mu(s)},$$

называется *апостериорной плотностью* параметра θ.

Из полученной плотности теперь можно смастерить оценку. Обычно берут среднее значение по плотности, что есть попросту

$$\mathsf{E}(\theta|X) = \int_\Theta t \cdot \rho_{\theta|X}(t|X)\, dt.$$

Обратите внимание, что так как УМО по определению является измеримым относительно X, то $\mathsf{E}(\theta|X) = \varphi(X)$ для некоторой борелевской φ, то есть она зависит только от элементов выборки.

Определение. Оценка $\widehat{\theta} = \mathsf{E}(\theta|X)$ называется *байесовской оценкой параметра* θ.

Польза полученной оценки проявляется в свете следующего определения.
Определение. Говорят, что оценка θ^* *лучше оценки* $\widehat{\theta}$ *в байесовском подходе с функцией потерь* g, если

$$\int_\Theta \mathsf{E}_t g(\theta^*, t)\, d\mu(t) < \int_\Theta \mathsf{E}_t g(\widehat{\theta}, t)\, d\mu(t).$$

Теорема. *Байесовская оценка является наилучшей в байесовском подходе с квадратичной функции потерь.*

7.2 Выбор априорного распределения

Звучит просто прекрасно. Но остаётся важный вопрос: а откуда нам брать это априорное распределение параметра? Как было сказано ранее, можно воспользоваться результатами прошлых наблюдений, но так можно сделать не всегда. Хочется иметь некоторый теоретический арсенал, позволяющий даже «вслепую» выбирать не очень уж плохие априорные распределения. Вот некоторые способы:

Метод I. Сопряжённые семейства

Было бы неплохо при переходе от априорного распределения к апостериорному получать не какую-то жесть, а что-то похожее на предыдущее распределение, хоть и с другими параметрами, что, к слову, поможет с дальнейшими вычислениями. Поэтому можно по распределению, которому подчиняется выборка, подобрать априорное распределение так, чтобы оно вместе с апостериорным лежали в одном семействе распределений.
Определение. В таком случае семейство распределений, которому принадлежит Q, называют *сопряжённым семейству* $\{\mathsf{P}_\theta \colon \theta \in \Theta\}$.

Пример. Рассмотрим выборку $X_1, \ldots, X_n$ из распределения $Bern(\theta)$. Её совместная плотность равна

$$\rho_\theta(\mathbf{x}) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}.$$

Чтобы получить апостериорную плотность, надо домножить $\rho_t(\mathbf{x})$ на априорную плотность и потом отнормировать это произведение, деля на некоторый интеграл. Таким образом, $\rho_{\theta|X}(t|\mathbf{x})$ пропорциональна $q(t)\rho_t(\mathbf{x})$, а стало быть, надо подобрать семейство для $q(t)$ таким образом, чтобы домножение на $\rho_t(\mathbf{x})$ не выкидывало нас за границы этого семейства. Внимательно смотря на табличку известных распределений и находя там что-то со степенями t и $(1-t)$, можно прийти к выводу, что следует взять в качестве априорного распределения $Beta(\alpha, \beta)$, то есть положить

$$q(t) = \frac{1}{B(\alpha,\beta)} t^{\alpha-1}(1-t)^{\beta-1} I(0 \leqslant t \leqslant 1).$$

В таком случае

$$\rho_{\theta|X}(t|\mathbf{x}) \sim t^{\alpha+\sum x_i - 1}(1-t)^{\beta+n-\sum x_i - 1} I(0 \leqslant t \leqslant 1),$$

поэтому эта плотность отвечает бета-распределению с параметрами $\alpha + \sum x_i$ и $\beta + n - \sum x_i$.

Заметьте, что нам не надо находить коэффициент пропорциональности, то есть тот самый интеграл в знаменателе апостериорной плотности, так как при фиксированной выборке это просто какая-то константа, служащая для нормировки (чтобы интеграл от плотности был равен единице), и тем самым определяющаяся однозначно. А мы уже знаем одно распределение, плотность которого с точностью до константы равна правой части – это и есть бета-распределение, а значит, именно ему равно апостериорное распределение. Ниже мы часто будем писать апостериорную плотность через значок $\sim$, забивая на все множители, которые не зависят от t.

Вспоминаем матожидание бета-распределения и находим байесовскую оценку

$$\widehat{\theta} = \mathsf{E}(\theta|X) = \int_\Theta t\rho_{\theta|X}(t|X)\,d\mu(t) = \frac{\alpha + \sum X_i}{(\alpha + \sum X_i) + (\beta + n - \sum X_i)} = \frac{\alpha + \sum X_i}{\alpha + \beta + n}.$$

Задача 7.1. Пусть $X_1, \ldots, X_n$ – выборка из распределения **(а)** $U(0,\theta)$, **(б)** $Pois(\theta)$, **(в)** $\mathcal{N}(\theta, 1)$, **(г)** $\mathcal{N}(0, \theta)$. Подберите сопряжённое распределение и найдите байесовскую оценку параметра θ. В качестве точечных оценок возьмите математическое ожидание апостериорных распределений и проверьте их на состоятельность.

Решение. **(а)** Имеем совместную плотность $\rho_t(\mathbf{x}) = t^{-n}I(0 < x_1, \ldots, x_n < t)$. Какое распределение имеет плотность от t, которая содержит степени t и индикатор с оценкой t снизу? Конечно же распределение Парето! Положим

$$q(t) = \frac{ka^k}{t^{k+1}}I(t > a).$$

В таком случае

$$\rho_{\theta|X}(t|\mathbf{x}) \sim \frac{1}{t^n} \cdot \frac{ka^k}{t^{k+1}}I(0 < x_1, \ldots, x_n, a < t) \sim \frac{1}{t^{n+k+1}}I(t > \max\{x_{(n)}, a\}).$$

Следовательно, апостериорным распределением является $Pareto(n + k, \max\{x_{(n)}, a\})$. Тогда искомая байесовская оценка равна

$$\mathsf{E}(\theta|X) = \int_{\max\{X_{(n)},a\}}^{+\infty} \frac{(n+k) \cdot \max\{X_{(n)}, a\}^{n+k}}{t^{n+k}}\,dt = \frac{(n+k)\max\{X_{(n)}, a\}}{n+k-1}.$$

При $\theta < a$ имеем плачевную ситуацию: элементы выборки не могут быть больше a, а значит, оценка не будет вообще зависеть от выборки, поэтому и состоятельности её не видать.

(б) Имеем совместную плотность

$$\rho_t(\mathbf{x}) = \frac{t^{\sum x_i}e^{-tn}}{\prod x_i!}.$$

Какое распределение имеет плотность от t, которая содержит степени t и экспоненту от $-t$? Конечно же гамма-распределение! Положим

$$q(t) = \frac{\lambda^\alpha t^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda t}I(t > 0).$$

В таком случае

$$\rho_{\theta|X}(t|\mathbf{x}) \sim t^{\alpha-1+\sum x_i}e^{-t(\lambda+n)}I(t > 0).$$

Следовательно, апостериорным распределением является $\Gamma\left(\alpha + \sum x_i, \lambda + n\right)$. Тогда искомая байесовская оценка равна

$$\mathsf{E}(\theta|X) = \frac{\alpha + \sum X_i}{\lambda + n}.$$

Какими бы ни были α и λ, с ростом n их «влияние» на оценку падает, и она будет асимптотически эквивалентна $\overline{X}$, что является состоятельной оценкой.

(в) Имеем совместную плотность

$$\rho_t(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum(x_i - t)^2\right) = \frac{1}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}\sum x_i^2 + t\sum x_i - \frac{nt^2}{2}\right).$$

Какое распределение имеет плотность от t, которая содержит экспоненту с t и t^2? Конечно же нормальное распределение! Положим

$$q(t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(t-a)^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}t^2 + \frac{a}{\sigma^2}t - \frac{a^2}{2\sigma^2}\right).$$

В таком случае

$$\rho_{\theta|X}(t|\mathbf{x}) \sim \exp\left[-\frac{1}{2}t^2\left(n+\frac{1}{\sigma^2}\right)+t\left(\sum x_i+\frac{a}{\sigma^2}\right)\right].$$

Следовательно, апостериорным распределением является $\mathcal{N}\left(\widehat{a},\widehat{\sigma^2}\right)$. Осталось только понять, чему равны $\widehat{a}$ и $\widehat{\sigma^2}$. Как видно из записи плотности $q(t)$, коэффициент перед t^2 в плотности нормального распределения должен быть равен $-1/2\widehat{\sigma^2}$, а перед t — $\widehat{a}/\widehat{\sigma^2}$. Это даёт нам следующую систему уравнений:

$$\begin{cases}\dfrac{\widehat{a}}{\widehat{\sigma^2}} = \sum X_i + \dfrac{a}{\sigma^2},\\ \dfrac{1}{\widehat{\sigma^2}} = n + \dfrac{1}{\sigma^2}.\end{cases}$$

Отсюда находим, что

$$\mathsf{E}(\theta|X) = \frac{\sum X_i + a/\sigma^2}{n+1/\sigma^2}.$$

Как и в пункте **(б)**, байесовская оценка асимптотически эквивалентна $\overline{X}$, которая состоятельна.

(г) Имеем совместную плотность

$$\rho_t(\mathbf{x}) = \frac{1}{(2\pi\theta)^{n/2}}\exp\left(-\frac{1}{2\theta}\sum x_i^2\right).$$

Какое распределение имеет плотность от t, которая содержит отрицательные степени t и экспоненту от $1/t$? Конечно же обратное гамма-распределение!.. А, ну да, тут уже не совсем очевидно. Будем говорить, что величина имеет *обратное гамма-распределение с параметрами* λ *и* α, если её плотность равна

$$q(t) = \frac{\lambda^\alpha t^{-\alpha-1}}{\Gamma(\alpha)}e^{-\lambda/t}I(t>0).$$

Его и возьмём за априорное распределение. В таком случае

$$\rho_{\theta|X}(t|\mathbf{x}) \sim t^{-\alpha-1-n/2}\exp\left[-\frac{1}{t}\left(\lambda+\frac{1}{2}\sum x_i^2\right)\right]I(t>0).$$

Следовательно, апостериорным распределением является $\text{Inv-Gamma}\left(\alpha+n/2, \lambda+\frac{1}{2}\sum x_i^2\right)$. Чтобы найти байесовскую оценку, для начала поймём, как выглядит матожидание у $\xi \sim \text{Inv-Gamma}(a,b)$:

$$\mathsf{E}\xi = \int_0^{+\infty} t\cdot\frac{b^a t^{-a-1}}{\Gamma(a)}e^{-b/t}\,dt =$$

$$= \int_0^{+\infty}\frac{b^a t^{-a+2}e^{-b/t}}{\Gamma(a)}\cdot\frac{1}{t^2}\,dt = \left[s=\frac{1}{t}\right] = \int_0^{+\infty}\frac{b^a s^{a-2}e^{-bs}}{\Gamma(a)}\,ds =$$

$$= \frac{b\Gamma(a-1)}{\Gamma(a)}\cdot\underbrace{\int_0^{+\infty}\frac{b^{a-1}s^{a-2}e^{-bs}}{\Gamma(a-1)}\,ds}_{\text{интеграл плотности }\Gamma(a-1,b)} = \frac{b\Gamma(a-1)}{\Gamma(a)} = \frac{b}{a-1}.$$

Значит, для $a=\alpha+n/2$ и $b=\lambda+\frac{1}{2}\sum x_i^2$ имеем

$$\mathsf{E}(\theta|X) = \frac{2\lambda+\sum X_i^2}{2\alpha+n-2}.$$

Состоятельность доказывается аналогично предыдущим пунктам. ■

Метод II. Распределение Джеффриса и снова информация Фишера

Возникает логичное желание задать на Θ равномерное распределение. Да, с неограниченным носителем так не выйдет (почти, см. задачу 7.2), но для ограниченных Θ это звучит вполне логично: если мы ничего не знаем о потенциальном параметре, то все возможные варианты равновероятны. Так делают, и это вполне допустимая практика, но этот способ имеет существенный недостаток.

Пример. Рассмотрим выборку из биномиального распределения с параметром $\sqrt{\theta}$:

$$\rho_\theta(k) = C_n^k \theta^{\frac{k}{2}} (1 - \sqrt{\theta})^{n-k}.$$

Хоть формально параметром является θ, ясно, что «главным героем» здесь выступает именно $\sqrt{\theta}$. И если мы бездумно зададим равномерное априорное распределение, то $\sqrt{\theta}$ будет распределена не равномерно, что уже не является вполне обоснованным. Если в данном игрушечном примере всё «ясно», то как поступать в общем случае (то есть какая именно функция от θ должна быть распределена равномерно) – совершенно не понятно.

Это является мотивацией к идее, что априорная плотность должна быть устойчива к замене переменной. Напомним, что если к случайной величине ξ с плотностью $\rho_\xi(x)$ применяется диффеоморфизм φ, то плотность пересчитывается как

$$\rho_{\varphi(\xi)}(y) = \frac{1}{|\varphi'(y)|} \cdot \rho_\xi(\varphi^{-1}(y)).$$

И тут на сцене появляется информация Фишера. Зададимся вопросом: как поменяется информация Фишера $I_X(\theta)$, если отныне параметром бы будем считать не θ, а некоторую $\varphi(\theta)$? Ответ неожиданный и приятный:

$$I_X(\varphi(\theta)) = \mathsf{E}_\theta \left(\frac{\partial \ln \rho_\theta(x)}{\partial \varphi(\theta)} \right)^2 = \mathsf{E}_\theta \left(\frac{\partial \ln \rho_\theta(x)}{\partial \theta} \cdot \frac{\partial \theta}{\partial \varphi(\theta)} \right)^2 = \left(\frac{\partial \theta}{\partial \varphi(\theta)} \right)^2 \mathsf{E}_\theta \left(\frac{\partial \ln \rho_\theta(x)}{\partial \theta} \right)^2 =$$
$$= \frac{1}{\left(\frac{\partial \varphi(\theta)}{\partial \theta} \right)^2} \cdot I_X(\theta) = \frac{1}{\varphi'(\theta)^2} \cdot I_X(\theta).$$

Вот те раз! Прямо как в формуле плотности при замене переменной появляется производная в знаменателе, правда на этот раз в квадрате. Поэтому для полного соответствия стоит взять от этого дела корень:

$$q(t) \sim \sqrt{I_X(t)},$$

где значок пропорциональности означает, что надо бы ещё нормировать эту штуку. Полученная априорная плотность в силу написанного выше будет инвариантна относительно замены переменной, что мы и хотели.

Определение. *Априорным распределением Джеффриса* называется распределение, плотность которого пропорциональна квадратному корню из информации Фишера (или в многомерном случае квадратному корню из определителя информационной матрицы).

Пример. Рассмотрим всю ту же выборку $X_1, \ldots, X_n$ из распределения $Bern(\theta)$. Посчитаем для неё информацию Фишера:

$$\rho_\theta(x) = \theta^x (1-\theta)^{1-x}, \quad \ln \rho_\theta(x) = x \cdot \ln \theta + (1-x) \cdot \ln (1-\theta),$$
$$\frac{\partial}{\partial \theta} \ln \rho_\theta(x) = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)},$$
$$i(\theta) = \mathsf{E}_\theta \left(\frac{X_1 - \theta}{\theta(1-\theta)} \right)^2 = (1-\theta) \cdot \left(\frac{0-\theta}{\theta(1-\theta)} \right)^2 + \theta \cdot \left(\frac{1-\theta}{\theta(1-\theta)} \right)^2 = \frac{1}{\theta(1-\theta)}.$$

Таким образом, $q(t)$ должна быть пропорциональна $\frac{1}{\sqrt{\theta(1-\theta)}}$, то есть априорным распределением является $Beta\left(\frac{1}{2},\frac{1}{2}\right)$, что не может не радовать, так как оно к тому же сопряженно распределению Бернулли.

Замечание. Иногда бывает так, что $I_X(\theta) \notin L_1(\Theta)$, и следовательно не пропорционально никакой плотности. В этом случае априорное распределение Джеффриса будет мерой на Θ (но не вероятностной), и мы можем лишь надеяться, что апостериорное распределение окажется вероятностным.

Определение. Невероятностные априорные распределения называют improper prior.

Посмотрим, как ведут себя такие распределения и насколько адекватными получаются из них оценки.

Задача 7.2. Пусть $X_1,\dots,X_n$ – выборка из распределения **(а)** $Pois(\theta)$, **(б)** $\mathcal{N}(\theta,1)$. Возьмите распределение Джеффриса в качестве априорного и найдите байесовскую оценку параметра θ. Сравните результат с первой задачей.

Решение. **(а)** Найдём информацию Фишера для пуассоновского распределения:

$$\rho_\theta(t)=\frac{\theta^t e^{-\theta}}{t!};\quad \ln\rho_\theta(t)=t\ln\theta-\theta-\ln t!;$$
$$\frac{\partial}{\partial\theta}\ln\rho_\theta(t)=\frac{t}{\theta}-1;\quad i(\theta)=\mathsf{D}_\theta\left(\frac{X_1}{\theta}-1\right)=\frac{1}{\theta^2}\mathsf{D}_\theta X_1=\frac{1}{\theta}.$$

Получается, что распределение Джеффриса имеет плотность

$$q(t)\sim\frac{1}{\sqrt{t}},$$

что не интегрируемо на $(0;+\infty)$, то есть мы получили improper prior. Но при этом

$$\rho_{\theta|X}(t|\mathbf{x})\sim t^{\sum x_i-1/2}e^{-tn},$$

то есть апостериорное распределение вполне себе определено, и равно $\Gamma\left(\sum x_i+1/2,n\right)$, и итого байесовская оценка равна

$$\mathsf{E}(\theta|X)=\frac{\sum X_i+1/2}{n}.$$

(б) Информацию Фишера позаимствуем из задачи 3.5: $i(\theta)=1/\sigma^2=1$, то есть распределение Джеффриса будет равномерным improper prior на $\mathbb{R}$ (то есть равномерное распределение на неограниченном Θ задать можно, хоть и не совсем легально). В таком случае

$$\rho_{\theta|X}(t|\mathbf{x})\sim\exp\left(-\frac{n}{2}\theta^2+\theta\sum x_i\right),$$

что есть $\mathcal{N}\left(\sum x_i/n,1/n\right)$, откуда

$$\mathsf{E}(\theta|X)=\frac{\sum X_i}{n}.$$

Как можно видеть, несмотря на то что затея с improper prior кажется неадекватной, она даёт нам довольно неплохие оценки, которые к тому же имеют не такое конское смещение в отличие от оценок из задачи 7.1. ■

Задача 7.3. Аня выиграла в акции от компании «Random Airlines» и совершенно бесплатно улетела в случайный город, в котором есть n автобусных маршрутов с номерами $1,2,\dots,n$. Выйдя из аэропорта, Аня увидела автобус номер 100. Оцените n.

Решение. Как видно, данных в условии не очень много, а значит, вариантов её понимания и решения масса, и однозначного правильного решения тут нет. Эта задача скорее творческая. Вот некоторые способы.

1. Будем считать, что наблюдение в виде номера автобуса имеет равномерное распределение от 1 до n, то есть $\mathsf{P}_n(X = t) = 1/n$ для $t \in \{1, \ldots, n\}$. В качестве априорного распределения удобно взять сопряжённое, что в нашем случае есть дискретный брат распределения Парето — $Zeta(s)$, при котором $\mathsf{P}_\theta(n = t) \sim \frac{1}{t^s \zeta(s)}$, где $\zeta(s)$ – дзета-функция Римана, или проще говоря какая-то константа для нормировки. Апостериорная плотность имеет вид

$$\mathsf{P}(n = t|X) = \frac{\frac{1}{t^{s+1}} I(t \geqslant X)}{\sum_{p=1}^{\infty} \frac{1}{p^{s+1}} I(p \geqslant X)}.$$

Отсюда байесовская оценка имеет вид

$$\mathsf{E}(n|X) = \sum_{t=1}^{\infty} \left(t \cdot \frac{\frac{1}{t^{s+1}} I(t \geqslant X)}{\sum_{p=1}^{\infty} \frac{1}{p^{s+1}} I(p \geqslant X)} \right) = \frac{\sum_{t=1}^{\infty} \frac{1}{t^s} I(t \geqslant X)}{\sum_{p=1}^{\infty} \frac{1}{p^{s+1}} I(p \geqslant X)} = \frac{\sum_{t=X}^{\infty} \frac{1}{t^s}}{\sum_{p=X}^{\infty} \frac{1}{p^{s+1}}}.$$

«Хвосты» рядов в числителе и знаменателе хорошо приближаются соответствующими интегралами, поэтому

$$\mathsf{E}(n|X) \approx \frac{\int_X^{+\infty} t^{-s}\, dt}{\int_X^{+\infty} t^{-s-1}\, dt} = X \cdot \frac{s}{s-1}.$$

2. Хотелось бы получить точную оценку, а не приближённую. Для этого можно считать, что и распределение номера автобуса, и распределение параметра n непрерывны (а почему бы и нет?). Так, будем считать, что номер автобуса X распределён равномерно на отрезке $[0; n]$, отчего его плотность равна $\rho_n(t) = 1/n$. Неплохим вариантом будет снова взять сопряжённое распределение, из задачи 7.1 мы знаем, что им является $Pareto(k, a)$, и даже в курсе, каковой будет байесовская оценка:

$$\mathsf{E}(n|X) = \frac{(1+k)\max\{X, a\}}{k} = X \cdot \frac{k+1}{k},$$

если a изначально выбрать достаточно маленьким.

3. Никто не говорил, что оценка должна быть обязательно построена исходя из байесовского подхода. Её можно построить и обычными методами. Например, если считать, что номер автобуса X распределён равномерно на $\{1, \ldots, n\}$, то по методу моментов

$$\mathsf{E}_n X = \sum_{k=1}^{n} \frac{k}{n} = \frac{n+1}{2} \Longrightarrow \widehat{n} = 2X - 1.$$

Или можно совсем угореть и рассмотреть ОМП:

$$\rho_n(x) = \frac{1}{n} \cdot I(x \in \{1, \ldots, n\}) \Longrightarrow \arg\max_{n \in \mathbb{N}} \rho_n(X) = X,$$

что, очевидно, будет весьма посредственной оценкой.

■

8 Линейная регрессия

На практике часто встречается ситуация, когда зависимость целевой величины от некоторых «фичей» можно приблизить чем-то линейным. В данном случае мы предполагаем, что истинная зависимость линейна и немного искажена каким-то шумом (ошибки измерения, выбросы и прочие вещи), который можно считать *случайным*. Отсюда полезно посмотреть на данную проблему с точки зрения теории вероятности. Давайте же приведём формальную постановку вопроса на языке статистики и установим некоторые приятные результаты.

Предположим, что i-ая целевая величина ($i \in \{1, \dots, n\}$) в своей первозданности есть линейная комбинация «фичей» Z_{ij} с некоторыми неизвестными параметрами $\theta_1, \dots, \theta_k$, то есть $\sum_{j=1}^{k} \theta_j Z_{ij}$. Но при её измерении появляется некоторый шум ε_i, поэтому наблюдение за этой величиной X_i можно представить как

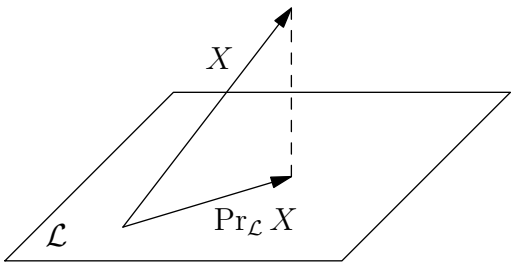
$$X_i = \sum_{j=1}^{k} \theta_j Z_{ij} + \varepsilon_i,$$

или, что эквивалентно,

$$X = Z\theta + \varepsilon,$$

где $Z = (Z_{ij})$ – матрица «фичей», $\theta = (\theta_1, \dots, \theta_k)^T$ – столбец из неизвестных параметров. Логично допустить, что случайные величины ε_i независимы в совокупности (наблюдения друг на друга не влияют), их матожидание $\mathsf{E}\varepsilon_i = 0$ (в среднем ошибки нет), а их дисперсия одинакова и равна некоторой неизвестной величине $\mathsf{D}\varepsilon_i = \sigma^2$, которую мы будем также считать за параметр в модели (то есть ковариационная матрица случайного вектора ε равна $\mathsf{D}\varepsilon = \sigma^2 E$). Как, надеюсь, стало понятно, наша задача — по вектору наблюдений X оценить вектор θ и дисперсию σ^2.

Для дальнейшего удобства также будет важно сделать допущение, что столбцы $\mathbf{z}_1, \dots, \mathbf{z}_k$ матрицы Z *линейно независимы*. Это позволяет интерпретировать задачу с позиций линейной алгебры: истинный вектор $l = Z\theta$ лежит в некотором подпространстве $\mathcal{L} = \langle \mathbf{z}_1, \dots, \mathbf{z}_k \rangle \subset \mathbb{R}^n$, образованном столбцами матрицы Z, в то время как наблюдаемый вектор X может в общем случае и не лежать в $\mathcal{L}$ (см. рис.). Отсюда логично в качестве «приближения» вектора X выбрать его проекцию $\Pr_{\mathcal{L}} X$ на это подпространство, так как она доставляет минимум расстояния между X и векторами из $\mathcal{L}$.



Осталось лишь найти оценку вектору параметров $\widehat{\theta}$, отвечающую проекции $\Pr_{\mathcal{L}} X = Z\widehat{\theta}$. Так как $\Pr_{\mathcal{L}} X$ – ортогональная проекция, то вектор $\delta = X - \Pr_{\mathcal{L}} X$ лежит в $\mathcal{L}^{\perp}$, а значит, он ортогонален любому вектору из $\mathcal{L}$, в частности, векторам $\mathbf{z}_1, \dots, \mathbf{z}_k$. Следовательно, вектор $Z^T\delta$, состоящий из скалярных произведений δ с $\mathbf{z}_j$, – нулевой, то есть

$$Z^T(X - \Pr_{\mathcal{L}} X) = 0 \Longrightarrow Z^T X = (Z^T Z)\widehat{\theta}.$$

Так как столбцы матрицы Z независимы, то матрица $Z^T Z$ будет невырожденной, поэтому у неё есть обратная, из чего получаем оценку

$$\widehat{\theta} = (Z^T Z)^{-1} Z^T X.$$

Определение. Полученная оценка называется *оценкой по методу наименьших квадратов* (неожиданно, правда?).

Сразу выделим полезные свойства полученной оценки.

Задача 8.1. Найдите математическое ожидание и матрицу ковариаций для $\widehat{\theta}$.

Решение. Линейность матожидания распространяется на многомерный случай:

$$\mathsf{E}(A\xi B) = A(\mathsf{E}\xi)B.$$

Значит,

$$\mathsf{E}_{\theta,\sigma^2}\widehat{\theta} = \mathsf{E}_{\theta,\sigma^2}((Z^TZ)^{-1}Z^TX) = (Z^TZ)^{-1}Z^T\mathsf{E}_{\theta,\sigma^2}X = (Z^TZ)^{-1}Z^T\mathsf{E}_{\theta,\sigma^2}(Z\theta+\varepsilon) =$$
$$= (Z^TZ)^{-1}Z^T\cdot Z\theta = \theta.$$

Менее очевидной является формула для ковариационной матрицы, но её легко вывести:

$$\mathsf{cov}(A\xi, B\eta) = A\,\mathsf{cov}(\xi, B\eta) = A(\mathsf{cov}(B\eta,\xi))^T = A(B\,\mathsf{cov}(\eta,\xi))^T = A\,\mathsf{cov}(\xi,\eta)B^T.$$

Теперь мы можем получить требуемое:

$$\mathsf{D}_{\theta,\sigma^2}\widehat{\theta} = \mathsf{D}_{\theta,\sigma^2}((Z^TZ)^{-1}Z^TX) = (Z^TZ)^{-1}Z^T\cdot\mathsf{D}_{\theta,\sigma^2}X\cdot((Z^TZ)^{-1}Z^T)^T =$$
$$= (Z^TZ)^{-1}Z^T\cdot\mathsf{D}_{\theta,\sigma^2}(Z\theta+\varepsilon)\cdot Z(Z^TZ)^{-1} = (Z^TZ)^{-1}Z^T\cdot\sigma^2E\cdot Z(Z^TZ)^{-1} = \sigma^2(Z^TZ)^{-1}.$$

■

Таким образом, полученная оценка $\widehat{\theta}$ является несмещённой оценкой вектора θ. Также можно показать, что несмещённой оценкой дисперсии σ^2 является

$$\widehat{\sigma^2} = \frac{1}{n-k}\left\|X - Z\widehat{\theta}\right\|^2$$

(мы по сути докажем этот факт в случае с нормально распределёнными ошибками в задаче 8.4).

Потренируемся в ~~глиномесии~~ нахождении МНК-оценки на следующей классической задаче.

Задача 8.2. Имеется 2 объекта с весами a и b. Мы взвесили с ошибками первый, второй и оба объекта вместе, причём дисперсия ошибки в последнем случае была в 4 раза больше. Сведите задачу к линейной регрессионной модели и найдите оценки для a и b.

Решение. Пусть наблюдения в первом, втором и третьем случае равнялись X_a, X_b и X_{ab} соответственно. Из условия имеем

$$\begin{pmatrix}X_a\\X_b\\X_{ab}\end{pmatrix} = \begin{pmatrix}a\\b\\a+b\end{pmatrix} + \begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\end{pmatrix} = \begin{pmatrix}1&0\\0&1\\1&1\end{pmatrix}\begin{pmatrix}a\\b\end{pmatrix} + \begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\end{pmatrix},$$

причём $\mathsf{D}\varepsilon_3 = 4\mathsf{D}\varepsilon_1 = 4\mathsf{D}\varepsilon_2 = 4\sigma^2$. Чтобы свести задачу к модели линейной регрессии выше, достаточно поделить на 2 третью строчку в формуле выше: тогда дисперсия ошибки по этой координате уменьшиться в 4 раза (так как $\mathsf{D}(\varepsilon_3/2) = (\mathsf{D}\varepsilon_3)/4$), чего бы нам и хотелось. Матрицей признаков и наблюдением тогда будут являться соответственно

$$Z = \begin{pmatrix}1&0\\0&1\\1/2&1/2\end{pmatrix},\quad X = \begin{pmatrix}X_a\\X_b\\X_{ab}/2\end{pmatrix}.$$

Теперь у нас есть всё, чтобы посчитать оценку:

$$Z^TZ = \begin{pmatrix}1&0&1/2\\0&1&1/2\end{pmatrix}\cdot\begin{pmatrix}1&0\\0&1\\1/2&1/2\end{pmatrix} = \begin{pmatrix}5/4&1/4\\1/4&5/4\end{pmatrix},\quad (Z^TZ)^{-1} = \begin{pmatrix}5/6&-1/6\\-1/6&5/6\end{pmatrix}$$

$$\widehat{\theta} = (Z^TZ)^{-1}Z^TX = \begin{pmatrix}5/6&-1/6\\-1/6&5/6\end{pmatrix}\begin{pmatrix}1&0&1/2\\0&1&1/2\end{pmatrix}\begin{pmatrix}X_a\\X_b\\X_{ab}/2\end{pmatrix} = \begin{pmatrix}5/6&-1/6&1/3\\-1/6&5/6&1/3\end{pmatrix}\begin{pmatrix}X_a\\X_b\\X_{ab}/2\end{pmatrix}\Rightarrow$$

$$\widehat{a} = \frac{5X_a}{6} - \frac{X_b}{6} + \frac{X_{ab}}{6},\quad \widehat{b} = -\frac{X_a}{6} + \frac{5X_b}{6} + \frac{X_{ab}}{6}$$

■

8.1 Гауссовская линейная модель

Всякие физики да химики из своих внутренних побуждений часто полагают ошибки нормальными, поэтому в дальнейшем под ε будем подразумевать гауссовский вектор $\mathcal{N}(0, \sigma^2 E)$. Данная модель называется *гауссовской линейной моделью*. Подобное допущение действительно бывает крайне полезным. Например, теперь можно утверждать следующий факт:

Теорема. *Статистика $(\widehat{\theta}, \|X - Z\widehat{\theta}\|^2)$ является полной достаточной в гауссовской линейной модели. Как следствие, оценки $\widehat{\theta}$ и $\widehat{\sigma^2}$ являются оптимальными.*

Задача 8.3. Взвешивание трёх грузов массами a, b, c на одних и тех же весах производится следующим образом: n_1 раз взвешиваются второй и третий груз вместе, n_2 раз взвешиваются первый и третий груз вместе и n_3 раз взвешиваются первый и второй груз вместе. В предположении, что все ошибки имеют распределение $\mathcal{N}(0, \sigma^2)$, сведите задачу к модели линейной регрессии и найдите оптимальные оценки для a, b, c.

Решение. Если в тупую перевести задачу на язык линейной регрессии, нам придётся иметь дело с матрицей признаков

$$Z' = \begin{pmatrix} 0 & \cdots & 0 & 1 & \cdots & 1 & 1 & \cdots & 1 \\ 1 & \cdots & 1 & 0 & \cdots & 0 & 1 & \cdots & 1 \\ 1 & \cdots & 1 & 1 & \cdots & 1 & 0 & \cdots & 0 \end{pmatrix}^T,$$

которая сама по себе выглядит неприятно, а ведь нужно ещё что-то обращать и много чего умножать — гадость одним словом. Куда проще здесь будет считать именно $\alpha = b + c$, $\beta = a + c$ и $\gamma = a + b$ параметрами модели, а через них потом выразить нужные. В таком случае вектор наблюдений можно выразить как

$$X = \begin{pmatrix} X_1^\alpha \\ \cdots \\ X_{n_1}^\alpha \\ X_1^\beta \\ \cdots \\ X_{n_2}^\beta \\ X_1^\gamma \\ \cdots \\ X_{n_3}^\gamma \end{pmatrix} = Z \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} + \varepsilon, \quad \text{где } Z = \begin{pmatrix} 1 & 0 & 0 \\ & \cdots & \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ & \cdots & \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ & \cdots & \\ 0 & 0 & 1 \end{pmatrix}.$$

Матрица Z намного проще в использовании, потому что её столбцы *ортогональны*: в таком случае матрица

$$Z^T Z = \begin{pmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{pmatrix}$$

будет диагональной, что в разы упрощает дальнейшую работу:

$$\widehat{\theta} = (Z^T Z)^{-1} Z^T X = \begin{pmatrix} 1/n_1 & 0 & 0 \\ 0 & 1/n_2 & 0 \\ 0 & 0 & 1/n_3 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix} X = \begin{pmatrix} \frac{1}{n_1} \sum X_i^\alpha \\ \frac{1}{n_2} \sum X_i^\beta \\ \frac{1}{n_3} \sum X_i^\gamma \end{pmatrix}$$

Далее выражаем a, b и c через α, β и γ и дело в шляпе: оптимальность оценок будет следовать из того, что они являются функциями от полных достаточных статистик.

$$\widehat{a} = \frac{\overline{X^\beta} + \overline{X^\gamma} - \overline{X^\alpha}}{2}, \quad \widehat{b} = \frac{\overline{X^\gamma} + \overline{X^\alpha} - \overline{X^\beta}}{2}, \quad \widehat{c} = \frac{\overline{X^\alpha} + \overline{X^\beta} - \overline{X^\gamma}}{2}$$

■

Рассмотрим внимательнее, как устроены найденные нами оценки в гауссовской линейной модели. В этом нам поможет

Теорема (об ортогональном разложении)**.** *Пусть $X \sim \mathcal{N}(a, \sigma^2 E_n)$ – гауссовский вектор, а $L_1 \oplus \ldots \oplus L_r$ – ортогональное разложение $\mathbb{R}^n$. Тогда $\Pr_{L_1} X, \ldots, \Pr_{L_r} X$ независимы и нормально распределены, и для всех $i \in \{1, \ldots, r\}$*

$$\frac{1}{\sigma^2} \| \Pr_{L_i} X - \mathsf{E} \Pr_{L_i} X \|^2 \sim \chi^2_{\dim L_i}.$$

Идея. Переводим X в о/н базис, согласованный с разложением $L_1 \oplus \ldots \oplus L_r$. Так как матрица S перехода от одного о/н базиса к другому – ортогональна, то по формуле пересчёта матожидания и матрицы ковариаций из задачи 8.1 в новом базисе вектор X будет иметь распределение $\mathcal{N}(Sa, \sigma^2 E_n)$. □

Применим теорему к нашей модели. По определению МНК-оценки, $\Pr_{\mathcal{L}} X = Z\widehat{\theta}$, а значит, $\Pr_{\mathcal{L}^\perp} X = X - Z\widehat{\theta}$, и по теореме об ортогональном разложении

$$\frac{1}{\sigma^2} \|X - Z\widehat{\theta}\|^2 \sim \chi^2_{\dim \mathcal{L}^\perp} = \chi^2_{n-k}. \tag{3}$$

Что же касается $\widehat{\theta}$, то она, как линейное преобразование гауссовского вектора X, имеет распределение $\mathcal{N}(\theta, \sigma^2 (Z^T Z)^{-1})$ (параметры мы нашли ранее в задаче 8.1). Следовательно, $\widehat{\theta}_i \sim \mathcal{N}(\theta_i, \sigma^2 \left[(Z^T Z)^{-1}\right]_{ii})$, или, что эквивалентно,

$$\frac{\widehat{\theta}_i - \theta_i}{\sigma \sqrt{\left[(Z^T Z)^{-1}\right]_{ii}}} \sim \mathcal{N}(0, 1).$$

Было бы неплохо избавиться от σ, чтобы оставить только один неизвестный θ_i. У нас уже есть одна статистика с известным распределением и торчащим σ – это (3). Если поделить одно на корень от другого, то получится от него избавиться, но непонятно, будет ли у полученной случайной величины конкретное распределение, не зависящее от параметров.

Оказывается, будет — из теоремы об ортогональном разложении следует, что $Z\widehat{\theta}$ и $X - Z\widehat{\theta}$ будут независимыми, а значит, $\widehat{\theta} = (Z^T Z)^{-1} Z^T \cdot Z\widehat{\theta}$ и $\widehat{\sigma^2} = \|X - Z\widehat{\theta}\|^2/(n-k)$ также независимы. Поэтому распределение

$$\frac{\widehat{\theta}_i - \theta_i}{\sigma \sqrt{\left[(Z^T Z)^{-1}\right]_{ii}}} \Bigg/ \sqrt{\frac{1}{(n-k)\sigma^2} \|X - Z\widehat{\theta}\|^2} = \frac{\widehat{\theta}_i - \theta_i}{\sqrt{\widehat{\sigma^2} \left[(Z^T Z)^{-1}\right]_{ii}}}$$

однозначно определено свёрткой, и не зависит от неизвестных параметров. Остаётся вопрос: зачем мы поделили на $n - k$? Для красоты? Не только:

Определение. Если случайные величины ξ и η независимы, причём $\xi \sim \mathcal{N}(0, 1)$, а $\eta \sim \chi^2_m$, то говорят, что случайная величина

$$\zeta = \frac{\xi}{\sqrt{\eta/m}}$$

имеет *распределение Стьюдента с m степенями свободы.* Обозначается как $\zeta \sim T_m$.

Из всего вышесказанного следует, что

$$\frac{\widehat{\theta}_i - \theta_i}{\sqrt{\widehat{\sigma^2} \left[(Z^T Z)^{-1}\right]_{ii}}} \sim T_{n-k}. \tag{4}$$

Полученные распределения приведённых статистик позволяют искать доверительные интервалы для θ_i и σ^2. Рассмотрим это на следующем примере.

Задача 8.4. Пусть $X_1, \dots, X_n$ – независимые случайные величины, где X_i распределена по нормальному закону $\mathcal{N}(a+bi, \sigma^2)$. Постройте точные доверительные интервалы для параметров a, b, σ^2.

Решение. В данной задаче матрица признаков имеет вид

$$Z = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & n \end{pmatrix}^T.$$

Вспоминаем формулу для суммы квадратов первых n натуральных чисел и считаем:

$$Z^TZ = \begin{pmatrix} n & \frac{n(n+1)}{2} \\ \frac{n(n+1)}{2} & \frac{n(n+1)(2n+1)}{6} \end{pmatrix}, \quad (Z^TZ)^{-1} = \begin{pmatrix} \frac{2(2n+1)}{n(n-1)} & -\frac{6}{n(n-1)} \\ -\frac{6}{n(n-1)} & \frac{12}{n(n^2-1)} \end{pmatrix},$$

$$\widehat{\theta} = (Z^TZ)^{-1}Z^TX = (Z^TZ)^{-1} \cdot \begin{pmatrix} \sum\limits_{i=1}^n X_i \\ \sum\limits_{i=1}^n iX_i \end{pmatrix} = \begin{pmatrix} \frac{2(2n+1)}{n(n-1)} \sum\limits_{i=1}^n X_i - \frac{6}{n(n-1)} \sum\limits_{i=1}^n iX_i \\ -\frac{6}{n(n-1)} \sum\limits_{i=1}^n X_i + \frac{12}{n(n^2-1)} \sum\limits_{i=1}^n iX_i \end{pmatrix} = \begin{pmatrix} \widehat{a} \\ \widehat{b} \end{pmatrix},$$

$$\widehat{\sigma^2} = \frac{1}{n-2}\|X - Z\widehat{\theta}\|^2 = \dots \text{ (лень досчитывать)}$$

Из соотношения (3) имеем доверительный интервал для σ^2:

$$\mathsf{P}_{\theta,\sigma^2}\left(\frac{(n-2)\widehat{\sigma^2}}{x_{(1+\gamma)/2}} < \sigma^2 < \frac{(n-2)\widehat{\sigma^2}}{x_{(1-\gamma)/2}}\right) = \mathsf{P}_{\theta,\sigma^2}\left(x_{(1-\gamma)/2} < \frac{(n-2)\widehat{\sigma^2}}{\sigma^2} < x_{(1+\gamma)/2}\right) = \gamma,$$

где x_p – p-квантиль распределения χ^2_{n-2}. В то же время из (4) имеем следующие ДИ для a и b:

$$\mathsf{P}_{\theta,\sigma^2}\left(\widehat{a} - y_{(1+\gamma)/2}\sqrt{\frac{2\widehat{\sigma^2}(2n+1)}{n(n-1)}} < a < \widehat{a} + y_{(1+\gamma)/2}\sqrt{\frac{2\widehat{\sigma^2}(2n+1)}{n(n-1)}}\right) = \gamma,$$

$$\mathsf{P}_{\theta,\sigma^2}\left(\widehat{b} - y_{(1+\gamma)/2}\sqrt{\frac{12\widehat{\sigma^2}}{n(n^2-1)}} < b < \widehat{b} + y_{(1+\gamma)/2}\sqrt{\frac{12\widehat{\sigma^2}}{n(n^2-1)}}\right) = \gamma,$$

где y_p – p-квантиль распределения T_{n-2}. Здесь мы не использовали квантиль $y_{(1-\gamma)/2}$, ибо распределение Стьюдента симметрично относительно нуля. ■

Последние две задачи дают теоретическое обоснование методу регуляризации, используемому в машинном обучении.

Задача 8.5. Применим байесовский подход к оценке параметра θ. Взяв в качестве априорного распределения $\mathcal{N}(0, \gamma^2 E_k)$ – гауссовский вектор с независимыми компонентами, найдите апостериорную плотность $\rho(\theta|X)$ (с точностью до константы) и байесовскую оценку. Решите задачу оптимизации $\rho(\theta|X) \to max_\theta$. Чем данный подход и полученная в нём оценка лучше, чем МНК?

Решение. Из условия мы полагаем, что априорное распределение задаётся плотностью

$$q(t) \sim \exp\left(-\frac{1}{2\gamma^2}\|t\|^2\right) = \exp\left(-\frac{1}{2\gamma^2}t^Tt\right).$$

Значит, апостериорному распределению соответствует

$$\rho_{\theta,X}(t|\mathbf{x}) \sim \exp\left(-\frac{1}{2\gamma^2}t^Tt - \frac{1}{2\sigma^2}(\mathbf{x} - Zt)^T(\mathbf{x} - Zt)\right).$$

Если мы попытаемся максимизировать эту плотность, то получим следующую задачу оптимизации:

$$\|X - Z\theta\|^2 + \frac{\sigma^2}{\gamma^2}\|\theta\|^2 \to \min_\theta,$$

что называется Ridge regression. Её преимущество в том, что мы «штрафуем» вектор θ за излишне большие координаты, что позволяет получать более стабильные решения. Особенно отчётливо это станет видно, когда мы найдём соответствующую байесовскую оценку. Это можно сделать напрямую, решив задачу выше, но мы поступим более интеллектуально, найдя параметры a и Σ многомерного нормального распределения, отвечающего $\rho_{\theta,X}(t|\mathbf{x})$.

$$\text{С одной стороны, } \rho_{\theta,X}(t|\mathbf{x}) \sim \exp\left[\frac{1}{\sigma^2}\mathbf{x}^T Z t - \frac{1}{2}t^T\left(\frac{1}{\gamma^2}E + \frac{1}{\sigma^2}Z^T Z\right)t\right].$$

$$\text{С другой – } \rho_{\theta,X}(t|\mathbf{x}) \sim \exp\left(-\frac{1}{2}(t-a)^T\Sigma^{-1}(t-a)\right) \sim \exp\left(a^T\Sigma^{-1}t - \frac{1}{2}t^T\Sigma^{-1}t\right).$$

$$\text{Получается, } \begin{cases} \Sigma^{-1} = \dfrac{1}{\gamma^2}E + \dfrac{1}{\sigma^2}Z^T Z, \\ a^T\Sigma^{-1} = \dfrac{1}{\sigma^2}\mathbf{x}^T Z. \end{cases}$$

Транспонируя второе равенство (благо Σ симметрична, и на неё это не повлияет), получаем оценку

$$\mathsf{E}(\theta|X) = \left(Z^T Z + \frac{\sigma^2}{\gamma^2}E\right)^{-1} Z^T X.$$

Получили практически решение задачи обычной линейной регрессии, но теперь к матрице $Z^T Z$ добавляется единичная с некоторой константой. Это и позволяет получать более адекватную оценку в случае, если эта матрица близка к вырожденной. Это происходит из-за того, что добавление такой матрицы сдвигает все собственные числа $Z^T Z$ на $\frac{\sigma^2}{\gamma^2}$ вправо, отчего определитель, как произведение собственных чисел, отдаляется от нуля. ■

Задача 8.6. Возьмите следующее априорное распределение: компоненты вектора θ независимы, $\theta_i \sim Laplace(\lambda)$. Найдите апостериорную плотность $\rho(\theta|X)$ (с точностью до константы). Сформулируйте задачу оптимизации $\rho(\theta|X) \to max_\theta$.

Решение. Теперь имеем

$$q(t) \sim \exp\left(-\lambda\sum_{i=1}^{k}|t_i|\right) = \exp\left(-\lambda\|t\|_1\right).$$

Следовательно, апостериорная плотность имеет вид

$$\rho_{\theta,X}(t|\mathbf{x}) \sim \exp\left(-\lambda\|t\|_1 - \frac{1}{2\sigma^2}(\mathbf{x}-Zt)^T(\mathbf{x}-Zt)\right).$$

Соответствующую задачу оптимизации можно сформулировать так:

$$\|X - Z\theta\|^2 + \frac{\sigma^2}{\lambda}\|\theta\|_1 \to \min_\theta$$

Здесь имеется похожая регуляризация, что и в задаче 8.5, но теперь мы пытаемся ограничить вектор параметров по $\|\cdot\|_1$-норме (так называемая Lasso regression). Такой подход помимо всего прочего обладает свойством «отбора признаков» (подробнее смотрите в курсе Машинного обучения). ■

9 Проверка статистических гипотез

В задачах выше мы предполагаем, что семейство распределений, откуда в теории могла прийти выборка, нам известно. Спрашивается: с какого перепугу мы его знаем? Да, можно до него догадаться, например, методом пристального взгляда («О, на гистограмме данные образуют красивый холмик, значит, это что-то нормальное»), но даже в таком случае у нас может возникнуть множество потенциальных кандидатов на роль семейства распределений (чаще всего мы будем рассматривать только два), и среди них надо откинуть самые бесперспективные. Отсюда хотелось бы иметь теоретически обоснованную возможность отвергать какие-либо предположения о распределении, которому подчиняется выборка. Распишем более формально.
Определение. *Статистической гипотезой* H называют предположение о принадлежности истинного распределения P некоторому классу $\mathcal{P}$. Обозначается как $H\colon \mathsf{P} \in \mathcal{P}$.

Предположим, что истинное распределение данных лежит в некотором семействе распределений $\mathcal{P}$, в котором имеются два непересекающихся подмножества $\mathcal{P}_0$ и $\mathcal{P}_1$ – это и есть наши догадки. Мы подвергаем сомнению, что имеет место принадлежность к классу $\mathcal{P}_0$, и в качестве противовеса берём класс $\mathcal{P}_1$.
Определение. В таком случае гипотеза $H_0\colon \mathsf{P} \in \mathcal{P}_0$ называется *основной гипотезой*, а гипотеза $H_1\colon \mathsf{P} \in \mathcal{P}_1$ – *альтернативой*. Обозначается как

$$H_0\colon \mathsf{P} \in \mathcal{P}_0 \quad \text{versus} \quad H_1\colon \mathsf{P} \in \mathcal{P}_1.$$

Часто приведённые классы распределений задаются некоторыми параметрами: $\mathcal{P} = \{\mathsf{P}_\theta\colon \theta \in \Theta\}$. В таком случае гипотезу можно сформулировать с терминах принадлежности некоторым подмножествам $\Theta_0, \Theta_1 \subset \Theta$:

$$H_0\colon \theta \in \Theta_0 \quad \text{versus} \quad H_1\colon \theta \in \Theta_1.$$

Так как принятие решения о том, отвергать ли H_0 или нет, зависит только от реализации выборки X, то критерий выбора можно задать некоторым измеримым множеством $R \subset \mathcal{X}$:

$$X \in R \Longrightarrow \text{ отвергаем } H_0$$
$$X \notin R \Longrightarrow \text{ не отвергаем } H_0.$$

Определение. Множество R, попадание в которое равносильно отвержению основной гипотезы, называется *критическим* или *критерием*.

Замечание. Важно подчеркнуть, что «не отвергаем» и «безоговорочно принимаем» H_0 – разные вещи. Если мы не смогли найти весомый довод против основной гипотезы, то это вовсе не значит, что она верна. Возможно, это не так, но из-за каких-то причин (плохой критерий, неудачная выборка и т. д.) мы не смогли её отвергнуть. Надо помнить, что *наша ключевая цель – найти весомые косвенные доказательства её неверности в пользу* H_1, а если таковых не нашлось, то мы либо принимаем гипотезу на веру (куда деваться?), либо подбираем другие критерии в надежде её опровергнуть. Удачное сравнение можно встретить в [2]:

> Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that he is guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0.

Хотя формально постановка задачи симметрична, часто мы подразумеваем неравнозначность гипотез. Это можно проиллюстрировать следующим хрестоматийным примером.

Пример. Предположим, что мы работаем в госпитале и проводим анализы на присутствие в организме раковых клеток. По сути, мы по реализации выборки из различных показателей (кровь, рентген, МРТ и т. д.) должны проверить гипотезу H_0: *пациент болен раком* против альтернативы H_1: *пациент здоров*. Если мы верно поставили диагноз, то всё ок. Иначе мы можем совершить одну из двух ошибок:

	Принимаем H_0	Отвергаем H_0
H_0 верна	Мы молодцы!	Ошибка I рода
H_1 верна	Ошибка II рода	Мы молодцы!

В случае *ошибки I рода* мы не окажем помощь больному человеку и обречём его на смерть, а в случае *ошибки II рода* мы начнём лечить здорового, и потеряем много денег. Обе ситуации неприятны, но с точки зрения морали первая куда хуже. Выбор гипотезы о том, что пациент болен, в качестве основной, а не наоборот, согласуется со сказанным в замечании: мы стараемся найти действительно убедительные свидетельства того, что пациент здоров (то есть неверна H_0), ибо в случае беспочвенного опровержения верной гипотезы мы буквально похороним пациента, и если таковых нет, то мы (может и с некоторым скепсисом) примем её.

Можно привести и такой пример: как известно, законы Ньютона не являются исчерпывающим описанием Вселенной и не работают корректно как в макро-, так и в микромире, то есть гипотеза H: *Выполняются законы Ньютона* неверна, при этом её часто принимают на веру. Это происходит не из-за того, что физики такие глупые (хотя...), а так как она вполне себе допустима для несложных физических моделей. Так и в общем случае: если гипотеза достаточно хорошо описывает происходящее, то её можно принять, даже несмотря на то, что в действительности она неверна. Таким образом, ошибка II рода не так опасна, в отличие от I рода, когда мы отвергаем верную гипотезу и больше к ней не возвращаемся.

Как же понять, когда критерий хороший, а когда не очень? Полезной можно найти следующую характеристику:

Определение. *Функцией мощности критерия* R называется функция

$$\beta(\mathsf{P}, R) = \mathsf{P}(X \in R).$$

Понятно, что в случае верности основной гипотезы H_0 (то есть когда $\mathsf{P} \in \mathcal{P}_0$) вероятность попадания в критическое множество должна быть низкой, а если верна H_1 – как можно больше. Возникает вопрос – как минимизировать одно и максимизировать другое? В контексте примера выше более верным представляется следующий подход: сначала надо поставить некое маленькое заранее оговоренное ограничение сверху на функцию мощности для $\mathsf{P} \in \mathcal{P}_0$, чтобы вероятность ошибки I рода была меньше фиксированного числа. В связи с этим важным является следующее

Определение. *Размером критерия* R называется

$$\sup_{\mathsf{P} \in \mathcal{P}_0} \beta(\mathsf{P}, R).$$

Говорят, что критерий R *имеет уровень значимости* α, если его размер не превышает α.

Отныне мы работаем с критериями, у которых можно явно задать уровень значимости α (например, часто берут 0.05). Среди таковых надо подобрать критерий с как можно меньшей ошибкой II рода, то есть максимизировать вероятность попадания в R при верности H_1. Тут, как это было при сравнении оценок, возникает проблема сравнения двух функций (как понять, какая лучше?). Возможное решение аналогично:

Определение. Говорят, что *критерий* R_1 *мощнее критерий* R_2, если $\forall \mathsf{P} \in \mathcal{P}_1 \colon \beta(\mathsf{P}, R_1) \geqslant \beta(\mathsf{P}, R_2)$. Критерий R называется *равномерно наиболее мощным (или сокращённо р. м. н. к.) уровня значимости* α, если он мощнее любого другого критерия уровня значимости α.

Также бывают полезным проверять потенциальный критерий на наличие следующих естественных свойств.

Определение. Критерий R для проверки

$$H_0 \colon \mathsf{P} \in \mathcal{P}_0 \quad \text{versus} \quad H_1 \colon \mathsf{P} \in \mathcal{P}_1$$

называется *несмещённым*, если

$$\sup_{\mathsf{P} \in \mathcal{P}_0} \beta(\mathsf{P}, R) \leqslant \inf_{\mathsf{P} \in \mathcal{P}_1} \beta(\mathsf{P}, R).$$

Последовательность критериев R_n для выборки $X = (X_1, \dots, X_n)$ называется *состоятельной*, если $\forall \mathsf{P} \in \mathcal{P}_1 \colon \beta(\mathsf{P}, R_n) \to 1$ при $n \to \infty$ (то есть ошибка II рода постепенно исчезает).

Пример. Рассмотрим модель сдвига $X_i \sim \mathcal{N}(\theta, 1)$. Предположим, в наших расчётах удобно полагать $\theta = \theta_0$, но нам хотелось бы убедиться, что это допущение состоятельно по сравнению с альтернативой $\theta > \theta_0$. Таким образом, проверим гипотезу

$$H_0 \colon \theta = \theta_0 \quad \text{versus} \quad H_1 \colon \theta > \theta_0.$$

Логично использовать критерий, основанный на статистике $T(X) = \overline{X}$ (тем более она достаточна, что как бы оправдывает такой выбор), а именно: если мы попадаем в множество $R = \{\mathbf{x} \colon T(\mathbf{x}) \geqslant c\}$ для некоторого c, то среднее слишком велико, и скорее всего предположение H_0 неверно, а иначе оно вполне допустимо. Подберём число c так, чтобы наш критерий имел уровень значимости α, то есть

$$\alpha = \mathsf{P}_{\theta_0}(\overline{X} \geqslant c) = \mathsf{P}_{\theta_0}(\underbrace{\sqrt{n}(\overline{X} - \theta_0)}_{\sim \mathcal{N}(0,1)} \geqslant \sqrt{n}(c - \theta_0)) \Longrightarrow \sqrt{n}(c - \theta_0) = x_{1-\alpha},$$

где x_p – p-квантиль для $\mathcal{N}(0, 1)$. Таким образом, $c = \theta_0 + x_{1-\alpha}/\sqrt{n}$ доставляет нам критерий с требуемым уровнем значимости. Посмотрим, как выглядит функция мощности для $\theta > \theta_0$:

$$\beta(\theta) = \mathsf{P}_\theta\left(\overline{X} \geqslant c\right) = \mathsf{P}_\theta\left(\sqrt{n}(\overline{X} - \theta) \geqslant \sqrt{n}(\theta_0 - \theta) + x_{1-\alpha}\right) = 1 - \Phi\left(\sqrt{n}(\theta_0 - \theta) + x_{1-\alpha}\right) \Subset$$

где Φ – функция распределения $\mathcal{N}(0, 1)$. Из её симметричности имеем

$$\Subset \Phi\left(\sqrt{n}(\theta - \theta_0) - x_{1-\alpha}\right).$$

В силу возрастания Φ функция мощности $\beta(\theta)$ будет также возрастать, поэтому $\forall \theta > \theta_0 \colon \beta(\theta) \geqslant \alpha$, и критерий R будет несмещённым. Также при $n \to \infty$ аргумент функции Φ стремится к $+\infty$, поэтому $\forall \theta > \theta_0 \colon \beta(\theta) \to 1$, а значит, критерий ещё и состоятелен.

9.1 Простые гипотезы

Как вы может быть заметили, в последнем примере мы не проверяли критерий на предмет р. м. н. к., потому что ~~мне лень~~ это довольно сложно сделать на практике. Чаще всего р. м. н. к. просто не существует. Но для игрушечных гипотез такой можно явно предъявить.

Определение. Гипотеза $H \colon \mathsf{P} \in \mathcal{P}$ называется *простой*, если множество предполагаемых распределений состоит из единственного кандидата: $\mathcal{P} = \{\mathsf{P}\}$. Иначе она называется *сложной*.

Предположим, нам надо столкнуть лбами две простые гипотезы:

$$H_0 \colon \mathsf{P} = \mathsf{P}_0 \quad \text{versus} \quad H_1 \colon \mathsf{P} = \mathsf{P}_1,$$

причём оба кандидата P_0 и P_1 абсолютно непрерывны относительно некоторой меры μ и имеют по ней плотности $p_0(t)$ и $p_1(t)$ соответственно.

Лемма (Нейман-Пирсон)**.** *Рассмотрим критерий $R_\lambda = \{x \in \mathcal{X} \colon p_1(x) - \lambda p_0(x) \geqslant 0\}$, где $\lambda > 0$. Если он обладает минимальным уровнем значимости α, то есть $\mathsf{P}_{\theta_0}(X \in R_\lambda) = \alpha$, то он является несмещённым р. м. н. к. уровня значимости α.*

Задача 9.1. Пусть $X_1, \dots, X_n$ – выборка из распределения $\mathcal{N}(0, \sigma^2)$. Постройте р.н.м.к. уровня значимости α для проверки гипотезы $H_0 \colon \sigma^2 = \sigma_0^2$ против альтернативы $H_1 \colon \sigma^2 = \sigma_1^2$.

Решение. По лемме выше для подходящего λ критерий

$$R_\lambda = \left\{\mathbf{x} \in \mathbb{R}^n \colon \frac{\rho_1(\mathbf{x})}{\rho_0(\mathbf{x})} \geqslant \lambda\right\} = \left\{\mathbf{x} \colon \left(\frac{\sigma_0^2}{\sigma_1^2}\right)^{n/2} \cdot \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\sum x_i^2\right] \geqslant \lambda\right\}$$

будет удовлетворять условию. Осталось сделать так, чтобы размер критерия был в точности равен α. Без потери общности скажем, что $\sigma_0^2 > \sigma_1^2$. Тогда

$$R_\lambda = \left\{ \mathbf{x} \in \mathbb{R}^n \colon \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \right) \sum x_i^2 \right] \geqslant \lambda \left(\frac{\sigma_1^2}{\sigma_0^2} \right)^{n/2} \right\} =$$

$$= \left\{ \mathbf{x} \colon -\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \right) \sum x_i^2 \geqslant \ln\lambda + \frac{n}{2}\ln\frac{\sigma_1^2}{\sigma_0^2} \right\} = \left\{ \mathbf{x} \colon \sum x_i^2 \leqslant \frac{\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2}\left(2\ln\lambda + n\ln\frac{\sigma_1^2}{\sigma_0^2} \right) \right\}.$$

В силу независимости элементов выборки $\sum X_i^2 \sim \chi_n^2$, поэтому λ и α связывает следующее соотношение:

$$\frac{\sigma_0^2\sigma_1^2}{\sigma_1^2 - \sigma_0^2}\left(2\ln\lambda + n\ln\frac{\sigma_1^2}{\sigma_0^2} \right) = x_\alpha,$$

где x_p – p-квантиль соответствующего распределения. Отсюда

$$\lambda = \left(\frac{\sigma_0^2}{\sigma_1^2} \right)^{n/2} \cdot \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2} \right) x_\alpha \right].$$

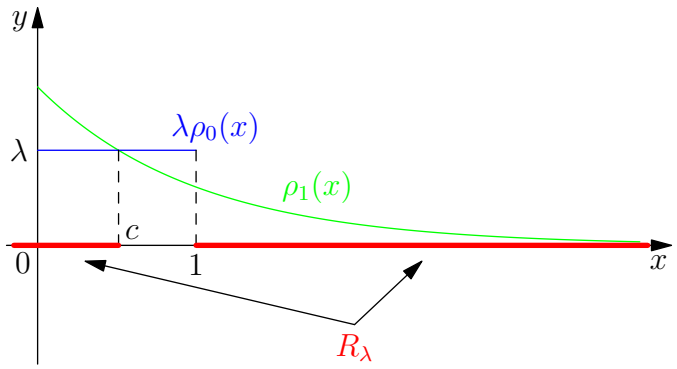
В случае $\sigma_0^2 < \sigma_1^2$ формула останется прежней за исключением замены x_α на $x_{1-\alpha}$ (подумайте, почему). ■

Задача 9.2. Пусть X_1 – выборка размера 1. Рассмотрим гипотезы

$$H_0 \colon X_1 \sim U(0;1) \quad \text{versus} \quad H_1 \colon X_1 \sim Exp(1).$$

Постройте р.н.м.к. для проверки H_0 против H_1 и вычислите его мощность.

Решение. Параметра в этой модели на прямую не дали, но это не значит, что мы не может применить лемму Неймана-Пирсона. Из монотонности $\rho_1(t) = e^{-t}$ (см. рис. ниже)



легко понять, что р.м.н.к. здесь будет

$$R_\lambda = \{ x \in \mathbb{R} \colon \rho_1(x) \geqslant \lambda\rho_0(x) \} = (-\infty; c] \cup [1; +\infty],$$

где c удовлетворяет равенствам $\lambda = e^{-c}$ и $\alpha = \mathsf{P}_0(R_\lambda) = c$, то есть $\lambda = e^{-\alpha}$. Отсюда также несложно посчитать мощность нашего критерия:

$$\beta(R_\lambda) = \mathsf{P}_1(R_\lambda) = 1 - \int_c^1 \rho_1(t)\,dt = 1 + e^{-t}\Big|_c^1 = 1 + e^{-1} - e^{-\alpha}.$$

■

9.2 Сложные гипотезы

Как и ранее, будет будем предполагать, что все потенциальные распределения P_θ (как из $\mathcal{P}_0$, так и из $\mathcal{P}_1$) имеют плотность ρ_θ относительно некоторой меры. Среди всевозможных сложных гипотез рассмотрим те, в которых мы делим множество параметров в виде прямой на два луча, но сначала нам надо ввести следующее

Определение. Говорят, что семейство $\{\mathsf{P}_\theta\colon \theta \in \Theta\}$ обладает *монотонным отношением правдоподобия по статистике* $T(\mathbf{x})$, если для всех θ_0 и θ_1 из Θ таких, что $\theta_0 < \theta_1$, функция $\dfrac{\rho_{\theta_1}(\mathbf{x})}{\rho_{\theta_0}(\mathbf{x})}$ является монотонной по $T(\mathbf{x})$ с одним и тем же типом монотонности (для уточнения этот тип монотонности добавляют в название, например, неубывающее/невозрастающее отношение правдоподобия).

Теорема (о монотонном отношении правдоподобия)**.** *Пусть* $\{\mathsf{P}_\theta\colon \theta \in \Theta\}$ *– семейство с неубывающим отношением правдоподобия по статистике* $T(\mathbf{x})$*. Поставим проблему проверки*

$$H_0\colon \theta \leqslant \theta_0 \quad \textit{versus} \quad H_1\colon \theta > \theta.$$

Если существует некоторое c *такое, что* $\mathsf{P}_{\theta_0}(T(X) \geqslant c) = \alpha$*, то критерий* $R = \{\mathbf{x}\colon T(\mathbf{x}) \geqslant c\}$ *является р. м. н. к. с уровнем значимости* α*.*

Замечание. В условии теоремы основную гипотезу можно поставить и как $H_0\colon \theta = \theta_0$.

Задача 9.3. Пусть $X_1, \ldots, X_n$ – выборка из распределения $Exp(\lambda)$. Постройте р.н.м.к. уровня значимости α для проверки гипотезы $H_0\colon \lambda = \lambda_0$ против альтернативы **(а)** $H_1\colon \lambda < \lambda_0$; **(б)** $H_1\colon \lambda > \lambda_0$.

Решение. **(а)** Сведём задачу к теореме выше введением иного параметра $\nu := -\lambda$. Тогда $\rho_\nu(t) = -\nu e^{\nu t}$, и гипотезы перепишутся как

$$H_0\colon \nu = \nu_0 \quad \text{versus} \quad H_1\colon \nu > \nu_0.$$

Рассмотрим отношение совместных плотностей для $\nu_2 > \nu_1$:

$$\frac{\rho_{\nu_2}(\mathbf{x})}{\rho_{\nu_1}(\mathbf{x})} = \left(\frac{-\nu_2}{-\nu_1}\right)^n \exp\Big[(\nu_2 - \nu_1)\sum x_i\Big],$$

что есть возрастающая функция от $\sum x_i$, то есть новое семейство распределений обладает неубывающим отношением правдоподобия. Из независимости X_i получаем, что если H_0 верна, то $\sum X_i \sim \Gamma(n, -\nu_0) = \Gamma(n, \lambda_0)$. Тогда положив $c = z_{1-\alpha}$, где z_p – p-квантиль $\Gamma(n, \lambda_0)$, по теореме о монотонном отношении правдоподобия критерий $R = \{\mathbf{x}\colon \sum x_i \geqslant c\}$ будет р.н.м.к.

(б) Теперь уже альтернатива выглядит ровно как в теореме выше, только теперь для $\lambda_2 > \lambda_1$

$$\frac{\rho_{\lambda_2}(\mathbf{x})}{\rho_{\lambda_1}(\mathbf{x})} = \left(\frac{\lambda_2}{\lambda_1}\right)^n \exp\Big[(\lambda_1 - \lambda_2)\sum x_i\Big],$$

и теперь семейство распределений обладает невозрастающим отношением правдоподобия по $\sum x_i$. Это легко чинится, если рассмотреть отношение плотностей как функцию от $-\sum x_i$. Тогда $R = \{\mathbf{x}\colon \sum x_i \leqslant c\}$ будет искомым критерием, где $c = z_\alpha$. ■

Задача 9.4. Пусть $X_1, \ldots, X_n$ – выборка из распределения $\mathcal{N}(\theta, 1)$. Постройте р.н.м.к. уровня значимости α для проверки гипотезы **(а)** $H_0\colon \theta \geqslant \theta_0$ против альтернативы $H_1\colon \theta < \theta_0$; **(б)** $H_0\colon \theta \leqslant \theta_0$ против альтернативы $H_1\colon \theta > \theta_0$.

Решение. **(а)** Введём параметр $\mu := -\theta$. Тогда плотность имеет вид

$$\rho_\mu(t) = \frac{1}{\sqrt{2\pi}} e^{-(t+\mu)^2/2}.$$

Отношение совместных плотностей для $\mu_2 > \mu_1$ есть

$$\frac{\rho_{\mu_2}(\mathbf{x})}{\rho_{\mu_1}(\mathbf{x})} = \exp\left[\frac{1}{2}\sum(x_i+\mu_1)^2 - \frac{1}{2}\sum(x_i+\mu_2)^2\right] = \exp\left[\frac{n}{2}(\mu_1^2-\mu_2^2) + (\mu_1-\mu_2)\sum x_i\right],$$

что является убывающей по $\sum x_i$ функцией. С учётом того, что $\sum X_i \sim \mathcal{N}(n\theta_0, n)$ при верности H_0, требуемым критерием будет являться $R = \{\mathbf{x}\colon \sum x_i \leqslant n\theta_0 + \sqrt{n}x_\alpha\}$, где x_p – p-квантиль распределения $\mathcal{N}(0,1)$.

(б) Тут никаких подводных камней нет, тупо теорема о монотонном отношении правдоподобия:

$$\frac{\rho_{\theta_2}(\mathbf{x})}{\rho_{\theta_1}(\mathbf{x})} = \exp\left[\frac{1}{2}\sum(x_i-\theta_1)^2 - \frac{1}{2}\sum(x_i-\theta_2)^2\right] =$$

$$= \exp\left[\frac{n}{2}(\theta_1^2-\theta_2^2) + (\theta_2-\theta_1)\sum x_i\right] \text{ — возрастает при } \theta_2 > \theta_1 \Longrightarrow$$

$$R = \left\{\mathbf{x}\colon \sum x_i \geqslant n\theta_0 + \sqrt{n}x_{1-\alpha}\right\} \text{ – р.н.м.к.}$$

■

Выйдем из мира розовых поней, где у нас имеется р.н.м.к., и вернёмся в суровую реальность.
Задача 9.5. $X_1, \ldots, X_n$ – выборка из распределения $Bern(\theta)$. Докажите, что не существует равномерного наиболее мощного критерия произвольного уровня значимости α для проверки гипотезы $H_0\colon \theta = \theta_0$ против альтернативы $H_1\colon \theta \neq \theta_0$.

Решение. На самом деле, утверждение задачи не совсем правда. Например, положим $\theta_0 = 1/2$, а $\alpha < 2^{-n}$. Тогда пустой критерий будет единственным с допустимым уровнем значимости, а значит, он автоматически р.м.н.к. Избавимся от сего нелепого случая, подразумевая под задачей, что для фиксированных θ_0 и α р.м.н.к. не существует для достаточного большого n.

Пусть существует р.м.н.к. R. Рассмотрим критерии вида $S_1 = \{\mathbf{x}\colon \sum x_i \geqslant c_1\}$ и $S_2 = \{\mathbf{x}\colon \sum x_i \leqslant c_2\}$, где константы c_1 и c_2 мы подберём «впритык» так, чтобы они имели уровень значимости α. Так как R – р.м.н.к., то его мощность должна быть больше мощностей этих критериев при любых $\theta \neq \theta_0$. Рассмотрим первый из них при $\theta \to 1-0$.

Выберем θ настолько близкой к единице, чтобы произвольное наблюдение с k единицами из n было вероятнее, чем все возможные наблюдения с меньшим количеством единиц, то есть

$$\mathsf{P}_\theta\left(X = (\underbrace{\ldots\ldots}_{k \text{ единиц}})\right) > \sum_{\mathbf{x}\colon \sum x_i < k} \mathsf{P}_\theta\left(X = \mathbf{x}\right).$$

Так сделать можно: в правой части не больше 2^n слагаемых с вероятностью не большей $\theta^{k-1}(1-\theta)^{n-k+1}$, а вероятность слева равна $\theta^k(1-\theta)^{n-k}$, то есть отношение левой части к правой не меньше $\frac{\theta}{1-\theta}\cdot 2^{-n}$, что при фиксированном n можно сделать сколь угодно большим.

Спрашивается: а зачем нам это всё? Из этого следует, что критерий R *обязан* содержать S_1 как подмножество. Действительно, выберем максимальный k такой, что оно не содержит какой-то вектор с $k \geqslant c_1$ единицами. Но тогда чтобы вероятность R была больше вероятности S_1 при выбранном θ, надо взять другие наблюдения с меньшим числом единиц, что всё равно не позволит получить нужную вероятность по выбору θ – противоречие. Аналогично $S_2 \subset R$.

А теперь вспомним, что константы для S_1 и S_2 мы выбирали так, чтобы они тютелька в тютельку были с нужным уровнем значимости. Поэтому если мы возьмём настолько большое n, чтобы $\mathsf{P}_{\theta_0}\left(\sum X_i = k\right) < \alpha/2$ для всех k, то S_1 и S_2 будут иметь уровень значимости больше $\alpha/2$. И вправду: если бы, например, $\mathsf{P}_{\theta_0}(S_1) \leqslant \alpha/2$, то c_1 можно было бы уменьшить на единичку, что не сильно бы увеличило уровень значимости по выбору n. Таким образом, так как $S_1 \cup S_2 \subset R$, то либо $S_1 \cap S_2 \neq \varnothing$, и R есть все исходы и, стало быть, имеет размер 1, либо $S_1 \cap S_2 = \varnothing$, и R имеет минимальный уровень значимости больше, чем α – противоречие. ■

Задача 9.6. $X_1, \ldots, X_n$ – выборка из равномерного распределения на отрезке $[0; \theta]$, $\theta > 0$. Постройте равномерно наиболее мощный критерий уровня значимости α для проверки гипотезы $H_0\colon \theta = \theta_0$ против альтернативы $H_1\colon \theta \neq \theta_0$.

Решение. Очевидно, что в случае, когда $X_{(n)} > \theta_0$, гипотеза H_0 однозначно неверна, поэтому такие выборки следует отнести в критическое множество. Также понятно, что слишком маленькое значение $X_{(n)}$ является серьёзным доводом для отвержения гипотезы H_0. Итого, давайте возьмём в качестве критерия множество

$$R = \{\mathbf{x}\colon x_{(n)} > \theta_0\} \cup \{\mathbf{x}\colon x_{(n)} \leqslant c\},$$

где c мы подберём так, чтобы размер R был в точности α:

$$\alpha = \mathsf{P}_{\theta_0}(R) = \mathsf{P}_{\theta_0}(X_{(n)} \leqslant c) = \mathsf{P}_{\theta_0}(X_1 \leqslant c)^n = \frac{c^n}{\theta_0^n} \Longrightarrow c = \theta_0 \sqrt[n]{\alpha}.$$

Докажем, что он и будет р.м.н.к.

Для $\theta \leqslant c$ всё очевидно: критерий полностью покрывает носитель плотности, то есть при таких θ у нас $\mathsf{P}_\theta(R) = 1$, и больше уже и не сделаешь.

Возьмём $c < \theta < \theta_0$. Пусть существует критерий S с большей мощностью для данной θ, то есть $\mathsf{P}_\theta(S) > \mathsf{P}_\theta(R)$. Но тогда

$$\mathsf{P}_{\theta_0}(S) = \int\limits_S \rho_{\theta_0}(\mathbf{x})\, d\mathbf{x} = \int\limits_{S \cap \{0 \leqslant x_i \leqslant \theta_0\}} \frac{d\mathbf{x}}{\theta_0^n} \geqslant \frac{\theta^n}{\theta_0^n} \int\limits_{S \cap \{0 \leqslant x_i \leqslant \theta\}} \frac{d\mathbf{x}}{\theta^n} = \frac{\theta^n}{\theta_0^n} \mathsf{P}_\theta(S) >$$

$$> \frac{\theta^n}{\theta_0^n} \mathsf{P}_\theta(R) = \frac{\theta^n}{\theta_0^n} \int\limits_R \rho_\theta(\mathbf{x})\, d\mathbf{x} = \frac{\theta^n}{\theta_0^n} \int\limits_{R \cap \{0 \leqslant x_i \leqslant \theta\}} \frac{d\mathbf{x}}{\theta^n} = \int\limits_{R \cap \{0 \leqslant x_i \leqslant \theta_0\}} \frac{d\mathbf{x}}{\theta_0^n} = \mathsf{P}_{\theta_0}(R) = \alpha,$$

то есть критерий S априори не может иметь требуемый уровень значимости.

Случай $\theta > \theta_0$ аналогичен предыдущему: мы захотим получить множество большей мощности по θ, но это непременно приведёт к увеличению мощности по θ_0, то есть размера критерия, в силу пропорциональности этих вероятностей и выбора множества R. ■

10 Тысяча и один критерий

В этом параграфе мы обсудим несколько полезных критериев. Часть из них строится для простых основных гипотез, то есть нам надо проверить, совпадает ли истинное распределение с некоторым данным. Такие критерии ещё называют *критериями согласия*. Также заметим, что некоторые критерии существенно используют тот факт, что исходная выборка достаточно большая, например, их размер может быть довольно большим при маленькой выборке и стремится к чему-то более адекватному при увеличении размера выборки. Поэтому полезным будет следующее

Определение. Критерий R для проверки гипотезы $H_0\colon \mathsf{P} = \mathsf{P}_0$ называется *асимптотическим критерием уровня значимости* α, если

$$\overline{\lim_{n\to\infty}}\, \mathsf{P}_0(R) \leqslant \alpha.$$

10.1 Критерий Колмогорова

Поставим на проверку гипотезу

$$H_0\colon \mathsf{P} = \mathsf{P}_0 \quad \text{versus} \quad H_1\colon \mathsf{P} \neq \mathsf{P}_0.$$

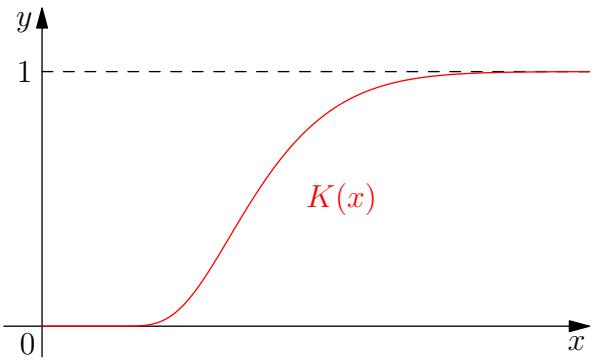
Как мы знаем из теоремы Гливенко-Кантелли, эмпирическая функция распределения $\widehat{F}_n$ равномерно сходится к истинной функции распределения F для почти всех выборок $X = (X_1, \ldots, X_n, \ldots)$, то есть

$$D_n = \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| \xrightarrow{\text{п.н.}} 0.$$

Таким образом, судить о выполнимости гипотезы H_0 можно исходя из того, насколько похожи $\widehat{F}_n$ и F. Оказывается, эта сходимость порядка $1/\sqrt{n}$, и при этом распределение $\sqrt{n}D_n$ отнюдь не рандомное.

Теорема (Колмогоров)**.** *Пусть F – непрерывная функция распределения. Тогда распределение $\sqrt{n}D_n$ не зависит от F и слабо сходится к распределению Колмогорова с функцией распределения*

$$K(t) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2t^2}, \quad t > 0.$$



Из теоремы следует, что если при большом n статистика D_n достаточно большая, то это является существенным доводом против H_0, то есть критерий для проверки этой гипотезы имеет вид

$$R = \left\{\sqrt{n}D_n \geqslant k_{1-\alpha}\right\},$$

где k_p – p-квантиль распределения Колмогорова (пощупать его можно здесь).

Слабая сходимость распределений из теоремы Колмогорова позволяет сказать, что сей критерий имеет асимптотический уровень значимости α. Что же насчёт других свойств?

Задача 10.1. Выведите состоятельность критерия Колмогорова из теоремы Колмогорова.

Решение. Сделаем допущение, что в качестве альтернативы мы берём непрерывные распределения, отличающиеся от взятого F (не непрерывные рассматривать нет смысла). Пусть

истинная функция распределения равна $G \neq F$, эмпирическую же обозначим как $\widehat{G}_n$. В таком случае статистику D_n можно оценить следующим образом:

$$D_n = \sup_{x\in\mathbb{R}} |\widehat{G}_n(x) - F(x)| \geqslant \sup_{x\in\mathbb{R}} \Big[|G(x) - F(x)| - |\widehat{G}_n(x) - G(x)|\Big] \geqslant$$
$$\geqslant \sup_{x\in\mathbb{R}} |G(x) - F(x)| - \sup_{x\in\mathbb{R}} |\widehat{G}_n(x) - G(x)| = c - D'_n,$$

причём $c = \sup_{x\in\mathbb{R}} |G(x) - F(x)| \neq 0$, а $D'_n = \sup_{x\in\mathbb{R}} |\widehat{G}_n(x) - G(x)|$ при домножении на $\sqrt{n}$ слабо сходится к распределению Колмогорова по одноимённой теореме. В таком случае

$$\mathsf{P}(\sqrt{n}D_n \geqslant k_{1-\alpha}) \geqslant \mathsf{P}(\sqrt{n}(c - D'_n) \geqslant k_{1-\alpha}) = \mathsf{P}(\sqrt{n}D'_n \leqslant c\sqrt{n} - k_{1-\alpha}).$$

Правая часть неравенства в скобках стремится к бесконечности, поэтому для любого ε найдётся N, что $\forall n \geqslant N\colon c\sqrt{n} - k_{1-\alpha} \geqslant 1/\varepsilon$, и для таких n:

$$\mathsf{P}(\sqrt{n}D'_n \leqslant c\sqrt{n} - k_{1-\alpha}) \geqslant \mathsf{P}(\sqrt{n}D'_n \leqslant 1/\varepsilon) \to K(1/\varepsilon).$$

Беря сколь угодно малый ε, мы устремим $K(1/\varepsilon)$ к единице, а значит, $\mathsf{P}(\sqrt{n}D_n \geqslant k_{1-\alpha}) \to 1$, и критерий состоятелен. ■

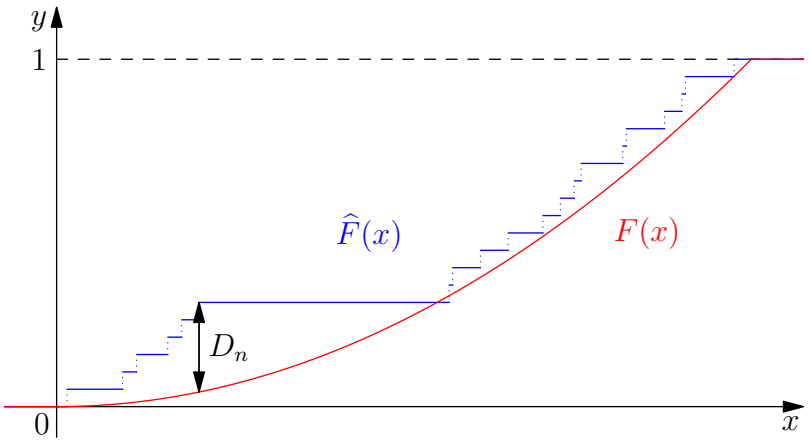
Задача 10.2. Дана выборка из неизвестного распределения: 1.63, 1.95, 1.14, 1.8, 0.19, 0.32, 1.3, 1.51, 0.03, 1.64, 1.75, 0.23, 0.36, 0.41, 1.49, 1.13, 1.81, 1.4, 1.45, 1.22. Проверьте гипотезу о том, что распределение, из которого взята выборка, имеет плотность $\rho(x) = \frac{x}{2} I(x \in [0;2])$. Уровень значимости выберете сами.

Решение. Для более простого подсчёта есть относительно простая формула для статистики D_n, которая следует из кусочно-постоянности $\widehat{F}_n$:

$$D_n = \max_{1\leqslant i\leqslant n} \left\{ \frac{i}{n} - F(X_{(i)}), F(X_{(i)}) - \frac{i-1}{n} \right\}.$$

С помощью формулы и пары строк на Python'е несложно убедиться, что для данной выборки и $F(x) = x^2/4$ $(x \in [0;2])$ мы имеем $D_n \approx 0.258$, а значит, $\sqrt{n}D_n \approx 1.1537$. Несложно загуглить, что $k_{0.85} \approx 1.138$, $k_{0.9} \approx 1.224$. Таким образом, минимальный уровень значимости, на котором мы должны отвергнуть основную гипотезу (см. раздел 10.6*), находится между 0.1 и 0.15, что весьма много (обычно берут 0.05 или 0.01), поэтому основания для отвержения H_0 нет.

Полезно оценить верность нашего решения на картинке:



Как можно видеть, рядом с нулём распределения уж необычно сильно различаются, но критерий Колмогорова это не смутило. ■

Задача 10.3. Докажите, что в критерии Колмогорова при справедливости нулевой гипотезы статистика D_n имеет некоторое фиксированное распределение одинаковое для всех $F(x)$.

Указание. Имеет смысл сделать замену $y = F(x)$, $x = F^{-1}(y)$.

Решение. Как таковой обратной функции у F может и не быть (она может быть постоянной на некотором промежутке), поэтому определим

$$F^{-1}(y) = \sup\{x\colon F(x) \leqslant y\}.$$

Статистику D_n можно переписать как

$$D_n = \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| = \sup_{y\in[0;1]} |\widehat{F}_n(F^{-1}(y)) - y|,$$

так как из отрезка постоянства $y = F(x)$ не может быть элементов выборки, поэтому на нём эмпирическая функция распределения также постоянна.

Посмотрим, как распределена $\widehat{F}_n(F^{-1}(y))$:

$$\widehat{F}_n(F^{-1}(y)) = \sum_{i=1}^{n} I(F^{-1}(y) \geqslant X_i) = \sum_{i=1}^{n} I(y \geqslant F(X_i)),$$

а $F(X_i)$, как известно из задачи 5.5, распределено равномерно на $[0;1]$. Таким образом, D_n выражается через независимые величины $F(X_i)$ с одним и тем же распределением, поэтому её распределение тоже однозначно определенно. ■

Задача 10.4*. Убедитесь, что $k_{1-\alpha} \sim \sqrt{-\frac{1}{2}\ln\frac{\alpha}{2}}$ при $\alpha \to 0$, где $k_{1-\alpha}$ – $(1-\alpha)$-квантиль распределения Колмогорова.

Решение. To be continued... ■

10.2 Критерий χ^2 Пирсона

Рассмотрим наблюдение из мультиномиального распределения с параметрами n, k и $\mathbf{p} = (p_1, \ldots, p_k)$, которое по сути является обобщением биномиального: проще говоря, у нас есть k-гранный кубик, выпадение i-ой грани которого происходит с вероятностью p_i; мы кидаем этот кубик n раз и записываем в вектор $X = (X_1, \ldots, X_k)$, что первая грань выпала X_1 раз, вторая – X_2 раз и т. д.. Можно также интерпретировать как

$$X_i = \sum_{j=1}^{n} I(B_j = i),$$

где B_j – результат j-ого броска, которые независимы между собой.

Пусть мы наблюдаем вектор $(X_1, \ldots, X_k)$ с таким распределением и хотим проверить гипотезу

$$H_0\colon \mathbf{p} = \mathbf{p}^0 = (p_1^0, \ldots, p_k^0) \quad \text{versus} \quad H_1\colon \mathbf{p} \neq \mathbf{p}^0.$$

По ЦПТ мы знаем, что при выполнимости гипотезы H_0

$$\frac{X - n\mathbf{p}^0}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

где Σ – какая-то там матрица ковариаций (если быть точнее, $\Sigma_{ii} = \mathsf{D}I(B_1 = i) = p_i^0(1-p_i^0)$, $\Sigma_{ij} = \mathsf{cov}\,(I(B_1 = i), I(B_1 = j)) = -p_i^0 p_j^0$ для $i \neq j$). Таким образом, компоненты вектора в левой части, то есть

$$\frac{X_i - np_i^0}{\sqrt{n}},$$

распределены почти что нормально. Тогда давайте в качестве меры отклонения от гипотезы H_0 возьмём взвешенную сумму квадратов компонент этого вектора:

Определение. *Статистикой хи-квадрат Пирсона* называется

$$\chi^2(X) = \sum_{i=1}^{k} \frac{(X_i - np_i^0)^2}{np_i^0}.$$

Логично предположить, что как сумма квадратов почти что нормально распределённых величин эта статистика стремится по распределению к хи-квадрат. Что ж, так оно и есть, хоть и доказывается это не тривиально.

Теорема. *Если H_0 верна, то $\chi^2(X) \xrightarrow{d} \zeta \sim \chi^2_{k-1}$.*

Идея. Главная задача – доказать утверждение из задачи 10.6, а остальное – дело техники. □

Итого, в качестве критерия проверки H_0 асимптотического уровня значимости α можно взять

$$R = \left\{\mathbf{x}\colon \chi^2(\mathbf{x}) > x_{1-\alpha}\right\},$$

где x_p – p-квантиль распределения χ^2_{k-1}. Следует помнить, что критерий этот – асимптотический, а значит, пользоваться им на малой выборки имеет мало смысла. Обычно критерий χ^2 используют при $n \geqslant 50$ и $np_i^0 \geqslant 5$ для всех $i = 1, \ldots, k$.

Пример *(третий закон Менделя).* Согласно наблюдениям, проведённым биологом Г. Менделем, разные признаки наследуются независимо друг от друга. Попробуем убедиться в этом статистически.

Предположим, у семейства гороха имеется два признака: цвет (жёлтый и зелёный) и форма (круглая или морщинистая). Скрещиваются два вида гороха: с доминантными признаками (жёлтые круглые горошины) и рецессивными (зелёные морщинистые горошины). По отдельности в результате селекции признаки распределяются в отношении 3 : 1 (по второму закону Менделя), поэтому если третий закон Менделя верен, то распределение двух признаков будет иметь вид 9 : 3 : 3 : 1.

Проведено $n = 556$ наблюдений. Посмотрим на эту статистику:

Тип горошин	Гипотетическая вероятность	Наблюдаемая частота
Желтые, круглые	9/16	315/556
Желтые, морщинистые	3/16	101/556
Зелёные, круглые	3/16	108/556
Зелёные, морщинистые	1/16	32/556

В наших обозначениях это значит, что вектор наблюдений равен $X = (315, 101, 108, 32)$, и проверяется гипотеза

$$H_0\colon \mathbf{p} = \mathbf{p}^0 = (9/16, 3/16, 3/16, 1/16).$$

Посчитаем статистику Пирсона:

$$\chi^2(X) = \frac{(315 - 556 \cdot 9/16)^2}{556 \cdot 9/16} + \frac{(101 - 556 \cdot 3/16)^2}{556 \cdot 3/16} + \frac{(108 - 556 \cdot 3/16)^2}{556 \cdot 3/16} + \frac{(32 - 556 \cdot 1/16)^2}{556 \cdot 1/16} \approx 0.47.$$

Если в качестве допустимого уровня значимости взять $\alpha = 0.05$, то пороговым значением для критерия Пирсона будет $(1 - \alpha)$-квантиль для χ^2_3, что есть примерно 7.815. Наблюдаемое значение гораздо меньше порогового значения, а значит, причин для отвержения гипотезы H_0 нет.

Задача 10.5. Среди первых 800 цифр числа π цифры 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 встречаются соответственно 74, 92, 83, 79, 80, 73, 77, 75, 76, 91 раз. Проверьте при помощи критерия хи-квадрат гипотезу о том, что различные цифры встречаются в числе π равновероятно. Рассмотрите уровни значимости $\alpha = 0.05$, $\alpha = 0.1$.

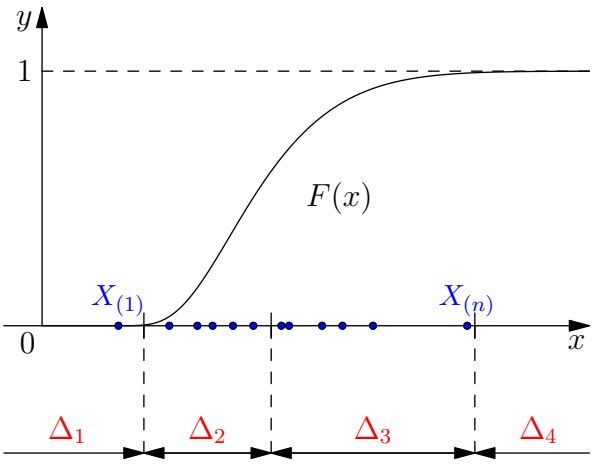
Решение. Python подсказывает, что для данной выборки статистика хи-квадрат равна $\chi^2(X) = 5.125$. Сверимся с квантилями для χ^2_9: $x_{0.95} \approx 16.92$, $x_{0.9} \approx 14.68$. Как видим, значение статистики и близко не подошло к границам критерия.

На самом деле несложно с помощью scipy.stats убедиться, что минимальный уровень значимости, на котором мы бы отвергли гипотезу, равен примерно 0.823 (а если точнее, $1 - F_{\chi^2_9}(5.125)$), то есть при проверке гипотезы мы позволяем себе ошибку I рода в размере 82.3% (!!!), что катастрофически много. ■

Отметим, что критерий χ^2 применяется далеко не только к модели выше. Он также позволяет проверять гипотезы о равенстве истинной функции распределения какой-то данной. Как же это происходит?

Пусть нам выборка $X_1, \dots, X_n$ из некоторого неизвестного нам распределения $F(x)$. Мы же в свою очередь хотим проверить, не является ли эта функция чем-то хорошим, то есть проверяем

$$H_0\colon F(x) = F_0(x).$$



Разобьём числовую прямую на k дизъюнктных множеств $\Delta_1, \dots, \Delta_k$ (чаще всего берут полуинтервалы $\Delta_i = (a_i; b_j]$, возможно и бесконечные). В данные интервалы как-то попали наши точки: пусть в i-ое множество Δ_i попало v_i точек. При этом в идеальном мире (и при верности H_0) вероятность попасть в Δ_i равна $p_i^0 = \int_{\Delta_i} dF_0(x) = F_0(b_i) - F_0(a_i)$. Это и сводит текущую задачу к задаче выше: для каждого элемента выборки на гранях k-гранного кубика написано, в какой полуинтервал оно попадёт, и гипотеза заключается в том, что вероятность выпадания определённой грани равна установленному числу. Получается, критерий имеет вид

$$R = \left\{ \sum_{i=1}^{k} \frac{(v_i - np_i^0)^2}{np_i^0} > x_{1-\alpha} \right\}.$$

Конечно же, если критерий χ^2 не отверг гипотезу, то нам это ровным счётом ни о чём не говорит. Мы могли разделить прямую как-то не очень удачно, из-за чего истинное распределение может легко мимикрировать под данное, имея одинаковые с ним вероятности промежутков Δ_i. Отсюда представляется логичным брать не слишком мало интервалов, чтобы мы смогли обнаружить различия между распределениями. Но и слишком маленькими их делать не следует, потому что тогда в некоторые интервалы может в теории не попасть ни одна точка, что на корню убивает предположение о нормальности $(v_i - np_i^0)/\sqrt{n}$. Обычно берут $k \approx \log_2 n$.

Пример. Решим задачку 10.2 с помощью критерия χ^2. Будем делить область значений на 4 ($\approx \log_2 20$) части. Встаёт вопрос: как именно разбить отрезок $[0; 2]$, откуда приходят значения выборки? Для начала рассмотрим самый простой вариант: брать равные по длине отрезки. Этот вариант имеет право на существование, однако следует помнить об ограничении $np_i^0 \geqslant 5$, без которого результаты нельзя назвать точными. У нас это и подавно не выполняется, поэтому продолжим.

В отрезки $[0; 0.5]$, $[0.5; 1]$, $[1; 1.5]$ и $[1.5; 2]$ попало соответственно 6, 0, 7 и 7 элементов выборки. При верности основной гипотезы их вероятности равны соответственно 1/16, 3/16, 5/16 и 7/16 (просто считаем интеграл плотности на этих отрезках). Статистика хи-квадрат равна

$$\chi^2(X) = \frac{(6 - 20 \cdot 1/16)^2}{20 \cdot 1/16} + \frac{(0 - 20 \cdot 3/16)^2}{20 \cdot 3/16} + \frac{(7 - 20 \cdot 5/16)^2}{20 \cdot 5/16} + \frac{(7 - 20 \cdot 7/16)^2}{20 \cdot 7/16} = 22.24,$$

в то время как 0.95-квантиль распределения χ^2_3 равен ≈ 7.81, а p-value (см. раздел 10.6*) равен $\approx 5.81 \cdot 10^{-5}$, поэтому на уровне значимости 0.05 гипотеза отвергается.

Рассмотрим другой способ: разобьём отрезок на равновероятные части. Такими отрезками будут $[0;1]$, $[1;\sqrt{2}]$, $[\sqrt{2};\sqrt{3}]$ и $[\sqrt{3};2]$ (их концы – прообразы точек 0.25, 0.5, 0.75 функции распределения $F_0(x) = x^2/4$). В эти отрезки попадёт соответственно 6, 5, 5 и 4 элементов выборки, и тогда $\chi^2(X) = 0.4$, что уже меньше 0.95-квантиля, поэтому гипотеза не отвергается. Но смею напомнить, что мощность такого критерия весьма низкая, поэтому этот результат не даёт нам никакой информации.

Напоследок приведём доказательство факта, из которого следует теорема выше.

Задача 10.6 *. Дан вектор $\mathbf{p} = (p_1, \dots, p_n)^T$, причём $\sum p_i = 1$. Определим матрицу

$$\Sigma = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n \end{pmatrix} - \mathbf{p}\mathbf{p}^T.$$

Рассмотрим гауссовский вектор $\xi \sim \mathcal{N}(0, \Sigma)$. Докажите, что

$$\sum_{i=1}^{n} \frac{\xi_i^2}{p_i} \sim \chi^2_{n-1}.$$

Решение. Нормальное решение лучше смотреть в [5], тут приведено решение автора сего конспекта.

Для начала нормируем векторы $\xi_i \mapsto \xi_i/\sqrt{p_i}$, чтобы из $\sum \xi_i^2/p_i$ сделать просто $\sum \xi_i^2$. Поэтому будем считать, что ковариационная матрица имеет вид

$$\Sigma = E_n - \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^T,$$

где $\sqrt{\mathbf{p}}$ означает результат поэлементного взятия корня из $\mathbf{p}$, E_n – единичная матрица размера n.

Как мы знаем из линала, существует ОНБ, в котором квадратичная форма принимает диагональный вид, то есть существует некоторая ортогональная матрица S такая, что $S\Sigma S^T = D$, где D – диагональная. На диагонали стоят собственные числа Σ, как их найти? Для это достаточно понять, что многочлен $P(t) = t(t-1) = t^2 - t$ является аннулирующим для Σ:

$$\Sigma^2 - \Sigma = (E_n - \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^T)^2 - E_n + \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^T = -\sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^T + \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^T\sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^T =$$
$$= \sqrt{\mathbf{p}}(\underbrace{\sqrt{\mathbf{p}}^T\sqrt{\mathbf{p}}}_{=\sum p_i = 1} - E_1)\sqrt{\mathbf{p}}^T = 0.$$

Легко также видеть, что $P(t)$ минимален (куда уж меньше?). Характеристический многочлен делится на минимальный, причём все корни характеристического автоматом являются корнями минимального, поэтому у Σ собственными числами будут только 0 и 1. Осталось понять, какой они кратности.

Ранг при домножении на невырожденную матрицу не меняется, поэтому $\operatorname{rk}\Sigma = \operatorname{rk} D =$ кол-во единиц в D. Мы знаем, что 0 точно собственное число Σ, то есть единиц не больше $n-1$. Оценка снизу берётся из неравенства $\operatorname{rk}(A+B) \leqslant \operatorname{rk} A + \operatorname{rk} B$:

$$n = \operatorname{rk} E_n \leqslant \operatorname{rk} \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^T + \operatorname{rk}\Sigma = 1 + \operatorname{rk}\Sigma \Longrightarrow \operatorname{rk}\Sigma \geqslant n-1.$$

Таким образом,

$$D = \begin{pmatrix} E_{n-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Зачем всё это было надо? А вот зачем: вектор $\eta = S\xi$ как линейное преобразование над гауссовским вектором имеет распределение $\mathcal{N}(0, D)$. Тогда его компоненты не коррелированы, а

значит, независимы (по свойству гауссовского вектора), причём одна из координат распределена как нуль из-за одного нуля на диагонали, а остальные – стандартно нормально. Ну и самое приятное: ортогональное преобразование не меняет норму вектора, поэтому

$$\sum_{i=1}^{n} \xi_i^2 = \sum_{i=1}^{n} \eta_i^2 = \sum_{i=1}^{n-1} \eta_i^2 \sim \chi^2_{n-1}$$

как сумма квадратов независимых величин с распределением $\mathcal{N}(0,1)$.

■

10.3 Линейные гипотезы в линейной регрессии

Вернёмся к модели гауссовской линейной регрессии:

$$X = Z\theta + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(0, \sigma^2 E)$, причём $\theta \in \mathbb{R}^k$ и σ^2 – неизвестные параметры, $X \in \mathbb{R}^n$ – наблюдение, $Z \in \mathbb{R}^{n\times k}$ – матрица признаков. Наша гипотеза будет состоять в предположении, что θ лежит в некоторой гиперплоскости, то есть

$$H_0 \colon T\theta = \tau,$$

где $T \in \mathbb{R}^{m\times k}$, $\tau \in \mathbb{R}^m$ – известные величины, причём будем допускать, что $\operatorname{rk} T = m \leqslant k$. Отсюда собственно и название гипотезы: мы накладываем некоторые линейные ограничения на параметр θ.

Напомним, что в предыдущих параграфах мы получили оценку

$$\widehat{\theta} = (Z^T Z)^{-1} Z^T X.$$

Из задачи 8.1 мы знаем матожидание и ковариационную матрицу у $\widehat{\theta}$, а значит, можем найти её и у $T\widehat{\theta}$:

$$\mathsf{E}_{\theta,\sigma^2} T\widehat{\theta} = T\mathsf{E}_{\theta,\sigma^2}\widehat{\theta} = T\theta,$$

$$\mathsf{D}_{\theta,\sigma^2} T\widehat{\theta} = T\left[\mathsf{D}_{\theta,\sigma^2}\widehat{\theta}\right] T^T = \sigma^2 \underbrace{T(Z^T Z)^{-1} T^T}_{=B} = \sigma^2 B.$$

$T\widehat{\theta}$, как линейное преобразование над нормально распределённым $\widehat{\theta}$, само нормально распределено, то есть

$$T\widehat{\theta} \sim \mathcal{N}(T\theta, \sigma^2 B).$$

Так как матрица B положительная определена, то у неё существует $\sqrt{B}$, а значит,

$$\frac{1}{\sigma}\sqrt{B}^{-1}(T\widehat{\theta} - T\theta) \sim \mathcal{N}(0, E) \Longrightarrow$$

$$\left\|\frac{1}{\sigma}\sqrt{B}^{-1}(T\widehat{\theta} - T\theta)\right\|^2 = \frac{1}{\sigma^2}(T\widehat{\theta} - T\theta)^T B^{-1}(T\widehat{\theta} - T\theta) \sim \chi^2_m.$$

При верности гипотезы $T\theta = \tau$, а значит, статистика от $\widehat{\theta}$

$$\frac{1}{\sigma^2}(T\widehat{\theta} - \tau)^T B^{-1}(T\widehat{\theta} - \tau) \sim \chi^2_m.$$

Вспомним, что у нас в запасе есть независящая от $\widehat{\theta}$ статистика

$$\frac{1}{\sigma^2}\|X - Z\widehat{\theta}\|^2 \sim \chi^2_{n-k}.$$

Поделив одно на другое, мы избавимся от неизвестной σ^2, да ещё и получим «хорошее распределение»:

Определение. Пусть независимые случайные величины ξ и η таковы, что $\xi \sim \chi^2_a$, $\eta \sim \chi^2_b$, где $a, b \in \mathbb{N}$. Тогда говорят, что случайная величина

$$\zeta = \frac{\xi/a}{\eta/b}$$

имеет *распределение Фишера со степенями свободы a и b*. Обозначается $\zeta \sim F_{a,b}$

Тогда при верности H_0 имеем

$$\frac{(T\widehat{\theta} - \tau)^T B^{-1} (T\widehat{\theta} - \tau)}{\|X - Z\widehat{\theta}\|^2} \cdot \frac{n-k}{m} \sim F_{m,n-k}.$$

Итоговый критерий записывается так:

$$R = \left\{ \frac{(T\widehat{\theta} - \tau)^T B^{-1} (T\widehat{\theta} - \tau)}{\|X - Z\widehat{\theta}\|^2} \cdot \frac{n-k}{m} > f_{1-\alpha} \right\},$$

где f_p – p-квантиль распределения Фишера со степенями свободы m и $n-k$.

Пример. Допустим, нам пришли две независимые выборки: $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_m$, элементы которых имеют распределение $\mathcal{N}(a, \sigma^2)$ и $\mathcal{N}(b, \sigma^2)$ соответственно. Хотелось бы проверить гипотезу

$$H_0 \colon a = b.$$

Выборки можно рассматривать как общую выборку из модели гауссовской линейной регрессии, а a и b – как координаты одного вектора параметров θ. Таким образом:

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \\ Y_1 \\ \vdots \\ Y_m \end{pmatrix} = Z\theta + \varepsilon = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(0, \sigma^2 E_{n+m})$. Стало быть, H_0 есть линейная гипотеза:

$$H_0 \colon T\theta = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = 0 = \tau.$$

Найдём величины, участвующие в критерии выше:

$$Z^T Z = \begin{pmatrix} n & 0 \\ 0 & m \end{pmatrix}, \quad \widehat{\theta} = \begin{pmatrix} \overline{X} \\ \overline{Y} \end{pmatrix}, \quad B = T(Z^T Z)^{-1} T^T = \frac{1}{n} + \frac{1}{m},$$

$$(T\widehat{\theta} - \tau)^T B^{-1} (T\widehat{\theta} - \tau) = \frac{nm(\overline{X} - \overline{Y})^2}{n+m}, \quad \|X - Z\widehat{\theta}\|^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 + \sum_{j=1}^{m} (Y_j - \overline{Y})^2.$$

Таким образом, критерий для проверки H_0 имеет вид

$$R = \left\{ (\mathbf{x}, \mathbf{y}) \colon \frac{nm(n+m-2)(\overline{\mathbf{x}} - \overline{\mathbf{y}})^2}{(n+m)\left(\sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2 + \sum_{j=1}^{m} (y_j - \overline{\mathbf{y}})^2\right)} > f_{1-\alpha} \right\},$$

где f_p – p-квантиль распределения Фишера со степенями свободы 1 и $n+m-2$.

10.4* Критерий Вальда (z-критерий)

Следующий материал необязателен, но его коснулись на семинарах. К тому же сей критерий является типичным примером *z-критерия*, то есть такого критерия, статистика которого сходится к чему-то нормальному.

В данном разделе мы рассмотрим, наверное, один из самых простых способов проверки *двусторонних* гипотез, то есть гипотез вида

$$H_0\colon \theta = \theta_0 \quad \text{versus} \quad H_1\colon \theta \neq \theta_0.$$

Для построения критерия нам понадобится асимптотически нормальная оценка $\widehat{\theta}$, то есть такая оценка, что

$$\sqrt{n} \cdot \frac{\widehat{\theta} - \theta}{\sigma(\theta)} \xrightarrow{d} \mathcal{N}(0,1),$$

где $\sigma^2(\theta)$ – асимптотическая дисперсия оценки $\widehat{\theta}$. Если мы имеем дело с какой-то сложной моделью, то получить точную формулу для $\sigma^2(\theta)$ может быть довольно сложно, поэтому вместо неё будем использовать состоятельную оценку $\widehat{\sigma}$ для $\sigma(\theta)$. В силу состоятельности отношение сих величин сходится по вероятности (а значит, и слабо) к 1, и по лемме Слуцкого:

$$T_\theta(X) = \sqrt{n} \cdot \frac{\widehat{\theta} - \theta}{\widehat{\sigma}} = \sqrt{n} \cdot \frac{\widehat{\theta} - \theta}{\sigma(\theta)} \cdot \frac{\sigma(\theta)}{\widehat{\sigma}} \xrightarrow{d} \mathcal{N}(0,1).$$

Вернёмся к проверке гипотезы. При верности H_0 имеем $T_{\theta_0}(X) \xrightarrow{d_{\theta_0}} \mathcal{N}(0,1)$, а значит, критерий

$$R = \{\mathbf{x}\colon |T_{\theta_0}(\mathbf{x})| > z_{1-\alpha/2}\}$$

будет иметь асимптотический уровень значимости α (здесь z_p – p-квантиль $\mathcal{N}(0,1)$). Действительно, если за Φ обозначить функцию распределения для $\mathcal{N}(0,1)$, то

$$\begin{gathered}
\mathsf{P}_{\theta_0}(|T_{\theta_0}(X)| > z_{1-\alpha/2}) = \\
= \mathsf{P}_{\theta_0}(T_{\theta_0}(X) > z_{1-\alpha/2}) + \mathsf{P}_{\theta_0}(T_{\theta_0}(X) < -z_{1-\alpha/2}) \xrightarrow[\text{из слаб. сх-ти}]{} (1 - \Phi(z_{1-\alpha/2})) + \Phi(-z_{1-\alpha/2}) = \\
= 1 - (1-\alpha/2) + (1 - \Phi(z_{1-\alpha/2})) = \alpha/2 + (1 - (1-\alpha/2)) = \alpha.
\end{gathered}$$

Теперь изучим критерий на предмет мощности. Предположим, что истинное значение θ не равно θ_0. Тогда

$$\begin{gathered}
\beta(\theta) = \mathsf{P}_\theta(|T_{\theta_0}(X)| > z_{1-\alpha/2}) = \\
= \mathsf{P}_\theta\left(\sqrt{n} \cdot \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}} > z_{1-\alpha/2}\right) + \mathsf{P}_\theta\left(\sqrt{n} \cdot \frac{\widehat{\theta} - \theta_0}{\widehat{\sigma}} < -z_{1-\alpha/2}\right) = \\
= \mathsf{P}_\theta\left(\sqrt{n} \cdot \frac{\widehat{\theta} - \theta}{\widehat{\sigma}} > z_{1-\alpha/2} + \sqrt{n} \cdot \frac{\theta_0 - \theta}{\widehat{\sigma}}\right) + \mathsf{P}_\theta\left(\sqrt{n} \cdot \frac{\widehat{\theta} - \theta}{\widehat{\sigma}} < -z_{1-\alpha/2} + \sqrt{n} \cdot \frac{\theta_0 - \theta}{\widehat{\sigma}}\right) \approx \\
\approx 1 - \Phi\left(z_{1-\alpha/2} + \sqrt{n} \cdot \frac{\theta_0 - \theta}{\widehat{\sigma}}\right) + \Phi\left(-z_{1-\alpha/2} + \sqrt{n} \cdot \frac{\theta_0 - \theta}{\widehat{\sigma}}\right).
\end{gathered}$$

Так как $\theta \neq \theta_0$, то содержимое в скобках стремится к $\pm\infty$, а значит, значения Φ либо примерно 1, либо примерно 0, отчего мощность близка к единице. Причём из написанного выше видно, что мощность тем больше, чем больше размер выборки и чем дальше от θ_0 находится рассматриваемый параметр из альтернативы.

Задача 10.7. (а) Пусть $X_1, \ldots, X_n$ – выборка на распределения $Bern(p)$. Предложите z-критерий (то есть критерий, распределение статистики которого сходится к нормальному распределению при $n \to \infty$) для проверки гипотезы

$$H_0\colon p = p_0 \quad \text{versus} \quad H_1\colon p \neq p_o.$$

(б) По данным Интернет-опроса за одного из кандидатов собирались проголосовать 3% избирателей. По официальным данным за этого кандидата в итоге проголосовали 4661075 из 5818955 избирателей. Нулевая гипотеза заключается в корректности данных Интернет-опроса. На каком уровне значимости можно её принять?

Решение. **(а)** Критерий Вальда тут подходит идеально. По ЦПТ имеется асимптотически нормальная оценка $\widehat{p} = \overline{X}$, асимптотическую дисперсию которой можно выразить точно и без оценивания: это просто дисперсия одного наблюдения, то есть $\sigma^2(p) = \mathsf{D}_p X_i = p(1-p)$. Итого, критерий имеет вид

$$R = \left\{ \mathbf{x} \colon \sqrt{n} \cdot \frac{\overline{\mathbf{x}} - p_0}{\sqrt{p_0(1-p_0)}} > z_{1-\alpha/2} \right\}$$

(б) Посчитаем статистику критерия:

$$T(X) = \sqrt{5818955} \cdot \frac{\frac{4661075}{5818955} - 0.03}{\sqrt{0.03 \cdot (1 - 0.03)}} \approx 10902.83.$$

Кажется, это больше, чем любой адекватный квантиль нормального распределения. Минимальный уровень значимости, на котором мы должны отвергнуть гипотезу, очень близок к нулю (Python при попытке его посчитать выдаёт просто 0), что является невероятно весомым доказательством против гипотезы о корректности Интернет-опроса (ох уж эти опросы!). ■

Критерий Вальда подходит и для проверки односторонних гипотез, например:

$$H_0 \colon \theta = \theta_0 \quad \text{versus} \quad H_1 \colon \theta > \theta_0.$$

В таких случаях критерий логично переформулировать так:

$$R_+ = \{ \mathbf{x} \colon T_{\theta_0}(\mathbf{x}) > z_{1-\alpha} \}.$$

На асимптотический уровень значимости это не повлияет, зато мы увеличим мощность: теперь мы можем забыть про «левый хвост» нормального распределения и больше уделить внимания правому, попадание в который более вероятно для $\theta > \theta_0$. Аналогично, если альтернатива имеет вид $H_1 \colon \theta < \theta_0$, то в такой ситуации лучше взять критерий

$$R_- = \{ \mathbf{x} \colon T_{\theta_0}(\mathbf{x}) < -z_{1-\alpha} \}.$$

Задача 10.8. Посетители ТРЦ Рио ходили по магазинам в среднем 3/4 часа, стандартное отклонение (a.k.a. корень из дисперсии) было равно 0.1. Потом на втором этаже появился детский паровозик, а на следующий день оказалось, что по выборке из 35 посетителей среднее время шоппинга составило 4/5 часа. Требуется проверить на уровне значимости 0.05 гипотезу о пользе паровозика. Какими будут H_0 и H_1? Придумайте критерий и проверьте гипотезу.

Решение. Выдвинем на проверку

$$H_0 \colon \text{Паровозик не повлиял} \quad \text{versus} \quad H_1 \colon \text{Паровозик помог}$$

Если верна H_0, то с появлением паровозика ничего не поменялось, поэтому среднее и отклонение распределения остались прежними, то есть 0.75 и 0.1 соответственно. Попробуем применить односторонний критерий Вальда (было бы странно в альтернативу, утверждающую, что паровозик помог, запихивать случай, когда среднее уменьшилось):

$$T(X) = \sqrt{35} \cdot \frac{0.8 - 0.75}{0.1} \approx 2.958.$$

Фактический уровень значимости будет примерно равен $1 - \Phi(2.958)$, что равняется ≈ 0.0015. Это меньше нашего уровня значимости, да и само по себе это весьма маленькое значение, что только больше побуждает отвергнуть H_0. С учётом сего факта и того, что средняя продолжительность шоппинга увеличилась, логично сделать вывод, что паровозик принёс пользу. ■

Задача 10.9. Пусть $X_1, \ldots, X_n$ – выборка из распределения $\mathcal{N}(a_1, \sigma_1^2)$, $Y_1, \ldots, Y_m$ – выборка из распределения $\mathcal{N}(a_2, \sigma_2^2)$, причём выборки независимы. Предложите критерий для проверки гипотезы $H_0 \colon \sigma_1^2 = \sigma_2^2$.

Решение. Гипотезу можно переформулировать так:

$$H_0\colon \delta = \sigma_1^2 - \sigma_2^2 = 0.$$

Таким образом, можно протестировать гипотезу о том, что параметр δ равен нулю. За оценку сего параметра логично взять $\widehat{\delta} = s^2(X) - s^2(Y)$, то есть разность выборочных дисперсий выборок X и Y (её асимптотическая нормальность следует из из теоремы о наследовании асимптотической нормальности). Осталось найти дисперсию данной оценки. Это сделать довольной просто, если вспомнить, что $\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}$ для выборки размера n. Таким образом, после гуглинга дисперсии распределения хи-квадрат получаем, что

$$\mathsf{D}\widehat{\delta} = \mathsf{D}s^2(X) + \mathsf{D}s^2(Y) = \frac{\sigma_1^4}{n^2}\cdot 2(n-1) + \frac{\sigma_2^4}{m^2}\cdot 2(m-1).$$

Сами параметры σ_1^2 и σ_2^2 мы не знаем, поэтому логично заменить их на их состоятельные оценки (для приличия возьмём их несмещённые оценки):

$$\widehat{\mathsf{D}\widehat{\delta}} = 2\frac{\left(s^2(X)\right)^2}{n-1} + 2\frac{\left(s^2(Y)\right)^2}{m-1}.$$

Возьмём от всего этого дела корень, чтобы получить стандартное отклонение, и запишем итоговый критерий:

$$R = \left\{ (X,Y)\colon \frac{s^2(X) - s^2(Y)}{\sqrt{2\frac{\left(s^2(X)\right)^2}{n-1} + 2\frac{\left(s^2(Y)\right)^2}{m-1}}} > z_{1-\alpha/2} \right\}.$$

■

10.5* Критерий омега-квадрат

Сего материала на семинарах не было, но он есть в некотором виде в конспектах [5], так что почему бы и нет.

Вновь поставим на проверку гипотезу H_0 о том, что истинное распределение равно некоторому непрерывному распределению F. У нас уже имеется критерий для проверки такой гипотезы: критерий Колмогорова. Но теперь мы используем другую парадигму: если раньше мы смотрели на отклонение в равномерной метрике, то сейчас мы будем оценивать его через интеграл (чем-то напоминает байесовский подход в сравнении оценок).

Определение. Пусть нам дана некоторая «весовая» функция $\psi(t)$ на $[0;1]$. *Статистикой омега-квадрат* называют

$$\omega^2(\psi) = \int_{\mathbb{R}} \left(\widehat{F}_n(x) - F(x)\right)^2 \psi(F(x))\, dF(x).$$

Среди многообразия весовых функций мы рассмотрим

$$\psi_1(t) \equiv 1 \quad \text{и} \quad \psi_2(t) = \frac{1}{t(1-t)}.$$

Выбор именно таких функций оправдывается их простотой и подходом в обнаружении отклонений. В [3, гл. 12, § 2] даётся такое описание:

> Первый из них хорошо улавливает расхождение между $\widehat{F}_n$ и F в области «типичных значений» случайной величины с функцией распределения F (часто он оказывается более чувствительным, чем критерий Колмогорова). Второй же, благодаря тому, что $\psi_2(y)$ быстро возрастает при $y \to 0$ и $y \to 1$, способен заметить различие «на хвостах» распределения F, которому придается дополнительный вес.

Как и в случае со статистикой критерия Колмогорова, статистика омега-квадрат, только домноженная уже на n, имеет некоторый предельный закон.

Теорема. *При верности гипотезы H_0 статистики $n\omega^2(\psi_1)$ и $n\omega^2(\psi_2)$ слабо сходятся к некоторым фиксированным распределениям F_1 и F_2 соответственно.*

У сих распределений также имеется разложение в ряд, но оно настолько ужасное, что мне не хотелось бы пугать им читателей. Если положить y_p и z_p за p-квантили F_1 и F_2 соответственно, то получатся два асимптотических критерия с уровнем значимости α:

$$R_1 = \left\{n\omega^2(\psi_1) > y_{1-\alpha}\right\} \text{ — } \textit{критерий Крамера — фон Мизеса — Смирнова}$$
$$R_2 = \left\{n\omega^2(\psi_2) > z_{1-\alpha}\right\} \text{ — } \textit{критерий Андерсона — Дарлинга}$$

Приведём некоторые квантили этих распределений, чтобы не ходить далеко искать:

α	0.5	0.15	0.1	0.05	0.025	0.01	0.001
$y_{1-\alpha}$	0.12	0.28	0.35	0.46	0.58	0.74	1.17
$z_{1-\alpha}$	0.77	1.62	1.94	2.49	3.08	3.88	5.97

Как на практике вычислять значение статистики омега-квадрат? Можно топорно вычислять интеграл с помощью, например, scipy.integrate.quad, но ответ будет приближённым, а нам хотелось бы точный. Благо в силу кусочно-постоянности $\widehat{F}_n$ можно упростить интеграл выше и получить следующие более приятные формулы:

$$n\omega^2(\psi_1) = \frac{1}{12n} + \sum_{i=1}^{n}\left[F(x_{(i)}) - \frac{2i-1}{2n}\right]^2,$$

$$n\omega^2(\psi_2) = -n - 2\sum_{i=1}^{n}\left[\frac{2i-1}{2n}\ln F(x_{(i)}) + \left(1 - \frac{2i-1}{2n}\right)\ln\left(1 - F(x_{(i)})\right)\right].$$

Пример. Рассмотрим данные из задачи 10.2.

Статистика из критерия КМС равна ≈ 0.1968, и судя по таблице, ей соответствует уровень значимости больше 0.15, что довольно много, и гипотезу мы не отвергаем.

Статистика для другой весовой функции показывает результат ≈ 2.67, и соответствующий уровень значимости уже будет меньше 0.05, что является доводом против H_0.

10.6* Немного про p-value

В задачах выше периодически возникали фразы по типу «минимальный уровень значимости, на котором мы должны отвергнуть гипотезу» и «фактический уровень значимости». Даже была оставлена ссылка на Википедию по этому поводу, но по мере составления конспекта стало понятно, что использование таких словечек вскользь скорее вредит, чем приносит пользу, и надо бы ввести некоторые разъяснения, что это, зачем оно надо и каких ложных выводов не стоит делать в этой связи. Опять же, на семинарах сего материала не было, и автор заранее приносит извинения за возможный дальнейший бред, но инструмент этот весьма полезный.

Определение. Пусть для проверки гипотезы $H_0\colon \mathsf{P} \in \mathcal{P}_0$ на уровне значимости α имеется критерий R_α. Назовём *p-value* или *фактическим уровнем значимости* следующую величину:

$$\text{p-value} = \inf\{\alpha\colon X \in R_\alpha\}.$$

Поясним, что тут происходит. Для каждого α у нас в рукаве имеется критерий R_α с размером α (логично допустить, что при $\alpha < \beta$ имеется вложение $R_\alpha \subset R_\beta$). Мы смотрим, при каких α реализация выборки X попала в критическое множество R_α, то есть при каких α нам следовало

бы отвергнуть гипотезу, и среди них берём инфимум. То есть p-value – это *минимальный уровень значимости, на котором мы должны отвергнуть гипотезу*. Таким образом,

$$H_0 \text{ отвергается} \iff X \in R_\alpha \iff \text{p-value} \leqslant \alpha.$$

Очень часто критерии имеют вид $R_\alpha = \{X\colon T(X) \geqslant c_\alpha\}$, где α – размер критерия. Предположим, что наблюдаемое значение статистики $T(X)$ равно t. Тогда p-value можно переписать так:

$$\text{p-value} = \inf\{\alpha\colon t \geqslant c_\alpha\} = \alpha(t),$$

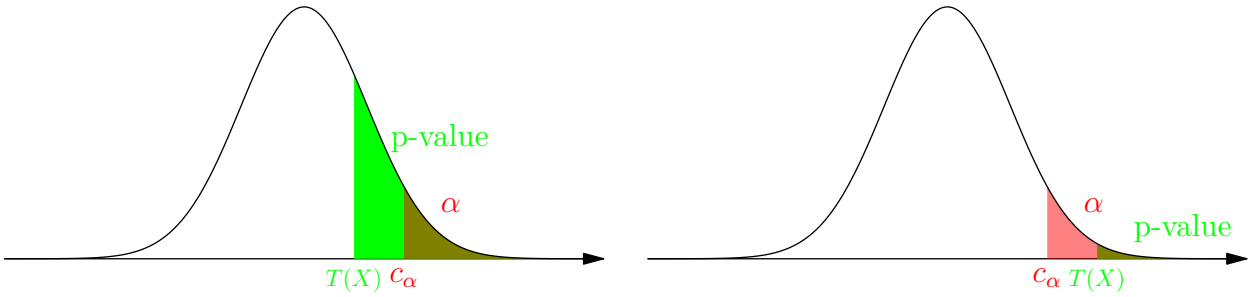
где $c_{\alpha(t)} = t$ (при уменьшении α граница c_α лишь увеличивается, поэтому инфимум достигается при $t = c_\alpha$). Вспоминая определение размера критерия, получаем, что

$$\text{p-value} = \alpha(t) = \sup_{\mathsf{P}\in\mathcal{P}} \mathsf{P}(T(X) \geqslant c_{\alpha(t)}) = \sup_{\mathsf{P}\in\mathcal{P}} \mathsf{P}(T(X) \geqslant t).$$

Если основная гипотеза простая, то супремум берётся по одному элементу – предполагаемому распределению из H_0. В таком случае можно переформулировать

Определение. p-value – это вероятность наблюдать статистику критерия такую же или даже более экстремальную, чем она есть на самом деле, при условии верности H_0.

Это можно проиллюстрировать следующими картинками, на которых изображена плотность статистики $T(X)$:



На левом рисунке значение статистики оказалось достаточно маленьким, и соответствующее p-value (что есть площадь под графиком, выделено зелёным) больше, чем заявленный уровень значимости α (выделен красным). Значит, гипотеза не отвергается. На правом же рисунке статистика $T(X)$ приняла весьма экстремальное значение и попала в «критическую зону». Отсюда делаем вывод о необходимости отвергнуть H_0.

Теперь следует сделать некоторые

Замечания. 1. **p-value не есть уровень значимости**, это разные вещи. Уровень значимости – это фиксированное число, величина позволяемой ошибки I рода, которую мы фиксируем до наблюдения. p-value же – функция от наблюдения, которое уже произошло. Существование p-value не освобождает от постановки уровня значимости в самом начале. У автора есть гипотеза, что первое определение p-value не любят использовать именно из-за подобной путаницы.

2. **p-value не есть вероятность того, что H_0 верна**. Гипотеза либо верна, либо нет, это не случайное событие, и байесовским подходом мы тут не занимаемся.

3. p-value можно рассматривать как степень уверенности в отклонении H_0. Если оно близко к нулю, то по версии H_0 произошло очень маловероятное событие, что и заставляет нас отклонить её. То есть чем меньше p-value, тем более мы спокойны о нашем решении в отвержении H_0. Например, в задаче 10.7 фактический уровень значимости не просто меньше заявленного уровня значимости, так ещё и чрезвычайно близок к нулю, поэтому сомнений в отвержении гипотезы должно быть минимум.

4. Высокое p-value не свидетельствует о верности H_0. Вполне возможно, что на самом деле верна

какая-нибудь альтернатива H_1, но мощность критерия оставляет желать лучшего, поэтому мы не попадаем в R_α и при больших α, что и означает большой p-value.

11 Коэффициенты корреляции

Часто на практике представляется весьма полезным проверить, зависимы ли какие-то две характеристики. Представим, что для некоторых n объектов есть признак X, их можно записать как вектор $(X_1, \dots, X_n)$, а также признак Y, их можно записать как вектор $(Y_1, \dots, Y_n)$. Таким образом, (X_i, Y_i) является случайным вектором для всех i, который описывает пару характеристик для i-ого объекта (выборки с таким свойством ещё называют *связанными*). Нас интересует, зависимы ли они, или, что эквивалентно, мы проверяем гипотезу о независимости:

$$H_0 \colon F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

Как известно, у независимости и корреляции есть некоторая связь (хотя между этими понятиями имеются и различия), поэтому логичным представляется исследовать корреляцию между элементами выборки, так как идейно и вычислительно это проще, чем проверять равенство выше для всех x, y. Для этого рассматривают всякие статистики с областью значений $[-1; 1]$, которые ознаменуют собой коррелированность выборок. Их обычно называют *коэффициентами корреляции*. Рассмотрим некоторые из них.

11.1 Коэффициент корреляции Пирсона

Самое простое, что можно придумать, - это взять «выборочную корреляцию».

Определение. *Коэффициентом корреляции Пирсона* называют следующую статистику:

$$\widehat{\rho} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}.$$

Выбор такого коэффициента корреляции оправдывается тем, что в силу УЗБЧ и теоремы о наследовании сходимости доказывается, что

$$\widehat{\rho} \xrightarrow{\mathsf{P}} \rho(X_1, Y_1) = \frac{\mathsf{cov}(X_1, Y_1)}{\sqrt{\mathsf{D}X_1 \mathsf{D}Y_1}} = \mathsf{corr}(X_1, Y_1), \quad n \to \infty.$$

В контексте проверки гипотезы H_0 выше особо важной является следующая

Теорема. *Если нормально распределённые выборки X, Y независимы и $n > 2$, то*

$$T := \widehat{\rho}\sqrt{\frac{n-2}{1-\widehat{\rho}^2}} \sim T_{n-2},$$

где T_m – распределение Стьюдента с m степенями свободы.

Проверка гипотезы проводится следующим образом: если коэффициент $\widehat{\rho}$ близок к границам отрезка $[-1; 1]$, то это является поводом отклонить H_0. В условиях теоремы выше критерий уровня значимости α проверки гипотезы можно формализовать как

$$R = [-1; 1] \setminus (t_{\alpha/2}; t_{1-\alpha/2}),$$

где t_p – p-квантиль распределения T_{n-2}.

11.2 Коэффициент корреляции Спирмэна

Какие минусы у коэффициента выше? Во-первых, конечно, не все рассматриваемые выборки нормальны, хотя такое допущение встречается довольно часто. Самое неприятное – низкая робастность, то есть неустойчивость статистики к выбросам, что особенно характерно для тех из них, которые основаны на выборочном среднем.

У нас уже встречались статистики, которые таким недостатком обладают в меньшей степени – это порядковые статистики. В этой связи давайте прибегнем к так называемым *ранговым критериям*, которые основываются на ранге — номере элементов выборки, расположенных в порядке возрастания.

Определение. Пусть R_i и S_j – место в вариационном ряду для X_i и Y_j соответственно. *Коэффициентом корреляции Спирмэна* называют следующую статистику:

$$\rho_S = \frac{\sum_{i=1}^n (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^n (R_i - \overline{R})^2 \sum_{i=1}^n (S_i - \overline{S})^2}}.$$

Если считать, что функции распределения F_X и F_Y непрерывны, то всё хорошо, вероятность того, что какие-либо два элемента выборки совпадут, равна нулю, поэтому почти наверное такое упорядочивание однозначно. Если же в выборке встречаются одинаковые значения, то обычно используют средние ранги. Например, если выборка представляет собой набор $2, 5, 5, 7$, то их средние ранги равны соответственно $1, 2.5, 2.5, 4$. Такой подход сохраняет сумму всех рангов, а вот с суммой квадратов будут проблемы, поэтому некоторые вещи ниже для такой модели неприменимы. Чтобы с этим всем не возиться, для простоты будем всё-таки подразумевать, что функции распределения непрерывны.

Оформим все свойства в одном утверждении.

Теорема. *Имеют место быть следующие свойства:*

1. *Коэффициент корреляции Спирмэна можно переписать в виде*
$$\rho_S = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (R_i - S_i)^2.$$
2. *При верности H_0 имеем $\mathsf{E}\rho_S = 0$, $\mathsf{D}\rho_S = \frac{1}{n-1}$, а также есть сходимость:*
$$\frac{\rho_S}{\sqrt{\mathsf{D}\rho_S}} \xrightarrow{d} \mathcal{N}(0, 1).$$
3. *Коэффициент корреляции и в самом деле отражает корреляцию между элементами выборки, то есть $-1 \leqslant \rho_S \leqslant 1$, причём крайние значения достигаются.*

Для приличия найдём матожидание и дисперсию сего коэффициента. Во-первых, сделаем витающее в воздухе замечание: $R_1, \ldots, R_n$ есть ничто иное, как перестановка чисел $1, \ldots, n$, поэтому если мы встретим какое-либо симметричное выражение, зависящее от R_i, то мы всегда в нём сможем сделать замену. Так, например, $\sum R_i = \frac{n(n+1)}{2}$, а $\sum R_i^2 = \frac{n(n+1)(2n+1)}{6}$. Во-вторых, при верности гипотезы H_0 величины R_i и S_j независимы при любых i, j, поэтому матожидание от их произведения раскладывается в произведение матожиданий. В свою очередь так как компоненты выборки независимы, то ранги могут образовывать любую перестановку равновероятно, откуда несложно посчитать $\mathsf{E}R_i = \mathsf{E}S_j = \frac{n+1}{2}$. С этими новыми знаниями посчитаем матожидание ρ_S:

$$\mathsf{E}\rho_S = \mathsf{E}\left(1 - \frac{6}{n^3 - n} \sum_{i=1}^n (R_i - S_i)^2\right) = \mathsf{E}\left(1 - \frac{6}{n^3 - n} \sum_{i=1}^n (R_i^2 - 2R_iS_i + S_i^2)\right) =$$

$$1 - \frac{6}{n^3 - n} \cdot 2 \cdot \frac{n(n+1)(2n+1)}{6} + \frac{12}{n^3 - n} \cdot n\mathsf{E}R_1S_1 = 1 - \frac{4n+2}{n-1} + \frac{12}{n^2-1} \cdot \left(\frac{n+1}{2}\right)^2 =$$

$$= \frac{-3n-3}{n-1} + 3 \cdot \frac{n+1}{n-1} = 0.$$

Отлично, теперь посмотрим на дисперсию. Так как прибавление константы на дисперсию не влияет, то оставим в формуле коэффициента только сумму произведений R_iS_i. Также нелишним

будет посчитать матожидание квадрата ранга:

$$\mathsf{E}R_1^2 = \sum_{i=1}^{n} i^2 \cdot \mathsf{P}(R_1 = i) = \frac{1}{n}\sum_{i=1}^{n} i^2 = \frac{(n+1)(2n+1)}{6},$$

и матожидание произведения двух разных рангов R_i и R_j: различных способов выбрать значения для них теперь равно $n(n-1)$, и они также равновероятны. Поэтому

$$\mathsf{E}R_iR_j = \sum_{i\neq j} \frac{1}{n(n-1)} \cdot ij = \sum_{i,j=1}^{n} \frac{1}{n(n-1)} \cdot ij - \sum_{i=1}^{n} \frac{1}{n(n-1)} \cdot i^2 =$$

$$= \frac{1}{n(n-1)} \cdot \frac{n^2(n+1)^2}{4} - \frac{1}{n(n-1)} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(3n+2}{12}.$$

Теперь можем начинать жёстко считать дисперсию:

$$\mathsf{D}\rho_S = \frac{144}{(n^3-n)^2}\mathsf{D}\sum R_iS_i = \frac{144}{(n^3-n)^2}\left[\mathsf{E}\left(\sum R_iS_i\right)^2 - \left(\mathsf{E}\sum R_iS_i\right)^2\right] =$$

$$= \frac{144}{(n^3-n)^2}\left[\sum \mathsf{E}R_i^2S_i^2 + \sum_{i\neq j}\mathsf{E}R_iS_iR_jS_j - \left(n\cdot\mathsf{E}R_1S_1\right)^2\right] =$$

$$= \frac{144}{(n^3-n)^2}\left[n\cdot\mathsf{E}R_1^2\cdot\mathsf{E}S_1^2 + n(n-1)\mathsf{E}R_1R_2\cdot\mathsf{E}S_1S_2 - \left(n\cdot\mathsf{E}R_1\cdot\mathsf{E}S_1\right)^2\right] =$$

$$= \frac{144}{n(n^2-1)^2}\left[\frac{(n+1)^2(2n+1)^2}{36} + (n-1)\cdot\frac{(n+1)^2(3n+2)^2}{144} - n\cdot\frac{(n+1)^4}{16}\right] =$$

$$= \frac{1}{n(n-1)^2}\left(4(2n+1)^2 + (n-1)(3n+2)^2 - 9n(n+1)^2\right) = \frac{1}{n-1}.$$

11.3 Коэффициент корреляции Кендалла

Схожую по идеологии ранжирования статистику ввёл М. Дж. Кендэлл. Только теперь мы смотрим на количество инверсий, которые образуются во второй выборке, если расположить их в порядке возрастания соответствующих элементов первой. То есть появляется некоторая мера неупорядоченности второй выборки относительно первой, и если выборки независимы, то логично предположить, что инверсий будет примерно столько же, сколько и правильно упорядоченных пар. Более формально:

Определение. Будем говорить, что пары (X_i, Y_i) и (X_j, Y_j) *согласованны* (считаем, что $1 \leqslant i < j \leqslant n$), если $X_i < X_j$ и $Y_i < Y_j$ или $X_i > X_j$ и $Y_i > Y_j$ (то есть $sign(X_j - X_i)(Y_j - Y_i) = 1$).

Пусть для выборок X и Y величина S есть число согласованных пар, а R – число несогласованных (по всем $1 \leqslant i < j \leqslant n$). При верности гипотезы они должны не слишком сильно отличаться, поэтому логично ввести следующую меру превышения согласованности над несогласованностью:

$$T = S - R = \sum_{i<j} sign(X_j - X_i)(Y_j - Y_i).$$

Понятное дело, что величина T может меняться от $-\frac{n(n-1)}{2}$ до $\frac{n(n-1)}{2}$ (второй вариант характерен для выборок с полным согласием порядка, а первый – наоборот, когда увеличение X означает уменьшение Y). Поэтому логично нормировать полученную статистику, чтобы она лежала на отрезке $[-1; 1]$, как и все коэффициенты корреляции.

Определение. *Коэффициентом корреляции Кендалла* называют следующую статистику:

$$\tau = \frac{2}{n(n-1)} \cdot T = \frac{2}{n(n-1)} \sum_{i<j} sign(X_i - X_j) \cdot sign(Y_i - Y_j)$$

Отметим следующие свойства:

1. При верности H_0 имеем $\mathsf{E}\tau = 0$, $\mathsf{D}\tau = \frac{2(2n+5)}{9n(n-1)}$, и есть сходимость
$$\frac{\tau}{\sqrt{\mathsf{D}\tau}} \xrightarrow{d} \mathcal{N}(0,1).$$
2. С учётом того, что $S + R = \frac{n(n-1)}{2}$, коэффициент корреляции можно переписать как
$$\tau = 1 - \frac{4}{n(n-1)}R.$$

Список литературы

[1] А. А. Боровков. Математическая статистика.

[2] L. Wasserman. All of Statictics.

[3] М. Б. Лагутин. Наглядная математическая статистика.

[4] М. Е. Жуковский, И. В. Родионов, Д. А. Шабанов. Введение в математическую статистику.

[5] М. П. Савёлов, В. Хаймоненко. Конспект лекций по математической статистике, осень 2021.