ФПМИ МФТИ

ВВЕДЕНИЕ В МАТСТАТЫ

Составитель: Андрей Куссев

Осень, 2022 год
Дата компиляции:
10.11.2025

Содержание

Примечание. Перед Вами учебное пособие по математической статистике, за основу которого взяты семинары уважаемого Алексея Сергеевича Волостнова, проведённые осенью 2022 года на 3 курсе ПМИ Физтеха. Целью сей методички является помощь в постижении этого не самого лёгкого курса, а также знакомство читателя с дополнительными, но не менее важными сведениями, которым не уделяют время на основном потоке по статистике.

Однако следует помнить: эта методичка написана ещё тем оболтусом и профаном. 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 автора для последующего исправления.

Код сего документа открыт и доступен в Overleaf и GitHub, там же можно найти актуальную версию PDF.

Часть I
Параметрический подход

1 Оценки и их свойства

В теории вероятности мы в основном работали над созданием инструментов для описания различных распределений случайных элементов. Мы фиксировали вероятностное пространство, притворяясь, что с ним знакомы, рассматривали случайные элементы на нём и с помощью некоторого арсенала описывали поведение случайной величины. Таким образом, теория вероятности отвечает на вопрос: если случайная величина распределена именно так, то какие свойства (например, асимптотические) она имеет?

Математическая статистика делает всё с точностью до наоборот: по свойствам того, что нам позволено лицезреть, необходимо определить, из какого распределения пришла наблюдаемая величина. 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 выборки.

1.1 Вероятностно-статистическая модель

Прежде всего необходимо провести некоторые косметические изменения в модели вероятностного пространства, с которым мы работали ранее. Для начала определимся с его измеримой частью, а именно с множеством элементарных исходов $\mathcal{X}$ и σ-алгеброй событий $\mathcal{F}$ на нём. Обычно в качестве первого берут непосредственно множество результатов наблюдения, а в качестве второго — борелевскую σ-алгебру $\mathcal{B}(\mathcal{X})$ на нём, которую ещё со времён теории вероятностей мы считаем вполне естественной. В частности, это могут быть (и чаще всего и будут):

- $\mathcal{X} = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, если наблюдение в ходе эксперимента представляет собой одну вещественную величину;
- $\mathcal{X} = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^n)$, если мы наблюдаем за несколькими величинами или целым случайным вектором. Напомним, что $\mathcal{B}(\mathcal{X})$ в данном случае можно породить, например, открытыми множествами в $\mathbb{R}^n$, открытыми брусами в том же пространстве или декартовыми произведениями $B_1 \times \ldots \times B_n$ борелевских множеств $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R})$, в зависимости от удобства в конкретном случае;
- $\mathcal{X} = \mathbb{R}^\infty$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^\infty)$, если наблюдения представляют собой последовательность случайных величин или векторов. Конечно, на практике мы не имеем бесконечного числа экспериментов, но такое пространство удобно при изучении асимптотических свойств оценок, когда количество наблюдаемых величин сколь угодно большое. Не будем вдаваться в подробности устройства этой σ-алгебры (как мы меры на ней), но отметим, что она также порождается конечными декартовыми произведениями $B_{n_1} \times \ldots \times B_{n_k}$, где $n_i \in \mathbb{N}$, а B_{n_i} могут быть как промежутками, так и вообще борелевскими множествами в $\mathbb{R}$.

Так как достоверно истинное распределение нам неизвестно, то и вероятностная мера на нашем пространстве определена не однозначно. Отсюда логично рассматривать семейство

$\mathcal{P}$ вероятностных мер на измеримом пространстве на $\mathcal{B}(\mathcal{X})$. Чаще всего это семейство можно описать одним или несколькими параметрами, то есть будет иметь вид $\mathcal{P} = \{\mathsf{P}_\theta \colon \theta \in \Theta\}$, где Θ — множество допустимых значений параметров. Например, нормальное распределение $\mathcal{N}(a, \sigma^2)$ можно описать двумя параметрами: его матожиданием a и дисперсией σ^2, таким образом, оно однозначно описывается параметром $(a, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. Так как распределение теперь не фиксировано, то часто в обозначения матожидания, сходимости по вероятности и т. д., которые используют конкретное распределение, мы будем писать индекс θ, чтобы подчеркнуть, какое распределение используется в данный момент.

Итого, мы построили *вероятностно-статистическую модель* $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathcal{P})$. Так как мы договорились в качестве $\mathcal{X}$ брать пространство значений наблюдаемых величин, то само *наблюдение* можно задать как отображение $X \colon \mathcal{X} \to \mathcal{X}$, $X(\omega) = \omega$, то есть величина X будет непосредственно показывать, какой элементарный исход был разыгран волей случая. Легко понять, что функция X является измеримой, и её распределение совпадает с истинным распределением $\mathsf{P} \in \mathcal{P}$. Также зачастую наблюдения на практике бывают независимыми (или мы слепо верим в то, что они независимы), поэтому в основном мы будем работать с семействами распределений $\mathcal{P}$ такими, при которых элементы выборки являются независимыми в совокупности случайными величинами. То есть если $\mathbf{X} = (X_1, X_2, \ldots)$ — наблюдение, то для любого $\mathsf{P} \in \mathcal{P}$ выполнено

$$\mathsf{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \prod_{i=1}^{n} \mathsf{P}(X_i \in B_i).$$

Читатель может резонно заметить, не умаляется ли общность введением выборочного пространства? Ведь на практике вероятностное пространство устроено куда сложнее, чем просто набор возможных значений эксперимента. Однако такое допущение не будет влиять на выводимые нами результаты. Во-первых, выборочное пространство порождается самим экспериментом. Грубо говоря, исходное пространство, которое слишком сложное либо вообще неизвестное, содержит в себе некоторый интересующий нас кусок, который и описывает распределение величины из эксперимента. Отсюда и логично использовать для дальнейших выкладок выборочное пространство, которое непосредственно связано с наблюдением. Во-вторых, все наши дальнейшие действия будут вестись не с самими наблюдениями, а с функциями от них. То есть мы не будем работать с самими элементами пространства, которое может быть устроено иначе, а с тем, что это пространство нам выдаёт.

Определение. Пусть $(\Omega, \mathcal{E})$ — измеримое пространство. Произвольная композиция $(\mathcal{B}(\mathcal{X})|\mathcal{E})$-измеримой функции $S \colon \mathcal{X} \to \Omega$ и наблюдения $\mathbf{X}$ называется *статистикой* (иногда под статистикой будет иметься в виду само отображение S). Обозначается как $S(\mathbf{X})$.

Пример 1.1. Рассмотрим некоторые широко применимые примеры статистик.

- Если интересующую функцию от параметра можно записать как матожидание от некоторой борелевской функции g, то логично в качестве оценки брать среднее арифметическое этой функции от элементов выборки, в широком наборе случаев оно будет стремиться к истинному матожиданию по ЗБЧ. Такая статистика

$$\overline{g(\mathbf{X})} = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

  называется *выборочной характеристикой функции* g. В частности, если $g(x) = x^k$,

то её выборочную характеристику

$$\overline{\mathbf{X}^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

называют *выборочным k-ым моментом*, для $k = 1$ её обычно называют просто *выборочным средним*.

- Мы научились приближать матожидание, а что с дисперсией? Её оценку можно получить, если усреднить отклонение наблюдений от выборочного среднего:

$$s^2(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{\mathbf{X}}\right)^2$$

Такую статистику называют *выборочной дисперсией*. Иногда её удобно перезаписать в следующем виде:

$$s^2 = \sum_{i=1}^{n}\left(\frac{X_i^2}{n} - \frac{2X_i\cdot\overline{\mathbf{X}}}{n} + \frac{\overline{\mathbf{X}}^2}{n}\right) = \frac{\sum X_i^2}{n} - 2\overline{\mathbf{X}}\cdot\frac{\sum X_i}{n} + \overline{\mathbf{X}}^2 = \overline{\mathbf{X}^2} - \overline{\mathbf{X}}^2.$$

- Полезно также рассмотреть *k-ую порядковую статистику* $X_{(k)}$ из вариационного ряда

$$X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)},$$

который получается упорядочиванием первоначальной выборки $\mathbf{X} = (X_1, \ldots, X_n)$.

■

Обратим внимание, что значения статистики по определению не зависят от параметра, он влияет лишь на её распределение. Требование весьма логичное, иначе получается странно: хотим оценить неизвестный параметр с помощью статистики, при этом она почему-то включает в себя этот параметр. Например, $X - \mathsf{E}_\theta X$ не является статистикой. Далее нам встретятся менее очевидные примеры, когда в определении функции используется значение параметра, но при этом как таковой зависимости от него нет.

Не все статистики подходят для оценивания неизвестного параметра, например, для оценки дисперсии логично использовать лишь неотрицательные функции. Какой бы хорошей статистикой $S(\mathbf{X})$ ни была, если она периодически выдаёт недопустимое значение параметра, то как оценка она неадекватна. Поэтому логично ввести следующее

Определение. Пусть $\mathcal{P} = \{\mathsf{P}_\theta \colon \theta \in \Theta\}$, а $\Omega = \Theta$, тогда статистика $S \colon \mathcal{X} \to \Omega$ называется *оценкой* параметра θ. Обычно их записывают с «крышечкой», как $\widehat{\theta}(\mathbf{X})$.

Периодически нам нужно будет оценивать не сам параметр, а некую функцию от него $\tau(\theta)$. Соответственно, оценка для этой функции должна лежать в множестве $\tau(\Theta)$. Например, пусть имеются наблюдения с распределением $\mathcal{N}(\theta_1, 1)$, а также наблюдения с распределением $\mathcal{N}(\theta_2, 1)$, и мы хотим понять, различаются ли они (то есть правда ли, что $\theta_1 = \theta_2$), в таком случае логично оценивать не сами параметры, а функцию $\tau(\theta) = \theta_1 - \theta_2$.

1.2 Основные свойства

Чтобы понимать, какие оценки хорошие, а какие — не очень, нужно выделить некоторые свойства оценок, которые было бы крайне желательно иметь. Многие из них звучат как «для всех $\theta \in \Theta$ выполнено...»: действительно, было бы обидно для одних параметров иметь

свойство, а для других — нет. Поначалу мы будем честно прописывать это, но далее всегда считаем, что рассматриваемое свойство выполнено для любого значения параметра.

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

Пример 1.2. Многие естественные оценки этим свойством обладают в силу линейности матожидания. Например, если $\mathcal{P} = \{\mathrm{Pois}(\lambda) \colon \lambda > 0\}$, то $\overline{\mathbf{X}}$ будет несмещённой оценкой λ, так как для любого i верно $\mathsf{E}_\lambda X_i = \lambda$. Впрочем, даже относительно простые оценки могут оказаться смещёнными, например, выборочная дисперсия s^2:

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

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

$$= \frac{n-1}{n} \cdot \mathsf{D}_\theta X_1 = \frac{n-1}{n} \cdot \sigma^2.$$

При больших n это отклонение будет невелико, однако если выборка мала, множитель $\frac{n-1}{n}$ может существенно поменять оценку, поэтому часто вместо обычной выборочной дисперсии берут её несмещённый аналог $S^2 = \frac{n}{n-1} s^2$ (обозначая её большой буквой, мы как бы указываем на её превосходство над обычной выборочной дисперсией). ■

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 с ростом размера выборки увеличилась бы и точность в предсказании параметра.

Определение. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка. Оценка $\theta_n^* = \theta_n^*(\mathbf{X})$ 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 оценка зависит от параметра n, то нижний индекс убирают.

Немаловажную роль играет и то, с какой скоростью оценка стремится к истинному значению параметра. Сходимость, конечно, хорошее свойство, но оно не говорит нам о том, как близко к параметру мы находимся и сколько нужно элементов выборки для достаточно точного приближения.

Определение. Оценка θ_n^* называется *асимптотически нормальной* оценкой пара-

метра $\tau(\theta)$, если для любого $\theta \in \Theta$

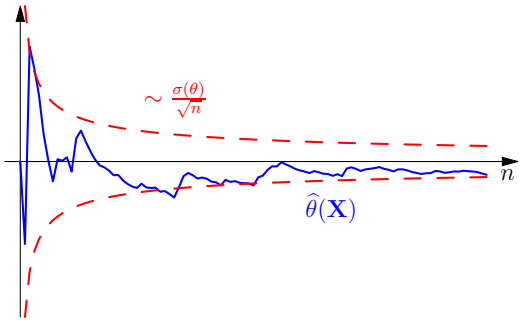
$$\sqrt{n}(\theta_n^* - \tau(\theta)) \xrightarrow{d_\theta} \mathcal{N}(0, \sigma^2(\theta)). \tag{1}$$

Величина $\sigma^2(\theta)$ называется *асимптотической дисперсией* (в многомерном варианте она представляет собой ковариационную матрицу).

Такая форма сходимости выбрана неслучайно: многие оценки будут асимптотическими нормальными по ЦПТ. Например, если $X_1, \ldots, X_n$ имеют конечные дисперсии, то оценка $\overline{\mathbf{X}}$ функции $\mathsf{E}_\theta X_1$ будет асимптотически нормальна, так как

$$\sqrt{n}(\overline{\mathbf{X}} - \mathsf{E}_\theta X_1) = \frac{X_1 + \ldots + X_n - n\mathsf{E}_\theta X_1}{\sqrt{n}} \xrightarrow{d_\theta} \mathcal{N}(0, \mathsf{D}_\theta X_1).$$

Также важен предельный закон в виде нормального распределения. Как известно, оно имеет очень «лёгкие» хвосты, и поэтому основная вероятностная масса сосредоточена вблизи матожидания. Такой результат ещё называют «правилом трёх сигм»: с вероятностью > 0.997 величина $\xi \sim \mathcal{N}(0, \sigma^2)$ принимает значения из интервала $(-3\sigma; 3\sigma)$, поэтому можно считать, что асимптотически нормальная оценка стремится к истинному значению параметра со скоростью порядка $\sigma(\theta)/\sqrt{n}$. Это также поясняет важность понятия асимптотической дисперсии: чем она меньше, тем точнее будет результат приближения, на таком соображении основан принцип сравнения оценок, описанный в разделе 3.1.



Конечно, оценки могут стремиться к параметру и с большей скоростью (см. пример 1.3), и тогда величина в левой части (1) стремится к нулевому распределению. В некоторых источниках такие оценки не считают асимптотически нормальными, так как их интересует нетривиальный предельный закон. Однако мы так делать не будем, ведь нам важна оценка на скорость сходимости, тем более мы считаем константу нормально распределённой с дисперсией нуль.

Утверждение 1.1. *Состоятельность оценки $\widehat{\theta}(\mathbf{X})$ следует из её сильной состоятельности или асимптотической нормальности.*

Доказательство. Первый факт очевиден, так как из сходимости почти всюду следует сходимость по вероятности. Докажем второе утверждение.

С одной стороны, по условию

$$\sqrt{n}\left(\widehat{\theta}(\mathbf{X}) - \theta\right) \xrightarrow{d_\theta} \mathcal{N}(0, \sigma^2(\theta)).$$

С другой, очевидно выполняется $1/\sqrt{n} \xrightarrow{d_\theta} 0$. Тогда по лемме Слуцкого $\widehat{\theta}(\mathbf{X}) - \theta \xrightarrow{d_\theta} 0$. Из сходимости по распределению к константе следует сходимость к ней по вероятности, значит, $\widehat{\theta}(\mathbf{X}) \xrightarrow{\mathsf{P}_\theta} \theta$. □

Пример 1.3. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из $U(0, \theta)$. Проверим на несмещённость, состоятельность, сильную состоятельность и асимптотическую нормальность следующие оценки параметра θ: $\widehat{\theta}(\mathbf{X}) = \overline{\mathbf{X}} + X_{(n)}/2$, $\theta^*(\mathbf{X}) = (n+1)X_{(1)}$ и $\widetilde{\theta}(\mathbf{X}) = X_{(1)} + X_{(n)}$.

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:

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

$$= n\theta \int_0^1 s(1-s)^{n-1}\, ds = n\theta \cdot B(2,n) = n\theta \cdot \frac{\Gamma(2)\Gamma(n)}{\Gamma(n+2)} = \frac{\theta}{n+1}.$$

Теперь мы готовы к решению задачи.

Несмещённость. Из линейности матожидания легко видеть, что несмещёнными будут $\theta^*(\mathbf{X})$ и $\widetilde{\theta}(\mathbf{X})$, а вот $\mathsf{E}_\theta\widehat{\theta}(\mathbf{X}) = \theta/2 + \frac{n\theta}{2(n+1)} \neq \theta$, поэтому $\widehat{\theta}(\mathbf{X})$ имеет небольшое, но всё-таки смещение.

Состоятельность. $\overline{\mathbf{X}} \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \theta/2$ 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

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

С другой стороны,

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

Таким образом, $X_{(n)} \xrightarrow{\mathsf{P}_\theta} \theta$, $X_{(1)} \xrightarrow{\mathsf{P}_\theta} 0$, но $(n+1)X_{(1)} \not\xrightarrow{\mathsf{P}_\theta} \theta$. Из всего этого получаем, что $\widehat{\theta}(\mathbf{X})$ будет состоятельной (сходимости по вероятности можно складывать), $\theta^*(\mathbf{X})$ не является состоятельной оценкой θ, а вот $\widetilde{\theta}(\mathbf{X})$ 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 $\widehat{\theta}(\mathbf{X})$ и $\widetilde{\theta}(\mathbf{X})$ будут сильно состоятельными. $\theta^*(\mathbf{X})$ же таковой не является, так как она даже не состоятельна.

Асимптотическая нормальность. Аналогично предыдущему пункту и по утверждению 1.1 оценка $\theta^*(\mathbf{X})$ 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 считаем нуль нормально распределённым), и по лемме Слуцкого $\sqrt{n}(\overline{\mathbf{X}}+X_{(n)}/2-\theta)\xrightarrow{d_\theta}\mathcal{N}(0,\mathsf{D}_\theta\overline{\mathbf{X}})$, $\sqrt{n}(X_{(1)}+X_{(n)}-\theta)\xrightarrow{d_\theta}0$. Таким образом, эти оценки будут ещё и асимптотически нормальными.

Как мы видим, скорость сходимости оценки $\widetilde{\theta}(\mathbf{X})$ ещё больше, чем $\sqrt{n}$, попробуем уточнить её и подберём δ так, чтобы распределение $n^\delta(\widetilde{\theta}(\mathbf{X})-X_{(n)})$ было чем-то нетривиальным.

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

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

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

что есть функция распределения для $\mathrm{Exp}\,(1/\theta)$. ■

1.3 Дельта-метод

Как мы могли видеть ранее, одних (У)ЗБЧ и ЦПТ (в том числе и их многомерных аналогов) уже достаточно для доказательства свойств простых оценок. К тому же сходимости почти наверное и по вероятности можно складывать, умножать или брать от них непрерывные функции. Но для доказательства асимптотической нормальности, и уж тем более нахождения асимптотической дисперсии этого мало, поэтому представляется крайне полезной следующая

Теорема 1.1 (дельта-метод).

Пусть $\boldsymbol{\xi}_n\xrightarrow{d}\boldsymbol{\xi}$, где $\boldsymbol{\xi},\boldsymbol{\xi}_n$ — m-мерные случайные векторы, $F(x)\colon\mathbb{R}^m\to\mathbb{R}^k$ — функция, дифференцируемая в точке $\mathbf{a}$, и $b_n\to 0$, $b_n\neq 0$. Тогда

$$\frac{F(\mathbf{a}+b_n\boldsymbol{\xi}_n)-F(\mathbf{a})}{b_n}\xrightarrow{d}d_{\mathbf{a}}F(\boldsymbol{\xi}).$$

В частности, для $m=k=1$ эта сходимость имеет вид

$$\frac{F(a+b_n\xi_n)-F(a)}{b_n}\xrightarrow{d}F'(a)\xi.$$

Доказательство. По определению дифференцируемости

$$F(\mathbf{a}+b_n\boldsymbol{\xi}_n) = F(\mathbf{a}) + d_{\mathbf{a}}F(b_n\boldsymbol{\xi}_n) + o(\|b_n\boldsymbol{\xi}_n\|),$$
$$\frac{F(\mathbf{a}+b_n\boldsymbol{\xi}_n) - F(\mathbf{a})}{b_n} = d_{\mathbf{a}}F(\boldsymbol{\xi}_n) + \boldsymbol{\alpha}(b_n\boldsymbol{\xi}_n)\|\boldsymbol{\xi}_n\|,$$

где $\|\boldsymbol{\alpha}(\mathbf{h})\| \to 0$ при $\|\mathbf{h}\| \to 0$, поэтому можно считать, что $\boldsymbol{\alpha}(\mathbf{h})$ непрерывна в нуле. Так как по лемме Слуцкого $b_n\boldsymbol{\xi}_n \xrightarrow{d} \mathbf{0}$, то по теореме о наследовании сходимости $\boldsymbol{\alpha}(b_n\boldsymbol{\xi}_n) \xrightarrow{d} \mathbf{0}$. Дважды применяя лемму Слуцкого, получаем сначала $\boldsymbol{\alpha}(b_n\boldsymbol{\xi}_n)\|\boldsymbol{\xi}_n\| \xrightarrow{d} \mathbf{0}$, а потом $d_{\mathbf{a}}F(\boldsymbol{\xi}_n) + \boldsymbol{\alpha}(b_n\boldsymbol{\xi}_n)\|\boldsymbol{\xi}_n\| \xrightarrow{d} d_{\mathbf{a}}F(\boldsymbol{\xi})$, что приводит нас к требуемой сходимости. □

Из этой теоремы следует замечательное утверждение, в англоязычной литературе именно его называют дельта-методом.

Следствие (о наследовании асимптотической нормальности)**.** *Пусть $\Theta \subset \mathbb{R}^m$, $\widehat{\boldsymbol{\theta}}$ — асимптотически нормальная оценка параметра $\boldsymbol{\theta}$ с асимптотической ковариационной матрицей $\Sigma(\boldsymbol{\theta})$, а $\tau\colon \mathbb{R}^m \to \mathbb{R}^k$ дифференцируема на Θ. Тогда $\tau\left(\widehat{\boldsymbol{\theta}}\right)$ асимптотически нормальная оценка $\tau(\boldsymbol{\theta})$ с асимптотической ковариационной матрицей $d_{\boldsymbol{\theta}}\tau \cdot \Sigma(\boldsymbol{\theta}) \cdot d_{\boldsymbol{\theta}}\tau^T$, то есть*

$$\sqrt{n}\left(\tau\left(\widehat{\boldsymbol{\theta}}\right) - \tau(\boldsymbol{\theta})\right) \xrightarrow{d_{\boldsymbol{\theta}}} \mathcal{N}(\mathbf{0}, d_{\boldsymbol{\theta}}\tau \cdot \Sigma(\boldsymbol{\theta}) \cdot d_{\boldsymbol{\theta}}\tau^T).$$

В частности, для $m = k = 1$, если исходная оценка имела асимптотическую дисперсию $\sigma^2(\theta)$, эта сходимость имеет вид

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

Этот результат можно понимать так: к $\widehat{\boldsymbol{\theta}}$, которая является «примерно» нормально распределённой при больших n, применяется преобразование, которое в малой окрестности $\boldsymbol{\theta}$ «почти что» линейное с матрицей $d_{\boldsymbol{\theta}}\tau$.

Доказательство. Зафиксируем произвольное $\boldsymbol{\theta} \in \Theta$. По условию имеется сходимость

$$\boldsymbol{\xi}_n = \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d_{\boldsymbol{\theta}}} \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \Sigma(\boldsymbol{\theta})).$$

Положим $b_n = 1/\sqrt{n}$, $\mathbf{a} = \boldsymbol{\theta}$. Тогда по теореме 1.1

$$\sqrt{n}\left(\tau\left(\widehat{\boldsymbol{\theta}}\right) - \tau(\boldsymbol{\theta})\right) \xrightarrow{d_{\boldsymbol{\theta}}} d_{\boldsymbol{\theta}}\tau(\boldsymbol{\xi}).$$

Так как $d_{\boldsymbol{\theta}}\tau$ — линейное преобразование, то $d_{\boldsymbol{\theta}}\tau(\boldsymbol{\xi})$ также является гауссовским вектором, среднее и ковариационная матрица которого равна $\mathbf{0}$ и $d_{\boldsymbol{\theta}}\tau \cdot \Sigma(\boldsymbol{\theta}) \cdot d_{\boldsymbol{\theta}}\tau^T$ соответственно. □

Пример 1.4. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из экспоненциального распределения с параметром θ, т.е. $p_\theta(t) = \theta e^{-\theta t} I(t > 0)$. Рассмотрим статистику $\left(k!/\overline{\mathbf{X}^k}\right)^{1/k}$, где k — некоторое фиксированное натуральное число, и докажем, что она является асимптотически нормальной оценкой θ.

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

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

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

сначала посчитав 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,

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

■

Пример 1.5. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из некоторого распределения с параметром $\theta = \sigma^2 = \mathsf{D}_\theta X_1$. Рассмотрим выборочную дисперсию $s^2 = \overline{\mathbf{X}^2} - \overline{\mathbf{X}}^2$. Несложно показать, что она является сильно состоятельной оценкой σ^2: действительно, по УЗБЧ $\overline{\mathbf{X}} \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \mathsf{E}_\theta X_1$, а $\overline{\mathbf{X}^2} \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \mathsf{E}_\theta X_1^2$. Значит, по теореме о наследовании сходимости почти наверное: $s^2 \xrightarrow{\mathsf{P}_\theta\text{-п.н.}} \mathsf{E}_\theta X_1^2 - (\mathsf{E}_\theta X_1)^2 = \mathsf{D}_\theta X_1 = \sigma^2$. Докажем асимптотическую нормальность в случае $\mathsf{E}_\theta X_1^4 < \infty$.

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

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

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

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

что и требовалось. ■

Пример 1.6. Рассмотрим ещё пару примеров статистических функционалов, которые носят гордое название *коэффициент асимметрии и эксцесса*:

$$\gamma_3 = \frac{\mathsf{E}(\xi - \mathsf{E}\xi)^3}{[\mathsf{D}\xi]^{3/2}}, \quad \gamma_4 = \frac{\mathsf{E}(\xi - \mathsf{E}\xi)^4}{[\mathsf{D}\xi]^2} - 3.$$

Как можно догадаться, первый из них показывает степень симметричности распределения величины ξ относительно своего матожидания. Второй же в некоторой степени отражает тяжесть хвостов распределения. Методом подстановки (см. раздел 2.1) несложно получить их выборочные аналоги:

$$\alpha_3 = \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{\mathbf{X}}\right)^3}{s^3(\mathbf{X})}, \quad \alpha_4 = \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{\mathbf{X}}\right)^4}{s^4(\mathbf{X})} - 3.$$

Данные коэффициенты используются для проверки нормальности: для гауссовской величины ξ они оба равны 0 (отсюда и вычитаемая тройка в γ_4), поэтому если для эмпирической функции распределения сии коэффициенты оказались близки к нулю, то допустимо сделать предположение о нормальности данных (более подробно см. в главе 10). Отсюда появляется потребность в рассмотрении их оценок и оценки их дисперсий. Докажем, что для нормальной модели α_3 является асимптотически нормальной оценкой γ_3 и найдём её асимптотическую дисперсию (аналогичная процедура для α_4 предлагается читателю в качестве упражнения, см. задачу 1.9).

Несложно убедиться, что значение α_3 не меняется при сдвиге и масштабировании выборки, поэтому будем считать, что $X_i \sim \mathcal{N}(0, 1)$. В таком случае удобнее считать

моменты:

$$\mathsf{E}X_1 = \mathsf{E}X_1^3 = \mathsf{E}X_1^5 = 0, \quad \mathsf{E}X_1^2 = 1, \quad \mathsf{E}X_1^4 = 3, \quad \mathsf{E}X_1^6 = 15.$$

По многомерной ЦПТ можно убедиться в асимптотической нормальности вектора, составленного из первых трёх выборочных моментов:

$$\sqrt{n}\left(\begin{pmatrix}\overline{\mathbf{X}}\\ \overline{\mathbf{X}^2}\\ \overline{\mathbf{X}^3}\end{pmatrix} - \begin{pmatrix}\mathsf{E}X_1\\ \mathsf{E}X_1^2\\ \mathsf{E}X_1^3\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}(0,\Sigma), \quad \text{где } \Sigma = \mathsf{D}\begin{pmatrix}X_1\\ X_1^2\\ X_1^3\end{pmatrix}. \tag{2}$$

Зная все моменты, посчитать элементы матрицы Σ не составляет особого труда:

$$\Sigma_{11} = \mathsf{D}X_1 = 1, \quad \Sigma_{22} = \mathsf{D}X_1^2 = \mathsf{E}X_1^4 - (\mathsf{E}X_1^2)^2 = 2, \quad \Sigma_{33} = \mathsf{D}X_1^3 = \mathsf{E}X_1^6 - (\mathsf{E}X_1^3)^2 = 15,$$
$$\Sigma_{12} = \Sigma_{21} = \mathsf{cov}(X_1, X_1^2) = \mathsf{E}X_1^3 - \mathsf{E}X_1 \cdot \mathsf{E}X_1^2 = 0,$$
$$\Sigma_{23} = \Sigma_{32} = \mathsf{cov}(X_1^2, X_1^3) = \mathsf{E}X_1^5 - \mathsf{E}X_1^2 \cdot \mathsf{E}X_1^3 = 0,$$
$$\Sigma_{13} = \Sigma_{31} = \mathsf{cov}(X_1, X_1^3) = \mathsf{E}X_1^4 - \mathsf{E}X_1 \cdot \mathsf{E}X_1^3 = 3.$$

Преобразовав формулу для α_3, выразим его через выборочные моменты:

$$\alpha_3 = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i^3 - 3X_i^2 \cdot \overline{\mathbf{X}} + 3X_i \cdot \overline{\mathbf{X}}^2 - \overline{\mathbf{X}}^3\right)}{\left(\overline{\mathbf{X}^2} - \overline{\mathbf{X}}^2\right)^{3/2}} = \frac{\overline{\mathbf{X}^3} - 3\overline{\mathbf{X}^2} \cdot \overline{\mathbf{X}} + 2\overline{\mathbf{X}}^3}{\left(\overline{\mathbf{X}^2} - \overline{\mathbf{X}}^2\right)^{3/2}}.$$

Таким образом,

$$\alpha_3 = \tau\left(\overline{\mathbf{X}}, \overline{\mathbf{X}^2}, \overline{\mathbf{X}^3}\right), \quad \text{где } \tau(x,y,z) = \frac{z - 3yx + 2x^3}{(y - x^2)^{3/2}},$$

и нам остаётся применить дельта-метод к сходимости (2) и функции τ. Для этого найдём градиент τ в точке $a = (\mathsf{E}X_1, \mathsf{E}X_1^2, \mathsf{E}X_1^3) = (0,1,0)$:

$$\frac{\partial\tau}{\partial x}(a) = \frac{\partial}{\partial x}\left.\frac{2x^3 - 3x}{(1-x^2)^{3/2}}\right|_a = -3, \quad \frac{\partial\tau}{\partial y}(a) = \frac{\partial}{\partial y}\left.\frac{0}{y^{3/2}}\right|_a = 0, \quad \frac{\partial\tau}{\partial z}(a) = \frac{\partial}{\partial z}\left.\frac{z}{(1-0)^{3/2}}\right|_a = 1.$$

Итого, получаем, что

$$\sqrt{n}\alpha_3 = \sqrt{n}\left(\tau\left(\overline{\mathbf{X}}, \overline{\mathbf{X}^2}, \overline{\mathbf{X}^3}\right) - \tau(a)\right) \xrightarrow{d} \mathcal{N}(0,\sigma^2), \quad \text{где } \sigma^2 = \nabla\tau^T \cdot \Sigma \cdot \nabla\tau =$$
$$= \begin{pmatrix}-3 & 0 & 1\end{pmatrix} \cdot \begin{pmatrix}1 & 0 & 3\\ 0 & 2 & 0\\ 3 & 0 & 15\end{pmatrix} \cdot \begin{pmatrix}-3\\ 0\\ 1\end{pmatrix} = 6.$$

■

Задачи

Задача 1.1. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из $\mathrm{Exp}(\theta)$. Для какой функции от параметра θ оценка $e^{-\overline{\mathbf{X}^2}}$ будет состоятельной? Является ли она при этом несмещённой? А асимптотически нормальной?

Задача 1.2. Найдите константу C такую, что статистика $C\sum_{i=1}^{n}|X_{2i-1} - X_{2i}|$ является несмещённой оценкой параметра σ, где $\mathbf{X} = (X_1, \ldots, X_{2n})$ — выборка из $\mathcal{N}(a, \sigma^2)$. Покажите, что она также будет асимптотически нормальной, и найдите её асимптотическую дисперсию.

Задача 1.3. Пусть выборка $\mathbf{X} = (X_1, \ldots, X_n)$ пришла из распределения со средним μ, дисперсией σ^2 и конечным четвёртым моментом. При каком условии предельное распределение

случайного вектора

$$\sqrt{n}\left(\begin{pmatrix}\overline{\mathbf{X}}\\ s^2(\mathbf{X})\end{pmatrix} - \begin{pmatrix}\mu\\ \sigma^2\end{pmatrix}\right)$$

имеет независимые компоненты?

Задача 1.4. Постройте несмещённую асимптотически нормальную оценку для параметра $e^{-\theta}$ с помощью выборки $\mathbf{X} = (X_1, \ldots, X_n)$ из распределения $\mathrm{Pois}(\theta)$, где $\theta > 0$, и найдите её асимптотическую дисперсию.

Задача 1.5. Оценка $\widehat{\theta}_n(\mathbf{X})$ параметра θ называется *асимптотически несмещённой*, если $\forall\theta\colon \mathsf{E}_\theta\widehat{\theta}_n \to \theta$. Докажите, что асимптотически несмещённая оценка со стремящейся к нулю дисперсией является состоятельной.

Задача 1.6. Предложите какую-нибудь модель и состоятельную оценку параметра в ней, которая не является асимптотически нормальной.

Замечание. Считаем константу нормально распределённой с дисперсией 0.

Задача 1.7. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из распределения со средним $\mu \neq 0$ и конечной дисперсией. Известно, что плотность распределения X_1 больше некоторого $c > 0$ в окрестности нуля. Покажите, что $\overline{\mathbf{X}}^{-1}$ является асимптотически нормальной оценкой $1/\mu$, но при этом $\mathsf{E}_\mu|\overline{\mathbf{X}}^{-1}| = \infty$.

Задача 1.8. Пусть $G(n, p)$ — случайный граф в модели Эрдеша-Реньи (каждое ребро берётся независимо от других с вероятностью p). Найдите какую-нибудь сильно состоятельную оценку параметра p как функцию от числа треугольников в графе.

Задача 1.9. Докажите, что выборочный коэффициент эксцесса α_4 (см. пример 1.6) для выборки из нормального распределения является асимптотически нормальной оценкой нуля, и найдите его асимптотическую дисперсию.

2 Методы нахождения оценок

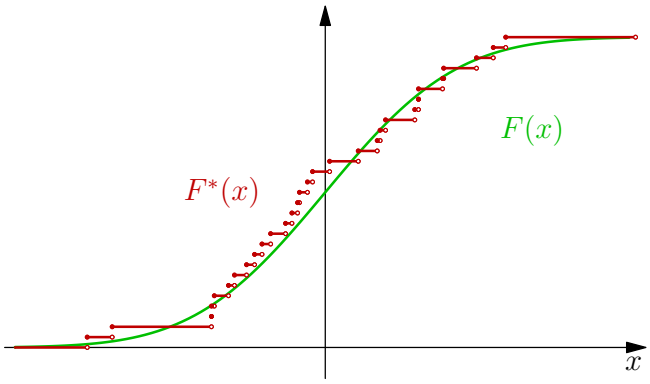
2.1 Метод подстановки и эмпирическая функция распределения

Идеологически большинство методов нахождения оценок устроены схожим образом и являются частными случаями *метода подстановки* (англ. plug-in estimator). Вот его основная идея. Часто интересующий нас параметр можно выразить как функционал от неизвестного распределения. Это может быть матожидание, медиана, мода и т. д. Причём обычно такие функционалы являются *достаточно хорошими* в плане непрерывности — для близких распределений значения функционала тоже будут близки. Отсюда появляется логичное желание подменить истинное распределение в функционале тем, что мы считаем хорошим приближением этого распределения.

В качестве такого приближения можно взять так называемую *эмпирическую функцию распределения*. Пусть $\mathbf{X} = (X_1, \dots, X_n)$ — выборка из независимых случайных величин из неизвестного распределения P с функцией распределения $F(x)$. Тогда эмпирическая функция распределения, построенная по сей выборке, имеет вид

$$F_n^*(\omega, x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i(\omega) \leqslant x)$$

Таким образом, эмпирическая функция есть случайный элемент, который по элементарному исходу выдаёт функцию, являющуюся, как несложно заметить, функцией распределения. Она представляет собой непрерывную справа кусочно-постоянную функцию со скачками в точках из выборки, а соответствующая вероятностная мера есть мера Дирака, расположенная в точках выборки, причём мера точки пропорциональна числу «выпаданий» этого значения.



Если нам нужно оценить функцию от параметра $\tau(\theta)$, которую можно записать как значение функционала G от истинного распределения P_θ, то оценкой по методу постановки будет случайная величина $G(\mathsf{P}_n^*)$, где P_n^* – эмпирическое распределение по выборке. Например, выборочную характеристику функции g можно представить в виде функционала G от эмпирического распределения, равного матожиданию функции $g(x)$, отчего её выбор в качестве оценки $\mathsf{E}_\theta g(X_i) = G(\mathsf{P}_\theta)$ становится ещё более оправданным:

$$\frac{1}{n} \sum_{i=1}^{n} g(X_i) = \sum_{x \in \{X_1, \dots, X_n\}} g(x) \cdot \frac{\#\ x \text{ в выборке}}{n} =$$

$$= \sum_{x \in \{X_1, \dots, X_n\}} g(x) \cdot \mathsf{P}_n^*(\{x\}) = \int_{\mathbb{R}} g(x)\, \mathsf{P}_n^*(dx) = G(\mathsf{P}_n^*).$$

Выбор именно такой аппроксимации истинной функции распределения объясняется её поточечной сходимостью к оной. Действительно, так как случайные величины $I(X_i \leqslant x)$ независимы, распределены одинаково и имеют конечный момент, то по УЗБЧ

$$F_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leqslant x) \xrightarrow{\text{п.н.}} \mathsf{E} I(X_1 \leqslant x) = \mathsf{P}(X_1 \leqslant x) = F(x).$$

Однако верен даже более сильный результат: сходимость почти наверное будет равномерной.

Теорема 2.1 (Гливенко, Кантелли).

Пусть $F_n^*(x)$ — эмпирическая функция распределения, построенная по выборке $\mathbf{X} = (X_1, \dots, X_n)$ неограниченного размера из независимых случайных величин с функцией распределения $F(x)$. Тогда

$$D_n(\omega) := \sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)| \xrightarrow{\text{п.н.}} 0, \quad n \to \infty.$$

2.2 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 $\mathsf{E}_\theta g_i(X_1) = m_i(\theta)$. Предположим, что у уравнения $m(\theta) = \overline{g(\mathbf{X})}$ 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 $\overline{g_i(\mathbf{X})}$ примерно равно $\mathsf{E} g_i(X_1)$, то логично положить решение уравнения выше за оценку параметра. Эту же оценку можно получить как оценку методом подстановки для функционала

$$G(\mathsf{P}) = m^{-1}\left(\int g_1(x)\,\mathsf{P}(dx), \dots, \int g_k(x)\,\mathsf{P}(dx)\right)$$

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

Пример 2.1. Рассмотрим некоторые примеры нахождения оценок по методу моментов.

- $X_i \sim \text{Pois}(\lambda)$. Так как $m(\lambda) = \mathsf{E}_\lambda X_1 = \lambda$ – тождественная, то в качестве оценки параметра можно взять $\overline{\mathbf{X}}$.
- $X_i \sim \text{Geom}(p)$. В данном случае $m(p) = \mathsf{E}_p X_1 = \frac{1-p}{p} = \frac{1}{p} - 1$, тогда $m^{-1}(t) = \frac{1}{t+1}$. Таким образом, оценка по методу моментов $\widehat{p} = \frac{1}{\overline{\mathbf{X}}+1}$.
- $X_i \sim \text{Beta}(\alpha, \beta)$. Тут уже надо оценивать двумерный параметр $\theta = (\alpha, \beta)$, поэтому найдём первый и второй моменты:

$$m_1(\alpha,\beta) = \mathsf{E}_\theta X_1 = \frac{\alpha}{\alpha+\beta} = x, \quad m_2(\alpha,\beta) = \mathsf{E}_\theta X_1^2 = \mathsf{D}_\theta X_1 + (\mathsf{E}_\theta X_1)^2 =$$

$$= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha^3+\alpha^2\beta+\alpha^2+\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} = y.$$

Тогда

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

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

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

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

- $X_i \sim \mathrm{U}(a, b)$. Тогда

$$m_1(a,b) = \mathsf{E}_\theta X_1 = \frac{a+b}{2} = x,$$

$$m_2(a,b) = \mathsf{E}_\theta X_1^2 = \mathsf{D}_\theta X_1 + (\mathsf{E}_\theta X_1)^2 = \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4} = \frac{a^2+b^2+ab}{3} = y.$$

Откуда $a + b = 2x$, $ab = a^2 + 2ab + b^2 - 3y = 4x^2 - 3y$, что есть коэффициенты квадратного уравнения 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, получаем следующую оценку по методу моментов:

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

■

При всей своей простоте у оценки по методу моментов имеются замечательные свойства.

Теорема 2.2 (сильная состоятельность оценки по методу моментов).

Пусть $m\colon \Theta \to m(\Theta)$ — биекция, и функция m^{-1} определена и непрерывна в каждой точке множества $m(\Theta)$. Также, $\mathsf{E}g_i(X_1) < \infty \ \forall i \in \{1, \ldots, k\}$, $\forall \theta \in \Theta$. Тогда оценка по методу моментов является сильно состоятельной оценкой параметра θ.

Доказательство. Следует из сохранения сходимости почти наверное под действием непрерывной функцией. □

Теорема 2.3 (асимптотическая нормальность оценки по методу моментов).

Если в условиях предыдущей теоремы функция m^{-1} дифференцируема на $m(\Theta)$, и $\mathsf{E}g_i(X_1)^2 < \infty \ \forall i \in \{1, \ldots, k\}$, $\forall \theta \in \Theta$, то оценка, полученная по методу моментов, асимптотически нормальна.

Доказательство. Так как вторые моменты $g_i(X_1)$ конечны, то по многомерной ЦПТ

$$\sqrt{n}\left(\begin{pmatrix}\overline{g_1(\mathbf{X})}\\ \ldots \\ \overline{g_k(\mathbf{X})}\end{pmatrix} - \begin{pmatrix}m_1(\theta)\\ \ldots \\ m_k(\theta)\end{pmatrix}\right) \xrightarrow{d_\theta} \mathcal{N}(0, \Sigma(\theta))$$

для некоторой ковариационной матрицы Σ. По следствию из теоремы 1.1 о сохранении асимптотической нормальности получаем

$$\sqrt{n}(\theta^*(\mathbf{X}) - \theta) = \sqrt{n}\left(m^{-1}\begin{pmatrix}\overline{g_1(\mathbf{X})}\\ \ldots \\ \overline{g_k(\mathbf{X})}\end{pmatrix} - m^{-1}\begin{pmatrix}m_1(\theta)\\ \ldots \\ m_k(\theta)\end{pmatrix}\right) \xrightarrow{d_\theta} \mathcal{N}\left(0, S \cdot \Sigma(\theta) \cdot S^T\right),$$

где $S = d_{m(\theta)}m^{-1}$. □

Для упрощения вычисления асимптотической дисперсии заметим, что если m есть диффеоморфизм Θ на $m(\Theta)$, то $S = d_{m(\theta)} m^{-1} = [d_\theta m]^{-1}$, то есть достаточно найти матрицу Якоби для функции m, а потом её обратить (или в одномерном случае просто поделить на квадрат производной).

Пример 2.2. Посчитаем для одной из оценок примера 2.1 асимптотическую дисперсию, а именно для выборки $X_i \sim \mathrm{Geom}(p)$. По ЦПТ мы знаем, что

$$\sqrt{n}\left(\overline{\mathbf{X}} - \frac{1-p}{p}\right) \xrightarrow{d_p} \mathcal{N}(0, \mathsf{D}_p X_1) = \mathcal{N}\left(0, \frac{1-p}{p^2}\right).$$

Мы получили, что $m(p) = \frac{1}{p} - 1$, поэтому $m'(p) = -\frac{1}{p^2}$. Из доказательства теоремы 2.3 следует, что

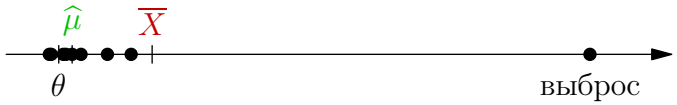
$$\sqrt{n}\left(\frac{1}{\overline{\mathbf{X}} + 1} - p\right) \xrightarrow{d_p} \mathcal{N}\left(0, \frac{1-p}{p^2} \Big/ \left(-\frac{1}{p^2}\right)^2\right) = \mathcal{N}\left(0, (1-p)p^2\right)$$

■

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. К тому же не факт, что они будут несмещёнными (хотя, как например в случае равномерного распределения выше, легко получить несмещённую оценку, используя несмещённую оценку дисперсии).

2.3 Выборочные квантили

Помимо недостатков, упомянутых ранее, у оценок по методу моментов (впрочем, как и у любых статистик, использующих выборочные характеристики) есть существенный минус — они крайне чувствительны к *выбросам*, то есть элементам выборки, которые сильно выбиваются из целевого распределения (они могут порождаться, например, ошибками измерения). Из-за них значение статистики может сильно поменяться, испортив результаты. В меньшей степени такой проблеме подвержены оценки, использующие порядковые статистики: даже с очень большим выбросом значение оценки поменяется незначительно (см. рис). Ещё говорят, что такие оценки *робастные*, то есть устойчивы к выбросам.



Самым простым примером оценки, построенной с помощью порядковых статистик, является выборочный квантиль. Для начала сформулируем определение обычного квантиля.

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

Таким образом, p-квантиль — некоторый функционал от распределения. Значит, его можно приблизить, взяв функционал от эмпирического распределения.

Определение. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — 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 2.4 (о выборочном квантиле).

Пусть $\mathbf{X} = (X_1, \ldots X_n)$ — выборка из распределения P с плотностью $\rho(x)$. Пусть z_p — p-квантиль распределения P, причем $\rho(x)$ непрерывна в точке 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 $p = 1/2$. Такой p-квантиль называют *медианой*, но выборочный аналог обычно определяют несколько иначе:

Определение. *Выборочной медианой* для выборки $\mathbf{X} = (X_1, \ldots, X_n)$ 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 $z_{n,1/2}$ заменяется на $\widehat{\mu}$.

Пример 2.3. Построим асимптотически нормальную оценку для параметра масштаба в модели распределения Коши:

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

Метод выборочного квантиля здесь оказывается весьма полезным и простым. Заметим, что 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, $\frac{3}{4}$-квантилем для данного семейства распределений будет в точности $z_{3/4} = \theta$. Тогда по теореме о выборочном квантиле $z_{n,3/4}$ будет асимптотически нормальной оценкой параметра θ:

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

Для сравнения 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

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

$$= \frac{\theta}{\pi(1-\theta^2)}\left(\frac{\pi}{\theta} - \pi\right) = \frac{1}{1+\theta} \Longrightarrow \widehat{\theta}(\mathbf{X}) = m^{-1}(\mathbf{X}) = 1\Big/\frac{1}{1+\mathbf{X}^2} - 1$$

Получилось весьма недурно. Но проверим, какова асимптотическая дисперсия полученной оценки:

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

Проведя несложные расчёты, получаем: $\mathsf{D}_\theta g(X_1) = \mathsf{E}_\theta g(X_1)^2 - (\mathsf{E}_\theta g(X_1))^2 = \frac{\theta+2}{2(\theta+1)^2} - \frac{1}{(\theta+1)^2} = \frac{\theta}{2(\theta+1)^2}$. Применяя дельта-метод, приходим к тому, 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). ■

Задачи

Задача 2.1. Найдите оценку вектора параметров (α, λ) по методу моментов для выборки $\mathbf{X} = (X_1, \ldots, X_n)$ из $\Gamma(\alpha, \lambda)$.

Задача 2.2. Рассмотрим модель сдвига для распределения Лапласа:

$$\rho_\theta(x) = \frac{1}{2} e^{-|x-\theta|}, \quad \theta \in \mathbb{R}.$$

Постройте оценки параметра θ по выборке $\mathbf{X} = (X_1, \ldots, X_n)$ с помощью метода моментов и метода квантилей. Сравните полученные оценки по их асимптотическим дисперсиям.

Задача 2.3. В этой задаче предлагается доказать теорему 2.4 об асимптотической нормальности оценки по методу квантилей.

(а) Пусть $\xi_1, \ldots, \xi_n \sim \mathrm{U}[0;1]$. Докажите, что

$$\xi_{(k)} \sim \frac{S_k}{S_{n+1}}, \quad S_m = \eta_1 + \ldots + \eta_m, \quad \eta_i \text{ — н.о.р. с распределением } \mathrm{Exp}(1).$$

(б) Докажите, что если $\alpha(n) - np = O(1)$, то

$$\sqrt{n}\left(\frac{S_{\alpha(n)}}{n} - p\right) \xrightarrow{d} \mathcal{N}(0, p);$$

(в) Докажите теорему 2.4 для выборки из равномерного распределения.

(г) Докажите теорему в общем случае. *Указание.* Вспомните из теорвера, как от произвольного распределения перейти к равномерному.

3 Сравнение оценок

3.1 Функция потерь и асимптотическая дисперсия

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*. 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^*.

Зачастую, в качестве такой функции потерь берут $g(x, y) = |x - y|$ или $g(x, y) = (x - y)^2$ – *квадратичную функцию потерь*. Последний вариант полезен и тем, что для несмещённых оценок функция риска представляет собой дисперсию (т. к. $\theta = \mathsf{E}_\theta\hat{\theta}$).

Ясно, что хорошая (в нашем понимании) оценка должна минимизировать функцию риска. Однако неясно, как сравнивать функции риска для разных оценок: для одних значений параметра одна оценка может иметь меньшую функцию риска, для других — уже другая. Самое простое решение этой проблемы — использовать *равномерный подход*. В нём мы умеем сравнивать только те функции риска, среди которых одна мажорирует другую.

Определение. 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*.

Пример 3.1. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из равномерного распределения на отрезке $U[0, \theta]$. Сравним следующие оценки параметра θ в среднеквадратичном подходе: $\widehat{\theta} = 2\overline{\mathbf{X}}$, $\theta^* = (n + 1)X_{(1)}$, $\widetilde{\theta} = \frac{n+1}{n}X_{(n)}$. Данные оценки являются несмещёнными (их матожидания были посчитаны ранее в примере 1.3), а значит, их функция риска есть дисперсия.

Для первой оценки получаем

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

Перед нахождением дисперсии второй оценки 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

$$R((n + 1)X_{(1)}, \theta) = \mathsf{D}_\theta(n + 1)X_{(1)} = \mathsf{E}_\theta(n + 1)^2X_{(1)}^2 - (\mathsf{E}_\theta(n + 1)X_{(1)})^2 =$$
$$= -\theta^2 + (n + 1)^2 \int_0^\theta t^2 \frac{n}{\theta}\left(1 - \frac{t}{\theta}\right)^{n-1} dt = -\theta^2 + n(n + 1)^2\theta^2 \int_0^1 s^2(1 - s)^{n-1} ds =$$
$$= -\theta^2 + n(n + 1)^2\theta^2 B(n, 3) = -\theta^2 + n(n + 1)^2\theta^2 \frac{2!(n - 1)!}{(n + 2)!} = \frac{\theta^2 n}{n + 2}.$$

Как видим, дисперсия не стремится к нулю при увеличении размера выборки, поэтому данная оценка весьма плохая в среднеквадратичном подходе. К тому же несложно проверить, что функция потерь оценки $\widehat{\theta}$ при любых значения θ будет не больше текущей.

Наконец, рассмотрим оценку $\widetilde{\theta}$. Тут распределение ищется попроще: 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,

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

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

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

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

Ещё один способ сравнения оценок заключается в рассмотрении их асимптотической дисперсии (то есть этот метод применим только для асимптотически нормальных оценок). Ранее мы поняли, что она характеризует скорость сходимости оценки к истинному значению параметра, поэтому логично брать те оценки, которые минимизируют эту функцию. Более точно,

Определение. Пусть $\widehat{\theta}$ и θ^* — асимптотически нормальные оценки функции $\tau(\theta)$, а $\sigma_1^2(\theta)$ и $\sigma_2^2(\theta)$ — асимптотические дисперсии $\widehat{\theta}$ и θ^* соответственно. Говорят, что оценка $\widehat{\theta}$ *лучше* оценки θ^*, если для любого $\theta \in \Theta$ выполнено неравенство $\sigma_1^2(\theta) \leqslant \sigma_2^2(\theta)$, причём для некоторого θ неравенство строгое.

Пример 3.2. Рассмотрим некоторые примеры оценок и сравним их в асимптотическом подходе.

- Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из распределения $\mathcal{N}(\theta, 1)$, где θ — неизвестный параметр сдвига. Мы уже знаем по крайней мере две оценки этого параметра. Первая, самая естественная, есть $\overline{\mathbf{X}}$, которая имеет асимптотическую дисперсию 1, так как
$$\sqrt{n}(\overline{\mathbf{X}} - \theta) \sim \mathcal{N}(0, 1).$$
Вторая же есть выборочная медиана $\widehat{\mu}(\mathbf{X})$, которая по теореме о выборочном квантиле является асимптотически нормальной с асимптотической дисперсией $1/4\rho'_\theta(0)^2 = \pi/2$. Таким образом, $\overline{\mathbf{X}}$ лучше $\widehat{\mu}$ в асимптотическом подходе.

- Иная ситуация обстоит с $X_1, \ldots, X_n \sim \text{Laplace}(\theta, 1)$. Хвосты этого распределения заметно тяжелее, чем у нормального закона, поэтому обычное среднее здесь менее эффективно. По ЦПТ асимптотическая дисперсия $\overline{\mathbf{X}}$ равна $\mathsf{D}_\theta X_1 = 2$, при этом в то же время по теореме о выборочном квантиле та же величина для $\widehat{\mu}$ равна $1/4\rho'_\theta(0)^2 = 1$.

- А вот оценки, полученные в примере 2.3, сравнить в данном подходе не получится:

для некоторых значений параметра одна оценка имеет меньшую асимптотическую дисперсию, а при других — вторая.

■

Помимо двух рассмотренных подходов существуют как минимум ещё два довольно популярных: *байесовский* и *минимаксный*. В отличие от равномерного и асимптотического подходов, которые умеют сравнивать не все оценки, эти два сводят функции риска к одной численной величине, что существенно расширяет наши возможности. Их определения и свойства приведены в главе 15.

Далее нас будет интересовать, какие оценки являются лучшими относительно тех метрик, которые мы ввели ранее. Про асимптотический подход мы поговорим позже в главе 4, а сейчас займёмся оптимизацией функции риска, а именно — функции риска для квадратичной функции потерь.

Нахождение наилучшей оценки среди всех возможных не имеет смысла, так как всегда можно взять оценку $\theta^* \equiv \theta_0$ для некоторого $\theta_0 \in \Theta$, и тогда функция риска $R(\theta^*, \theta_0)$ будет равна нулю. И если мы хотим найти оценку, которая будет не хуже любой другой, ей придётся «обогнать» и такую тривиальную оценку, тогда её функция риска будет тождественно нулевой, что, конечно, не представляется возможным. Поэтому обычно в среднеквадратичном подходе рассматривают несмещённые оценки: в отличие от тривиальных оценок выше они «что-то знают», к тому же для них рассматриваемая функция потерь равна дисперсии. Впрочем, смещённые оценки нельзя просто так сбрасывать со счетов, так как они бывают получше некоторых несмещённых.

Пример 3.3 (*парадокс Штейна*)**.** Рассмотрим модель, в которой элементы выборки представляют собой гауссовский вектор с независимыми компонентами, которые имеют единичную дисперсию и неизвестный параметр сдвига, свой для каждой компоненты, то есть они имеют распределение $\mathcal{N}(\boldsymbol{\theta}, E_k)$, где $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ — неизвестный вектор параметров, который необходимо оценить. Для простоты будем считать, что выборка состоит из одного единственного вектора $\mathbf{X} = (X_1, \ldots, X_k)$. Для $k = 1$ это будет обычная нормальная модель сдвига, и в ней имеется очевидная оценка $\widehat{\boldsymbol{\theta}} = \mathbf{X}$, которая, как выяснится позже, будет наилучшей в среднеквадратичном подходе.

Логично предположить, что такими же «хорошими» будут оценки $\widehat{\boldsymbol{\theta}} = \mathbf{X}$, то есть каждую θ_i мы оценим соответствующей компонентой вектора X_i. Однако это неправда в смысле среднеквадратичной функции риска $R(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathsf{E}_\theta \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$ при $k \geqslant 3$. Оказывается, имеется оценка равномерно лучше оценки X, а именно

$$\widehat{\boldsymbol{\theta}} = \mathbf{X} - \frac{k-2}{\|\mathbf{X}\|^2} \cdot \mathbf{X}.$$

Несмотря на, то что она имеет очевидное смещение, её функция риска всегда меньше функции риска, казалось бы, вполне естественной оценки $\mathbf{X}$. Читателю предлагается доказать этот факт.

Этот пример является типичной иллюстрацией соотношения между смещением и дисперсией оценки (*англ.* bias-variance tradeoff). Дело в том, что среднеквадратичную ошибку оценки $\widehat{\theta}$ параметра θ можно разложить на её дисперсию и квадрат смещения:

$$R(\widehat{\theta}, \theta) = \mathsf{E}_\theta(\widehat{\theta} - \theta)^2 = \mathsf{D}_\theta(\widehat{\theta} - \theta) + \Big(\mathsf{E}_\theta[\widehat{\theta} - \theta]\Big)^2 = \underbrace{\mathsf{D}_\theta\widehat{\theta}}_{\text{дисперсия}} + \Big(\underbrace{\mathsf{E}_\theta[\widehat{\theta} - \theta]}_{\text{смещение}}\Big)^2$$

Таким образом, среднеквадратичная ошибка зависит не только от того, насколько близко к истине предсказываемое значение параметра, но и от разброса этого значения, нашей уверенности в оценке. Зачастую добавление систематического сдвига помогает

уменьшить дисперсию оценки, что может улучшить рассматриваемую функцию риска. В примере выше значения, которые выдаёт оценка, становятся более кучными, что и уменьшает среднеквадратичную ошибку. Помимо данной модели можно вспомнить так называемую «гребневую» регрессию, которая, несмотря на смещение, позволяет строить более устойчивые оценки параметров (см. пример 15.1). ■

Возникает закономерный вопрос: насколько сильно можно уменьшить среднеквадратичное отклонение оценки? Как быстро может убывать дисперсия с ростом размера выборки? Чтобы ответить на эти вопросы, нам придётся ввести некоторые ограничения на нашу вероятностно-статистическую модель, которые называют *условиями регулярности*.

3.2 Условия регулярности

Перед тем, как их озвучить, нам понадобится обобщение понятия плотности: по аналогии с непрерывным случаем хотелось бы и для дискретных распределений считать матожидание как интеграл по какой-то одной конкретной мере. Только по какой?

Определение. *Считающей мерой* на $\mathbb{Z}^k$ называется функция $\mu\colon \mathcal{B}(\mathbb{R}^k) \to \mathbb{N}\cup\{+\infty\}$, определённая как

$$\mu(B) = \sum_{\mathbf{x}\in\mathbb{Z}^k} I(\mathbf{x} \in B).$$

Ясно, что это σ-конечная мера, и она равна числу целочисленных точек, которое попадает в данное множество, откуда собственно и пошло название. Несложно также понять, как считается интеграл произвольной измеримой функции f по этой мере:

$$\int_{\mathbb{R}^k} f(\mathbf{x})\,\mu(d\mathbf{x}) = \sum_{\mathbf{x}\in\mathbb{Z}^k} f(\mathbf{x}),$$

если, конечно, этот ряд сходится абсолютно. Такое представление позволяет записывать матожидание от дискретных случайных векторов $\xi\colon \Omega \to \mathbb{Z}^k$ через интеграл с плотностью по считающей мере:

$$\mathsf{E}g(\xi) = \sum_{\mathbf{x}\in\mathbb{Z}^k} g(\mathbf{x})\cdot \mathsf{P}(\xi = \mathbf{x}) = \int_{\mathbb{R}^k} g(\mathbf{x})\cdot\rho(\mathbf{x})\,\mu(d\mathbf{x}),$$

где $\rho(\mathbf{x}) = \mathsf{P}(\xi = \mathbf{x})$ — *плотность по мере* μ. Отныне под словами «плотность по мере μ» мы будем подразумевать либо обычную плотность, которая у нас была ранее, по классической мере Лебега, либо дискретную по считающей мере. Семейства распределений, у которых есть плотность по одной и той же мере, мы будем называть *доминируемым*.

Данные соображения можно распространить на случай произвольной σ-конечной меры μ с помощью теоремы Радона-Никодима.

Определение. Пусть μ — некоторая σ-конечная мера на $\mathbb{R}^n$. Семейство вероятностных мер $\mathcal{P}$ называется *доминируемым* относительно меры μ, если $\forall \mathsf{P} \in \mathcal{P}\colon \mathsf{P} \ll \mu$ (напомним, что $\nu \ll \mu$, если $\forall B \in \mathbb{R}^n\colon \mu(B) = 0 \Rightarrow \nu(B) = 0$). Производную Радона-Никодима $\frac{d\mathsf{P}}{d\mu}$ называют *обобщённой плотностью*.

Несложно показать, что выполнена *формула пересчёта*. Она позволяет перейти от интеграла по неизвестной мере к интегралу плотности по известной мере, которая и доминирует семейство:

$$\int_{\mathbb{R}^n} f(\mathbf{x})\,\mathsf{P}(d\mathbf{x}) = \int_{\mathbb{R}^n} f(\mathbf{x})\cdot\frac{d\mathsf{P}}{d\mu}\,\mu(d\mathbf{x}).$$

В различных источниках условия регулярности формулируют по-разному, мы остановимся на следующем варианте.

Условия регулярности

C1 Распределения P_θ имеют плотность $\rho_\theta(\mathbf{x})$ по некоторой мере μ, носитель которой (т.е. множество $\{x\colon \rho_\theta(\mathbf{x}) > 0\}$) не зависит от θ;

C2 Θ — открытое связное множество в $\mathbb{R}$;

C3 Для любого $\theta \in \Theta$ и для любой статистики $S(\mathbf{X})$ с локально ограниченным вторым моментом (то есть $\forall \theta_0\, \exists \varepsilon > 0, c > 0\, \forall \theta \in B_\varepsilon(\theta_0)\colon \mathsf{E}_\theta S(\mathbf{X})^2 < c$) выполнено

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

C4 Функция $I_\mathbf{X}(\theta) = \mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{X})\right)^2$ конечна и положительна.

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

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

Но если остальные условия проверяются относительно легко, выполнение условия C3 установить «в лоб» довольно проблематично. Удобным представляется следующее достаточное условие, которое выполнено для широкого класса моделей.

Теорема 3.1.

Пусть выполнены условия C1, C2 и C4, а также известно, что:

1. Для μ-п.в. $\mathbf{x}$ функция $\sqrt{\rho_\theta(\mathbf{x})}$ непрерывно дифференцируема;
2. Функция $I_\mathbf{X}(\theta)$ непрерывна по θ.

Тогда выполнено условие C3.

Остаётся лишь убедиться в локальной ограниченности второго момента взятой статистики, что обычно весьма просто: например, достаточным условием будет непрерывность второго момента по θ. Далее мы не будем тратить время на то, чтобы удостовериться в выполнении условий регулярности, но для любопытных читателей дан необходимый арсенал для непосредственной проверки.

3.3 Информация Фишера и эффективность

Сейчас же важно разобраться с условием C4, а конкретно с введённой там функцией. Она является одной из важнейших характеристик выборки, которая будет встречаться

нам ещё не раз.

Определение. Величина $I_{\mathbf{X}}(\theta)$ называется *информацией Фишера* выборки $\mathbf{X}$, а величина

$$u_\theta(\mathbf{X}) = \frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{X})$$

— *вкладом выборки*. Также удобно использовать информацию Фишера для выборки, состоящей из одного элемента, её обычно обозначают $i(\theta)$.

Далее мы постепенно поймём, что эта функция и вправду отражает наше интуитивное понимание термина «информация». Для начала отметим некоторые простые свойства этой величины.

Утверждение 3.1. *1. Для информации Фишера выполнено тождество*

$$\mathsf{E}_\theta\frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{X}) = 0.$$

2. Информация Фишера аддитивна: для любых независимых выборок $\mathbf{X}$ и $\mathbf{Y}$ верно $I_{(\mathbf{X},\mathbf{Y})}(\theta) = I_{\mathbf{X}}(\theta) + I_{\mathbf{Y}}(\theta)$. Как следствие, $I_{\mathbf{X}}(\theta) = ni(\theta)$, где $i(\theta)$ – информация, содержащаяся в одном элементе выборки.

Доказательство. 1. Это можно показать, воспользовавшись пунктом С3 из условий регулярности и взяв статистику $S(\mathbf{X}) \equiv 1$:

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

Из этого следует, что информацию Фишера можно также записать как дисперсию:

$$I_{\mathbf{X}}(\theta) = \mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{X})\right)^2 - 0 = \mathsf{D}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{X})\right).$$

2. Воспользуемся линейностью дисперсии для независимых случайных величин:

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

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

□

Перейдём к менее очевидным свойствам. Информация Фишера фигурирует в следующем неравенстве на дисперсию оценки.

Теорема 3.2 (неравенство Рао-Крамера).

Для любой несмещённой оценки θ^* функции $\tau(\theta)$ с локально ограниченным $\mathsf{E}_\theta\theta^{*2}$ и для любого $\theta\in\Theta$ справедливо неравенство

$$\mathsf{D}_\theta(\theta^*(\mathbf{X})) \geqslant \frac{(\tau'(\theta))^2}{I_{\mathbf{X}}(\theta)} = \frac{(\tau'(\theta))^2}{ni(\theta)}.$$

Доказательство. Из условия следует применимость условия С3 регулярности для ста-

тистики $\theta^*(\mathbf{X})$:

$$\tau'(\theta) = \frac{\partial}{\partial\theta}\mathsf{E}_\theta\theta^*(\mathbf{X}) \overset{\text{C3}}{=} \mathsf{E}_\theta\left(\theta^*(\mathbf{X})\frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{X})\right) \Subset .$$

По утверждению 3.1 верно $\mathsf{E}_\theta\frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{X}) = 0$. Прибавим эту величину, домноженную на $\tau(\theta)$, к правой части равенства сверху:

$$\Subset\mathsf{E}_\theta\left(\theta^*(\mathbf{X})\frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{X})\right) - \tau(\theta)\cdot\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{X})\right) = \mathsf{E}_\theta\left([\theta^*(\mathbf{X}) - \tau(\theta)]\frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{X})\right).$$

Это было сделано для того, чтобы применить неравенство Коши-Буняковского.

$$\tau'(\theta)^2 \leqslant \mathsf{E}_\theta\left(\theta^*(\mathbf{X}) - \tau(\theta)\right)^2\cdot\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\ln\rho_\theta(\mathbf{X})\right)^2 = \mathsf{D}_\theta\theta^*(\mathbf{X})\cdot I_{\mathbf{X}}(\theta),$$

что и требовалось. □

С одной стороны, неравенство Рао-Крамера приводит нас к пессимистичному выводу: в условиях регулярности наилучший порядок убывания дисперсии несмещённой оценки есть $1/n$. Но с другой стороны, оно даёт нам чёткий пример для подражания, на который нам стоит ориентироваться при составлении оценок. В некоторых случаях удаётся выжать максимум в этом направлении.

Определение. Оценка θ^* называется *эффективной*, если для неё выполнено равенство в неравенстве Рао-Крамера.

Появляется резонное желание понять, а когда наша оценка имеет наименьшую дисперсию, которую позволяет нам неравенство выше. Его удовлетворяет следующая

Теорема 3.3 (критерий эффективности).

Оценка θ^* эффективна тогда и только тогда, когда для любого $\theta\in\Theta$ выполнено

$$\theta^*(\mathbf{X}) - \tau(\theta) = \frac{\tau'(\theta)(\ln\rho_\theta(\mathbf{X}))'_\theta}{I_{\mathbf{X}}(\theta)}.$$

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

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

Пример 3.4. Экспоненциальному классу распределений принадлежат многие семейства, которые мы рассматривали ранее. Приведём лишь некоторые из них.

- $X_i\sim\mathcal{N}(a,\sigma^2)$.

  $$\rho_{(a,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) = \exp\Bigg(\underbrace{\left(\frac{1}{2}\ln\frac{1}{2\pi\sigma^2} - \frac{a^2}{2\sigma^2}\right)}_{=a_0(a,\sigma^2)} - \frac{1}{2\sigma^2}\cdot x^2 + \frac{a}{\sigma^2}\cdot x\Bigg).$$

  Таким образом, $T_1(x) = x$, $T_2(x) = x^2$.

- $X_i \sim \mathrm{Exp}(\lambda)$.
$$\rho_\lambda(x) = \lambda e^{-\lambda x} I(x > 0) = I(x > 0) \cdot \exp(\ln \lambda - \lambda x),$$
и тогда $T_1(x) = x$.

- $X_i \sim \Gamma(\alpha, \lambda)$.
$$\rho_{(\alpha,\lambda)}(x) = \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} I(x > 0) = I(x > 0) \cdot \exp\big((\alpha \ln \lambda - \ln \Gamma(\alpha)) - \lambda x + (\alpha - 1) \ln x\big).$$
В данном случае $T_1(x) = x$, $T_2(x) = \ln x$.

- $X_i \sim \mathrm{Beta}(\alpha, \beta)$.
$$\rho_{(\alpha,\beta)}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} I(0 < x < 1) =$$
$$= I(0 < x < 1) \cdot \exp\big(-\ln B(\alpha, \beta) + (\alpha - 1) \ln x + (\beta - 1) \ln (1 - x)\big)$$
Получаем $T_1(x) = \ln x$, $T_2(x) = \ln (1 - x)$

- Рассмотрим теперь дискретные распределения. $X_i \sim \mathrm{Bern}(p)$. В данном случае обобщённая плотность имеет вид
$$\rho_p(x) = \begin{cases} p, & x = 0 \\ 1 - p, & x = 1 \end{cases}$$
Для удобства её лучше представить как
$$\rho_p(x) = p^x (1 - p)^{1-x} = \exp(x \ln p + (1 - x) \ln (1 - p)) = \exp\left(\ln (1 - p) + x \ln \frac{p}{1 - p}\right)$$
Тогда $T_1(x) = x$.

- $X_i \sim \mathrm{Pois}(\lambda)$.
$$\rho_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} \exp(-\lambda + x \ln \lambda),$$
и тогда $T_1(x) = x$.

■

Теорема 3.4.

Пусть $\theta \subset \mathbb{R}$. Тогда эффективная оценка существует только для экспоненциальных семейств. Более того, в этом случае $\theta^*(\mathbf{X}) = \overline{T(\mathbf{X})}$ является эффективной оценкой для $-a_0'(\theta)/a_1'(\theta)$, а любая другая эффективная оценка будет линейной функцией от $\theta^*(\mathbf{X})$.

Пример 3.5. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из $\mathrm{Pois}(\lambda)$. Информацию Фишера для неё можно найти напрямую, взяв интеграл из определения $i(\lambda)$ (и тем самым мы честно покажем, что выполняются условия регулярности).
$$i(\lambda) = \mathsf{E}_\lambda \left(\frac{\partial}{\partial \lambda} \ln \rho_\lambda(X_1)\right)^2 = \sum_{k=0}^{\infty} \left(\frac{\rho_\lambda'(k)}{\rho_\lambda(k)}\right)^2 \cdot \rho_\lambda(k) = \sum_{k=0}^{\infty} \left(\frac{k}{\lambda} - 1\right)^2 \cdot \frac{e^{-\lambda} \cdot \lambda^k}{k!} =$$
$$= 1 + e^{-\lambda} \cdot \sum_{k=1}^{\infty} \left(\frac{k\lambda^{k-2}}{(k-1)!} - \frac{2\lambda^{k-1}}{(k-1)!}\right) = 1 - 2 + e^{-\lambda} \cdot \sum_{k=1}^{\infty} \left(\frac{(k-1)\lambda^{k-2}}{(k-1)!} + \frac{\lambda^{k-2}}{(k-1)!}\right) = \frac{1}{\lambda}.$$

Поэтому если мы хотим несмещённо оценить $\tau(\lambda) = \lambda$, то нижняя оценка на дисперсию окажется равной

$$\frac{\tau'(\lambda)}{ni(\lambda)} = \frac{\lambda}{n}.$$

К счастью, естественная оценка $\lambda^*(\mathbf{X}) = \overline{\mathbf{X}}$ имеет ровно такую дисперсию:

$$\mathsf{D}_\theta \overline{\mathbf{X}} = \frac{1}{n^2} \sum \mathsf{D}_\theta X_i = \frac{\lambda}{n}.$$

Таким образом, $\overline{\mathbf{X}}$ является эффективной оценкой.

Однако можно поступить и проще. Из примера выше мы знаем, что это семейство принадлежит экспоненциальному классу распределений с $T(\mathbf{x}) = \sum x_i$, $a_0(\lambda) = -\lambda$ и $a_1(\lambda) = \ln \lambda$. Тогда по теореме выше оценка $\lambda^*(\mathbf{X}) = \overline{\mathbf{X}}$ является эффективной оценкой функции

$$-\frac{a_0'(\lambda)}{a_1'(\lambda)} = -\frac{-1}{1/\lambda} = \lambda.$$

Более того, по критерию эффективности можно легко найти информацию Фишера.

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

■

3.4 Многомерный случай

Посмотрим теперь, как можно обобщить введённую нами теорию на многомерный случай. Пусть теперь Θ — открытое множество в $\mathbb{R}^k$, и $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$. Информация Фишера есть дисперсия вклада выборки, а, как мы знаем из теории вероятности, аналогом дисперсии в многомерном пространстве служит ковариационная матрица.

Определение. Матрица ковариаций вклада выборки

$$I_{ij}(\theta) = \mathsf{E}_\theta \left(\frac{\partial}{\partial \theta_i} \ln \rho_\theta(\mathbf{X}) \cdot \frac{\partial}{\partial \theta_j} \ln \rho_\theta(\mathbf{X}) \right)$$

называется *информационной матрицей*.

Полученные нами результаты довольно естественно обобщаются на случай многомерного параметра. Однако для их корректности нужно уточнить условие C4, потребовав $\det I_{\mathbf{X}}(\theta) > 0$, чтобы матрицу можно было обращать.

Теорема 3.5 (неравенство Рао-Крамера в многомерном случае).

Для любой несмещённой оценки θ^* для $\tau(\theta)$ с локально ограниченным $\mathsf{E}_\theta \theta_i^{*2}$ и для любого $\theta \in \Theta$ справедливо неравенство

$$\mathsf{D}_\theta \theta^*(\mathbf{X}) \geqslant \frac{\partial \tau}{\partial \theta} I_{\mathbf{X}}^{-1}(\theta) \left(\frac{\partial \tau}{\partial \theta} \right)^T.$$

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

Теорема 3.6 (критерий эффективности в многомерном случае).

Оценка θ^* эффективна тогда и только тогда, когда для любого $\theta \in \Theta$

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

Пример 3.6. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из нормального распределения с параметрами $\theta = (a, \sigma^2)$. Для начала рассмотрим одномерные случаи, когда один из параметров известен, а второй — нет.

Пусть параметр σ^2 известен, а неизвестный параметр a необходимо оценить (то есть рассматривается модель сдвига). Имеем

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

Следовательно, наша модель принадлежит экспоненциальному семейству, а значит, по теореме 3.4 $\overline{T_1(\mathbf{X})} = \overline{\mathbf{X}}$ является эффективной оценкой для

$$-\frac{a'_0(a)}{a'_1(a)} = -\left(\frac{1}{2}\ln\frac{1}{2\pi\sigma^2} - \frac{a^2}{2\sigma^2}\right)'_a \Big/ \left(\frac{a}{\sigma^2}\right)'_a = a.$$

Для разнообразия посчитаем информацию одного наблюдения по определению, это нам пригодится позднее:

$$i(a) = \mathsf{D}_a\left(\frac{\partial}{\partial a}\ln\rho_a(X_1)\right) = \mathsf{D}_a\left(\frac{X_1 - a}{\sigma^2}\right) = \mathsf{D}_\theta\frac{X_1}{\sigma^2} = \frac{1}{\sigma^4}\mathsf{D}_a X_1 = \frac{1}{\sigma^2}.$$

Теперь будем считать, что a известно, а дисперсия σ^2 — нет (такая модель называется моделью масштаба). Она 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 $\widehat{\sigma^2} = \overline{(\mathbf{X}-a)^2}$ 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

$$i(\sigma^2) = \frac{\tau'(\sigma^2)\cdot(\ln\rho_{\sigma^2}(X_1))'_{\sigma^2}}{\widehat{\sigma^2} - \tau(\sigma^2)} = \frac{1\cdot\left(\frac{-1}{2\sigma^2} + \frac{(X_1-a)^2}{2\sigma^4}\right)}{(X_1-a)^2 - \sigma^2} = \frac{1}{2\sigma^4}.$$

Теперь попробуем найти эффективные оценки для различных функций от вектора параметров. Начнём с функции $\tau(a, \sigma^2) = (a, a^2 + \sigma^2)$. Во-первых, найдём вклад выборки:

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

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

$$\mathsf{cov}\left(\frac{\partial}{\partial a}\ln\rho(\mathbf{X}), \frac{\partial}{\partial\sigma^2}\ln\rho(\mathbf{X})\right) =$$

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

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

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

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

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

Несложно посчитать, что $\mathsf{E}_\theta X_1^3 = a^3 + 3a\sigma^2$. Поэтому

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

Находим матрицу Якоби для $\tau(a, \sigma^2) = (a, a^2 + \sigma^2)$, обращаем информационную матрицу $I_{\mathbf{X}}(\theta)$ (благо это нетрудно) и считаем ответ, воспользовавшись критерием эффективности:

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

Однако нам, конечно, хотелось бы найти эффективную оценку для вектора (a, σ^2). К сожалению, такой просто не существует. Действительно, предположим, что такая оценка θ^* существует. Тогда по критерию эффективности

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

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

3.5 Информация Фишера для статистик

Пусть $S(\mathbf{X})$ — статистика, у которой обобщённая плотность по мере μ равна $g_\theta(s)$.

Определение. Информацией Фишера статистики $S(\mathbf{X})$ называется величина

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

Для удобства потребуем выполнения условия регулярности

$$\frac{\partial}{\partial\theta}\mathsf{E}_\theta T(\mathbf{X}) = \mathsf{E}_\theta\left(T(\mathbf{X})\frac{\partial}{\partial\theta}\ln g_\theta(S(\mathbf{X}))\right)$$

для любой $S(\mathbf{X})$-измеримой статистики $T(\mathbf{X})$ с ограниченным вторым моментом.

Рассмотрим ещё одно свойство информации Фишера, которое дополнительно оправдывает её пафосное название, а также подводит нас к теме одного из следующих параграфов. Давайте поймём, как связана информация Фишера введённой нами статистики и информация всей выборки.

Теорема 3.7.

В условиях регулярности для статистики $S(\mathbf{X})$ выполняется неравенство $I_S(\theta) \leqslant I_{\mathbf{X}}(\theta)$ для любого $\theta \in \Theta$.

Доказательство. Ключевым наблюдением здесь является то, что

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

Действительно, данное тождество можно проверить по определению УМО. Статистика $\frac{\partial}{\partial \theta} \ln g_\theta(S(\mathbf{X}))$ будет $S(\mathbf{X})$-измеримой как функция от $S(\mathbf{X})$, и по условиям регулярности для любого $C \in \sigma(S(\mathbf{X}))$:

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

С помощью полученного равенства перепишем информацию Фишера для статистики $S(\mathbf{X})$:

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

По свойству УМО можно занести $S(\mathbf{X})$-измеримую $\frac{\partial}{\partial \theta} \ln g_\theta(S(\mathbf{X}))$ под знак УМО, а потом воспользоваться формулой полного матожидания:

$$\circledcirc \mathsf{E}_\theta \left[\mathsf{E}_\theta \left(\frac{\partial}{\partial \theta} \ln g_\theta(S(\mathbf{X})) \cdot \frac{\partial}{\partial \theta} \ln \rho_\theta(\mathbf{X}) \middle| S(\mathbf{X}) \right) \right] = \mathsf{E}_\theta \left(\frac{\partial}{\partial \theta} \ln g_\theta(S(\mathbf{X})) \cdot \frac{\partial}{\partial \theta} \ln \rho_\theta(\mathbf{X}) \right) \leqslant$$

Наконец, неравенство Коши-Буняковского даёт нам искомый результат:

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

□

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

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

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

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

$$\ln \rho_\theta(\mathbf{X}) = \ln g_\theta(S(\mathbf{X})) + c(\mathbf{X}),$$

$$\rho_\theta(\mathbf{X}) = g_\theta(S(\mathbf{X})) \cdot h(\mathbf{X}). \tag{3}$$

Итог: только статистики $S(\mathbf{X})$, которые удовлетворяют равенству (3), сохраняют информацию при редуцировании данных. Из-за подобного свойства при «сжатии данных» эти статистики представляют особый интерес с практической точки зрения. Мы рассмотрим их подробнее в параграфе 5.

3.6 Геометрический смысл

Закончим обсуждение информации Фишера новым взглядом на сию величину. Ранее мы множество раз убедились, что она характеризует количество информации, которая в пределе даёт нам выборка. Теперь же посмотрим на неё с точки зрения геометрии пространства нашей модели.

Как можно было бы мерить расстояние между распределениями взятого нами параметрического семейства? Первое, что приходит в голову, — использовать расстояние между соответствующими параметрами. Однако данный подход имеет множество недостатков: он зависит от конкретной параметризации и в принципе не всегда отражает наши представления о метрике. Например, явственно видно, что степень расхождения между распределениями $\mathcal{N}(0,1)$ и $\mathcal{N}(1,1)$ гораздо выше, чем между $\mathcal{N}(0,100)$ и $\mathcal{N}(1,100)$, хотя в смысле параметров сдвига и масштаба (μ,σ) расстояния для данных пар равны.

Другой подход — использовать расстояния непосредственно между самими распределениями. Распространённым примером такого «расстояния» является дивергенция Кульбака-Лейблера, которая для распределений с плотностями $p(\mathbf{x})$ и $q(\mathbf{x})$ по мере μ равна

$$KL(p||q) = \int_{\mathbb{R}^n} p(\mathbf{x}) \cdot \ln\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) \mu(d\mathbf{x}).$$

К сожалению, в прямом смысле данная величина метрикой не является, но она всё равно довольно хорошо показывает разницу между распределениями, что видно уже в случае нормального семейства (см. задачу 3.3). Посмотрим, как она ведёт себя локально, когда распределения отличаются не слишком сильно. Для этого дважды продифференцируем KL-дивергенцию двух распределений $\rho_{\boldsymbol{\theta}}$ и $\rho_{\boldsymbol{\theta}'}$, где первый параметр $\boldsymbol{\theta}$ фиксирован, а по второму, $\boldsymbol{\theta}'$, мы и будем дифференцировать:

$$\nabla_{\boldsymbol{\theta}'} KL(\rho_{\boldsymbol{\theta}}||\rho_{\boldsymbol{\theta}'}) = \nabla_{\boldsymbol{\theta}'} \int_{\mathbb{R}^n} \rho_{\boldsymbol{\theta}}(\mathbf{x}) \cdot \ln \rho_{\boldsymbol{\theta}}(\mathbf{x})\, \mu(d\mathbf{x}) - \nabla_{\boldsymbol{\theta}'} \int_{\mathbb{R}^n} \rho_{\boldsymbol{\theta}}(\mathbf{x}) \cdot \ln \rho_{\boldsymbol{\theta}'}(\mathbf{x})\, \mu(d\mathbf{x}) =$$

$$= -\int_{\mathbb{R}^n} \rho_{\boldsymbol{\theta}}(\mathbf{x}) \cdot \nabla_{\boldsymbol{\theta}'} \ln \rho_{\boldsymbol{\theta}'}(\mathbf{x})\, \mu(d\mathbf{x})$$

$$\nabla^2_{\boldsymbol{\theta}'} KL(\rho_{\boldsymbol{\theta}}||\rho_{\boldsymbol{\theta}'}) = -\int_{\mathbb{R}^n} \rho_{\boldsymbol{\theta}}(\mathbf{x}) \cdot \nabla^2_{\boldsymbol{\theta}'} \ln \rho_{\boldsymbol{\theta}'}(\mathbf{x})\, \mu(d\mathbf{x}).$$

Для $\boldsymbol{\theta}' = \boldsymbol{\theta}$ первое выражение, градиент KL-дивергенции, равно нулю в силу условий регулярности, что вообще-то не сильно впечатляет — в этой точке KL-дивергенция равна минимальному значению, нулю, ибо два распределения совпадают. Второе же выражение, гессиан KL-дивергенции, более интересно, потому что согласно задаче 3.2 оно в точности равно информационной матрице Фишера в точке $\boldsymbol{\theta}$. Иначе говоря, информация Фишера описывает локальную кривизну пространства распределений, в котором в качестве метрики взята KL-дивергенция:

$$KL(\rho_{\boldsymbol{\theta}}||\rho_{\boldsymbol{\theta}+\Delta}) = \Delta^T I_{\mathbf{X}}(\boldsymbol{\theta})\Delta + o(\|\Delta\|^2).$$

В частности, её можно рассматривать как I квадратичную форму, которая превращает параметрическое семейство распределений в риманово многообразие, что, однако, уже выходит за рамки сего текста. Отметим лишь пример применения наблюдений выше — модификацию градиентного спуска при нахождения максимума правдоподобия (см. раздел 4.2).

Задачи

Задача 3.1. В модели сдвига-масштаба стандартного распределения Коши,

$$\{\mathrm{Cauchy}(a,\sigma)\colon a\in\mathbb{R},\sigma\in\mathbb{R}_+\},$$

найдите информационную матрицу Фишера.

Задача 3.2. Докажите, что если в дополнение к условиям регулярности добавить равенства

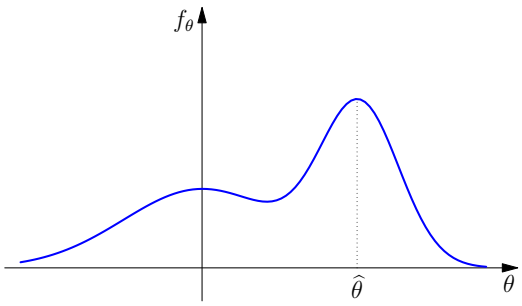
$$\int \frac{\partial}{\partial\theta_i\partial\theta_j}\rho''_{\boldsymbol{\theta}}(\mathbf{x})\,d\mathbf{x}=0$$

(в частности, это верно, если интеграл плотности дважды дифференцируем по параметру), то информацию Фишера можно представить как

$$I_{\mathbf{X}}(\boldsymbol{\theta})_{ij}=-\mathsf{E}_{\boldsymbol{\theta}}\frac{\partial^2\ln\rho_{\boldsymbol{\theta}}(\mathbf{X})}{\partial\theta_i\partial\theta_j}.$$

Задача 3.3. Посчитайте KL-дивергенцию между двумя нормальными распределениями $\mathcal{N}(\mu_1,\sigma_1^2)$ и $\mathcal{N}(\mu_2,\sigma_2^2)$.

4 Оценка максимального правдоподобия



Данный метод был широко популяризирован известным статистиком Рональдом Фишером и обладает множеством любопытных свойств. Но для них нужно потребовать выполнения некоторых условий в нашей модели, поэтому сначала поговорим о мотивации такой оценки и рассмотрим примеры.

Идея ОМП проста: давайте среди всех возможных распределений возьмём то, при котором вероятность лицезреть именно такую выборку наиболее высокая, то есть наблюдение наиболее правдоподобно.

Определение. Пусть $\mathcal{P} = \{\mathsf{P}_\theta \colon \theta \in \Theta\}$ — доминируемое семейство распределений P_θ с плотностью $\rho_\theta(x)$. *Функцией правдоподобия* выборки $X_1, \ldots, X_n$ называется плотность их совместного распределения, взятая от элементов выборки:

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

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

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

Заметим, что точки максимума функции правдоподобия и её логарифмического братаблизнеца совпадают в силу монотонности логарифма, поэтому максимизировать можно любую из этих функций. Второй вариант обычно предпочтительнее, так как минимизировать сумму гораздо проще, чем произведение.

Может показаться, что такая оценка кардинально отличается от предыдущих, получаемых методом подстановки. Однако это не так, и ОМП можно задать как функционал от эмпирического распределения, который минимизирует «расстояние» между распределениями. Вероятно, это «расстояние» покажется знакомым, так как оно тесно связано с расстоянием Кульбака-Лейблера, которое имеет широкое применение в статистике и машинном обучении.

Определение. Пусть P — вероятностная мера с обобщённой плотностью p, а Q — произвольное распределение. Введём величину, называемую *кросс-энтропией*

$$d(\mathsf{P}, \mathsf{Q}) = -\int \ln p(x)\, \mathsf{Q}(dx).$$

Мы называем её «расстоянием», потому что эта функция не является метрикой (она даже не симметрична), однако она хорошо показывает похожесть распределений. Посему логично рассмотреть функционал

$$G(\mathsf{Q}) = \arg\min_{\theta \in \Theta} d(\mathsf{P}_\theta, \mathsf{Q})$$

Если $\mathsf{Q} = \mathsf{P}_\theta$, то функционал будет равен θ, что гарантирует следующее

Утверждение 4.1. *Если* P, Q — *два доминируемых относительно* μ *распределения с плотностями* p *и* q *соответственно, то выполнено неравенство*

$$d(\mathsf{P},\mathsf{P}) \leqslant d(\mathsf{Q},\mathsf{P})$$

причём равенство достигается тогда и только тогда, когда μ*-п.н. выполнено* $p = q$.

Доказательство. Как известно, $\ln(1+x) \leqslant x$ при $x > -1$. Тогда выполнено неравенство

$$\log\frac{q}{p} = \log\left(1 + \left(\frac{q}{p} - 1\right)\right) \leqslant \frac{q}{p} - 1.$$

Значит, домножая его на p и навешивая знак интеграла, получаем:

$$d(\mathsf{P},\mathsf{P}) - d(\mathsf{Q},\mathsf{P}) = \int p \log\frac{q}{p}\,d\mu \leqslant \int p\left(\frac{q}{p} - 1\right)\,d\mu = \int p\,d\mu - \int q\,d\mu = 1 - 1 = 0.$$

Причём по свойству интеграла Лебега равенство верно, когда μ-п.н. $\log q/p \geqslant q/p - 1$, то есть $q/p = 1$, что и требовалось. □

По методу подстановки оценка для сего функционала может быть получена как

$$G(\mathsf{P}_n^*) = \arg\max_{\theta\in\Theta} \int \ln\rho_\theta(x)\,\mathsf{P}_n^*(dx) = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ln\rho_\theta(X_i),$$

что есть точка максимума логарифмической функции правдоподобия.

Пример 4.1. Найдём ОМП для некоторых известных распределений. Чаще всего мы будем делать это через нахождение стационарных точек, в которых производная функции равна нулю. Проверки, что такие точки действительно доставляют непременно максимум функции правдоподобия, будут опускаться и оставляться читателю в качестве упражнения.

- $X_i \sim \mathrm{Geom}(p)$.

$$f_p(X_1,\dots,X_n) = \prod (1-p)^{X_i} p, \quad L_p(X_1,\dots,X_n) = n\ln p + \sum X_i \ln(1-p),$$
$$\frac{\partial}{\partial p} L_p(X_1,\dots,X_n) = \frac{n}{p} - \sum \frac{X_i}{1-p} = 0, \quad \frac{n(1-p) - p\sum X_i}{p(1-p)} = 0,$$
$$p\left(n + \sum X_i\right) = n \Longrightarrow \widehat{p} = \frac{n}{n + \sum X_i} = \frac{1}{1+\overline{\mathbf{X}}}.$$

- $X_i \sim \mathrm{U}(0,a)$. В данном случае тупо взять и продифференцировать не выйдет, так как плотность разрывна. Функция правдоподобия выглядит так:

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

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.

- $X_i \sim \mathcal{N}(a, \sigma^2)$.

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

$$L_\theta(X_1, \ldots, X_n) = -\frac{n}{2} \ln 2\pi\sigma^2 - \sum \frac{(X_i - a)^2}{2\sigma^2}$$

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

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

- $X_i \sim \Gamma(\alpha, \lambda)$, где α — известная величина, а λ — неизвестный параметр.

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

Для $X_1, \ldots, X_n > 0$ имеем

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

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

- $X_i \sim \text{Pareto}(k, a)$.

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

Для фиксированного k максимум f_θ достигает при $\widehat{a} = X_{(1)}$ (аналогично пункту с $\mathrm{U}(0, a)$). Тогда если принять a равным первой порядковой статистике, получаем

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

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

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

■

Пример 4.2. Рассмотрим также очень полезный и поучительный пример для $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \mathcal{N}(a, \Sigma)$ — выборки из независимых гауссовских векторов, где $a \in \mathbb{R}^k$, $\Sigma \in \mathbb{S}^k_{++}$. Напомним, что плотность гауссовского вектора размерности k равна

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

Найдём оценку максимального правдоподобия для вектора средних a и ковариационной матрицы Σ.

Для начала найдём логарифмическую функцию правдоподобия:

$$f_\theta(\mathbf{X}_1,\dots,\mathbf{X}_n) = (2\pi)^{-nk/2}(\det\Sigma)^{-n/2}\exp\left(\sum_{i=1}^n -\frac{1}{2}(\mathbf{X}_i-a)^T\Sigma^{-1}(\mathbf{X}_i-a)\right),$$

$$\ln f_\theta(\mathbf{X}_1,\dots,\mathbf{X}_n) = -\frac{nk}{2}\ln 2\pi - \frac{n}{2}\ln\det\Sigma - \frac{1}{2}\sum_{i=1}^n(\mathbf{X}_i-a)^T\Sigma^{-1}(\mathbf{X}_i-a).$$

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

$$d_a\ln f_\theta(\mathbf{X}_1,\dots,\mathbf{X}_n) = -\frac{1}{2}d_a\left(\sum_{i=1}^n(\mathbf{X}_i-a)^T\Sigma^{-1}(\mathbf{X}_i-a)\right) =$$

$$= -\frac{1}{2}\sum_{i=1}^n\left(d_a(\mathbf{X}_i-a)^T\Sigma^{-1}(\mathbf{X}_i-a) + (\mathbf{X}_i-a)^T\Sigma^{-1}d_a(\mathbf{X}_i-a)\right) =$$

$$= \frac{1}{2}\sum_{i=1}^n\left(d_aa^T\Sigma^{-1}(\mathbf{X}_i-a) + (\mathbf{X}_i-a)^T\Sigma^{-1}d_aa\right) =$$

$$= \frac{1}{2}\sum_{i=1}^n\left(\langle d_aa,\Sigma^{-1}(\mathbf{X}_i-a)\rangle + \langle\Sigma^{-T}(\mathbf{X}_i-a),d_aa\rangle\right) =$$

$$= \left\langle\sum_{i=1}^n\Sigma^{-1}(\mathbf{X}_i-a),d_aa\right\rangle = 0 \Longrightarrow \sum_{i=1}^n\Sigma^{-1}(\mathbf{X}_i-a) = 0 \Longrightarrow \widehat{a} = \overline{\mathbf{X}}.$$

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:

$$\ln f_\theta(\mathbf{X}_1,\dots,\mathbf{X}_n) = -\frac{nk}{2}\ln 2\pi - \frac{n}{2}\ln\det\Sigma - \frac{1}{2}\operatorname{tr}\sum_{i=1}^n(\mathbf{X}_i-a)^T\Sigma^{-1}(\mathbf{X}_i-a).$$

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

$$\ln f_\theta(\mathbf{X}_1,\dots,\mathbf{X}_n) = -\frac{nk}{2}\ln 2\pi - \frac{n}{2}\ln\det\Sigma - \frac{1}{2}\sum_{i=1}^n\operatorname{tr}(\mathbf{X}_i-a)(\mathbf{X}_i-a)^T\Sigma^{-1} =$$

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

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 курса алгебры, 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:

$$d_\Xi \ln f_\theta(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \frac{n}{2} d_\Xi(\ln\det\Xi) - \frac{1}{2} d_\Xi \left(\sum_{i=1}^n \langle (\mathbf{X}_i - a)(\mathbf{X}_i - a)^T, \Xi \rangle \right) =$$

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

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

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

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

■

Конечно, у этого метода сразу видно несколько недостатков. Во-первых, нельзя быть точно уверенным, что стационарные точки функции правдоподобия будут доставлять максимум. Во-вторых, не у всех функций правдоподобия можно легко вычислить стационарные точки. Первая проблема либо решается нахождением второго дифференциала, либо игнорируется (мы, как можно заметить, пошли вторым путём).

4.1 Асимптотическая эффективность

Так же, как и в предыдущем разделе, нас интересует наилучшая оценка, только теперь та, у которой наименьшая возможная асимптотическая дисперсия. Оказывается, что при некоторых ограничениях таковой является ОМП.

Условия сильной регулярности

Помимо С1-С4 также предполагаем выполнимость следующих постулатов:

С5 Для $\theta_1 \neq \theta_2 \in \Theta$ распределения P_{θ_1} и P_{θ_2} различны;

С6 Для μ-п.в. $\mathbf{x}$ плотность $\rho_\theta(\mathbf{x})$ трижды непрерывно дифференцируема;

С7 $\int \rho_\theta(\mathbf{x}) d\mathbf{x}$ можно дважды дифференцировать под знаком интеграла;

С8 Существует функция $H(\mathbf{x})$ такая, что $\mathsf{E}_\theta H(\mathbf{X}) < \infty$ и при всех $\mathbf{x}$ для любого $\theta \in \Theta$ верно $\left|\frac{\partial^3}{\partial\theta^3} \ln \rho_\theta(\mathbf{x})\right| \leqslant H(\mathbf{x})$.

При данных условиях можно утверждать следующее.

Теорема 4.1.

Пусть для любого $n \in \mathbb{N}$ и любой реализации выборки $x_1, \ldots, x_n$ существует и единственно решение θ^* уравнения правдоподобия

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln \rho_\theta(x_i) = 0.$$

Тогда при условиях C1-C8 оценка θ^* является асимптотически нормальной с асимптотической дисперсией $i(\theta)^{-1}$.

Но не менее потрясающий результат заключается в том, что оценка выше в некотором смысле не улучшаема.

Теорема 4.2.

При условиях C1-C8 для любой асимптотически нормальной оценки $\widehat{\theta}$ с непрерывной асимптотической дисперсией $\sigma^2(\theta)$ выполнено $\sigma^2(\theta) \geqslant i(\theta)^{-1}$.

Таким образом, в текущих ограничениях ОМП не хуже любой другой оценки с непрерывной асимптотической дисперсией в асимптотическом подходе.

Пример 4.3. Условие на непрерывность асимптотической дисперсии существенно. Можно построить такую оценку, которая для некоторых значений параметра будет оценивать его чуть лучше, чем обычно, порождая разрыв.

Рассмотрим нормальную модель сдвига: $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$, где $\theta \in \mathbb{R}$. В ней имеется естественная ОМП $\widehat{\theta} = \overline{\mathbf{X}}$, которая по теореме выше будет асимптотически эффективной оценкой параметра θ с асимптотической дисперсией 1. Её можно в некоторой степени «усовершенствовать», взяв оценку

$$\theta^* = \begin{cases} \overline{\mathbf{X}}, & |\overline{\mathbf{X}}| \geqslant n^{-1/4}, \\ \overline{\mathbf{X}}/2, & |\overline{\mathbf{X}}| < n^{-1/4}. \end{cases}$$

Идея проста: если среднее достаточно близко к нулю, то скорее всего параметр сдвига нулевой, и в таком случае можно искусственно уменьшить значение оценки, тем самым уменьшив разброс.

Для удобства перепишем сию оценку в следующем виде:

$$\theta^* = \overline{\mathbf{X}} \cdot \left(\frac{1}{2} + \frac{1}{2} \cdot I(|\overline{\mathbf{X}}| \geqslant n^{-1/4}) \right)$$

При $\theta = 0$ распределение $\sqrt{n}\overline{\mathbf{X}}$ одинаково, поэтому $\mathsf{P}(|\overline{\mathbf{X}}| \geqslant n^{-1/4}) = \mathsf{P}(\sqrt{n}|\overline{\mathbf{X}}| \geqslant n^{1/4}) \to 0$, откуда $I(|\overline{\mathbf{X}}| \geqslant n^{-1/4}) \xrightarrow{d} 0$, и по лемме Слуцкого

$$\sqrt{n}\theta^* = \underbrace{\sqrt{n}\overline{\mathbf{X}}}_{\sim \mathcal{N}(0,1)} \cdot \underbrace{\left(\frac{1}{2} + \frac{1}{2} \cdot I(|\overline{\mathbf{X}}| \geqslant n^{-1/4}) \right)}_{\xrightarrow{d} 1/2} \xrightarrow{d} \mathcal{N}(0, 1/4).$$

Пусть теперь $\theta \neq 0$. В таком случае $\overline{\mathbf{X}} \xrightarrow{d} \theta \neq 0$, поэтому $\mathsf{P}(|\overline{\mathbf{X}}| \geqslant n^{-1/4}) \to 1$, и

соответственно $\sqrt{n}I(|\overline{\mathbf{X}}| < n^{-1/4}) \xrightarrow{d} 0$. Всё по той же лемме Слуцкого имеем

$$\sqrt{n}(\theta^* - \theta) = \sqrt{n}(\overline{\mathbf{X}} - \theta) + \underbrace{\frac{\sqrt{n}}{2} \cdot I(|\overline{\mathbf{X}}| < n^{-1/4})}_{\xrightarrow{d} 0} \xrightarrow{d} \mathcal{N}(0, 1).$$

Таким образом, асимптотическая дисперсия оценки равна

$$\sigma^2(\theta) = \begin{cases} 1/4, & \theta = 0; \\ 1, & \theta \neq 0, \end{cases}$$

что, вообще говоря, меньше дисперсии ОМП. ■

Не всегда корни уравнения правдоподобия можно легко найти в явном виде, как это было ранее. Вместо трудозатратного аналитического нахождения максимума правдоподобия можно попытаться найти его численно. Стандартным способом является градиентный спуск или его модификации по типу SGD, Momentum и прочих. В следующих разделах мы обсудим некоторые техники, которые позволяют находить ОМП более эффективно.

4.2 Натуральный градиентный спуск

Вернёмся к понятию информации Фишера, введённой в прошлом параграфе. В этом разделе мы обсудим её геометрический смысл, который подскажет нам более эффективный способ нахождения максимума правдоподобия. Из задачи 3.2 читатель поймёт, что при дополнительных условиях регулярности информационную матрицу Фишера можно найти иным способом: как матожидание матрицы Гессе логарифма правдоподобия:

$$I_{\mathbf{X}}(\boldsymbol{\theta})_{ij} = -\mathsf{E}_{\boldsymbol{\theta}} \frac{\partial^2 \ln \rho_{\boldsymbol{\theta}}(\mathbf{X})}{\partial \theta_i \partial \theta_j}.$$

Как известно из курса матанализа, гессиан показывает характер выпуклости функции. С учётом того, что математическое ожидание приближается выборочным средним, за знак которого можно вынести знак производной:

$$-i(\boldsymbol{\theta})_{ij} = \mathsf{E}_{\boldsymbol{\theta}} \frac{\partial^2 \ln \rho_{\boldsymbol{\theta}}(X_1)}{\partial \theta_i \partial \theta_j} \xleftarrow{\boldsymbol{\theta}\text{-п.н.}} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln \rho_{\boldsymbol{\theta}}(X_i)}{\partial \theta_i \partial \theta_j} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \sum_{i=1}^{n} \frac{\ln \rho_{\boldsymbol{\theta}}(X_i)}{n},$$

информация Фишера характеризует выпуклость оптимизируемой функции правдоподобия. Чем выше информация, тем более ярко будет выражен искомый максимум правдоподобия; чем ниже информация, тем более размыто он будет выделяться среди своей окрестности.

4.3 Одношаговые оценки

Вместо численной максимизации правдоподобия можно (так же численно) находить корень уравнения правдоподобия. Здесь мы остановимся на так называемом *методе Ньютона*. Если перед нами стоит задача найти решение уравнения $\varphi(\theta) = 0$, где φ — какая-то дифференцируемая функция, то корень можно найти итеративно: начинаем с некоторого начального приближения θ_0 и далее вычисляем последующие значения θ_n:

$$\theta_{n+1} = \theta_n - \frac{\varphi(\theta_n)}{\varphi'(\theta_n)}.$$

Мотивация такого алгоритма проста: если $\widehat{\theta}$ есть корень сего уравнения, то при разложении в ряд Тейлора функции φ в окрестности точки θ, близкой к $\widehat{\theta}$, имеем

$$0 = \varphi(\widehat{\theta}) \approx \varphi(\theta) + (\widehat{\theta} - \theta) \cdot \varphi'(\theta),$$

откуда можно выразить $\widehat{\theta}$:

$$\widehat{\theta} \approx \theta - \frac{\varphi(\theta)}{\varphi'(\theta)}.$$

В нашем случае вместо φ хочется взять производную логарифмической функции правдоподобия $\frac{\partial}{\partial\theta}L_\theta(\mathbf{X})$. Данный метод легко обобщается и на многомерный случай:

$$\theta_{n+1} = \theta_n - [d^2 L_{\theta_n}]^{-1} \cdot dL^T_{\theta_n},$$

где $dL^T_{\theta_n}$ — столбец из частных производных $L_{\theta_n}(\mathbf{X})$, а $d^2L_{\theta_n}$ — матрица из вторых производных (так называемая матрица Гессе).

Неочевидным остаётся то, на сколько шагов запускать сей алгоритм. Оказывается, при некоторых условиях достаточно совершить всего лишь один шаг, чтобы получить не просто хорошую оценку, а асимптотически эффективную, то есть в такой же асимптотической дисперсией, как у и ОМП.

Теорема 4.3.

В условиях регулярности C1-C8 если θ^* — асимптотически нормальная оценка параметра θ, то одношаговая оценка

$$\widehat{\theta} = \theta^* - \frac{\frac{\partial}{\partial\theta}L_{\theta^*}(\mathbf{X})}{\frac{\partial^2}{\partial\theta^2}L_{\theta^*}(\mathbf{X})}$$

будет иметь асимптотическую дисперсию $1/i(\theta)$.

Пример 4.4. Рассмотрим модель сдвига распределения Коши:

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

У нас уже имеется довольно неплохая оценка $\theta^* = \widehat{\mu}$, у которой мы даже знаем асимптотическую дисперсию $\sigma^2(\theta) = \pi^2/4$. Однако для ОМП она будет равна $1/i(\theta) = 2$, так как

$$i(\theta) = \mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}L_\theta(X_1)\right)^2 = \mathsf{E}_\theta\left(\frac{-2(X_1-\theta)}{1+(X_1-\theta)^2}\right)^2 = \int_{\mathbb{R}}\frac{4x^2}{\pi(1+x^2)^3}\,dx = \frac{1}{2},$$

то есть ещё есть куда стремиться. Но благо есть одношаговая оценка, построенная по $\widehat{\mu}$, и она уже будет иметь ту же самую асимптотическую дисперсию:

$$\widehat{\theta} = \widehat{\mu} + \sum_{i=1}^n \frac{X_i - \widehat{\mu}}{1+(X_i-\widehat{\mu})^2} \Bigg/ \sum_{i=1}^n \frac{1-(X_i-\widehat{\mu})^2}{\left(1+(X_i-\widehat{\mu})^2\right)^2}.$$

Проверим, что дисперсия и вправду такая. Смоделируем 100000 выборок размера $n = 100$ и прикинем, какую дисперсию имеют обычная выборочная медиана и одношаговая оценка.

```
def compare_variance(n):
    num = 100000
    x = sps.cauchy.rvs(size=(n, num))
    med = np.median(x, axis=0)
    one_step = med +\
        ((x - med) / (1 + (x - med) ** 2)).sum(axis=0) /\
        ((1 - (x - med) ** 2) / (1 + (x - med) ** 2) ** 2).sum(axis=0)
```

```
    print('Variance for median -', med.var() * n)
    print('Variance for one-step estimator -', one_step.var() * n)

compare_variance(100)
```

```
Variance for median - 2.527736079325076
Variance for one-step estimator - 2.0560138435262205
```

Как можно лицезреть, асимптотические дисперсии получились довольно близкими к теоретическим значениям. Однако следует помнить, что асимптотическая дисперсия на то и асимптотическая, чтобы показывать поведение оценки лишь для большого размера выборки. Для малого количества наблюдений, к примеру, $n = 20$ медиана более выигрышна, так как одношаговая оценка чувствительнее относится к выбросам.

```
compare_variance(20)
```

```
Variance for median - 2.7872805561964147
Variance for one-step estimator - 45.95128078716877
```

■

4.4 EM-алгоритм

Рассмотрим другой распространённый метод, который не только значительно упрощает нахождение решений уравнения правдоподобия, но и позволяет «выжимать» из данных больше информации. Данная тема заслуживает отдельной главы, однако мы затронем её лишь вскользь; более подробное её описание и больше примеров можно найти, например, в [5].

Ключевой идеей является введение так называемых *латентных величин* в нашу модель. Проще говоря, мы предполагаем, что помимо исходных данных в природе существуют некоторые скрытые, недоступные нам наблюдения. Если ввести достаточно удачные латентные переменные, то их совместная с видимой выборкой плотность распределения окажется «достаточно хорошей» для максимизации, например, будет принадлежать экспоненциальному классу распределений, в то время как исходная функция правдоподобия может быть тяжела в плане оптимизации.

Пример 4.5. Мультимодальные данные, в которых имеются два и более ярковыраженных пика, зачастую удобно объяснять с помощью модели *смеси нормальных распределений*. Для простоты рассмотрим смесь двух гауссиан:

$$\rho_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}}(x) = \pi_1 \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + \pi_0 \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right), \qquad (4)$$

где $\pi_1 + \pi_0 = 1$, $\pi_1, \pi_0 > 0$.

У смеси распределений есть естественная интерпретация: перед генерацией очередного элемента выборки бросается монетка, которая падаем орлом вверх с вероятностью π_1. Если выпал орёл, то величина генерируется из распределения $\mathcal{N}(\mu_1, \sigma_1^2)$, иначе — из $\mathcal{N}(\mu_0, \sigma_0^2)$. Это отражает тот факт, что объекты выборки могут иметь разную природу (например, пол, расу и т.д.), определяющую вероятностный закон, которому подчиняется целевая величина.

Модель эта, конечно, хорошая, но имеет существенный недостаток, критичный в теку-

щем контексте: у выборки из такого параметрического семейства нет ОМП (см. задачу 4.4). Ничего страшного, решение уравнения правдоподобия всё ещё может существовать и обладать разумными свойствами по типу состоятельности, однако вскрывается другой нюанс: правдоподобие выборки крайне неприятно максимизировать. Так как плотность равна некоторой сумме, то логарифмическую функцию правдоподобия не получится удобно «причесать», и градиенты у неё будут паршивыми.

Однако всё меняется, если помимо исходной выборки рассматривать те самые результаты выпадений монетки: если обозначить за z_i индикатор выпадения орла для i-ого элемента выборки, то совместная плотность будет иметь вид

$$f_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}}(\mathbf{x},\mathbf{z}) = \prod_{i=1}^{n} (\pi_1 \varphi_{\mu_1,\sigma_1^2}(x_i))^{z_i} \cdot (\pi_0 \varphi_{\mu_0,\sigma_0^2}(x_i))^{1-z_i},$$

где для краткости мы обозначили за φ_{μ,σ^2} плотность $\mathcal{N}(\mu,\sigma^2)$. Логарифмическая функция правдоподобия в таком случае равна

$$L_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}}(\mathbf{x},\mathbf{z}) = \sum_{i=1}^{n} \Big[z_i \ln(\pi_1 \varphi_{\mu_1,\sigma_1^2}(x_i)) + (1-z_i)\ln(\pi_0 \varphi_{\mu_0,\sigma_0^2}(x_i)) \Big],$$

что максимизировать одно удовольствие. ■

Однако мы вообще говоря не знаем значений латентных переменных, чтобы за просто так оптимизировать совместную плотность. Можно было бы взять её маргинальную версию, беря интеграл по латентным переменным, но их распределение, очевидно, зависит от неизвестных параметров, которые мы хотели оценить посредством латентных переменных — замкнутый круг. Выход из него заключается в том, чтобы попеременно обновлять то значения параметров, то распределение на латентных переменных. Сделаем это следующим образом.

Рассмотрим общую задачу максимизации логарифмической функции правдоподобия $L_{\mathbf{x}}(\boldsymbol{\theta}) = \ln \rho_{\boldsymbol{\theta}}(\mathbf{x})$, и на текущем t-ом шаге мы взяли в качестве параметра $\boldsymbol{\theta}_t$. Используем его в условном распределении $\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})$, по которому возьмём матожидание полной плотности:

$$L_{\mathbf{x}}(\boldsymbol{\theta}) = \ln \int \rho_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\,dz = \ln \int \frac{\rho_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}{\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})} \rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})\,dz = \ln \mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})} \frac{\rho_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}{\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})}.$$

Теперь, когда на сцене появилось матожидание, можно применить неравенство Йенсена и перейти к нижней оценки лог. функции правдоподобия, которую удобнее максимизировать.

$$L_{\mathbf{x}}(\boldsymbol{\theta}) = \ln \mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})} \frac{\rho_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}{\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})} \geqslant \mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})} \left(\ln \frac{\rho_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}{\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})} \right).$$

Матожидание справа сокращённо называют ELBO (англ. evidence lower bound). Эта нижняя оценка обладает важным свойством: при $\boldsymbol{\theta} = \boldsymbol{\theta}_t$ в неравенстве выше достигается равенство. Действительно,

$$\mathrm{ELBO}(\boldsymbol{\theta}_t) = \mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})} \left(\ln \frac{\rho_{\boldsymbol{\theta}_t}(\mathbf{x},\mathbf{z})}{\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})} \right) =$$

$$= \mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})} \left(\ln \frac{\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x}) \cdot \rho_{\boldsymbol{\theta}_t}(\mathbf{x})}{\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})} \right) = \mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})} \ln \rho_{\boldsymbol{\theta}_t}(\mathbf{x}) = \ln \rho_{\boldsymbol{\theta}_t}(\mathbf{x}) = L_{\mathbf{x}}(\boldsymbol{\theta}_t).$$

Такие функции ещё называют *вариационной нижней оценкой*. При её наличии оптимизацию исходной функции $L_{\mathbf{x}}(\boldsymbol{\theta})$ можно свести к оптимизации оценки, которую зачастую проводить легче. Тогда если $\boldsymbol{\theta}_{t+1} = \arg\max_{\boldsymbol{\theta}} \mathrm{ELBO}(\boldsymbol{\theta})$, то

$$L_{\mathbf{x}}(\boldsymbol{\theta}_{t+1}) \geqslant \mathrm{ELBO}(\boldsymbol{\theta}_{t+1}) \geqslant \mathrm{ELBO}(\boldsymbol{\theta}_t) = L_{\mathbf{x}}(\boldsymbol{\theta}_t),$$

то есть по прошествии шага алгоритма правдоподобие не уменьшилось. Таким образом, задача свелась к максимизации какой-то не очень сложной функции. Обычно к ELBO дополнительно добавляют константу $\mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})}\ln\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})$, которая не влияет на $\arg\max$, но при этом оптимизируемая функция принимает более приятный вид:

$$J(\theta)=\mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})}\left(\ln\frac{\rho_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}{\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})}\right)+\mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})}\ln\rho_{\theta_t}(\mathbf{z}|\mathbf{x})=\mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})}\ln\rho_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}).$$

Итак, приведём окончательный вариант так называемого *EM-алгоритма*:

1. В качестве начального приближения параметров выбираем некоторое θ_0;
2. Для t от 0 до k повторяем следующее:
   - Берём текущее значение параметров $\boldsymbol{\theta}_t$;
   - (E-шаг, от слова Expectation) Считаем матожидание
   $$J(\theta)=\mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})}\ln\rho_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}),$$
   по латентным переменным $\mathbf{z}$, распределённым в соответствии с $\rho_{\boldsymbol{\theta}_t}(\mathbf{z}|\mathbf{x})$;
   - (M-шаг, от слова Maximization) Ищем точку максимума функции выше:
   $$\boldsymbol{\theta}_{t+1}=\arg\max_{\boldsymbol{\theta}}J(\theta).$$

Пример 4.6. Вернёмся к задаче разделения смесей из примера 4.5. Полная функция правдоподобия уже была найдена ранее, с помощью неё легко найти условное распределение на $\mathbf{z}$:

$$\mathsf{P}_{\boldsymbol{\mu}_t,\boldsymbol{\sigma}_t,\pi_t}(Z_i=j|X_i=x_i)\sim\mathsf{P}_{\boldsymbol{\mu}_t,\boldsymbol{\sigma}_t,\pi_t}(Z_i=j,X_i=x_i)=\pi_{t,j}\cdot\varphi_{\mu_{t,j},\sigma^2_{t,j}}(x_i),\quad j\in\{0,1\}.$$

Так как эти вероятности в сумме должны дать 1, то

$$\mathsf{P}_{\boldsymbol{\mu}_t,\boldsymbol{\sigma}_t,\pi_t}(Z_i=j|X_i=x_i)=\frac{\pi_{t,j}\cdot\varphi_{\mu_{t,j},\sigma^2_{t,j}}(x_i)}{\pi_{t,0}\cdot\varphi_{\mu_{t,0},\sigma^2_{t,0}}(x_i)+\pi_{t,1}\cdot\varphi_{\mu_{t,1},\sigma^2_{t,1}}(x_i)}=:p_{ij},$$

откуда

$$J(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi})=\mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\mu}_t,\boldsymbol{\sigma}_t,\boldsymbol{\pi}_t}(\mathbf{z}|\mathbf{x})}\ln\rho_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi}}(\mathbf{x},\mathbf{z})=$$

$$=\sum_{i=1}^n\Big[p_{i,1}\cdot(\ln\pi_1+\ln\varphi_{\mu_1,\sigma_1^2}(x_i))+p_{i,0}\cdot(\ln\pi_0+\ln\varphi_{\mu_0,\sigma_0^2}(x_i))\Big]=\sum_{i=1}^n(p_{i,1}\ln\pi_1+p_{i,0}\ln\pi_0)-$$

$$-\frac12\sum_{i=1}^n\left(p_{i,1}\cdot\left[\frac{(x_i-\mu_1)^2}{\sigma_1^2}+\ln(2\pi\sigma_1^2)\right]+p_{i,0}\cdot\left[\frac{(x_i-\mu_0)^2}{\sigma_0^2}+\ln(2\pi\sigma_0^2)\right]\right).$$

E-шаг осуществлён, осталось максимизировать $J(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi})$ для получения новых приближений параметров. При нежелании продолжать выкладки это можно сделать программно, но мы всё же найдём ответ аналитически, к тому же параметры $\boldsymbol{\pi}$ и $(\boldsymbol{\mu},\boldsymbol{\sigma})$ разнесены по разным слагаемым.

Первое слагаемое в $J(\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi})$ достигает своего максимума при $\pi_{t+1,j}=\sum_{i=1}^n p_{ij}/n$, а второе — при

$$\mu_{t+1,j}=\frac{\sum_{i=1}^n p_{ij}x_i}{\sum_{i=1}^n p_{ij}},\quad\sigma^2_{t+1,j}=\frac{\sum_{i=1}^n p_{ij}(x_i-\mu_{t+1,j})^2}{\sum_{i=1}^n p_{ij}}.$$

■

Не всегда латентные переменные представляют собой что-то естественное, как в примере выше. Порой их вводят искусственно, но от этого EM-алгоритм не становится менее эффективным.

Пример 4.7 (*робастная регрессия*)**.** В задаче регрессии, которая будет подробно рассмотрена в главе 14, предполагается, что каждое наблюдение равно некоторой линейной комбинации набора признаков с неизвестными коэффициентами $\boldsymbol{\theta}$, к которому под воздействием некоторых факторов прибавляется случайный шум с неизвестным коэффициентом разброса σ:

$$Y_i = \mathbf{x}_i^T\boldsymbol{\theta} + \sigma\varepsilon_i, \quad \mathsf{E}\varepsilon_i = 0, \quad i \in \{1, \ldots, n\}.$$

Обычно величину ошибки ε_i предполагают нормально распределённой, то есть $Y_i \sim \mathcal{N}(\mathbf{x}_i^T\boldsymbol{\theta}, \sigma^2)$. Однако такая модель плохо объясняет данные с выбросами, поэтому в таком случае в качестве распределения ε_i берут что-то с более тяжёлыми хвостами, например, распределение Стьюдента с k степенями свободы. Но всплывает другая проблема: оно не принадлежит экспоненциальному семейству распределению, что затрудняет нахождение ОМП для параметров $(\boldsymbol{\theta}, \sigma)$.

Чтобы применить EM-алгоритм к этой задаче, вспомним, как определяется распределение Стьюдента:

$$\varepsilon_i \stackrel{d}{=} \frac{\xi_i}{\sqrt{\eta_i/k}}, \quad \xi_i \sim \mathcal{N}(0,1), \quad \eta_i \sim \chi^2_k = \Gamma(k/2, 1/2).$$

Таким образом, ошибку можно представить как нормально распределённую $\sigma\xi_i \sim \mathcal{N}(0, \sigma^2)$, которую дополнительно поделили на случайный масштаб $\sqrt{\eta_i/k}$. Тогда давайте возьмём $Z_i = \eta_i/k \sim \Gamma(k/2, k/2)$ за латентную переменную. Полное правдоподобие такой модели будет иметь вид

$$\rho_{\boldsymbol{\theta},\sigma}(\mathbf{y}, \mathbf{z}) = \prod_{i=1}^n \rho(z_i) \underbrace{\rho_{\boldsymbol{\theta},\sigma}(y_i|z_i)}_{\sim\mathcal{N}(\mathbf{x}_i^T\boldsymbol{\theta}, \sigma^2/z_i)} = \prod_{i=1}^n \frac{(k/2)^{k/2} z_i^{k/2-1} e^{-kz_i/2}}{\Gamma(k/2)} \cdot \frac{e^{-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2}{2\sigma^2/z_i}}}{\sqrt{2\pi\sigma^2/z_i}}.$$

Согласно E-шагу, нам нужно найти матожидание его логарифма по $\mathbf{z} \sim \rho_{\boldsymbol{\theta}_t,\sigma_t}(\mathbf{z}|\mathbf{y})$ при фиксированных видимых переменных $\mathbf{Y}$. По формуле Байеса,

$$\rho_{\boldsymbol{\theta}_t,\sigma_t}(\mathbf{z}|\mathbf{y}) \sim \rho_{\boldsymbol{\theta}_t,\sigma_t}(\mathbf{y}, \mathbf{z}) \sim \prod_{i=1}^n \underbrace{z_i^{k/2-1/2} \exp\left[-z_i \cdot \left(\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\theta}_t)^2}{2\sigma_t^2} + \frac{k}{2}\right)\right]}_{\sim\Gamma((k+1)/2, (y_i - \mathbf{x}_i^T\boldsymbol{\theta}_t)^2/(2\sigma_t^2) + k/2)}.$$

Находим искомое матожидание, одновременно скидывая в константу C всё, что не зависит от значений $(\boldsymbol{\theta}, \sigma)$:

$$\mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t,\sigma_t}(\mathbf{z}|\mathbf{x})} \ln \rho_{\boldsymbol{\theta},\sigma}(\mathbf{x}, \mathbf{z}) = \underbrace{\sum_{i=1}^n \mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t,\sigma_t}(\mathbf{z}|\mathbf{x})} \ln \rho(z_i)}_{=\text{const}} + \sum_{i=1}^n \mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t,\sigma_t}(\mathbf{z}|\mathbf{x})} \ln \rho_{\boldsymbol{\theta},\sigma}(y_i|z_i) =$$

$$= C + \sum_{i=1}^n \mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t,\sigma_t}(\mathbf{z}|\mathbf{x})} \left(\ln\sqrt{z_i} - \ln\sqrt{2\pi\sigma^2} - z_i \cdot \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2}{2\sigma^2}\right) =$$

$$= C' - n\ln\sqrt{2\pi\sigma^2} - \sum_{i=1}^n \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2}{2\sigma^2} \cdot \mathsf{E}_{\mathbf{z}\sim\rho_{\boldsymbol{\theta}_t,\sigma_t}(\mathbf{z}|\mathbf{x})} z_i =$$

$$= C' - n\ln\sqrt{2\pi\sigma^2} - \sum_{i=1}^n \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2}{2\sigma^2} \cdot \frac{k+1}{(y_i - \mathbf{x}_i^T\boldsymbol{\theta}_t)^2/\sigma_t^2 + k}.$$

Проведя максимизацию по $(\boldsymbol{\theta}, \sigma)$ или вспоминая формулу оценки для взвешенного МНК из будущего раздела 14.2, находим новые значения параметров:

$$\boldsymbol{\theta}_{t+1} = (X^T D X)^{-1} X^T D Y, \quad \sigma^2_{t+1} = \frac{\|D(Y - \boldsymbol{\theta}_{t+1} X)\|^2}{n},$$

где D — диагональная матрица с $D_{ii} = \frac{k+1}{(y_i - \mathbf{x}_i^T \boldsymbol{\theta}_t)^2/\sigma_t^2 + k}$.

Несмотря на некоторую искусственность, которая пропитана введением Z_i, в результате получается предельно интерпретируемый и логичный результат: обычная линейная регрессия превращается в взвешенную, причём размер веса зависит от того, насколько сильно выбивается наблюдение из общей тенденции. Чем больше разница между текущим предсказанием $\mathbf{x}_i^T \boldsymbol{\theta}_t$ и наблюдаемым значением y_i, тем меньше вес D_{ii}, поэтому и вклад этого наблюдения приуменьшается, а степень этой регуляризации определяется гиперпараметром k. Грубо говоря, модель сама понимает, какие наблюдения являются выбросами, а какие нет. ■

Задачи

Задача 4.1. *Категориальное распределение* является обобщением распределение Бернулли: для него с. в. принимает не 2, а уже k значений:

$$\mathsf{P}_{p_1,\ldots,p_k}(X_1 = i) = p_i, \quad \text{где } p_i > 0,\ i \in \{1, \ldots, k\},\ \sum p_i = 1.$$

Для выборки $\mathbf{X} = (X_1, \ldots, X_n)$ из категориального распределения постройте ОМП для вектора параметров $(p_1, \ldots, p_k)$.

Задача 4.2. Рассмотрим выборку $X_1, Y_1, \ldots, X_n, Y_n$, где $X_i, Y_i \sim \mathcal{N}(\theta_i, \sigma^2)$, то есть в модели $2n$ наблюдений и $n+1$ параметр — n параметров сдвига и один параметр масштаба. Найдите оценку максимального правдоподобия для параметра σ^2 и покажите, что она смещена и несостоятельна.

Задача 4.3. Докажите, что в примере 4.2 полученная стационарная точка действительно является точкой максимума функции правдоподобия.

Задача 4.4. Докажите, что у выборки $\mathbf{X} = (X_1, \ldots, X_n)$ из распределения с плотностью (4) не существует оценки максимального правдоподобия для вектора параметров $(\mu_1, \mu_0, \sigma_1^2, \sigma_0^2, \pi)$.

Задача 4.5. Реализуйте EM-алгоритм для робастной регрессии на вашем любимом языке программирования. Убедитесь, что правдоподобие модели по видимым переменным не убывает. Как ведёт себя правдоподобие и оценки параметров при изменении гиперпараметра k (например, когда данные приходят из модели с $k = 3$, а алгоритм применяется для $k = 1; 10; \ldots$)?

5 Достаточные статистики

5.1 Улучшение оценок

Рассмотрим подробнее статистики, сохраняющие информацию о неизвестном параметре при редукции данных, которых мы коснулись ранее в разделе 3.5.

Определение. σ-алгебра $\mathcal{G} \subset \mathcal{B}(\mathcal{X})$ называется *достаточной* для семейства распределений $\mathcal{P}_\theta$, если существует вариант условного распределения $\mathsf{P}_\theta(X \in B \,|\, \mathcal{G})$, которое не зависит от θ. Статистика $S(\mathbf{X})$ называется *достаточной*, если достаточна σ-алгебра $\sigma(S(\mathbf{X}))$, или, что эквивалентно, существует вариант условного распределения $\mathsf{P}_\theta(X \in B \,|\, S(\mathbf{X}))$, которое не зависит от θ.

Неформально смысл такого определения можно понимать так. Представим себе машину-генератор случайных чисел, который выдаёт нам вектор в виде выборки. С одной стороны, «розыгрыш» значения вектора выборки происходит в соответствии совместной функции распределения наблюдения, а с другой — его можно разделить на два этапа: сначала выбираем значение t для достаточной статистики $S(\mathbf{X})$, а после этого – само значение $\mathbf{X}$ в соответствии с условным распределением $\mathsf{P}_\theta(\mathbf{X} \in B | S(\mathbf{X}) = t)$. Так как оно не зависит от параметра θ, то вся информация о нём хранится в первом этапе. Таким образом, то, какое именно значение $\mathbf{X}$ доставляет равенство $S(\mathbf{X}) = t$, нас не особо интересует в отличие от самого $S(\mathbf{X})$.

Проверять статистику на достаточность по определению весьма затруднительно, поэтому нам на выручку приходит следующая

Теорема 5.1 (критерий факторизации).

Пусть $\mathcal{P}_\theta$ — доминируемое семейство распределений с обобщённой плотностью ρ_θ. Тогда $S(\mathbf{X})$ — 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 в конце раздела 3.5).

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

$$\rho_\theta(\mathbf{x}) = h(\mathbf{x}) \exp\left(a_0(\theta) + \sum_{i=1}^{k} a_i(\theta) T_i(\mathbf{x})\right), \tag{5}$$

и из критерия факторизации получаем требуемое.

Пример 5.1. Рассмотрим пример достаточной статистики не для экспоненциального семейства. Введём модель с равномерным распределением $\mathrm{U}(0, \theta)$, 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

$$\rho_\theta(\mathbf{x}) = \frac{1}{\theta^n} I(0 < x_i < \theta,\ i = 1, \ldots, n) = I(0 < x_{(1)}) \cdot \frac{I(x_{(n)} < \theta)}{\theta^n} = h(\mathbf{x}) \cdot g(T(\mathbf{x}), \theta),$$

где $h(\mathbf{x}) = I(0 < x_{(1)})$, $g(t, \theta) = \frac{I(t<\theta)}{\theta^n}$, $T(\mathbf{x}) = x_{(n)}$. Таким образом, $X_{(n)}$ является

достаточной статистикой по критерию факторизации. ■

Один из главных плюсов достаточных статистик заключается в том, что они позволяют «улучшать» оценки в среднеквадратичном подходе.

Теорема 5.2 (Колмогоров, Блэкуэлл, Рао).

Пусть $S(\mathbf{X})$ является достаточной статистикой для семейства $\mathcal{P} = \{\mathsf{P}_\theta \colon \theta \in \Theta\}$, и $\theta^*(\mathbf{X})$ — несмещённая оценка $\tau(\theta)$. Тогда $T(\mathbf{X}) = \mathsf{E}_\theta(\theta^*(\mathbf{X}) \,|\, S(\mathbf{X}))$ является статистикой и несмещённой оценкой $\tau(\theta)$, причём она «лучше» исходной в среднеквадратичном смысле, то есть

$$\mathsf{E}_\theta(T - \tau(\theta))^2 \leqslant \mathsf{E}_\theta(\theta^* - \tau(\theta))^2, \quad \theta \in \Theta.$$

Доказательство. Первое заключение теоремы дано не зря: совершенно не очевидно, что УМО будет вообще статистикой — вдруг её значение зависит от неизвестного параметра? В общем случае это, вообще говоря, вполне возможно (см. задачу 5.8), но оказывается, для достаточных статистик всё хорошо. Действительно, УМО можно записать как матожидание по условной мере (см. §7, гл. II, [11]):

$$\mathsf{E}_\theta(\theta^*(\mathbf{X}) \,|\, S(\mathbf{X})) = \int \theta^*(x)\, \mathsf{P}_\theta(dx \,|\, S(\mathbf{X})).$$

Распределение, по которому берётся матожидание, постоянно и не зависит от параметра, поэтому и УМО от него не зависит и, следовательно, является статистикой.

Несмещённость $T(\mathbf{X})$ очевидна из свойств УМО. Последнее заключение можно доказать с помощью неравенства Йенсена. Напомним его: если $g(x)$ — выпуклая вниз функция и $\mathsf{E}|g(\xi)| < \infty$, то почти наверное выполнено неравенство

$$g\left(\mathsf{E}(\xi|\eta)\right) \leqslant \mathsf{E}(g(\xi)|\eta).$$

Возьмём произвольный $\theta \in \Theta$, функцию $g(x) = (x - \tau(\theta))^2$ и $\xi = \theta^*(\mathbf{X})$. Тогда по неравенству Йенсена

$$\left(\mathsf{E}_\theta(\theta^* \,|\, S(\mathbf{X})) - \tau(\theta)\right)^2 \leqslant \mathsf{E}_\theta\left[(\theta^* - \tau(\theta))^2 \,|\, S(\mathbf{X})\right]$$

Беря от обеих частей неравенства матожидание, получаем требуемое. □

5.2 Оптимальные оценки

Оказывается, при некоторых дополнительных ограничениях достаточная статистика может дать нам не просто оценку лучше прежней, так ещё и *лучшую* в среднеквадратичном подходе.

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

Сразу выделим важное свойство таких оценок.

Утверждение 5.1. *Если $\widehat{\theta}(\mathbf{X})$ и $\theta^*(\mathbf{X})$ — две оптимальные оценки, то они равны P_θ-п.н. для любого θ.*

Доказательство. 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 $4\mathsf{E}_\theta\,(\theta^*)^2 \leqslant \mathsf{E}_\theta(\widehat{\theta}+\theta^*)^2$, что при сложении даёт $2\mathsf{E}_\theta\widehat{\theta}^2 + 2\mathsf{E}_\theta\,(\theta^*)^2 \leqslant 4\mathsf{E}_\theta\widehat{\theta}\theta^*$. 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. □

Таким образом, оптимальная оценка не более чем единственна. Существование оной можно показать с помощью следующего понятия.

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

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

Определение по сути говорит, что статистика $S(\mathbf{X})$ выражает параметр единственным образом, то есть вы не можете двумя разными способами несмещённо оценить функцию от параметра с помощью $S(\mathbf{X})$, так как иначе матожидание их разности даст нуль, который по определению полной статистики несмещённо оценивается лишь нулём. Весьма полезной для нахождения полных достаточных статистик оказывается следующая

Теорема 5.3.

Пусть $\theta\in\Theta\subset\mathbb{R}^k$ и для семейства $\mathcal{P}_\theta$ выполняется (5). Пусть кроме того множество значений $(a_1(\theta),\dots,a_k(\theta))$ для $\theta\in\Theta$ содержит внутреннюю точку. Тогда $T(\mathbf{X})$ является полной достаточной статистикой.

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

Теорема 5.4 (Леман, Шеффе).

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

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

Пример 5.2. По выборке из распределения $\mathcal{N}(a,\sigma^2)$ построим оптимальную оценку для вектора параметров $\theta=(a,\sigma^2)$. Распишем более подробно совместную плотность 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 $T(\mathbf{X}) = (\sum X_i^2, \sum X_i)$ является достаточной, а следовательно

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

будут являться оптимальными оценками. ■

Пример 5.3. По выборке размера $n \geqslant 2$ из распределения $\mathrm{Exp}(\lambda)$ найдём оптимальную оценку для параметра λ. Так как $\mathsf{E}_\theta \overline{\mathbf{X}} = \frac{1}{\lambda}$, то логично предположить, что $1/\overline{\mathbf{X}}$ даст нам что-то подходящее. Проверим эту догадку. Так как $X_i \sim \mathrm{Exp}(\lambda) \sim \Gamma(1, \lambda)$, то $\sum X_i \sim \Gamma(n, \lambda)$ (в силу независимости X_i). Это в свою очередь означает, что

$$\mathsf{E}_\theta \left(\frac{1}{\sum X_i} \right) = \int_0^{+\infty} \frac{1}{x} \frac{\lambda^n x^{n-1}}{\Gamma(n)} e^{-\lambda x}\, dx =$$

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

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

Возникает логичный вопрос: а зачем условие $n \geqslant 2$? Что будет, если рассмотреть случай $n = 1$? Оказывается, что оптимальной оценки в данном случае просто нет. Действительно, пусть существует оптимальная $T(X_1)$ такая, что $\mathsf{E}_\lambda T(X_1) = \lambda$. В данной модели выполняются условия регулярности, причём так как $T(X_1)$ — оптимальна, то у неё существует второй момент, поэтому можно применить свойство С3 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. ■

Как можно видеть, теорема не всегда даёт положительный результат, так как несмещённой оценки может просто не быть. Но есть и другая проблема — для некоторых семейств распределений нет полной достаточной статистики. От такого и вправду никто не застрахован: чтобы статистика была достаточной, она должна быть достаточно «жирной» в плане информации, в то время как полная статистика наоборот — скромной, чтобы вдруг не появилось других способов несмещённо оценить нуль.

Пример 5.4. Рассмотрим семейство распределений $\mathcal{N}(\theta, \theta^2)$, где $\theta > 0$. 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 $(\sum X_i^2, \sum X_i)$ выводится из выражения параметра θ двумя способами:

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

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

при этом выражение под знаком матожидания не равно нулю почти наверное.

Пусть нашлась достаточная статистика $T(\mathbf{X})$, то есть по критерию факторизации

$\rho_\theta(\mathbf{x}) = g(T(\mathbf{x}), \theta) \cdot h(\mathbf{x})$. Также мы знаем, что

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

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 $h(\mathbf{x})$, и убрать, например, $\sum x_i$, оставив наедине $T(\mathbf{x})$ и $\sum \mathbf{x}_i^2$:

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

$$\sum x_i^2 = -\ln\left(\frac{g(T(\mathbf{x}),1)\cdot g(T(\mathbf{x}),1/3)}{Cg(T(\mathbf{x}),1/2)\cdot g(T(\mathbf{x}),1/2)}\right)$$

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

Это и показывает тот факт, что $T(\mathbf{X})$ 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, § 23 в [7]).

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

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

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

откуда и получаем заветное противоречие. ■

5.3 Статистика помогает теории вероятностей

Занятно, что полученные результаты применимы в задачах по классической теории вероятностей, в которых нет вероятностно-статистической модели и выборочного пространства, но из доказательства фактов для нашей модели будет следовать их справедливость вообще.

Первое приложение заключается в новом способе нахождения УМО. Идея в следующем: пусть нам нужно найти УМО некоторой случайной величины ξ при условии полной достаточной статистики $T(\mathbf{X})$. Тогда по теореме Лемана-Шеффе это УМО будет оптимальной оценкой матожидания ξ, причём, как известно, УМО есть борелевская функция от условия $\varphi(T(\mathbf{X}))$. Таким образом, если мы подберём φ так, чтобы $\mathsf{E}\varphi(T(\mathbf{X})) = \mathsf{E}\xi$, то из единственности оптимальной оценки будет следовать, что $\varphi(T(\mathbf{X}))$ есть искомое УМО.

Пример 5.5. Пусть $X_1, \ldots, X_n$ — н.о.р.с.в. из распределения $\mathrm{Pois}(\lambda)$. Найдём $\mathsf{E}(X_1^2 | X_1 + \ldots + X_n)$. Заметим, что никакой статистики тут нет — вероятностное пространство, на котором заданы величины, произвольно, а параметр λ фиксирован, и его не надо оценивать. Но УМО однозначно определяется совместным распределением аргумента и условия, поэтому достаточно найти его в модели масштаба, когда семейство распределений есть $\mathcal{P} = \{\mathrm{Exp}(\lambda) \colon \lambda \in \mathbb{R}_+\}$.

Из сказанного выше легко понять, что $\sum X_i$ является полной достаточной статистикой,

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:

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

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

■

Но по-настоящему крутой и полезный результат, который будет использован далее, сформулирован в следующем утверждении.

Теорема 5.5 (Басу).

Пусть $S(\mathbf{X})$ — полная достаточная статистика, $A(\mathbf{X})$ — статистика, распределение которой одинаково при всех $\theta \in \Theta$ (*англ.* `ancillary statistic`). Тогда статистики $A(\mathbf{X})$ и $S(\mathbf{X})$ независимы при любом $\theta \in \Theta$.

Доказательство. Проведём доказательство в лоб: покажем независимость событий из σ-алгебр, порождённых S и A. Пусть $C \in \sigma(A)$, то есть $\exists B \in \mathcal{B}(\mathbb{R}^n)$: $A^{-1}(B) = C$. Рассмотрим $I_B \circ A(\mathbf{X}) = I(\mathbf{X} \in C)$. Её распределение также не зависит от θ, так как определяется распределением $A(\mathbf{X})$. Это значит, что $\mathsf{E}_\theta I(\mathbf{X} \in C)$ является некоторой константой, независящей от θ. То есть $I(\mathbf{X} \in C)$ является несмещённой оценкой константы $\mathsf{E}_\theta I(\mathbf{X} \in C)$. 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

$$\mathsf{E}_\theta(I(\mathbf{X} \in C) | S(\mathbf{X})) = \mathsf{E}_\theta I(\mathbf{X} \in C).$$

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

$$\int_D I(\mathbf{x} \in C)\, \mathsf{P}_\theta(d\mathbf{x}) = \int_D \mathsf{E}_\theta I(\mathbf{X} \in C)\, d\mathsf{P}_\theta.$$

Но первый интеграл равен $\int I(\mathbf{x} \in C \cap D)\, \mathsf{P}_\theta(d\mathbf{x}) = \mathsf{P}_\theta(\mathbf{X} \in C \cap D)$, а второй — $\mathsf{E}_\theta I(\mathbf{X} \in C) \cdot \int I(\mathbf{x} \in D)\, \mathsf{P}_\theta(d\mathbf{x}) = \mathsf{P}_\theta(\mathbf{X} \in C) \cdot \mathsf{P}_\theta(\mathbf{X} \in D)$, что и требовалось. □

Продемонстрируем работу данной теоремы.

Пример 5.6. Докажем, что статистики $\overline{\mathbf{X}}$ и s^2, построенные по выборке из нормального распределения, независимы. Рассмотрим модель сдвига $\mathcal{N}(a, \sigma^2)$, где a – параметр, а σ^2 – *известная* величина. Распишем плотность, как в примере 5.2:

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

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

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

где $T(\mathbf{x}) = \sum x_i$. Применяем критерий факторизации и понимаем, что $\overline{\mathbf{X}}$ – достаточная статистика. Она же будет полной, так как модель принадлежит экспоненциальному семейству, и функция перед $T((\mathbf{X})X)$ 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 $a+\sigma Y_i$, где Y_i имеет стандартное нормальное распределение. Но тогда

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

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

Пример 5.7. Пусть X_1, X_2, X_3 — н.о.р.с.в. из распределения $\mathrm{Exp}(\lambda)$. Докажем, что

$$X_1 + X_2 + X_3 \perp\!\!\!\perp \frac{X_1}{X_1 + X_2 + X_3}.$$

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. ■

Задачи

Задача 5.1. Покажите, что достаточная статистика из примера 5.1 полна.

Задача 5.2. Найдите достаточную статистику в модели сдвига-масштаба для экспоненциального распределения, то есть для семейства

$$\mathcal{P} = \left\{\mathsf{P}_{\alpha,\beta} \colon F_{\mathsf{P}_{\alpha,\beta}}(t) = F_0\left(\frac{t-\alpha}{\beta}\right)\right\}, \quad F_0(t) = 1 - e^{-t}.$$

Будет ли эта статистика полной?

Задача 5.3. Докажите теорему 5.4.

Задача 5.4. В условиях задачи 1.4 при помощи полученной в ней несмещённой оценки найти оптимальную оценку функции $e^{-\theta}$.

Задача 5.5. Пусть $X_1, \ldots, X_n$ — н.о.р.с.в. из распределения $\mathrm{U}(0, \theta)$. Найдите $\mathsf{E}(X_{(k)}|X_{(1)})$.

Задача 5.6 (*Я.Профи*)**.** Пусть $\xi_1, \ldots, \xi_{10}$ — н.о.р.с.в. из распределения $\mathrm{U}(0, 11)$. Обозначим $X = \min(\xi_1, \ldots, \xi_{10})$, $Y = \max(\xi_1, \ldots, \xi_{10})$. Найдите $\mathsf{D}(X - 4Y)$.

Задача 5.7. Пусть $X_1, \ldots, X_n$ — выборка из одномерного нормального распределения с

неизвестными параметрами сдвига и масштаба. Напомним, что величина

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{\mathbf{X}}\right)^k, \quad k \in \mathbb{N}$$

называется выборочным центральным моментом. Выборочные асимметрия и эксцесс — это статистики μ_3/s^3 и $\mu_4/s^4 - 3$, где s^2 — выборочная дисперсия. Доказать, что каждая из них не зависит от s^2.

Задача 5.8. Приведите пример параметрического семейства $\mathcal{P} = \{\mathsf{P}_\theta \colon \theta \in \Theta\}$ и оценок $T(\mathbf{X})$ и $\widehat{\theta}(\mathbf{X})$ параметра θ таких, что $\mathsf{E}_\theta\left(\widehat{\theta}(\mathbf{X})\middle| T(\mathbf{X})\right)$ не является статистикой.

Задача 5.9. Докажите, что если $T(\mathbf{X})$ — достаточная статистика, то оценка максимального правдоподобия является функцией от $T(\mathbf{X})$.

6 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 $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ такая, что для любого $\theta \in \Theta$

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

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

Такой подход удобен и тем, что длина интервала будет показывать нашу уверенность в оценке, что совершенно не делает обычная точечная оценка. К тому же мы сами вольны выбирать, с каким уровнем доверия рассматривать доверительный интервал. Обычно на практике берут $\gamma = 0.9$, 0.95 или 0.99.

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

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

то $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ 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, $T_2^{(n)}(\mathbf{X}) - T_1^{(n)}(\mathbf{X})$ должно стремиться к 0 для всех θ (если это возможно).

Пример 6.1. Реализуем способ построения интервалов, который первый приходит в голову. Нам по сути нужно ограничить снизу вероятность нахождения неизвестного параметра внутри какой-то окрестности, или, что эквивалентно, ограничить сверху вероятность слишком сильного расхождения. У нас уже есть инструмент из курса теории вероятностей, который помогает в такой оценке — неравенство Чебышёва.

Рассмотрим выборку $X_1, \ldots, X_n \sim \mathrm{Bern}(p)$, где $p \in (0; 1)$ — неизвестный параметр. Так как $\mathsf{E}_p\overline{\mathbf{X}} = p$, то по неравенству Чебышёва

$$\mathsf{P}_p\left(|\overline{\mathbf{X}} - p| \geqslant \varepsilon\right) \leqslant \frac{\mathsf{D}_\theta\overline{\mathbf{X}}}{\varepsilon^2} = \frac{\mathsf{D}_p X_1}{n\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \leqslant \frac{1}{4n\varepsilon^2}.$$

Если нам нужно найти доверительный интервал уровня доверия γ, то полученное ограничение на вероятность ошибки должно равняться $1-\gamma$, то есть $\varepsilon = \frac{1}{2\sqrt{n(1-\gamma)}}$. Таким образом,

$$\mathsf{P}_p\left(\overline{\mathbf{X}} - \frac{1}{2\sqrt{n(1-\gamma)}} < p < \overline{\mathbf{X}} + \frac{1}{2\sqrt{n(1-\gamma)}}\right) \geqslant \gamma.$$

Мы получили интервал длины порядка $1/\sqrt{n}$, и, как будет ясно далее, эта асимптотика неулучшаема.

Впрочем, для более сложных распределений ситуация не столь благоприятная: выскакивающая $\mathsf{D}_\theta X_1$ скорее всего зависит от неизвестного параметра θ, от которого избавиться будет не так просто, в отличие от данного примера. Это мотивирует нас искать более удачные подходы в нахождении ДИ.

■

6.1 Методы построения интервалов

Метод I. Центральная функция

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

1. распределение $G(\mathbf{X}, \theta)$ 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 $G(\mathbf{X}, \theta)$ через x_p и возьмём $0 \leqslant p_1 < p_2 \leqslant 1$ такие, что $p_2 - p_1 = \gamma$. Определим $T_1(\mathbf{x})$ и $T_2(\mathbf{x})$ как решения относительно θ соответственно уравнений $G(\mathbf{x}, \theta) = x_{p_1}$ и $G(\mathbf{x}, \theta) = x_{p_2}$. Наличие и единственность решения следует из условия 2 определения выше. Тогда из монотонности по θ функции $G(\mathbf{x}, \theta)$ получаем, что

$$\mathsf{P}_\theta\big(T_1(\mathbf{X}) < \theta < T_2(\mathbf{X})\big) = \mathsf{P}_\theta\big(x_{p_1} < G(\mathbf{X}, \theta) < x_{p_2}\big) = p_2 - p_1 = \gamma.$$

Удобно брать $p_2 = \frac{1+\gamma}{2}$ и $p_1 = \frac{1-\gamma}{2}$ (в таком случае интервал называется *центральным*), особенно когда распределение $G(\mathbf{X}, \theta)$ симметрично относительно начала координат.

Пример 6.2. Пусть $X_1, \ldots, X_n \sim \Gamma(\alpha, \lambda)$, где α — известная величина. Построим точный доверительный интервал уровня доверия γ для параметра λ.

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 $\Gamma(n\alpha, 1)$. ■

Пример 6.3. Пусть теперь $X_1, \ldots, X_n \sim \mathrm{U}(0, \theta)$. Построим доверительный интервал для параметра θ.

Рассмотрим функцию $G(\mathbf{X}, \theta) = \frac{X_{(n)}}{\theta}$. Она будет центральной функцией, поскольку

$$\forall t \in [0, 1]\colon \mathsf{P}_\theta(G(\mathbf{X}, \theta) \leqslant t) = \mathsf{P}_\theta(X_{(n)} \leqslant t\theta) = \frac{(t\theta)^n}{\theta^n} = t^n,$$

значит, её распределение не зависит от θ, а выполнение условия 2 очевидно. Из этого представления легко найти квантиль распределения: $x_p = \sqrt[n]{p}$. Таким образом,

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

Немаловажно будет посмотреть на асимптотику длины полученного интервала. Примем

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

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

■

Не всегда очевидно, как найти хоть какую-нибудь центральную функцию, а вдруг её вообще нет? Однако в широком наборе случаев такая функция всё-таки существует, причём часто она имеет относительно приемлемый вид. Для её нахождения нам потребуется следующее вспомогательное

Утверждение 6.1. *Пусть ξ — случайная величина с непрерывной функцией распределения $F(x)$. Тогда $F(\xi) \sim \mathrm{U}(0,1)$.*

Доказательство. Для $t \in (0,1)$ определим $F^{-1}(t)$ как $\sup\{x\colon F(x) = t\}$ (множество, по которому берётся супремум, не пусто в силу непрерывности). Тогда из неравенства $F(\xi) \leqslant t$ следует $\xi \leqslant F^{-1}(t)$. В обратную сторону импликация выполняется почти наверное: вероятность события

$$\inf\{x\colon F(x) = t\} < \xi \leqslant \sup\{x\colon F(x) = t\}$$

равна нулю, так как функция распределения ξ одинакова при левой и правой части сего неравенства. Значит,

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

что есть функция распределения $\mathrm{U}(0,1)$. □

Таким образом, если для любого θ функция распределения элементов выборки $F_\theta(x)$ непрерывна, то вне зависимости от значения θ статистики $F(X_i)$ распределены равномерно. Смастерим из них одну статистику. Складывать равномерные величины — дело неблагодарное, поэтому предварительно возьмём от них функцию $\varphi(t) = -\ln t$. Под действием φ распределение становится экспоненциальным. Действительно, если раньше плотность равнялась $\rho(t) = I(0 < t < 1)$, то теперь она равна

$$\tilde\rho(x) = |(\varphi^{-1}(x))'| \cdot \rho(\varphi^{-1}(x)) = e^{-x} I(0 < e^{-x} < 1) = e^{-x} I(x > 0),$$

то есть $-\ln F_\theta(X_i) \sim \mathrm{Exp}(1) = \Gamma(1,1)$. Отсюда $G(\mathbf{X},\theta) = -\sum \ln F_\theta(X_i)$ как сумма независимых случайных величин распределена как $\Gamma(n,1)$. Если к тому же нам известно, что при фиксированном x функция $F_\theta(x)$ строго монотонна и непрерывна по θ, то полученная функция будет центральной.

Пример 6.4. Пусть $X_1, \ldots, X_n \sim \mathrm{Pareto}(\theta, 1)$, где $\theta > 0$. Построим точный доверительный интервал уровня доверия γ для параметра θ. Для распределения Парето функция распределения имеет вид $F_\theta(t) = 1 - t^{-\theta}$, $t \geqslant 1$. Для упрощения последующих выкладок заметим, что $1 - F_\theta(X_i)$ в силу симметричности распределена также равномерно на $[0,1]$, поэтому статистика

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

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

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

■

Метод II. Использование оценки дисперсии Допустим на нас с неба свалилась асимптотически нормальная оценка θ^*, то есть $\sqrt{n}(\theta^* - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta))$ 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 z_p p-квантиль для $\mathcal{N}(0,1)$, то из сходимости по распределению следует

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

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 записать интервал как

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

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

Пример 6.5. Рассмотрим модель сдвига $X_1, \ldots, X_n \sim \text{Cauchy}(\theta, 1)$. 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:

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

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

Пример 6.6. Теперь пусть $X_i \sim \text{Pois}(\lambda)$, где λ – неизвестный параметр, для которого нужно построить ДИ. По ЦПТ имеем

$$\sqrt{n} \cdot \frac{\overline{\mathbf{X}} - \lambda}{\lambda} \xrightarrow{d_\lambda} \mathcal{N}(0,1),$$

и после замены дисперсии λ на её состоятельную оценку $\widehat{\lambda} = \overline{\mathbf{X}}$ получаем асимптотический интервал

$$\left(\overline{\mathbf{X}} - z_{(1+\gamma)/2}\sqrt{\frac{\overline{\mathbf{X}}}{n}},\ \overline{X} + z_{(1+\gamma)/2}\sqrt{\frac{\overline{\mathbf{X}}}{n}}\right).$$

■

Метод III. Стабилизация дисперсии Наконец, рассмотрим немного усовершенствованный способ, который можно применять в некоторых частных случаях, когда асимптотическая дисперсия является функцией от оцениваемого параметра:

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

Для начала отметим главный недостаток предыдущего метода — он дополнительно загрубляет доверительный интервал при подстановке оценки дисперсии, что непредсказуемо меняет вероятность накрытия интервалом истинного значения параметра. Хотелось бы избавиться от зависимости асимптотической дисперсии от параметра, заменив её на что-то константное. Благо у нас есть инструмент, который позволяет видоизменять дисперсию — это дельта-метод. Идея заключается в том, чтобы перейти от исходной оценки к некоторой хорошей функции от неё, а именно первообразной от обратного стандартного отклонения:

$$\psi(t)=\int\frac{dt}{\sigma(t)}.$$

В таком случае, применяя дельта-метод, приходим к следующей сходимости:

$$\sqrt{n}\cdot(\psi(\theta_n^*)-\psi(\theta))\xrightarrow{d_\theta}\mathcal{N}(0,\sigma^2(\theta)\cdot\psi'(\theta)^2)=\mathcal{N}\left(0,\sigma^2(\theta)\cdot\left(\frac{1}{\sigma(\theta)}\right)^2\right)=\mathcal{N}(0,1).$$

Отсюда получаем асимптотический ДИ для $\psi(\theta)$, который посредством взятия обратной функции можно превратить в ДИ для θ:

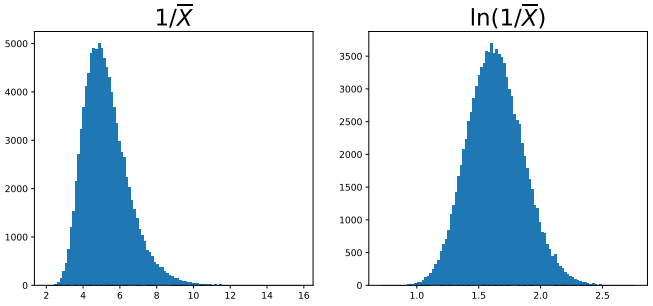
$$\gamma=\lim_{n\to\infty}\mathsf{P}_\theta\left(\psi(\theta_n^*)-\frac{z_{(1+\gamma)/2}}{\sqrt{n}}<\psi(\theta)<\psi(\theta_n^*)+\frac{z_{(1+\gamma)/2}}{\sqrt{n}}\right)=$$
$$=\lim_{n\to\infty}\mathsf{P}_\theta\left(\psi^{-1}\left[\psi(\theta_n^*)-\frac{z_{(1+\gamma)/2}}{\sqrt{n}}\right]<\theta<\psi^{-1}\left[\psi(\theta_n^*)+\frac{z_{(1+\gamma)/2}}{\sqrt{n}}\right]\right).$$

В заключение скажем, что последний переход корректен в силу строгого возрастания функции ψ, ведь её производная положительна.

Пример 6.7. Рассмотрим выборку $X_1,\ldots,X_n\sim\mathrm{Exp}(\theta)$. Из примера 1.4 нам известная асимптотически нормальная оценка

$$\sqrt{n}\left(\frac{1}{\overline{\mathbf{X}}}-\theta\right)\xrightarrow{d_\theta}\mathcal{N}(0,\theta^2),$$

с помощью которой легко построить асимптотический ДИ по методу II:



$$\lim_{n\to\infty}\mathsf{P}_\theta\left(\frac{1}{\overline{\mathbf{X}}}-\frac{z_{(1+\gamma)/2}}{\sqrt{n}\overline{\mathbf{X}}}<\theta<\frac{1}{\overline{\mathbf{X}}}+\frac{z_{(1+\gamma)/2}}{\sqrt{n}\overline{\mathbf{X}}}\right)=\gamma.$$

Усовершенствуем её, применив стабилизацию дисперсии. В данном случае

$$\psi(t)=\int\frac{dt}{\sigma(t)}=\int\frac{dt}{t}=\ln(t).$$

Отсюда получаем заветный интервал

$$\lim_{n\to\infty}\mathsf{P}_\theta\left(\exp\left[-\ln\overline{\mathbf{X}}-\frac{z_{(1+\gamma)/2}}{\sqrt{n}}\right]<\theta<\exp\left[-\ln\overline{\mathbf{X}}+\frac{z_{(1+\gamma)/2}}{\sqrt{n}}\right]\right)=\gamma.$$

■

Отметим важную особенность, свойственную стабилизованным оценками. Несмотря на сходимость к нормальному распределению, для небольших размеров выборки оценки могут слабо походить на что-то нормальное: зачастую их распределение асимметрично, и приближение нормальным законом будет несостоятельным. При стабилизации дисперсии же распределение выравнивается и становится более куполообразным. Для иллюстрации этого эффекта изобразим гистограммы обычной и стабилизованной оценки из предыдущего примера для $\theta = 5$. На первой картинке левый хвост распределения гораздо легче правого, в то время как на второй картинке всё куда приятнее.

6.2 Интервалы для нормального распределения

Ввиду распространённости нормального распределения среди реальных данных особенно полезно уметь строить доверительные интервалы в нормальной модели, когда элементы выборки имеют распределение $\mathcal{N}(a, \sigma^2)$. В принципе, методов выше хватает для построения асимптотических доверительных интервалов, однако в этом разделе мы попробуем построить их точные аналоги, а также познакомимся с некоторыми важными распределениями, которые непосредственно связаны с нормальным и будут встречаться нам ещё не раз.

Модель сдвига Будем считать, что дисперсия элементов выборки известна и равна σ^2, и нашей задачей будет построить доверительный интервал для неизвестного параметра a, который отвечает за сдвиг распределения. Данная задача легко решается методом центральной функции, который был рассмотрен выше. Действительно, если $X_i \sim \mathcal{N}(a, \sigma^2)$, то

$$G(\mathbf{X}, a) = \sqrt{n} \cdot \frac{\overline{\mathbf{X}} - a}{\sigma} \sim \mathcal{N}(0, 1),$$

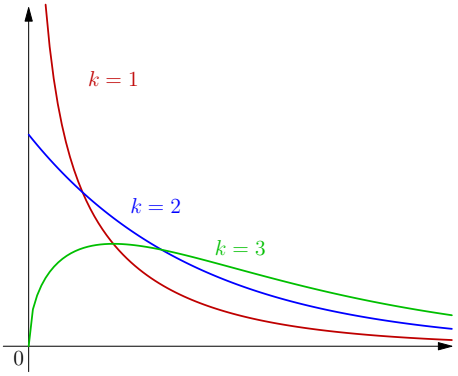
поэтому в качестве границ интервала можно взять статистики $\overline{\mathbf{X}} \pm \frac{\sigma z_{(1+\gamma)/2}}{\sqrt{n}}$, где здесь и далее z_p — p-квантиль стандартного нормального распределения:

$$\mathsf{P}_a\left(\overline{\mathbf{X}} - \frac{\sigma z_{(1+\gamma)/2}}{\sqrt{n}} < a < \overline{\mathbf{X}} + \frac{\sigma z_{(1+\gamma)/2}}{\sqrt{n}}\right) = \mathsf{P}_a\left(-z_{(1+\gamma)/2} < \sqrt{n} \cdot \frac{\overline{\mathbf{X}} - a}{\sigma} < z_{(1+\gamma)/2}\right) = \gamma.$$

Модель масштаба Теперь перейдём к обратному случаю, когда сдвиг a известен заранее, а дисперсию σ^2 предстоит оценить. Тут также не представляет труда найти центральную функцию, на сей раз это будет

$$G(\mathbf{X}, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - a)^2 = \sum_{i=1}^{n}\left(\frac{X_i - a}{\sigma}\right)^2.$$

Заметим, что $(X_i - a)/\sigma\mathcal{N}(0, 1)$, а также слагаемые в сумме выше независимы в совокупности. Значит, распределение $G(\mathbf{X}, \sigma^2)$ фиксировано и не зависит от σ^2. Оно имеет специальное название.



Определение. Пусть $\xi_1, \ldots, \xi_k$ — независимые одинаково распределённые случайные величины, и $\xi_i \sim \mathcal{N}(0, 1)$. Тогда распределение $\eta = \xi_1^2 + \ldots + \xi_k^2$ называют *распределением хи-квадрат* (χ_k^2) *с* k *степенями свободы.*

Не будем подробно останавливаться на свойствах этого распределения, но стоит отметить, что распределение χ^2_k является частным случаем гамма-распределения. И вправду: непосредственно проверяется, что если $\xi \sim \mathcal{N}(0,1)$, то $\xi^2 \sim \Gamma(1/2, 1/2)$, поэтому по свойству гамма-распределения $\chi^2_k = \Gamma(k/2, 1/2)$. Плотности распределения χ^2_k для некоторых k даны на рисунке.

Итак, из вышесказанного следует, что $G(\mathbf{X}, \sigma^2) \sim \chi^2_n$. За $\chi^2_{k,p}$ возьмём p-квантиль распределения χ^2_k. Тогда получаем следующий интервал:

$$\mathsf{P}_{\sigma^2}\left(\frac{1}{\chi^2_{n,(1+\gamma)/2}}\sum_{i=1}^n (X_i - a)^2 < \sigma^2 < \frac{1}{\chi^2_{n,(1-\gamma)/2}}\sum_{i=1}^n (X_i - a)^2\right) =$$
$$= \mathsf{P}_{\sigma^2}\left(\chi^2_{n,(1-\gamma)/2} < G(\mathbf{X}, \sigma^2) < \chi^2_{n,(1+\gamma)/2}\right) = \gamma.$$

Замечание. Подумайте, почему центральная функция из модели сдвига не пригодна для построения доверительного интервала в данном случае.

Модель сдвига-масштаба Перейдём к наименее тривиальному примеру, когда неизвестны оба параметра распределения. Здесь нам здорово поможет результат из примера 5.6, где было показано с помощью теоремы Басу, что $\overline{\mathbf{X}} \perp\!\!\!\perp s^2$. Но также ключевую роль в этой задаче будет играть следующая

Теорема 6.1 (Фишер).

Пусть s^2 — выборочная дисперсия в нормальной модели сдвига-масштаба $X_1, \ldots, X_n \sim \mathcal{N}(a, \sigma^2)$. Тогда

$$\frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Доказательство. Для дальнейшего удобства нормируем наблюдения, а именно представим $X_i = a + \sigma Y_i$, где $Y_i \sim \mathcal{N}(0,1)$. Перепишем выборочную дисперсию:

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

Способ I. Заметим, что $\eta = \left(\sum \frac{Y_i}{\sqrt{n}}\right)^2 \sim \chi^2_1$, так как $\sum \frac{Y_i}{\sqrt{n}} \sim \mathcal{N}(0,1)$, и в то же время

$$\zeta = ns^2/\sigma^2 + \eta = \sum Y_i^2 \sim \chi^2_n.$$

Мы знаем, что $ns^2/\sigma^2 \perp\!\!\!\perp \eta$, так как левая и правая части есть функции от выборочных дисперсии и среднего соответственно. Из свойства хар. функций: $\varphi_\eta(t) \cdot \varphi_{ns^2/\sigma^2}(t) = \varphi_\zeta(t)$, поэтому

$$\varphi_{ns^2/\sigma^2}(t) = \frac{\varphi_\zeta(t)}{\varphi_\eta(t)}.$$

Но так как $\chi^2_n = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$, то в равенство выше можно подставить хар. функцию для гамма-распределения:

$$\varphi_{ns^2/\sigma^2}(t) = \frac{(1-2it)^{-\frac{n}{2}}}{(1-2it)^{-\frac{1}{2}}} = (1-2it)^{-\frac{n-1}{2}},$$

что есть хар. функция для $\Gamma\left(\frac{n-1}{2}, \frac{1}{2}\right) = \chi^2_{n-1}$. Значит, по теореме о единственности ns^2/σ^2 имеет в точности распределение χ^2_{n-1}.

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 $\mathbf{Z} = A\mathbf{Y}$ является гауссовым вектором с нулевым матожиданием и матрицей ковариаций $AEA^T = E$, то есть Z_i — независимые и стандартно нормально распределены. При этом в силу ортогональности длина вектора не меняется, то есть $\sum Z_i^2 = \sum Y_i^2$. Таким образом, выборочная дисперсия выше перепишется как

$$\frac{ns^2}{\sigma^2} = \sum_{i=1}^n Y_i^2 - \left(\sum_{i=1}^n \frac{Y_i}{\sqrt{n}}\right)^2 = \sum_{i=1}^n Z_i^2 - Z_1^2 = \sum_{i=2}^n Z_i^2 \sim \chi^2_{n-1}.$$

□

Отсюда можно легко выписать доверительный интервал для σ^2:

$$\mathsf{P}_{a,\sigma^2}\left(\frac{ns^2}{\chi^2_{n-1,(1+\gamma)/2}} < \sigma^2 < \frac{ns^2}{\chi^2_{n-1,(1-\gamma)/2}}\right) = \mathsf{P}_{a,\sigma^2}\left(\chi^2_{n-1,(1-\gamma)/2} < \frac{ns^2}{\sigma^2} < \chi^2_{n-1,(1+\gamma)/2}\right) = \gamma.$$

Теперь построим интервал для a. Вспомним, с чего мы начинали:

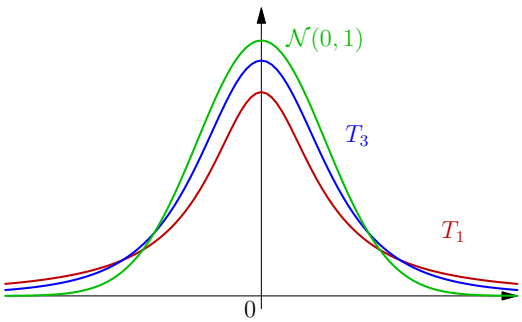
$$\sqrt{n} \cdot \frac{\overline{\mathbf{X}} - a}{\sigma} \sim \mathcal{N}(0,1).$$

Таким образом, если мы поделим эту функцию на корень из ns^2/σ^2, мы избавимся от неизвестной σ^2, оставив лишь неизвестный параметр a, причём в силу независимости выборочных дисперсий и среднего распределение полученной величины будет фиксированным.

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 $\zeta \sim T_m$.



Деление на m обусловлено следующим свойством: $T_m \xrightarrow{d} \mathcal{N}(0,1)$ при $m \to \infty$. Действительно, если $\eta_m = \xi_1^2 + \ldots + \xi_m^2 \sim \chi^2_m$, где $\xi_i \sim \mathcal{N}(0,1)$, то по ЗБЧ $\eta_m/m \xrightarrow{\mathsf{P}} 1$, откуда несложно вывести требуемое по лемме Слуцкого. Однако важно помнить, что распределение Стьюдента имеет более «тяжёлые» хвосты, которые при малых n дают существенные различия в квантилях нормального распределения и Стьюдента. При $m = 1$ это распределение и вовсе будет распределением Коши, у которого чрезвычайно тяжёлые хвосты, поэтому T_m можно считать гладким переходом между этими двумя крайностями.

Возвращаясь к нашей задаче, получаем, что

$$\sqrt{n-1} \cdot \frac{\overline{\mathbf{X}} - a}{s} \sim T_{n-1},$$

откуда находим доверительный интервал для a:

$$\mathsf{P}_{a,\sigma^2}\left(\overline{\mathbf{X}} - \frac{sT_{n-1,(1+\gamma)/2}}{\sqrt{n-1}} < a < \overline{\mathbf{X}} + \frac{sT_{n-1,(1+\gamma)/2}}{\sqrt{n-1}}\right) =$$
$$= \mathsf{P}_{a,\sigma^2}\left(-T_{n-1,(1+\gamma)/2} < \sqrt{n-1}\cdot\frac{\overline{\mathbf{X}}-a}{s} < T_{n-1,(1+\gamma)/2}\right) = \gamma.$$

Задачи

Задача 6.1. По выборке $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \theta^2)$ постройте точный доверительный интервал уровня доверия γ для параметра $\theta \in \mathbb{R}$.

Задача 6.2. По выборке $X_1, \ldots, X_n \sim \chi^2_m$ постройте асимпт. доверительный интервал уровня доверия γ для параметра $m > 0$.

Задача 6.3. По выборке $X_1, \ldots, X_n \sim \mathrm{Bern}(p)$ постройте асимптотический доверительный интервал уровня доверия γ и сравните его с интервалом из примера 6.1.

Задача 6.4. В примере 6.3 найденный интервал, конечно, хороший, но ещё далёк от идеала. С помощью статистики $X_{(n)}$ постройте точный доверительный интервал *наименьшей длины* уровня доверия γ для параметра θ.

Задача 6.5. С помощью стабилизации дисперсии постройте доверительный интервал для p уровня доверия γ по выборке из распределения $\mathrm{Bern}(p)$.

Задача 6.6. Найдите точную *доверительную область* уровня доверия γ для вектора (a, σ^2) в модели сдвига-масштаба для нормального распределения $\mathcal{N}(\mu, \sigma^2)$, то есть такое борелевское множество $B(X_1, \ldots, X_n)$ в $\mathbb{R}^2$, что

$$\forall a \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+ \colon \mathsf{P}_{a,\sigma^2}((a,\sigma^2) \in B) = \gamma,$$

и найдите асимптотику её площади.

Задача 6.7. Рассмотрим одноэлементную выборку из распределения $\mathcal{N}(a, \sigma^2)$ (оба параметра неизвестны). Приведите пример нетривиального доверительного интервала для параметра σ^2 уровня доверия γ.

Часть II
Проверка статистических гипотез

7 Введение в теорию проверки гипотез

На практике часто возникает необходимость делать выводы о неизвестном распределении, из которого пришла наблюдаемая выборка. Рассмотрим некоторые примеры.

- Вы исследуете на работоспособность некоторое лекарство. Для этого вы предлагаете больным принять его и смотрите, как поменялись показатели здоровья у испытуемых (температура, давление и т.д.). Нужно понять, есть ли эффект от лекарства: достаточно ли сильно поменялись эти показатели, чтобы судить о действенности средства?

- Вы разработчик приложения, который хочет привнести некоторое нововведение, и вам нужно понять, не ломает ли оно его работу. Для измерения качества приложения имеется множество метрик (например, среднее время сессии, частота клика на определённый элемент и т.д.), которые крайне желательно не ухудшать. Для этого проводится так называемое *AB-тестирование*, при котором часть пользователей (тестовая, экспериментальная выборка) получает новую функциональность, а у другой части (контрольной выборки) всё остаётся по-старому. Вам известны значения метрик у обеих выборок, и необходимо понять, есть ли значимые отличия между ними (в худшую сторону), или можно предположить, что разница не так велика?

- Вы аналитик поиска, который борется с так называемым «фродом». Мошенники могут с помощью ботов задавать одни и те же запросы в поисковике, чтобы увеличить трафик своего сомнительного сайта. Можно предположить, что распределения по времени у обычного частотного запроса и спама отличаются: первый задаётся равномерно, а второй – примерно в одно и то же время залпом (крайне идеализированная картина, взято для примера). Отсюда появляется потребность для каждого частотного запроса в проверке гипотезы о том, что распределение по времени запроса равномерное.

Во всех этих ситуациях имеется некоторое предположение относительно неизвестного распределения, и ему зачастую в противовес даётся альтернатива, относительно которой нужно проверить состоятельность основного предположения: либо оно допустимо (модель из нулевой гипотезы достаточно хорошо описывает данные), либо есть серьёзные основания считать его неверным (значения наблюдений нетипичны для такой модели). Задача сего раздела — математически строго формализовать процесс принятия решения о том, отвергать ли выдвигаемую гипотезу или нет.

Определение. *Статистической гипотезой* H называют предположение о принадлежности истинного распределения P некоторому классу $\mathcal{P}$. Обозначается как $H\colon \mathsf{P} \in \mathcal{P}$.

Часто класс распределений задаётся некоторыми параметрами: $\mathcal{P} = \{\mathsf{P}_\theta\colon \theta \in \Theta\}$. В таком случае гипотезу можно сформулировать в терминах принадлежности некоторому подмножеству $\Theta_0 \subset \Theta$, что записывается как $H_0\colon \theta \in \Theta_0$.

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 (нулевой) гипотезой*, а гипотеза $H_1\colon \mathsf{P} \in \mathcal{P}_1$ — *альтернативой*. Обозначается как

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

Замечание. Далеко не всегда стоит брать гипотезы такие, что $\mathcal{P}_0 \sqcup \mathcal{P}_1 = \mathcal{P}$. Это может быть вызвано отсутствием интереса к некоторым альтернативам или желанием найти более оптимальные способы проверки гипотезы: во многих ситуациях лучше хорошо отклонять основную гипотезу в частных случаях, чем отклонять средне при всех возможных.

Очевидно, принятие решения о том, отвергать ли H_0 или нет, должно зависеть только от реализации выборки $\mathbf{X}$, поэтому выбор определяется некоторым измеримым множеством $R \subset \mathcal{X}$, при попадании в которое мы должны отвергнуть основную гипотезу:

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

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

Зачастую, это множество можно задать как прообраз луча для некоторой статистики:

$$R = \{\mathbf{x} \in \mathbb{R}^n\colon T(\mathbf{x}) \geqslant c\}.$$

В таком случае $T(\mathbf{X})$ называется *статистикой критерия*, а порог c — *критическим значением*. Тогда отвержение основной гипотезы равносильно выполнению $T(\mathbf{X}) \geqslant c$, то есть принятию статистикой критерия слишком экстремального значения, не свойственного основной гипотезе.

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 неверности H_0 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 книге [6]: основная гипотеза своего рода «подсудимый», и по презумпции невиновности она считается невиновной, то есть верной. Тогда проверка гипотезы есть судебный процесс, на котором мы играем роль прокурора. Наша задача — найти доказательства против H_0 в лице «потерпевшей» H_1. Если мы их не обнаружили, это не значит, что H_0 и вправду «невиновна».

Отметим ещё одну особенность процедуры проверки гипотез. Хотя формально постановка задачи симметрична, часто мы подразумеваем неравнозначность гипотез. Это можно проиллюстрировать следующим хрестоматийным примером.

Пример 7.1. Предположим, что вы работаете в госпитале и проводите анализы на присутствие в организме раковых клеток. По сути, вы 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 вы верно поставили диагноз, то всё хорошо. Иначе вы можете 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* вы не окажете помощь больному человеку и обречёте его на смерть, а в случае *ошибки II рода* вы будете лечить здорового и потеряете 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 выше: 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. Впрочем, в других задачах может представляться логичным минимизировать ошибку II рода, что зависит от предметной области.

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 глупые, а потому что она вполне 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. ■

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

Определение. *Функцией мощности критерия* R называется функция

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

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.

Обычно в качестве α берут число 0.05, то есть в таком случае мы позволяем себе ошибку I рода в 5%, однако в разных отраслях используют и другие, меньшие уровни значимости в зависимости от того, насколько катастрофичны последствия ошибки I рода.

Отныне мы работаем с критериями, у которых можно явно задать уровень значимости α. Среди таковых надо подобрать критерий с как можно меньшей ошибкой II рода, то

есть с максимальной функцией мощности. Тут, как это было при сравнении оценок, возникает проблема сравнения двух функций (как понять, какая лучше?). Возможное решение аналогично: уметь сравнивать только те критерии, мощность одного из которых мажорирует мощность другого.

Определение. Говорят, что *критерий* R_1 *мощнее критерия* R_2, если $\forall \mathsf{P} \in \mathcal{P}_1 \colon \beta(\mathsf{P}, R_1) \geqslant \beta(\mathsf{P}, R_2)$.

Также бывает 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 $\mathbf{X} = (X_1, \ldots, X_n)$ 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 7.2. 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, перед нами встала проблема проверки *односторонней* гипотезы

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

Логично использовать критерий, основанный на статистике $T(\mathbf{X}) = \overline{\mathbf{X}}$, 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, иначе оно вполне допустимо. Подберём число c так, чтобы наш критерий имел уровень значимости α, то есть

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

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

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

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

$$\circledminus\, \Phi\left(\sqrt{n}(\theta - \theta_0) - z_{1-\alpha}\right).$$

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. ■

7.1 Критерий Вальда

В данном разделе мы рассмотрим, наверное, один из самых простых способов проверки *двусторонних* гипотез, то есть гипотез вида

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

Другой его особенностью является тот факт, что точного распределения статистики критерия мы знать не будем, мы в курсе лишь её предельного распределения, отчего и уровень значимости будет устанавливаться лишь в пределе.

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

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

Это вызывает некоторые проблемы, так как настоящие размер и мощность критерия могут отличаться от теоретических, особенно при малом размере выборки, однако обычно такие критерии просты, и искать их гораздо проще точных.

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:

$$T_\theta(\mathbf{X}) = \sqrt{n}\cdot\frac{\widehat{\theta}-\theta}{\widehat{\sigma}} = \sqrt{n}\cdot\frac{\widehat{\theta}-\theta}{\sigma(\theta)}\cdot\frac{\sigma(\theta)}{\widehat{\sigma}} \xrightarrow{d} \mathcal{N}(0,1).$$

Вернёмся к проверке гипотезы. При верности H_0 имеем $T_{\theta_0}(\mathbf{X}) \xrightarrow{d_{\theta_0}} \mathcal{N}(0,1)$, 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

$$\mathsf{P}_{\theta_0}(|T_{\theta_0}(\mathbf{X})| > z_{1-\alpha/2}) =$$
$$= \mathsf{P}_{\theta_0}(T_{\theta_0}(\mathbf{X}) > z_{1-\alpha/2}) + \mathsf{P}_{\theta_0}(T_{\theta_0}(\mathbf{X}) < z_{\alpha/2}) \xrightarrow[\text{из слаб. сх-ти}]{} (1-\Phi(z_{1-\alpha/2})) + \Phi(z_{\alpha/2}) =$$
$$= 1-(1-\alpha/2)+\alpha/2 = \alpha.$$

Теперь изучим критерий на предмет мощности. Предположим, что истинное значение θ не равно θ_0. Тогда

$$\beta(\theta) = \mathsf{P}_\theta(|T_{\theta_0}(\mathbf{X})| > z_{1-\alpha/2}) =$$
$$= \mathsf{P}_\theta\left(\sqrt{n}\cdot\frac{\widehat{\theta}-\theta_0}{\widehat{\sigma}} > z_{1-\alpha/2}\right) + \mathsf{P}_\theta\left(\sqrt{n}\cdot\frac{\widehat{\theta}-\theta_0}{\widehat{\sigma}} < -z_{1-\alpha/2}\right) =$$
$$= \mathsf{P}_\theta\left(\sqrt{n}\cdot\frac{\widehat{\theta}-\theta}{\widehat{\sigma}} > z_{1-\alpha/2} + \sqrt{n}\cdot\frac{\theta_0-\theta}{\widehat{\sigma}}\right) + \mathsf{P}_\theta\left(\sqrt{n}\cdot\frac{\widehat{\theta}-\theta}{\widehat{\sigma}} < -z_{1-\alpha/2} + \sqrt{n}\cdot\frac{\theta_0-\theta}{\widehat{\sigma}}\right) \approx$$

$$\approx 1 - \Phi\left(z_{1-\alpha/2} + \sqrt{n} \cdot \frac{\theta_0 - \theta}{\widehat{\sigma}}\right) + \Phi\left(-z_{1-\alpha/2} + \sqrt{n} \cdot \frac{\theta_0 - \theta}{\widehat{\sigma}}\right).$$

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.

Пример 7.3. 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 согласованности этих данных, которую мы хотим проверить на уровне значимости $\alpha = 0.01$ (для надёжности).

Каждому избирателю с номером i можно поставить в соответствие случайную величину $X_i \sim \mathrm{Bern}(p)$, которая равна 1, если избиратель проголосовал за данного кандидата, и 0 иначе. Гипотезу в таком случае можно записать как

$$H_0\colon p = p_0 = 0.03.$$

По ЦПТ имеется асимптотически нормальная оценка $\widehat{p} = \overline{\mathbf{X}}$, 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

$$R = \left\{\mathbf{x}\colon \sqrt{n} \cdot \frac{\overline{\mathbf{x}} - p_0}{\sqrt{p_0(1-p_0)}} > z_{1-\alpha/2}\right\}$$

Посчитаем статистику критерия для приведённой реализации выборки $\mathbf{x}_0$:

$$T(\mathbf{x}_0) = \sqrt{5818955} \cdot \frac{\frac{4661075}{5818955} - 0.03}{\sqrt{0.03 \cdot (1 - 0.03)}} \approx 10902.83.$$

Критическое значение при данном уровне значимости равняется $z_{1-0.01/2} \approx 2.58$, то есть статистика значительно опережает этот порог, посему гипотезу H_0 следует отвергнуть. ■

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

$$R_- = \{\mathbf{x}\colon T_{\theta_0}(\mathbf{x}) < -z_{1-\alpha}\}.$$

Пример 7.4. Посетители ТРЦ Рио ходили по магазинам в среднем 1 час, стандартное отклонение равнялось 0.5. Потом на втором этаже появился детский паровозик, а на следующий день оказалось, что по выборке из 35 посетителей среднее время шоппинга составило 6/5 часа. Требуется проверить на уровне значимости 0.01 гипотезу о пользе паровозика.

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 1 и 0.5 соответственно. Попробуем применить односторонний критерий Вальда (было бы странно в альтернативу, утверждающую, что паровозик помог, запихивать случай, когда среднее уменьшилось):

$$T(\mathbf{x}_0) = \sqrt{35} \cdot \frac{1.2 - 1}{0.5} \approx 2.366.$$

В то же время критическое значение равняется $z_{1-0.01} \approx 2.326$, что меньше значения статистики, поэтому основная гипотеза отвергается, то есть паровозик статистически значимо увеличил среднюю продолжительность покупок. Заметим, что если бы мы применяли двусторонний критерий, то критическое значение бы равнялось $z_{1-0.01/2} \approx 2.58$, и поэтому гипотеза H_0 бы не отвергалась. ■

Следует понимать, что односторонний критерий берётся только в случае достоверного понимания, что «вторая сторона» не может реализоваться. На практике такое встречается довольно редко, да и исследователям важно отслеживать изменение в обе стороны. Однако недобросовестные аналитики могут воспользоваться этим трюком, искажая уровень значимости критерия (см. задачу 7.3).

Критерий Вальда также очень удобен для построения *двухвыборочных критериев*, которые более подробно будут изучены в главе 12. Обычно они проверяют, нет ли каких-либо общих свойств у распределений двух выборок. Покажем пример работы критерия Вальда в случае проверки равенства средних, что можно использовать для проверки наличия эффекта.

Пример 7.5. Пусть $X_1, \ldots, X_n$ — выборка из распределения $\mathrm{Pois}(\lambda_1)$, $Y_1, \ldots, Y_m$ — выборка из распределения $\mathrm{Pois}(\lambda_2)$, причём выборки независимы. Проверим гипотезу $H_0\colon \lambda_1 = \lambda_2$. Её можно переформулировать так:

$$H_0\colon \delta := \lambda_1 - \lambda_2 = 0.$$

Таким образом, можно протестировать гипотезу о том, что параметр δ равен нулю. За оценку сего параметра логично взять $\widehat{\delta} = \overline{\mathbf{X}} - \overline{\mathbf{Y}}$, то есть разность выборочных средних X и Y (её асимптотическая нормальность следует из теоремы о наследовании асимптотической нормальности). В силу независимости выборок дисперсия данной оценки равна

$$\mathsf{D}\widehat{\delta} = \mathsf{D}\overline{\mathbf{X}} + \mathsf{D}\overline{\mathbf{Y}} = \frac{\lambda_1}{n} + \frac{\lambda_2}{m},$$

Сами параметры λ_1 и λ_2 мы не знаем, поэтому придётся заменить честные значения на их состоятельные оценки:

$$\widehat{\mathsf{D}\widehat{\delta}}(\mathbf{X}, \mathbf{Y}) = \frac{\overline{\mathbf{X}}}{n} + \frac{\overline{\mathbf{Y}}}{m}.$$

Возьмём от всего этого дела корень, чтобы получить стандартное отклонение, и запишем итоговую статистику критерия, которая с ростом n и m сходится к $\mathcal{N}(0, 1)$:

$$W(\mathbf{X}, \mathbf{Y}) = \frac{\overline{\mathbf{X}} - \overline{\mathbf{Y}}}{\sqrt{\frac{\overline{\mathbf{X}}}{n} + \frac{\overline{\mathbf{Y}}}{m}}}.$$

Для общей альтернативы $H_1\colon \delta \neq 0$ подойдёт критерий

$$R_\alpha = \left\{(\mathbf{x}, \mathbf{y})\colon\ |W(\mathbf{x}, \mathbf{y})| > z_{1-\alpha/2}\right\}.$$

■

7.2 p-value

Как было видно по примерам выше, превышение критического значения могло быть разным: где-то оно было небольшим, а где-то — многократным. Однако распределение статистики критерия каждый раз разное, поэтому не всегда очевидно, насколько сильное отклонение от нулевой гипотезы мы получили в очередной раз. В этом контексте крайне удобна следующая характеристика, которая показывает, насколько сильно мы можем быть уверены в отклонении гипотезы.

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 статистику:

$$\text{p-value}(\mathbf{X}) = \inf\{\alpha\colon \mathbf{X} \in R_\alpha\}.$$

Поясним, что тут происходит. Для каждого α у нас в рукаве имеется критерий R_α с размером α. С уменьшением α мы становимся более консервативными и боимся отвергать гипотезу, поэтому критическое множество уменьшается. На рисунке 1 выделены три критерия с размерами $\alpha_1 > \alpha_2 > \alpha_3$. При α_3 критерий достаточно мал, и туда выборка не попадает, а при α_1 и α_2 критерии довольно жирные, отчего выборка там лежит. Таким образом, мы смотрим, при каких α реализация выборки X попала в критическое множество R_α, то есть при каких α нам следовало бы отвергнуть гипотезу, а потом берём по ним инфимум (например, на картинке p-value будет равняться α_2).

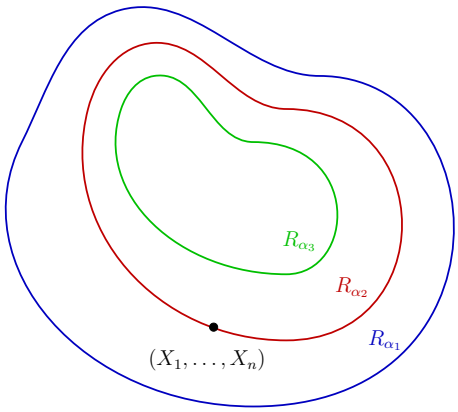


Рис. 1: Критерии для разных уровней значимости

То есть p-value – это *минимальный уровень значимости, на котором мы должны отвергнуть гипотезу*. Таким образом,

$$H_0 \text{ отвергается} \iff \mathbf{X} \in R_\alpha \iff \text{p-value} \leqslant \alpha.$$

Более наглядно смысл p-value виден в случае, когда критерий задаётся некоторой статистикой: $R_\alpha = \{\mathbf{x}\colon T(\mathbf{x}) \geqslant c_\alpha\}$, где α – размер критерия. Предположим, что реализовалась значение выборки $\mathbf{x}$ и наблюдаемое значение статистики критерия $T(\mathbf{x})$ равно t. Тогда p-value можно переписать так:

$$\text{p-value}(\mathbf{x}) = \inf\{\alpha\colon t \geqslant c_\alpha\} = \alpha(t),$$

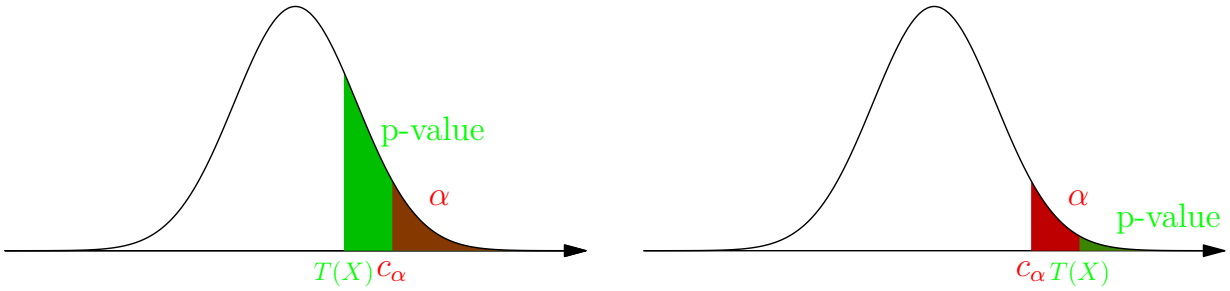
где $c_{\alpha(t)} = t$ (при уменьшении α граница c_α лишь увеличивается, поэтому инфимум достигается при $t = c_\alpha$). Вспоминая определение размера критерия, получаем, что

$$\text{p-value}(\mathbf{x}) = \alpha(t) = \sup_{\mathsf{P}\in\mathcal{P}_0} \mathsf{P}(T(\mathbf{X}) \geqslant c_{\alpha(t)}) = \sup_{\mathsf{P}\in\mathcal{P}_0} \mathsf{P}(T(\mathbf{X}) \geqslant t) = \sup_{\mathsf{P}\in\mathcal{P}_0} \mathsf{P}(T(\mathbf{X}) \geqslant T(\mathbf{x})).$$

Если распределение статистики одно и то же при любом $\mathsf{P} \in \mathcal{P}_0$ (в частности, когда гипотеза простая), то супремум берётся по одному элементу. 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 $T(\mathbf{X})$, если H_0 верна.



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 $T(\mathbf{X})$ приняла весьма экстремальное значение и попала в «критическую зону». Отсюда делаем вывод о необходимости отвергнуть H_0.

Заметим, что p-value очень просто считается: достаточно знать функцию распределения статистики критерия:

$$\text{p-value}(\mathbf{x}) = \mathsf{P}_0(T(\mathbf{X}) \geqslant T(\mathbf{x})) = 1 - F_{T(\mathbf{X})}(T(\mathbf{x})).$$

Также следует выделить следующее важное свойство p-value, придающее этой величине универсальный характер.

Теорема 7.1.

Пусть статистика $T(\mathbf{X})$ имеет непрерывное распределение, которое одинаково при всех $\mathsf{P} \in \mathcal{P}_0$. Тогда p-value$(\mathbf{X})$ распределено равномерно на отрезке $[0;1]$.

Доказательство. Так как функция распределения F статистики $T(\mathbf{X})$ непрерывна, то по утверждению 6.1 величина $F(T(\mathbf{X}))$ распределена равномерно на отрезке $[0;1]$. Тогда по выведенному выше, если $\mathbf{x}$ — наблюдаемое значение выборки, то

$$\text{p-value}(\mathbf{x}) = \mathsf{P}(T(\mathbf{X}) \geqslant T(\mathbf{x})) = \mathsf{P}\big[\underbrace{F(T(\mathbf{X}))}_{\sim \mathrm{U}[0;1]} \geqslant F(T(\mathbf{x}))\big] = 1 - F(T(\mathbf{x})),$$

то есть p-value$(\mathbf{X}) = 1 - F(T(\mathbf{X})) \sim \mathrm{U}[0;1]$. □

Таким образом, 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. С другой стороны, высокое p-value не свидетельствует о верности H_0. Вполне возможно, что на самом деле верна альтернатива H_1, но критерий оказался недостаточно мощным для обнаружения несоответствия с основной гипотезой.

Задачи

Задача 7.1 (*The Permanent Illusion, [2]*)**.** Случайно равновероятно возьмём M — произвольного человека с планеты Земля и поставим на проверку гипотезу H_0: M — американец. Критерий предлагается построить на основе его профессии, а именно возьмём множество $R = \{m \in \text{Земля}\colon m \text{ — конгрессмен}\}$. Очевидно, при верности H_0 вероятность $\mathsf{P}_0(M \in R)$

крайне мала, поэтому данный критерий обладает разумным уровнем значимости, например, $\alpha = 0.01$. Таким образом, если случайный человек оказался конгрессменом, то в соответствии с критерием мы должны отклонить гипотезу H_0. Всё ли корректно в данной процедуре? Стоит ли её применять в реальной жизни?

Задача 7.2. Партия в преферанс предназначена для трёх игроков, каждому из которых раздаётся случайным образом по 10 карт, а остальные 2 карты скидываются в прикуп (итого, 32 карты — от семёрок до тузов). Двое игроков заметили, что третьему за 100 партий на руки выпал хотя бы один туз 87 раз. На уровне значимости 0.01 проверьте гипотезу о том, что он играет честно, против альтернативы, что он подмешивает себе тузов.

Задача 7.3. Имеется выборка $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$. Проверяется гипотеза $H_0\colon \mu = 0$ с помощью стандартной статистики Вальда $T(\mathbf{X}) = \overline{\mathbf{X}}$. Чтобы повысить мощность критерия, аналитик решил схитрить: сначала он смотрит на знак полученного среднего, и если значение получилось положительным, то он берёт правосторонний критерий $\{T(\mathbf{x}) \geqslant c_\alpha\}$ уровня α против альтернативы $H_1\colon \mu > 0$. В ином случае он проверяет гипотезу левосторонним критерием $\{T(\mathbf{x}) \leqslant c_\alpha\}$, который обычно используют при альтернативе $H_2\colon \mu < 0$. Насколько корректна данная процедура? Какая ошибка I рода может достигаться при таком алгоритме проверки?

Задача 7.4. Пусть $X_1, \ldots, X_n$ — выборка из $\mathrm{Bern}(\theta)$, и ставится на проверку гипотеза $H_0\colon \theta = 1/2$. Представим, что данные не имеются сразу на руках, а подаются последовательно. Нетерпеливый аналитик не хочет долго ждать, поэтому он осуществляет проверку следующим образом: после получения очередного элемента X_k он строит критерий Вальда уровня значимости α для выборки $X_1, \ldots, X_k$ и отвергает H_0, если для этого критерия имеет место отвержение. Если же после получения всей выборки отвержений так и не было, то H_0 принимается. Контролируется ли в данной процедуре ошибка I рода на уровне α? К чему будет стремиться ошибка I рода при $n \to \infty$?

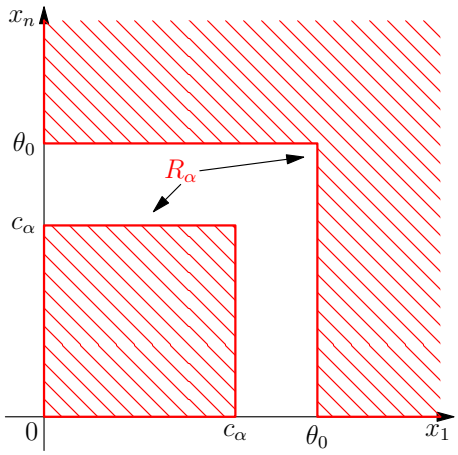
Задача 7.5. Пусть $X_1, \ldots, X_n$ — выборка из распределения $\mathcal{N}(a_1, \sigma_1^2)$, $Y_1, \ldots, Y_m$ — выборка из распределения $\mathcal{N}(a_2, \sigma_2^2)$, причём выборки независимы. Предложите асимптотический критерий для проверки гипотезы $H_0\colon \sigma_1^2 = \sigma_2^2$.

8 Равномерно наиболее мощные критерии

Среди всевозможных критериев с заданным уровнем значимости α, как было замечено выше, разумнее всего брать те, что имеют как можно большую мощность. Во многих случаях нельзя точно сказать, какой критерий лучше: при разных альтернативах предпочтительнее могут оказаться разные критерии. Поэтому ситуации, когда можно выделить один критерий, умеющий переплюнуть любой другой, представляют особый интерес.

Определение. Критерий R уровня значимости α называется *равномерно наиболее мощным (или сокращённо р.н.м.к.) уровня значимости* α, если он мощнее любого другого критерия уровня значимости α.

Пример 8.1. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из равномерного распределения на отрезке $[0; \theta]$, $\theta > 0$. Нам предстоит голыми руками построить р.н.м.к. уровня значимости α для проверки гипотезы $H_0\colon \theta = \theta_0$ против альтернативы $H_1\colon \theta \neq \theta_0$.



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

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

где $c = c_\alpha$ мы подберём так, чтобы размер R был в точности α:

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

Докажем, что он и будет р.н.м.к.

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

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

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

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

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. ■

Однако чаще всего р.н.м.к. просто нет, и в зависимости от ситуации нужно подбирать наиболее подходящий критерий. Правда, доказательство отсутствия р.н.м.к. иногда представляет собой непростую задачу.

Пример 8.2. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из распределения $\mathrm{Bern}(\theta)$. Будем проверять выполнимость гипотезы о равенстве параметра фиксированному числу:

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

Вообще говоря, в некоторых вырожденных случаях р.н.м.к. существует. Действительно, например, положим $\theta_0 = 1/2$, а $\alpha < 2^{-n}$. Тогда пустой критерий будет единственным с допустимым уровнем значимости, а значит, он автоматически р.н.м.к. Докажем, что для фиксированных θ_0 и α р.н.м.к. не существует для достаточного большого n.

Пусть существует р.н.м.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.н.м.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

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

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 R 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, $\mathsf{P}_{\theta_0}(\mathbf{X} \in S_1) \leqslant \alpha/2$, 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. ■

8.1 Простые гипотезы и лемма Неймана-Пирсона

В подавляющем большинстве ситуаций р.н.м.к. просто не существует, особенно если речь идёт о *двусторонних* гипотезах, которые проверяют равенство параметра определённому значению против альтернативы неравенства. 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 $\rho_0(t)$ и $\rho_1(t)$ соответственно.

Теорема 8.1 (лемма Неймана-Пирсона).

Рассмотрим критерий $R_\lambda = \{\mathbf{x} \in \mathcal{X}\colon \rho_1(\mathbf{x}) - \lambda\rho_0(\mathbf{x}) \geqslant 0\}$, где $\lambda > 0$. Если его размер равен α, то есть $\mathsf{P}_0(X \in R_\lambda) = \alpha$, то он является несмещённым р.н.м.к. уровня значимости α.

Доказательство. Пусть S — произвольный критерий уровня значимости α, то есть с $\mathsf{P}_0(\mathbf{X} \in S) \leqslant \alpha = \mathsf{P}_0(\mathbf{X} \in R_\lambda)$. Рассмотрим функцию $f\colon \mathcal{X} \to \mathbb{R}$, определённую как

$$f(\mathbf{x}) = (I(\mathbf{x} \in R_\lambda) - I(\mathbf{x} \in S)) \cdot (\rho_1(\mathbf{x}) - \lambda\rho_0(\mathbf{x})).$$

Заметим, что она неотрицательна на всём $\mathcal{X}$: для $\mathbf{x} \in R_\lambda$ обе скобки неотрицательные, а для $\mathbf{x} \notin R_\lambda$ — неположительные. Тогда

$$0 \leqslant \int_{\mathcal{X}} f(\mathbf{x})\,\mu(d\mathbf{x}) = \mathsf{P}_1(\mathbf{X} \in R_\lambda) - \mathsf{P}_1(\mathbf{X} \in S) - \lambda(\underbrace{\mathsf{P}_0(\mathbf{X} \in R_\lambda) - \mathsf{P}_0(\mathbf{X} \in S)}_{\geqslant 0}) \leqslant$$

$$\leqslant \mathsf{P}_1(\mathbf{X} \in R_\lambda) - \mathsf{P}_1(\mathbf{X} \in S),$$

откуда $\mathsf{P}_1(\mathbf{X} \in R_\lambda) \geqslant \mathsf{P}_1(\mathbf{X} \in S)$, то есть критерий R_λ оказался мощнее. Теперь покажем несмещённость. Обозначим ошибку II рода за γ. С одной стороны,

$$\alpha = \int_{R_\lambda} \rho_0(\mathbf{x})\,\mu(d\mathbf{x}) \leqslant \frac{1}{\lambda}\int_{R_\lambda} \rho_1(\mathbf{x})\,\mu(d\mathbf{x}) = \frac{1-\gamma}{\lambda}.$$

С другой стороны, аналогично получаем

$$1 - \alpha = \int_{\overline{R_\lambda}} \rho_0(\mathbf{x})\,\mu(d\mathbf{x}) \geqslant \frac{1}{\lambda}\int_{\overline{R_\lambda}} \rho_1(\mathbf{x})\,\mu(d\mathbf{x}) = \frac{\gamma}{\lambda}.$$

Таким образом,

$$\frac{\gamma}{1-\alpha} \leqslant \lambda \leqslant \frac{1-\gamma}{\alpha},$$

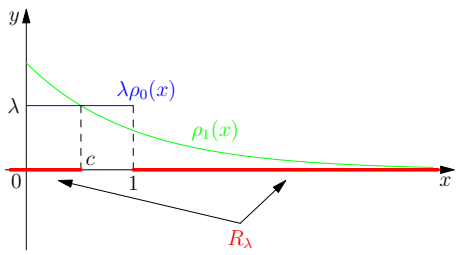
откуда $\mathsf{P}_0(\mathbf{X} \in R_\lambda) = \alpha \leqslant 1 - \gamma = \mathsf{P}_1(\mathbf{X} \in R_\lambda)$, что и требовалось. □

Пример 8.3. Пусть X_1 – выборка размера 1. Рассмотрим гипотезы

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

Построим р.н.м.к. для проверки H_0 против H_1.

Из монотонности $\rho_1(t) = e^{-t}$ (см. рис.) легко понять, что р.н.м.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 $\alpha = \mathsf{P}_0(\mathbf{X} \in R_\lambda) = c$, то есть $\lambda = e^{-\alpha}$. Отсюда также несложно посчитать мощность нашего критерия:

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

■

Пример 8.4. Пусть $X_1, \ldots, X_n$ — выборка из распределения $\mathcal{N}(0, \sigma^2)$. Построим р.н.м.к. уровня значимости α для проверки гипотезы $H_0\colon \sigma^2 = \sigma_0^2$ против альтернативы $H_1\colon \sigma^2 = \sigma_1^2$.

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

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

В силу независимости элементов выборки при верности H_0 выполнено $\sum X_i^2 \sim \sigma_0^2 \chi_n^2$, поэтому λ и α связывает следующее соотношение:

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

где $\chi_{n,p}^2$ — p-квантиль соответствующего распределения. Отсюда

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

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

8.2 Сложные гипотезы и монотонное отношение правдоподобия

Как и ранее, 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. Введём 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 8.2 (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

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

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

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

Пример 8.5. Пусть $X_1, \dots, X_n$ — выборка из распределения $\mathcal{N}(\theta, 1)$. Построим р.н.м.к. уровня значимости α для проверки следующих гипотез:

- $H_0\colon \theta \leqslant \theta_0$ versus $H_1\colon \theta > \theta_0$. Распишем отношение совместных плотностей:
$$\frac{\rho_{\theta_2}(\mathbf{x})}{\rho_{\theta_1}(\mathbf{x})} = \exp\left[\frac{1}{2}\sum(x_i-\theta_1)^2 - \frac{1}{2}\sum(x_i-\theta_2)^2\right] =$$
$$= \exp\left[\frac{n}{2}(\theta_1^2-\theta_2^2) + (\theta_2-\theta_1)\sum x_i\right] \text{ — возрастает по } \sum x_i \text{ при } \theta_2 > \theta_1$$
Так как $\sum X_i \sim \mathcal{N}(n\theta_0, n)$ при верности H_0, то $\left(\sum X_i - n\theta_0\right)/\sqrt{n} \sim \mathcal{N}(0,1)$, и требуемым критерием будет являться
$$R = \left\{\mathbf{x}\colon \left(\sum x_i - n\theta_0\right)\Big/\sqrt{n} \geqslant z_{1-\alpha}\right\} = \left\{\mathbf{x}\colon \sum x_i \geqslant n\theta_0 + \sqrt{n}z_{1-\alpha}\right\},$$
где z_p – p-квантиль распределения $\mathcal{N}(0,1)$.

- $H_0\colon \theta \geqslant \theta_0$ versus $H_1\colon \theta < \theta_0$. Введём параметр $\mu := -\theta$. Тогда плотность будет иметь 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 возрастающей по $-\sum x_i$ функцией. Опять же, имеем $\sum X_i \sim \mathcal{N}(n\theta_0, n)$ при верности H_0, откуда получаем р.м.н.к.
$$R = \left\{\mathbf{x}\colon -\left(\sum x_i - n\theta_0\right)\Big/\sqrt{n} \geqslant z_{1-\alpha}\right\} = \left\{\mathbf{x}\colon \sum x_i \leqslant n\theta_0 + \sqrt{n}z_\alpha\right\}.$$

■

Пример 8.6. Пусть $X_1, \dots, X_n$ — выборка из распределения $\mathrm{Exp}(\lambda)$. Будем строить р.н.м.к. для основной гипотезы $H_0\colon \lambda = \lambda_0$ на уровне значимости α против односторонних альтернатив.

- $H_1\colon \lambda > \lambda_0$. Для $\lambda_2 > \lambda_1$ имеем
$$\frac{\rho_{\lambda_2}(\mathbf{x})}{\rho_{\lambda_1}(\mathbf{x})} = \left(\frac{\lambda_2}{\lambda_1}\right)^n \exp\left[(\lambda_1-\lambda_2)\sum x_i\right],$$
поэтому семейство распределений обладает неубывающим отношением правдоподобия по $-\sum x_i$. Из независимости X_i получаем, что если H_0 верна, то $\sum X_i \sim \Gamma(n, \lambda_0)$. Тогда по теореме выше р.н.м.к. будет критерий
$$R = \left\{\mathbf{x}\colon -\sum x_i \geqslant -x_\alpha\right\} = \left\{\mathbf{x}\colon \sum x_i \leqslant y_\alpha\right\},$$
где y_p — p-квантиль распределения $\Gamma(n, \lambda_0)$.

- $H_1\colon \lambda < \lambda_0$. 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 $\nu_2 > \nu_1$:
$$\frac{\rho_{\nu_2}(\mathbf{x})}{\rho_{\nu_1}(\mathbf{x})} = \left(\frac{-\nu_2}{-\nu_1}\right)^n \exp\left[(\nu_2-\nu_1)\sum x_i\right],$$

что есть возрастающая функция от $\sum x_i$, то есть новое семейство распределений обладает неубывающим отношением правдоподобия. Таким образом, по теореме 8.2 $R = \{\mathbf{x}\colon \sum x_i \geqslant y_{1-\alpha}\}$ будет р.н.м.к.

■

Задачи

Задача 8.1. Найдите р.н.м.к. в модели сдвига $X_1, \ldots, X_n \sim \mathrm{Exp}(1, \theta)$ для проверки гипотезы $H_0\colon \theta = \theta_0$ против альтернативы $H_1\colon \theta \neq \theta_0$.

Задача 8.2. Предложите достаточное условие, при котором р.н.м.к. из леммы Неймана-Пирсона будет единственным с точностью до μ-п.н.

Задача 8.3. Докажите, что в модели сдвига $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$, не существует р.н.м.к. для проверки гипотезы

$$H_0\colon \theta = 0 \quad \text{versus} \quad H_1\colon \theta \neq 0.$$

9 Критерии согласия

В этом параграфе мы рассмотрим самые популярные критерии, которые проверяют, согласуются ли наблюдения с некоторым конкретным теоретическим распределением. Отсюда пошло название *критериев согласия*. В англоязычной литературе распространён другой термин — goodness of fit tests, который, вообще говоря, охватывает несколько больший класс критериев. Как можно догадаться из перевода, это критерии, проверяющие качество объяснения данных выбранной статистической моделью. Помимо простых гипотез они проверяют, например, принадлежность выбранному семейству распределений (нормальному, экспоненциальному и т.д.). О них речь пойдёт в следующей главе, а пока остановимся на вышеуказанном частном случае.

9.1 Критерий Колмогорова

Поставим на проверку гипотезу о том, что наблюдения поступают нам из какого-то непрерывного распределения P_0.

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

Как мы знаем из теоремы Гливенко-Кантелли, эмпирическая функция распределения $\widehat{F}_n$ равномерно сходится к истинной функции распределения F для почти всех выборок $\mathbf{X} = (X_1, \ldots, X_n, \ldots)$, 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 близко к нулю значение D_n. Оказывается, сходимость этой величины к нулю имеет порядок $1/\sqrt{n}$, притом у $\sqrt{n}D_n$ имеется предельное распределение.

Теорема 9.1 (Колмогоров).

Пусть F – непрерывная функция распределения. Тогда случайная величина $\sqrt{n}D_n$ распределена одинаково вне зависимости от F и слабо сходится к *распределению Колмогорова* с функцией распределения

$$K(t) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2t^2}, \quad t > 0. \tag{6}$$

Доказательство. Докажем лишь инвариантность распределения $\sqrt{n}D_n$. Статистику D_n можно переписать как

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

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 утверждения 6.1, 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 определено однозначно. □

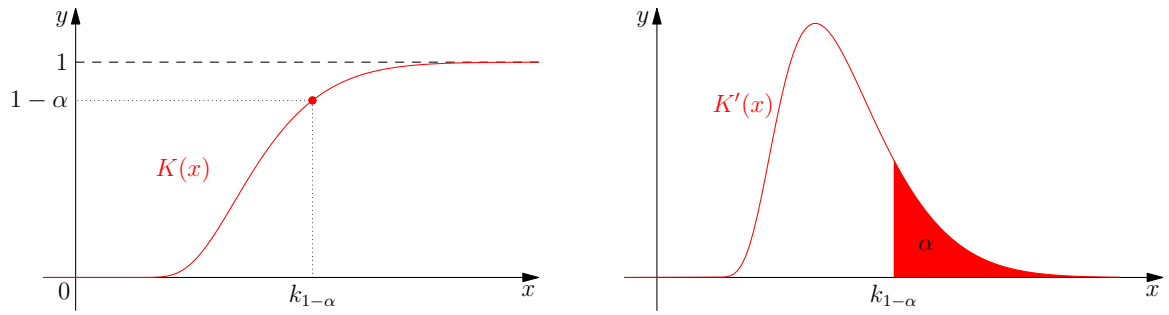


Рис. 2: CDF и PDF распределения Колмогорова

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. Для его нахождения можно либо воспользоваться таблицами с распространёнными квантилями, либо использовать приближение

$$k_{1-\alpha} \sim \sqrt{-\frac{1}{2}\ln\frac{\alpha}{2}} \text{ при } \alpha \to 0.$$

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

Теорема 9.2.

Критерий Колмогорова состоятелен против общей альтернативы H_1.

Доказательство. Пусть истинная функция распределения равна $G \neq F$. 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 $D'_n = \sup_{x\in\mathbb{R}} |\widehat{G}_n(x) - G(x)| \xrightarrow{\text{п.н.}} 0$ по теореме Гливенко-Кантелли, поэтому и $D'_n + k_{1-\alpha}/\sqrt{n} \xrightarrow{\text{п.н.}} 0$. В таком случае

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

□

Осталось только понять, как на практике находить статистику критерия и проверять основную гипотезу. Ключевое наблюдение заключается в том, что теоретическая функция распределения не убывает и по условию непрерывна, а эмпирическая — кусочно-постоянна. Из этого следует, что супремум в определении D_n достигается либо в точке разрыва $\widehat{F}_n$, либо в левостороннем пределе к точке разрыва. Поэтому справедлива следующая формула:

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

Ещё проще осуществить проверку можно с помощью функции `scipy.stats.kstest`, которая выведет значение статистики D_n и соответствующий p-value.

Пример 9.1. Одним из существенных недостатков данного критерия является его низкая мощность, то есть он далеко не всегда хорошо улавливает отличия теоретической и эмпирической функций распределения. Особенно это проявляется для «хвостов» распределений, различия которых равномерная метрика почти не чувствует. Проиллюстрируем это следующей программой, которая применяет критерий Колмогорова для распределения $\mathcal{N}(0,1)$, хотя истинное распределение иное.

```
import scipy.stats as sps
n_iter = 10000
size = 100
alpha = 0.05
distr_set = [sps.norm(scale=2**0.5).cdf, sps.laplace.cdf, sps.cauchy.cdf]
distr_name = ["N(0, 2)", "Laplace", "Cauchy"]
print(f"Type II error at significance level {alpha} for:")
for i, cdf in enumerate(distr_set):
    rvs = sps.norm.rvs(size=(n_iter, size))
    reject_cnt = 0
    for j in range(n_iter):
        result = sps.kstest(rvs[j], cdf)
        reject_cnt += (result.pvalue < alpha)
    print(f"- {distr_name[i]} distribution = {1 - reject_cnt / n_iter}")
```

```
Type II error at significance level 0.05 for:
- N(0, 2) distribution = 0.5734
- Laplace distribution = 0.9443
- Cauchy distribution = 0.073
```

■

9.2 Критерий ω^2

Попробуем подойти к проблеме иначе, применив другую меру различия двух функций. В критерии Колмогорова используется равномерная метрика, недостатки которой мы уже обсудили. Теперь испробуем L_2-метрику, которая рассматривает интеграл квадрата разности функций по некоторой мере.

Определение. Пусть нам дана некоторая «весовая» функция $\psi(t)$ на $[0;1]$. *Статистикой омега-квадрат* называют

$$\omega^2(\psi) = \int_{\mathbb{R}} \Big(\widehat{F}_n(x) - F(x)\Big)^2 \psi(F(x))\, dF(x).$$

Как можно заметить, интеграл берётся по теоретическому распределению, соответствующему $F(x)$. Сделано это для того, чтобы распределение статистики не зависело от $F(x)$ при верности основной гипотезы. Действительно, осуществим замену переменной $y = F(x)$:

$$\omega^2(\psi) = \int_0^1 \Big(\widehat{F}_n(F^{-1}(y)) - y\Big)^2 \psi(y)\, d\mu(y),$$

где μ — классическая мера Лебега на $[0;1]$. Из доказательства теоремы 9.1 мы знаем, что $\widehat{F}_n(F^{-1}(y))$ распределена одинаково вне зависимости от F, а значит и распределение

статистики ω^2 одно и то же, что и требовалось

Среди многообразия весовых функций обычно берут следующие:

$$\psi_1(t) \equiv 1 \quad \text{и} \quad \psi_2(t) = \frac{1}{t(1-t)}.$$

Выбор именно таких функций оправдывается их простотой и подходом в обнаружении отклонений. В [10, 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 уже домноженная на n, имеет некоторый предельный закон.

Теорема 9.3.

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 весьма ужасное, чтобы приводить его здесь. 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:

$$R_1 = \left\{n\omega^2(\psi_1) > y_{1-\alpha}\right\} \quad \text{— } \textit{критерий Крамера — фон Мизеса — Смирнова}$$
$$R_2 = \left\{n\omega^2(\psi_2) > z_{1-\alpha}\right\} \quad \text{— } \textit{критерий Андерсона — Дарлинга}$$

Некоторые квантили этих распределений приведены в таблице.

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? Как и в случае критерия Колмогорова, в силу кусочно-постоянности $\widehat{F}_n$ можно упростить интеграл выше и получить следующие более приятные формулы:

$$n\omega^2(\psi_1) = \frac{1}{12n} + \sum_{i=1}^{n}\left[F(x_{(i)}) - \frac{2i-1}{2n}\right]^2,$$

$$n\omega^2(\psi_2) = -n - 2\sum_{i=1}^{n}\left[\frac{2i-1}{2n}\ln F(x_{(i)}) + \left(1 - \frac{2i-1}{2n}\right)\ln\left(1 - F(x_{(i)})\right)\right].$$

В случае критерия Крамена-фон Мизеса-Смирнова имеется реализация проверки гипотезы `scipy.stats.cramervonmises`. К сожалению, для критерия Андерсона-Дарлинга реализации не предусмотрено.

Пример 9.2. Сравним рассмотренные ранее критерии на примере выборки из распределения Коши, которую мы будем проверять на нормальность. На рисунке 3 приведены графики теоретической функции распределения $\mathcal{N}(0,1)$ и эмпирической функция распределения, построенной по сгенерированной выборке.

Чтобы ещё лучше «обмануть» критерии, мы взяли распределение Коши с коэффициентом масштаба 0.5. Даже визуально видно значимое отличие полученных функций, но посмотрим, что скажут критерии на уровне значимости, скажем, 0.05.

```
n = 100
rvs = sps.cauchy(scale=0.5).rvs(size=n, random_state=73)
cdf = sps.norm.cdf
# Применим критерий Колмогорова
pvalue = sps.kstest(rvs, cdf).pvalue
print(f"Kolmogorov test's pvalue - {pvalue}")
# Применим критерий Крамера-фон Мизеса-Смирнова
pvalue = sps.cramervonmises(rvs, cdf).pvalue
print(f"Cramer-von Mises-Smirnov's pvalue - {pvalue}")
# Применим критерий Андерсона-Дарлинга
ordered = np.sort(rvs)
stat = -n
for i in range(1, n + 1):
    stat -= 2 * (2 * i - 1) * np.log(cdf(ordered[i - 1])) / (2 * n)
    stat -= 2 * (1 - (2 * i - 1) / (2 * n)) * np.log(1 - cdf(ordered[i - 1]))
print(f"Anderson-Darling test's staticstic - {stat}")
```

■

```
Kolmogorov test's pvalue - 0.18241574592222012
Cramer-von Mises-Smirnov's pvalue - 0.1267607929221075
Anderson-Darling test's staticstic - inf
```

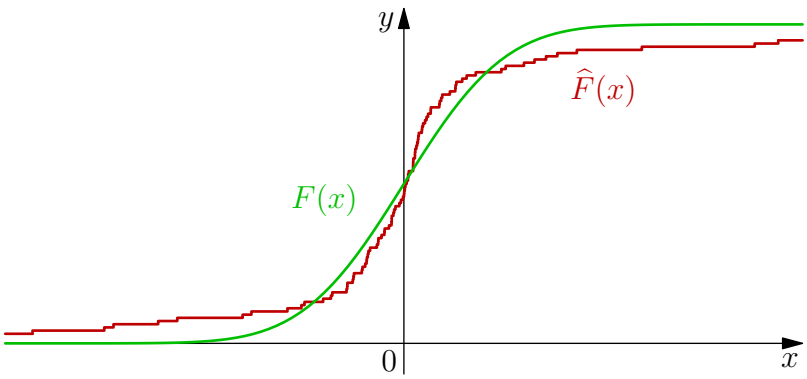


Рис. 3: Распределение из нулевой гипотезы, отличное от настоящего

Первые два критерия не отвергают основную гипотезу. Критерий Колмогорова не видит, что функции распределения, вообще говоря, отличаются «много где», он лишь видит максимальное отклонение, которое его устраивает. Критерий Крамена-фон Мизеса-Смирнова справляется не лучше: в нём мы берём интеграл по мере, соответствующей нормальной $F(x)$, которая чрезвычайно мала на хвостах, из-за чего различие там имеет малый вес. В то же время критерий Андерсона-Дарлинга с его специальной весовой функцией это различие обнаружил, отчего значение статистики получилось чрезвычайно большим, то есть при данном критерии гипотеза отвергается.

9.3 Критерий χ^2 Пирсона

Рассмотрим наблюдение из мультиномиального распределения с параметрами n, k и $\mathbf{p} = (p_1, \dots, p_k)$, 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 $\mathbf{X} = (X_1, \dots, X_k)$, что первая грань выпала X_1 раз, вторая — X_2 раз и т. д. Компоненты этого вектора можно записать как

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

где B_j – результат j-ого броска, причём величины $B_1, \dots, B_n$ независимы в совокупности.

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

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

По ЦПТ мы знаем, что при выполнимости гипотезы H_0

$$\frac{\mathbf{X} - n\mathbf{p}^0}{\sqrt{n}} \xrightarrow{d_{\mathbf{P}_0}} \boldsymbol{\zeta} \sim \mathcal{N}(0, \Sigma), \tag{7}$$

где Σ — матрица ковариаций вектора

$$\begin{pmatrix} I(B_1 = 1) \\ \dots \\ I(B_1 = k) \end{pmatrix}$$

Её элементы несложно высчитать:

$$\begin{aligned} \Sigma_{ii} &= \mathsf{D} I(B_1 = i) = p_i^0(1 - p_i^0) \\ \Sigma_{ij} &= \mathsf{cov}\,(I(B_1 = i), I(B_1 = j)) = -p_i^0 p_j^0, \quad i \neq j \end{aligned} \Longrightarrow \Sigma = \begin{pmatrix} p_1^0 & 0 & \cdots & 0 \\ 0 & p_2^0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n^0 \end{pmatrix} - \mathbf{p}^0 {\mathbf{p}^0}^T.$$

Таким образом, компоненты вектора в левой части (7), 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

$$\chi^2(\mathbf{X}) = \sum_{i=1}^{k} \frac{(X_i - np_i^0)^2}{np_i^0}. \tag{8}$$

Логично предположить, что как сумма квадратов почти что нормально распределённых величин эта статистика стремится по распределению к хи-квадрат.

Теорема 9.4.

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

Доказательство. Положим $\xi_i = \zeta_i / \sqrt{p_i^0}$. Применим к сходимости (7) непрерывную

функцию $f(x_1, \ldots, x_k) = \sum_{i=1}^{k} \frac{x_i^2}{p_i^0}$. По теореме о наследовании сходимости получим, что

$$\chi^2(\mathbf{X}) \xrightarrow{d_{\mathbf{P}_0}} f(\boldsymbol{\zeta}) = \sum_{i=1}^{n} \xi_i^2.$$

Следовательно, достаточно показать, что если $\boldsymbol{\zeta} \sim \mathcal{N}(0, \Sigma)$ (где Σ была посчитана ранее), то $f(\boldsymbol{\zeta}) \sim \chi^2_{k-1}$.

Несложно посчитать, какова будет ковариационная матрица для случайного вектора $\boldsymbol{\xi}$:

$$\Sigma' = E_n - \mathbf{r}\mathbf{r}^T,$$

где $r_i = \sqrt{p_i^0}$, E_n – единичная матрица размера n.

Так как вектор $\mathbf{r}$ имеет единичную длину, то существует ортогональное отображение S, переводящее $\mathbf{r}$ в базисный вектор $\mathbf{e}_n = (0, \ldots, 0, 1)^T$. Тогда

$$S\Sigma' S^T = SE_nS^T - (S\mathbf{r})(S\mathbf{r})^T = E_n - \mathbf{e}_n\mathbf{e}_n^T = \operatorname{diag}(\underbrace{1, 1, \ldots, 1}_{n-1 \text{ единиц}}, 0) =: D.$$

Вектор $\boldsymbol{\eta} = S\boldsymbol{\xi}$ как линейное преобразование над гауссовским вектором имеет распределение $\mathcal{N}(0, S\Sigma'S^T) = \mathcal{N}(0, D)$. 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. 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 $\mathcal{N}(0, 1)$. □

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

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

где $\chi^2_{k-1,p}$ – p-квантиль распределения χ^2_{k-1}. Следует помнить, что этот критерий — асимптотический, а значит, его использование при малой выборке не имеет смысла. Обычно критерий χ^2 используют при $n \geqslant 50$ и $np_i^0 \geqslant 5$ для всех $i = 1, \ldots, k$.

Пример 9.3 (*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 реализация выборки равна $\mathbf{x} = (315, 101, 108, 32)$, и проверяется гипотеза

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

Посчитаем статистику Пирсона:

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

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. ■

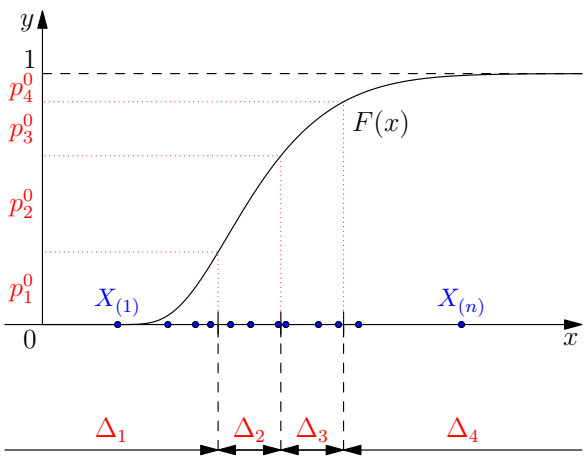
Пример 9.4. Среди первых 800 цифр числа π цифры 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 встречаются соответственно 74, 92, 83, 79, 80, 73, 77, 75, 76, 91 раз. Проверим при помощи критерия хи-квадрат гипотезу о том, что различные цифры встречаются в числе π равновероятно, на уровне значимости $\alpha = 0.05$.

Решим задачу с помощью функции `scipy.stats.chisquare` на языке Python. В качестве аргументов она принимает массив наблюдаемых частот $(X_1, \ldots, X_k)$ и массив ожидаемых частот $(np_1^0, \ldots, np_k^0)$ (по умолчанию $p_1^0 = \ldots = p_k^0 = 1/n$).

```
obs = np.array([74, 92, 83, 79, 80, 73, 77, 75, 76, 91])
result = sps.chisquare(obs)
alpha = 0.05
threshold = sps.chi2(len(obs)-1).ppf(1 - alpha) # Считаем квантиль chi^2_9
print(f"Chi2 statistic = {result.statistic}")
print(f"pvalue = {result.pvalue}")
print(f"Threshold = {threshold}")
```

```
Chi2 statistic = 5.125
pvalue = 0.8232783432788753
Threshold = 16.918977604620448
```

Как можно видеть, значение статистики критерия мало по сравнению с критическим значением, да и p-value весьма велико, поэтому гипотеза не отвергается. ■



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 $X_1, \ldots, X_n$ из некоторого неизвестного нам распределения $F(x)$. Мы же в свою очередь хотим проверить, не равна ли она чему-то хорошему, то есть проверяем

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

Разобьём числовую прямую на k дизъюнкт-

ных множеств $\Delta_1, \ldots, \Delta_k$ (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 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

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

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 9.5. Рассмотрим выборку из примера 9.2, которая на самом деле имеет распределение Cauchy$(0, 0.5)$, но мы проверяем гипотезу H_0: $F = \Phi$, то есть что данные распределены стандартно нормально. Разобьём носитель распределения на четыре равновероятные (по теоретическому распределению) части, что можно сделать с помощью функции `scipy.stats.norm.ppf`, находящей квантили. Для простоты будем использовать функцию `pandas.cut`, которая сама разделит выборку по «бинам». Как можно видеть, даже немощный χ^2-критерий смог найти отклонения в отличие от критерия Колмогорова.

```
n = 100
rvs = sps.cauchy(scale=0.5).rvs(size=n, random_state=73)
bins_count = 4
quantiles = sps.norm.ppf(np.linspace(0, 1, num=bins_count+1))
partition = pd.cut(rvs, quantiles).__array__()
intervals, obs = np.unique(partition, return_counts=True)
for interval, count in zip(intervals, obs):
    print(f"There is {count} dots in {interval}")
print("P-value is", sps.chisquare(obs).pvalue)
```

```
There is 19 dots in (-inf, -0.674]
There is 27 dots in (-0.674, 0.0]
There is 36 dots in (0.0, 0.674]
There is 18 dots in (0.674, inf]
P-value is 0.0384293188578885
```

■

Задачи

Задача 9.1. Согласно *закону Бенфорда*, первая цифра ξ_1 случайного числа $\xi = \overline{\xi_1 \dots \xi_n}$ из достаточно широко диапазона имеет распределение

$$\mathsf{P}(\xi_1 \leqslant d) = \log_{10}(d+1), \quad d \in \{1, \dots, 9\}.$$

Для выборки из стран мира (данные можно взять, например, отсюда) и уровня значимости 0.05 проверить гипотезу о том, что численность населения подчиняется закону Бенфорда.

Задача 9.2. Статистика χ^2 по сути своей является некоторой мерой расхождения между теоретическим распределением и эмпирическим, полученным по наблюдаемым частотам. В качестве расстояния между распределениями можно использовать и другие аналоги, которые, впрочем, будут в некотором смысле эквивалентны исходному. Вам предлагается это проверить.

Напомним, что *дивергенцией Кульбака-Лейблера* двух дискретных распределений $\mathsf{P} = (p_1, \dots, p_k)$ и $\mathsf{Q} = (q_1, \dots, q_k)$ называется величина

$$D(\mathsf{P} \parallel \mathsf{Q}) = \sum_{i=1}^{k} p_i \cdot \log \frac{p_i}{q_i}.$$

Пусть k-гранный кубик с вероятностями выпадения i-ой грани p_i^0 подкидывают n раз, $(\nu_1, \dots, \nu_k)$ — наблюдаемые частоты, а $p_i = \nu_i / n$. Докажите, что

$$2n \cdot D(\mathsf{P} \parallel \mathsf{P}^0) \xrightarrow{d} \chi^2_{k-1}, \quad n \to \infty.$$

Что можно сказать про сходимость статистики $2n \cdot D(\mathsf{P}^0 \parallel \mathsf{P})$? Какую из этих статистик лучше использовать в качестве основы критерия?

Задача 9.3. Модифицируйте критерий Колмогорова таким образом, чтобы, во-первых, в качестве основной гипотезы $H_0 \colon \mathsf{P} = \mathsf{P}_0$ можно было выбрать произвольное распределение P_0 (необязательно непрерывное), и, во-вторых, критерий остался состоятельным против альтернативы $H_1 \colon \mathsf{P} \neq \mathsf{P}_0$.

10 Goodness of fit критерии для сложных гипотез

Теперь разберёмся, как проверять принадлежность распределения некоторому семейству. Например, может возникнуть необходимость проверить нормальность распределения, то есть когда основная гипотеза имеет вид

$$H_0\colon \mathsf{P} \in \{\mathcal{N}(a, \sigma^2)\colon a \in \mathbb{R}, \sigma^2 > 0\}.$$

Далее нам часто будут встречаться критерии, в которых подразумевается нормальность данных, и обычно такие критерии мощнее тех, что работают в общем случае.

Вообще говоря, эта задача не такая простая: во-первых, крайне желательно, чтобы статистика критерия не зависела от истинных значений параметров при верности гипотезы, так как в таком случае проще считать p-value, а во-вторых, критерий должен иметь высокую мощность на широком классе распределений. Понятное дело, мы не можем идеально отслеживать отклонение от принадлежности рассматриваемому семейству. Чаще всего каждый критерий смотрит на какую-то одну характеристику распределения, отклонение от которой важно проверить в данной ситуации. Приведём пример такого критерия.

Что мы любим в нормальном распределении? Во-первых, оно симметрично относительно своего матожидания, а во-вторых, оно имеет довольно лёгкие хвосты. К счастью, есть характеристики, которые в некоторой степени отражают наличие этих двух свойств. Они основаны на так называемых *центральных моментах*:

$$\mu_k(\xi) = \mathsf{E}(\xi - \mathsf{E}\xi)^k.$$

Как функционал от распределения, центральный момент обладает «выборочным» аналогом, который получается методом подстановки:

$$m_k(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{\mathbf{X}}\right)^k.$$

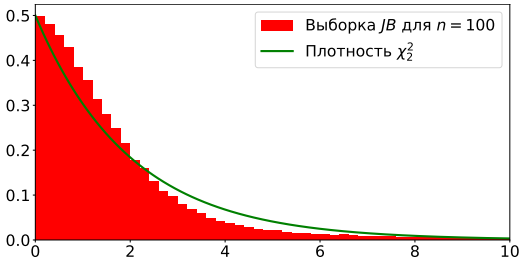
Определение. *Коэффициентом асимметрии* называется величина $\mu_3/\mu_2^{3/2}$. *Коэффициентом эксцесса* называется величина $\mu_4/\mu_2^2 - 3$. Как следствие, числа $\alpha_3 = m_3/m_2^{3/2}$ и $\alpha_4 = m_4/m_2^2 - 3$ называют *выборочными коэффициентами асимметрии и эксцесса* соответственно.

Из-за того, что мы используем центрированные моменты, а также делим на степень дисперсии, распределение этих коэффициентов не зависит от параметров сдвига и масштаба, а вычитание тройки в коэффициенте эксцесса позволяет нам сказать, что оба коэффициента для нормального распределения равны нулю. Посредством дельта-метода можно найти асимптотические дисперсии α_3 и α_4, которые равны 6 и 24 соответственно. Уже сейчас можно на их основе построить критерий Вальда (см. раздел 7.1), но можно контролировать и симметричность, и вес хвостов, взяв комбинированный тест со статистикой

$$JB(\mathbf{X}) = \left(\frac{\sqrt{n}\alpha_3(X)}{\sqrt{6}}\right)^2 + \left(\frac{\sqrt{n}\alpha_4(X)}{\sqrt{24}}\right)^2 = \frac{n}{6}\left(\alpha_3^2 + \frac{1}{4}\alpha_4^2\right),$$

который равен сумме квадратов нормированных коэффициентов. Известно, что $JB \xrightarrow{d} \chi^2_2$. Таким образом, построен асимптотический *критерий Харке-Бера*

$$R = \{\mathbf{x}\colon JB(\mathbf{x}) \geqslant \chi^2_{2,1-\alpha}\}.$$



Имеется функция scipy.stats.jarque_bera, которая вычисляет значение статистики и p-value. Однако приближение этим предельным распределением является достаточно точным лишь при $n > 2000$, что связано с плохой сходимостью распределения α_4 к нормальному и зависимостью коэффициентов (см. рис.). Поэтому зачастую используют более мощный K^2-*критерий*, который к коэффициентам асимметрии и эксцесса применяет дополнительные нормализующие преобразования, что позволяет достаточно хорошо приближать предельное распределение уже при малых n (более подробно описано в [8, 3.2.2.16] или в [3]). Чтобы не считать статистику критерия самостоятельно, можно воспользоваться реализацией K^2-критерия в функции scipy.stats.normaltest.

Пример 10.1. Посмотрим, как K^2-критерий справляется с альтернативами, которые по-разному отличаются от нормального распределения.

```
size = 100
samples = {
    'Laplace': sps.laplace.rvs(size=size),
    'Uniform': sps.uniform(loc=-1, scale=2).rvs(size=size),
    'Skewed': sps.norm.rvs(size=size) + sps.expon.rvs(size=size),
    'With two hills': sps.laplace(loc=1.5).rvs(size=size) * \
        np.random.choice([1, -1], size=size)
}
print("P-value for distribution:")
for name, sample in samples.items():
    print(f"- {name}: {sps.normaltest(sample).pvalue}")
```

```
P-value for distribution:
- Laplace: 0.01217731862212881
- Uniform: 8.63304887582185e-07
- Skewed: 0.0016579199197891447
- With two hills: 0.19779708143850236
```

Первые два распределения симметричны, но имеют хвосты, отличные от нормального. Третье распределение есть свёртка нормального распределения с экспоненциальным, что в результате даёт «скошенное» распределение. В нашем случае все три альтернативы успешно отвергаются на уровне значимости 0.05. Последний пример интереснее: в нём выборка из сдвинутого распределения Лапласа берётся со случайным знаком, из-за чего плотность распределения образует «два холма», что даже отдалённо не похоже на нормальное распределение. Однако сдвиг подобран так, чтобы коэффициент эксцесса был примерно нулевой, что в купе с симметричностью не даёт критерию отвергнуть гипотезу о нормальности. Впрочем, критерий будет маломощен на любом сколь угодно плохом распределении, у которого $\alpha_3, \alpha_4 \approx 0$ (см. задачу 10.1). ■

10.1 QQ-plot и критерий Шапиро-Уилка

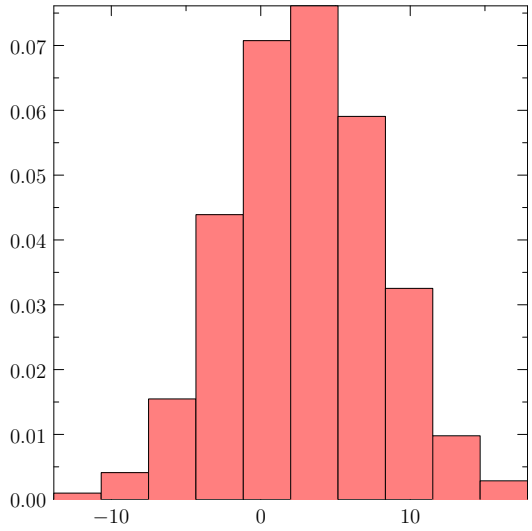
Множество потенциальных кандидатов на роль семейства распределений, из которого пришла выборка, можно значительно сузить, если провести некоторый визуальный анализ данных. Например, если выборка одномерная, то полезно для начала построить гистограмму, чтобы определить носитель распределения и его характерные особенности, или Box-plot, чтобы обнаружить выбросы. Другим неформальным инструментом описатель-

ной статистики является так называемый QQ-plot, который позволяет обнаружить явно выраженные отклонения от заданного семейства распределений.

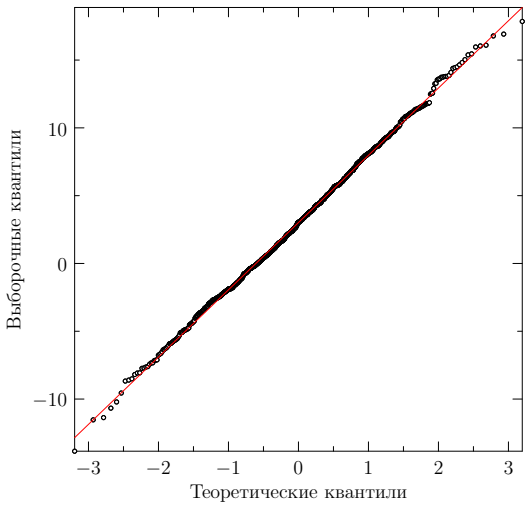
Предположим, что задано некоторое распределение с функцией распределения $F_0(x)$ и плотностью $\rho(x)$, непрерывной на носителе. Зададим семейство распределений, принадлежность к которому мы хотим проверить, посредством параметров сдвига и масштаба:

$$\mathcal{P}_0 = \left\{ \mathsf{P} : F_{\mathsf{P}}(x) = F(x; a, \sigma) := F_0\left(\frac{x-a}{\sigma}\right),\ a \in \mathbb{R}, \sigma > 0 \right\},$$

или, что эквивалентно, $F_0^{-1}\left[F_{\mathsf{P}}(x)\right] = \frac{x-a}{\sigma}$, то есть при верности основной гипотезы $F_0^{-1} \circ F_{\mathsf{P}}$ является какой-то линейной функцией. С помощью метода подстановки можно заменить истинную функцию распределения F_{P} на её выборочный аналог $\widehat{F}$ и оценить, насколько сильно полученная функция похожа на линейную. Чтобы не считать эту композицию напрямую, достаточно найти её значения в точках разрыва $X_{(i)}$, а именно в точках $(X_{(i)}, F_0^{-1}(i/n))$ для $i \in \{1, \ldots, n\}$, однако обычно берут точки $(X_{(i)}, F_0^{-1}((i-0.5)/n))$, дабы не было проблем с прообразом F_0 при $i = n$. Отсюда и название QQ-plot: он показывает соотношение между теоретическими квантилями, которые откладываются по оси абсцисс, и выборочными квантилями, которые откладываются по оси ординат.



(a) Выборка из $\mathcal{N}(3, 25)$



(b) QQ-plot для $F_0 = \Phi$

Несложно показать, что с ростом n и при верности гипотезы эти точки действительно будут описывать нужную прямую. Если $\mathbf{X}$ — выборка из распределения $F(x; a, \sigma)$, то $Y_i = (X_i - a)/\sigma$ имеет распределение $F_0(x)$, и тогда по теореме 2.4 о выборочном квантиле

$$Y_{(i)} - F_0^{-1}\left(\frac{i}{n}\right) \xrightarrow{\mathsf{P}} 0$$

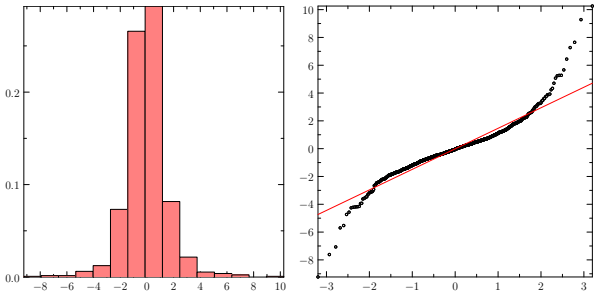
при $n \to \infty$ и $i/n \to p \in (0, 1)$, поэтому

$$X_{(i)} = a + \sigma Y_{(i)} \approx F_0^{-1}\left(\frac{i}{n}\right) \approx a + \sigma F_0^{-1}\left(\frac{i-0.5}{n}\right).$$

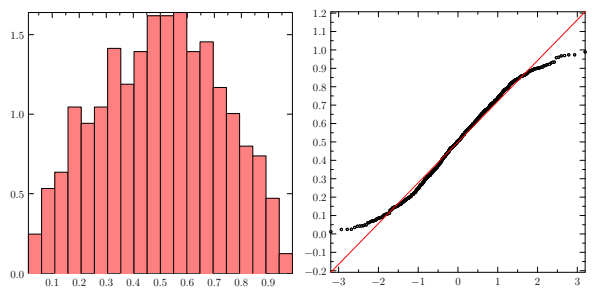
Однако некоторые выводы об истинном распределении можно сделать и в случае альтернативы, то есть QQ-plot даже в случае плохой аппроксимации прямой может подсказать, какую очередную F_0 следует взять. Посмотрим на рис. 5, как выглядят QQ-plot с $F_0 = \Phi$ (такие графики ещё называют Normal QQ-plot). На каждый график нанесена прямая, которая наиболее точно подгоняется под данные точки (как это делается, см. в главе

14). Можно заметить, что для тяжёлых хвостов график начинает закручиваться против часовой стрелки (как, например, для T_3 на рис. 5a), а для лёгких — по часовой (проще всего взять распределение с ограниченным носителем, например, бета с рис. 5b). Из-за такой особенности распределения с положительным коэффициентом асимметрии образуют график, выпуклый вниз, а с отрицательным — вверх.

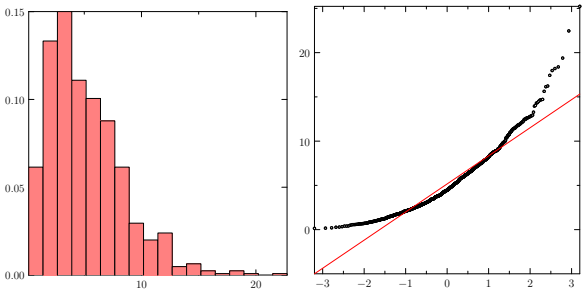
Рис. 5: Normal QQ-plot для различных распределений



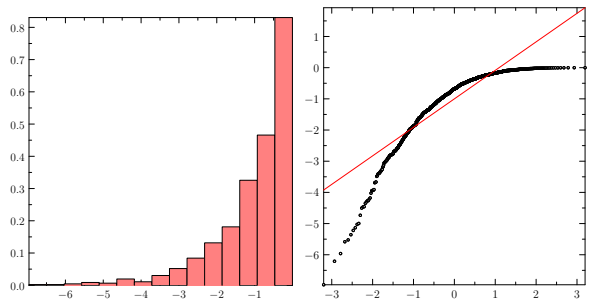
(a) Выборка из T_3



(b) Выборка из $B(2,2)$



(c) Выборка из χ^2_5



(d) Выборка из $-\mathrm{Exp}(1)$

Следует помнить, что QQ-plot не является критерием: он не может доказать, что основная гипотеза верна, равно как и сделать заключение о неверности гипотезы — всё делается «на глаз». Если на графике наблюдается значимое отклонение от прямой, мы можем не проверять формально гипотезу, однако гарантий, что при проверке она будет отвержена, никаких нет. Впрочем, можно формализовать понятие «отклонения от прямой», введя характеристику

$$W = \frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \overline{\mathbf{X}})^2}, \tag{9}$$

причём

$$\mathbf{a} = \frac{\mathbf{m}^T V^{-1}}{\sqrt{\mathbf{m}^T V^{-2} \mathbf{m}}}, \quad m_i = \mathsf{E} Y_{(i)}, \quad V_{ij} = \mathsf{cov}(Y_{(i)}, Y_{(j)}),$$

где $(Y_1, \ldots, Y_n)$ — выборка из распределения F_0.

Оказывается, при верности основной гипотезы $H_0 \colon \mathsf{P} \in \mathcal{P}_0$ распределение W зависит разве что от n. Критерий, основанный на статистике W, называют *критерием Шапиро-Уилка*. Ввиду нетривиальности определения сей статистики W обычно не задумываются ни о её виде, ни о вычислении вектора $\mathbf{a}$, а просто пользуются готовой реализацией, например, `scipy.stats.shapiro`. На практике такой критерий оказывается наиболее мощным для проверки гипотезы нормальности, и если отклонение от нормальности в данной ситуации существенно, то используют именно этот критерий. Подробный вывод формулы можно встретить в примере 14.2.

10.2 Подстановка неизвестного параметра

Напоследок рассмотрим ещё один способ проверки принадлежности семейству распределений с параметрами сдвига и масштаба, которая заключается в использовании критериев согласия из предыдущей главы. Идея следующая: вместо какой-то фиксированной функции распределения давайте использовать ту, которая получается подстановкой на место неизвестных параметров их состоятельных оценок, то есть среди всех функций распределения из основной гипотезы возьмём самую правдоподобную.

Посмотрим на примере модификации критерия Колмогорова. Напомним, что он имеет вид

$$R = \{\sqrt{n}D_n \geqslant k_{1-\alpha}\}, \quad D_n = \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)|.$$

Поставим на проверку гипотезу о принадлежности истинного распределения модели сдвига-масштаба

$$H_0\colon F(x) = F_0(x; a, \sigma) = F_0\left(\frac{x-a}{\sigma}\right).$$

В качестве функции F, которую мы подставим в статистику D_n, возьмём $F(x; \widehat{a}, \widehat{\sigma})$, где $\widehat{a}, \widehat{\sigma}$ — некоторые состоятельные оценки параметров a и σ. Например, пусть $F_0 = \Phi$ (то есть проверяется гипотеза о нормальности), и в качестве оценок параметров возьмём выборочные среднее и дисперсию. Тогда полученная модифицированная статистика

$$D'_n = \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_0(x; \overline{\mathbf{X}}; s(\mathbf{X}))|,$$

домноженная на $\sqrt{n}$, будет иметь какое-то предельное распределение. Критерий, основанный на статистике $\sqrt{n}D'_n$, называется *критерием Лиллиефорса*.

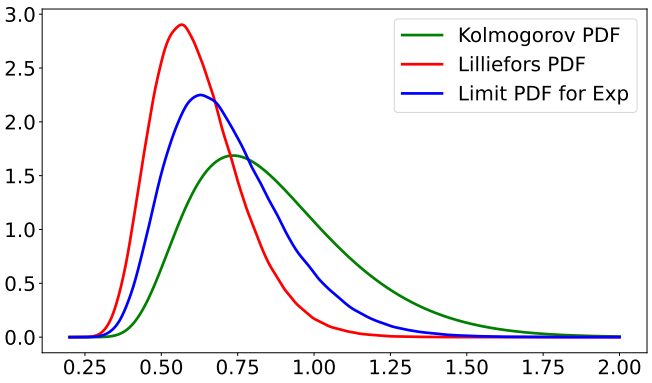


Рис. 6: Плотности предельных распределений

Подвох заключается в том, что, вообще говоря, это распределение отличается от распределения Колмогорова, которое было предельным в обычном критерии Колмогорова и не зависело от истинного F. В общем случае статистики модифицированных критериев согласия, во-первых, могут зависеть от F_0 (то есть для каждого конкретного семейства предельное распределение, а стало быть и критические значения, будет своими), и во-вторых, могут зависеть от вида оценок $\widehat{a}, \widehat{\sigma}$.

На рис. 6 изображены плотности предельных распределений статистик всяких критериев: Колмогорова, Лиллиефорса и модифицированного критерия Колмогорова с $F_0(x) = 1 - e^{-x}$ (проверяющего экспоненциальность). Как можно видеть, модифицированные критерии, несмотря на увеличение основной гипотезы, оказываются куда более требовательными, и для них критические значения будут меньше, однако при $n < 30$ следует пользоваться более точными значениями (их можно смоделировать самостоятельно или воспользоваться таблицами, например, этой). Модифицированный критерий Колмогорова для нормальной и экспоненциальной моделей реализован в функции `statsmodels.stats.diagnostic.lilliefors`.

Помимо критерия Колмогорова можно модифицировать и другие критерии согласия,

например, критерий Андерсона-Дарлинга, статистика которого будет иметь вид

$$n\omega_n'^2 = \int_{\mathbb{R}} \frac{\left(\widehat{F}_n(x) - F_0(x;\widehat{a},\widehat{\sigma})\right)^2}{F_0(x;\widehat{a},\widehat{\sigma})(1-F_0(x;\widehat{a},\widehat{\sigma}))}\, dF_0(x;\widehat{a},\widehat{\sigma}).$$

Как и ранее, предельное распределение сей статистики отличается от того, что было ранее, и может существенно зависеть от F_0. Данный критерий реализован в функции `scipy.stats.anderson` для множества семейств распределений, правда, вместо p-value он возвращает критические значения для различных уровней значимости.

10.3 Критерий отношения правдоподобий

Попробуем обобщить подход из леммы Неймана-Пирсона, когда отвержение гипотезы основывается на чрезмерно большом правдоподобии альтернативы по сравнению с правдоподобием нулевой гипотезы. В случае сложных гипотез

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

предлагается брать супремум функции правдоподобия $f_\theta(\mathbf{x})$ на соответствующем множестве параметров: это своего рода лучшее, что может предложить каждая из гипотез:

$$\widetilde{\lambda}(\mathbf{x}) = \frac{\sup\limits_{\theta\in\Theta_1} f_\theta(\mathbf{x})}{\sup\limits_{\theta\in\Theta_0} f_\theta(\mathbf{x})}, \quad \mathbf{x} \in \mathcal{X}.$$

Однако зачастую берут статистику

$$\lambda(\mathbf{x}) = \frac{\sup\limits_{\theta\in\Theta} f_\theta(\mathbf{x})}{\sup\limits_{\theta\in\Theta_0} f_\theta(\mathbf{x})}, \quad \mathbf{x} \in \mathcal{X}, \tag{10}$$

в которой максимизация в числителе производится по всему множеству параметров Θ. Обычно это более удобно для общей альтернативы $\Theta_1 = \Theta \setminus \Theta_0$, при том что разницы в значениях λ и $\widetilde{\lambda}$ практически нет. Таким образом, большое значение этих статистик свидетельствует о недостаточной правдоподобности нулевой гипотезы по сравнению с альтернативой, отчего разумным представляется взять критерий вида $R_c = \{\mathbf{x} \in \mathcal{X}\colon \lambda(\mathbf{x}) > c\}$. Если вдруг нам удастся выбрать c так, что $\mathsf{P}_\theta(R_c) \leqslant \alpha$ для всех $\theta \in \Theta_0$, то мы получим критерий уровня значимости α, который называется *критерием отношения правдоподобий* (англ. likelihood ratio test) или сокращённо КОП. Уже на данном этапе он весьма полезен, так как даёт лёгкий способ строить критерии. Однако далеко не всегда удаётся найти распределение статистики критерия и уж тем более убедиться, что оно одинаковое при нулевой гипотезе. Тем удивительнее, что в некоторых моделях асимптотическое поведение статистики одно и то же, причём имеет относительно приемлемый вид.

Теорема 10.1 (Уилкс).

Пусть $\Theta \subset \mathbb{R}^d$ — многообразие размерности k, в котором выделено подмногообразие Θ_0 размерности l. Тогда в условиях регулярности при всех $\theta \in \Theta_0$ имеется сходимость

$$2\ln\lambda(\mathbf{X}) \xrightarrow{d_\theta} \chi^2_{k-l}.$$

Количество степеней свободы у предельного распределения можно интерпретировать как коразмерность подповерхности Θ_0. Таким образом, при условиях теоремы имеет место

асимптотический критерий уровня значимости α, который можно записать в виде

$$R_\alpha = \{\mathbf{x} \in \mathcal{X} \colon 2\ln\lambda(\mathbf{x}) \geqslant \chi^2_{k-l,1-\alpha}\}.$$

Пример 10.2. Рассмотрим нормальную модель сдвига-масштаба $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, где оба параметра неизвестны. Поставим на проверку гипотезу $H_0 \colon \mu = \mu_0$ против общей альтернативы, используя критерий отношения правдоподобий. Для этого нам нужно найти две ОМП: в общем случае и при фиксированном среднем μ_0. Первый случай разбирался ранее в примере 4.1, где ОМП имели вид

$$\widehat{\mu} = \overline{\mathbf{X}}, \quad \widehat{\sigma^2} = s^2(\mathbf{X}).$$

Во втором случае несложно убедится, что ОМП для дисперсии равна

$$\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2.$$

Значит, отношение правдоподобий равно

$$\lambda(\mathbf{X}) = \frac{f_{\widehat{\mu},\widehat{\sigma^2}}(\mathbf{X})}{f_{\mu_0,\widetilde{\sigma^2}}(\mathbf{X})} = \frac{\left(\dfrac{1}{\sqrt{2\pi\widehat{\sigma^2}}}\right)^n \exp\left[-\dfrac{1}{2\widehat{\sigma^2}}\displaystyle\sum_{i=1}^n (X_i - \widehat{\mu})^2\right]}{\left(\dfrac{1}{\sqrt{2\pi\widetilde{\sigma^2}}}\right)^n \exp\left[-\dfrac{1}{2\widetilde{\sigma^2}}\displaystyle\sum_{i=1}^n (X_i - \mu_0)^2\right]} = \left(\frac{\widetilde{\sigma^2}}{\widehat{\sigma^2}}\right)^{n/2},$$

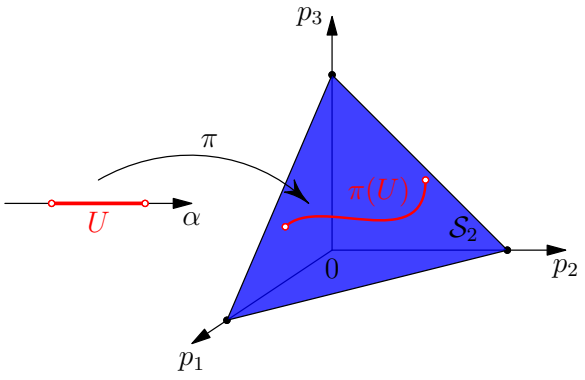
откуда статистика критерия равна

$$2\ln\lambda(\mathbf{X}) = n\ln\frac{\widetilde{\sigma^2}}{\widehat{\sigma^2}} = n\ln\left(\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sum_{i=1}^n (X_i - \overline{\mathbf{X}})^2}\right).$$

Сам критерий уровня значимости α будет иметь вид $\{\mathbf{x} \colon 2\ln\lambda(\mathbf{X}) \geqslant \chi^2_{1,1-\alpha}\}$. ■

В более сложных моделях, где максимизация правдоподобия в явном виде затруднительна, можно использовать численные методы оптимизации, такие как градиентный спуск. Потренироваться можно в задаче 10.2.

Наконец рассмотрим следующий частный случай применения КОП, имеющий множество приложений, которые будут снова и снова встречаться далее. Рассмотрим выборку $\mathbf{X} = (X_1, \ldots, X_n)$ категориальных признаков, как мы уже делали при рассматрении критерия χ^2 в разделе 9.3: для простоты скажем, что $X_i \in \{1, \ldots, k\}$, причём $p_i = \mathsf{P}(X_1 = i)$. Параметры модели лежат в $(k-1)$-мерном симплексе



$$\mathcal{S}_{k-1} = \left\{(x_1, \ldots, x_k) \colon \sum_{i=1}^k x_i = 1, \quad x_i \geqslant 0\right\},$$

а сама модель параметризуется $k-1$ числами $p_1, \ldots, p_{k-1}$ (p_k восстанавливается однозначно). Поставим на проверку гипотезу о том, что истинное значение вектора $\mathbf{p} = (p_1, \ldots, p_k)$ лежит на некоторой подповерхности в $\mathcal{S}_{k-1}$, которую для простоты можно описать $l < k-1$ параметрами $(\alpha_1, \ldots, \alpha_l)$, лежащими в открытом $U \subset \mathbb{R}^l$. Формально,

$$H_0 \colon \mathbf{p} \in \pi(U) = \{(p_1, \ldots, p_k) \mid \exists(\alpha_1, \ldots, \alpha_l) \in U \colon p_i = \pi_i(\alpha_1, \ldots, \alpha_l), \quad i = 1, \ldots, k\},$$

где $\pi \colon U \to S_{k-1}$ — инъективная функция, задающая параметризацию посредством $\alpha_1, \ldots, \alpha_l$. Такая гипотеза может символизировать собой выражение одних вероятностей

через другие или функциональные ограничения, которым подчиняются вероятности p_i. Для её проверки воспользуемся КОП, который требует нахождения ОМП на множестве $\Theta_0 = \pi(U)$ и $\Theta = \mathcal{S}_{k-1}$. Первая оценка, очевидно, зависит от природы функции π и в каждом случае вычисляется отдельно, обозначим её за $\widehat{\alpha}(\mathbf{X})$, то есть $\sup_{\mathbf{p}\in\Theta_0} f_{\mathbf{p}}(\mathbf{X}) = f_{\pi(\widehat{\alpha})}(\mathbf{X})$. Теперь найдём вторую оценку. Рассмотрим реализацию выборки $\mathbf{x} = (x_1, \ldots, x_n)$, в которой ν_i раз выпало значение i, таким образом, $\nu_1 + \ldots + \nu_k = n$. Тогда правдоподобие в точке $\mathbf{x}$ равно

$$f_{\mathbf{p}}(\mathbf{x}) = p_1^{\nu_1} \cdot \cdots \cdot p_k^{\nu_k}.$$

Для получения ОМП максимизируем логарифмическую функцию правдоподобия со следующими условиями:

$$\begin{cases} \nu_1 \ln p_1 + \ldots + \nu_k \ln p_k \longrightarrow \max\limits_{\mathbf{p}} \\ p_1 + \ldots + p_k = 1. \end{cases}$$

Решая задачу условной оптимизации, получаем ожидаемое решение $p_i = \frac{\nu_i}{n}$. Подставляя к формулу (10), находим статистику критерия:

$$T(\mathbf{x}) = 2\sum_{i=1}^{k} \nu_i \ln \frac{\nu_i}{n\pi(\widehat{\alpha})}.$$

Хотелось бы применить теорему Уилкса, чтобы сделать вывод о асимптотическом поведении статистики, однако для этого нужно потребовать некоторые условия регулярности, о которых мы заблаговременно умолчали в формулировке теоремы. Применительно к нашей задаче они будут иметь следующий вид:

1. Функции π_i дважды непрерывно дифференцируемы (модель достаточно гладкая);
2. Для всех $\alpha \in U$ и всех i справедлива оценка $\pi_i(\alpha) > c > 0$ (в условиях регулярности часто требуется компактность множества параметров, при этом p_i нельзя быть нулевыми, отсюда и такое требование);
3. Матрица $\left(\frac{\partial \pi_i}{\partial \alpha_j}\right)$ полного ранга l (тогда $\pi(U)$ будет гладкой поверхностью).

Стоит отметить, что второе условие формально не выполняется в большинстве случаев, но на практике оно не столь обременительно: можно считать, что c достаточно мало и на деле значения ниже него не реализуются. Поэтому либо про это условие не вспоминают вовсе, либо заявляют в самом начале, а потом ни разу не упоминают (мы пойдём по второму пути).

Теорема 10.2.

В условиях регулярности выше при любом $\alpha \in U$ статистика $T(\mathbf{X})$ стремится по распределению к χ^2_{k-1-l}.

10.4 Параметрический критерий χ^2

Полученный выше критерий уже вполне рабочий и позволяет решать множество задач, однако в литературе чаще всего встречается его аналог, который является обобщением

критерия χ^2 Пирсона. Чтобы понять, как его получить, заметим, что статистика критерия выше записывается через дивергенцию Кульбака-Лейблера:

$$T(\mathbf{x}) = 2\sum_{i=1}^{k} \nu_i \ln \frac{\nu_i}{n\pi_i(\widehat{\alpha})} = 2n \sum_{i=1}^{k} \frac{\nu_i}{n} \ln \frac{\nu_i}{n\pi_i(\widehat{\alpha})} = 2n \cdot D(\mathsf{P} \parallel \widehat{\mathsf{P}}),$$

где $\mathsf{P} = (\nu_1/n, \ldots, \nu_k/n)$ — эмпирическое распределение, а $\widehat{\mathsf{P}} = (\pi_1(\widehat{\alpha}), \ldots, \pi_k(\widehat{\alpha}))$ — наиболее правдоподобное распределение с точки зрения H_0, а величина ν_i, напомним, равна количеству выпадений i в выборке. При решении задачи 9.2 читатель уже убедился, что KL-дивергенция и χ^2-расстояние, равное

$$\chi^2(\mathsf{P} \parallel \mathsf{Q}) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{q_i},$$

асимптотически эквивалентны (с точностью до константы 2), то есть при $\mathsf{P} \to \mathsf{Q}$ отношение KL и χ^2 от этих распределений стремится к 2. Применяя те же рассуждения к нашей статистике, получаем предельное распределение для статистики χ^2 при верности H_0:

$$\chi^2(\mathbf{X}) = n \cdot \chi^2(\mathsf{P} \parallel \mathsf{Q}) = n \sum_{i=1}^{k} \frac{(\nu_i/n - \pi_i(\widehat{\alpha}))^2}{\pi_i(\widehat{\alpha})} = \sum_{i=1}^{k} \frac{(\nu_i - n\pi_i(\widehat{\alpha}))^2}{n\pi_i(\widehat{\alpha})} \xrightarrow{d} \chi^2_{k-1-l}.$$

Полученная статистика крайне похожа на аналогичную статистику хи-квадрат (8), только тут вместо конкретных значений p_i^0 мы подставляем их оценку максимального правдоподобия. Таким образом, обычная статистика χ^2 является частным случаем текущей, когда основная гипотеза говорит о принадлежности подмногообразию размерности 0, то есть точке.

Подытожим всё вышесказанное в одной теореме, непосредственное доказательство которой для любознательных читателей предлагается в книге [9, § 30.3].

Теорема 10.3.

Рассмотрим категориальную модель, в которой параметром является вектор $\mathbf{p} = (p_1, \ldots, p_k)$ с $p_i > 0$ и $\sum p_i = 1$, причём $X_1 \in \{1, \ldots, k\}$, $\mathsf{P}_{\mathbf{p}}(X_1 = i) = p_i$. Множество допустимых $\mathbf{p}$ обозначим за $\mathcal{S}_{k-1}$, а количество i в выборке за ν_i.
Пусть задано натуральное $l < k - 1$, открытое множество $U \subset \mathbb{R}^l$ и функции $\pi_i \colon U \to \mathcal{S}_{k-1}$, $i = 1, \ldots, k$, которые удовлетворяют следующим условиям:

- Найдётся $c > 0$ такое, что для всех $\alpha \in U$ и $i = 1, \ldots, k$ верно $\pi_i(\alpha) > c$;
- Функции π_i дважды непрерывно дифференцируемы;
- Матрица $\left(\frac{\partial \pi_i}{\partial \alpha_j}\right)$, $i = 1, \ldots, k$, $j = 1, \ldots, l$, имеет ранг l.

Тогда для всех $\alpha \in U$ статистика

$$\chi^2(\mathbf{X}) = \sum_{i=1}^{k} \frac{(\nu_i - n\pi_i(\widehat{\alpha}))^2}{n\pi_i(\widehat{\alpha}(\mathbf{X}))}, \tag{11}$$

где $\widehat{\alpha}(\mathbf{X})$ — ОМП для параметра $\alpha \in U$, стремится к χ^2_{k-1-l} при $n \to \infty$. Таким образом, критерий

$$R_\alpha = \left\{ \mathbf{x} \colon \sum_{i=1}^{k} \frac{(\nu_i - n\pi_i(\widehat{\alpha}))^2}{n\pi_i(\widehat{\alpha}(\mathbf{x}))} > \chi^2_{k-1-l,\,1-\gamma} \right\}$$

является асимптотическим критерием уровня значимости γ для проверки гипотезы $H_0\colon \mathbf{p} \in \pi(U)$ против общей альтернативы.

Полученный критерий называют *параметрическим или обобщённым критерием* χ^2. Звучит всё довольно сложно, поэтому предлагается сразу рассмотреть следующий хрестоматийный

Пример 10.3 (*пример 1 из § 57 [7]*)**.** Как известно, существует 4 группы крови, 0, A, B и AB. Они определяются наличием генов 0, A и B, причём первый из них является рецессивным. Значит, если вероятности получить от одного из родителей ген 0, A и B равны p, q и $r = 1 - p - q$ соответственно, то вероятности появления групп крови у ребёнка можно вычислить, как указано в таблице ниже. Проверим гипотезу о том, что такой механизм наследования имеет место.

Группа крови	Комбинации, приводящие к группе	Вероятность
0	00	r^2
A	0A, AA	$2pr + p^2$
B	0B, BB	$2qr + q^2$
AB	AB	$2pq$

Пусть нам дана выборка с частотами $\nu_1 = 121$, $\nu_2 = 120$, $\nu_3 = 79$ и $\nu_4 = 33$, которые равны количеству людей с группами 0, A, B и AB соответственно. При верности нулевой гипотезы вероятности появления конкретной группы определяется таблицей выше, а значит, функция правдоподобия имеет следующий вид:

$$f_{p,q,r}(\nu) = (r^2)^{\nu_1} \cdot (2pr + p^2)^{\nu_2} \cdot (2qr + p^2)^{\nu_3} \cdot (2pq)^{\nu_4}.$$

Для подсчёта ОМП параметров (p, q, r) потребуется помощь численных методов. С помощью простого градиентного спуска мы можем максимизировать логарифмическую функцию правдоподобия и получить оценки $\widehat{p} \approx 0.246$, $\widehat{q} \approx 0.173$, $\widehat{r} \approx 0.58$. Оценки самих вероятностей, $\pi(\alpha)$ в старых обозначениях, равны

$$\widehat{p_1} \approx 0.337, \quad \widehat{p_2} \approx 0.347, \quad \widehat{p_3} \approx 0.231, \quad \widehat{p_4} \approx 0.085.$$

Подставим полученные значения в статистику (11), что даст нам $\chi^2(\nu) \approx 0.437$.

Разберёмся со степенями свободы предельного распределения статистики. Изначально было $k = 4$ групп, при этом гипотеза заключалась в том, что модель параметризуется двумя числами — p и q (r восстанавливается однозначно), то есть Θ_0 имеет размерность $l = 2$. Стало быть, при верности нулевой гипотезы статистика критерия распределена примерно как $\chi^2_{k-1-l} = \chi^2_1$. Для уровня значимости $\alpha = 0.05$ пороговое значение равно $\chi^2_{1,1-\alpha} \approx 3.841$, а соответствующее p-value равно ≈ 0.5084, то есть оснований для отвержения гипотезы о виде наследования группы крови нет. ■

Задачи

Задача 10.1. Подберите **(а)** дискретное; **(б)** сингулярное распределение, на котором критерий Харке-Бера будет несостоятельным.

Задача 10.2. Рассмотрим модель сдвига-масштаба с распределением Коши:

$$\rho_{\theta_1,\theta_2}(x) = \frac{1}{\pi\theta_2\left[1+\left(\frac{x-\theta_1}{\theta_2}\right)^2\right]}.$$

С помощью численных методов (например, функции `scipy.optimize.minimize`) научитесь считать статистику КОП для проверки гипотезы $H_0\colon \theta_1 = 0$. Зафиксировав параметр масштаба и перебрав разные размеры выборки, сгенерируйте несколько выборок при верности гипотезы H_0 и проверьте распределение статистики критерия и pvalue. При разных параметрах сдвига и размерах выборки изучите критерий на предмет мощности.

11 Корреляционный анализ

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:

$$H_0 \colon F_{X,Y}(x, y) = F_X(x) F_Y(y).$$

11.1 Коэффициенты корреляции

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.

11.1.1 Коэффициент корреляции Пирсона

Самое простое, что можно придумать, — это взять «выборочную корреляцию», то есть воспользоваться методом подстановки для функционала в лице корреляции.

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

$$\widehat{\rho} \xrightarrow{\mathsf{P}} \rho(X_1, Y_1) = \frac{\mathsf{cov}(X_1, Y_1)}{\sqrt{\mathsf{D}X_1 \mathsf{D}Y_1}} = \mathsf{corr}(X_1, Y_1), \quad n \to \infty.$$

Обычно этот коэффициент используют в случае нормальности выборок: при таком допущении можно построить достаточно мощный критерий проверки некоррелируемости, который даёт следующая

Теорема 11.1.

Если нормально распределённые выборки X, Y независимы и $n > 2$, то

$$P(\widehat{\rho}) := \widehat{\rho} \sqrt{\frac{n-2}{1-\widehat{\rho}^2}} \sim T_{n-2}.$$

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 подставить в виде

$$R = [-1; 1] \setminus (t_{\alpha/2}; t_{1-\alpha/2}),$$

где t_p — p-квантиль прообраза распределения T_{n-2} при действии отображения P.

Для удобства подсчёта можно пользоваться реализацией `scipy.stats.pearsonr`.

11.1.2 Коэффициент корреляции Спирмэна

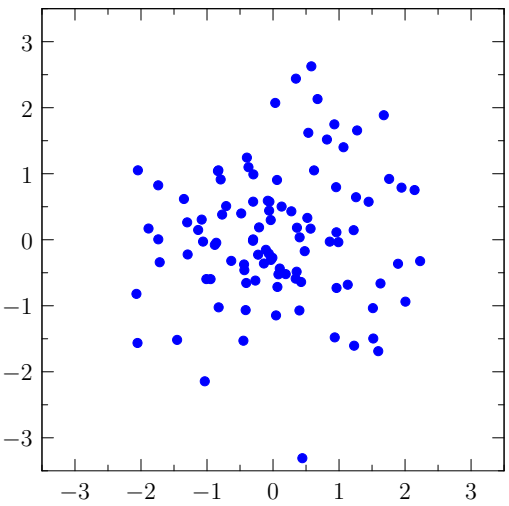
Какие минусы у коэффициента выше? Во-первых, конечно, не все рассматриваемые выборки нормальны, хотя такое допущение встречается довольно часто. Самое неприятное – низкая робастность, то есть неустойчивость статистики к выбросам, что особенно характерно для тех из них, которые основаны на выборочном 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.

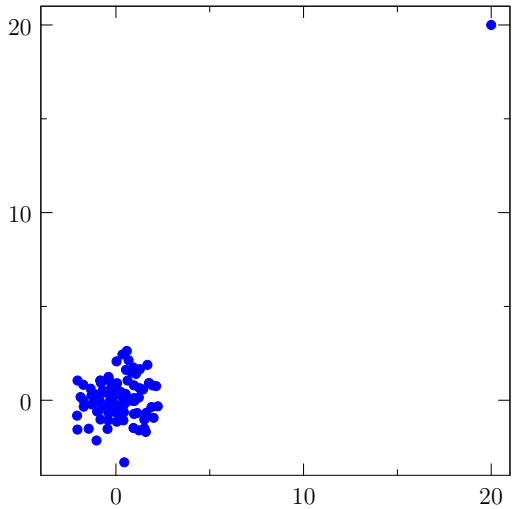
Определение. Пусть R_i и S_j – место в вариационном ряду для X_i и Y_j соответственно. *Коэффициентом корреляции Спирмэна* называют следующую статистику:

$$\rho_S = \frac{\sum_{i=1}^n (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^n (R_i - \overline{R})^2 \sum_{i=1}^n (S_i - \overline{S})^2}}.$$

Пример 11.1. Посмотрим, как данный коэффициент борется с выбросами. Рассмотрим выборку $(X_i, Y_i) \sim \mathcal{N}(0, E)$, в которой закрался выброс:



(a) Выборка из независимых компонент



(b) То же самое, но с выбросом

Для левой выборки значения коэффициентов Пирсона и Спирмэна примерно равны 0.08 и 0.06 соответственно, что довольно мало и вполне отражает действительность. С правой выборкой всё несколько хуже: значение коэффициента Пирсона для неё равно $\widehat{\rho} \approx 0.81$, что катастрофически много (Т-критерий выше явно отклоняет гипотезу о независимости). Но на коэффициент Спирмэна добавление выброса повлияет незначительно: ранги просто немного сдвинутся, что даст нам значение $\rho_S \approx 0.087$. Таким образом, сей коэффициент можно использовать для распределений, далёких от нормального. ■

Может возникнуть вопрос, а как вычислять ранги для выборки с повторяющимися элементами? 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 11.2.

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.

Доказательство. Первое утверждение проверяется непосредственно. Третье является следствием неравенства КБШ. Осталось найти 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:

$$\mathsf{D}\rho_S = \frac{144}{(n^3-n)^2}\mathsf{D}\sum R_iS_i = \frac{144}{(n^3-n)^2}\left[\mathsf{E}\left(\sum R_iS_i\right)^2 - \left(\mathsf{E}\sum R_iS_i\right)^2\right] =$$

$$= \frac{144}{(n^3-n)^2}\left[\sum \mathsf{E}R_i^2S_i^2 + \sum_{i\neq j}\mathsf{E}R_iS_iR_jS_j - (n\cdot \mathsf{E}R_1S_1)^2\right] =$$

$$= \frac{144}{(n^3-n)^2}\left[n\cdot \mathsf{E}R_1^2\cdot \mathsf{E}S_1^2 + n(n-1)\mathsf{E}R_1R_2\cdot \mathsf{E}S_1S_2 - (n\cdot \mathsf{E}R_1\cdot \mathsf{E}S_1)^2\right] =$$

$$= \frac{144}{n(n^2-1)^2}\left[\frac{(n+1)^2(2n+1)^2}{36} + (n-1)\cdot\frac{(n+1)^2(3n+2)^2}{144} - n\cdot\frac{(n+1)^4}{16}\right] =$$

$$= \frac{1}{n(n-1)^2}\left(4(2n+1)^2 + (n-1)(3n+2)^2 - 9n(n+1)^2\right) = \frac{1}{n-1}.$$

□

11.1.3 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 увеличе-

ние 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:

$$\tau = \frac{2}{n(n-1)} \cdot T = \frac{2}{n(n-1)} \sum_{i<j} sign(X_i - X_j) \cdot sign(Y_i - Y_j)$$

С учётом того, что $S + R = \frac{n(n-1)}{2}$, коэффициент корреляции можно переписать как

$$\tau = 1 - \frac{4}{n(n-1)} R.$$

Как и ранее, у неё можно найти среднее и дисперсию, по которым можно составить предельный закон. Мы приведём лишь численные значения, прийти к которым предлагается читателю в качестве упражнения.

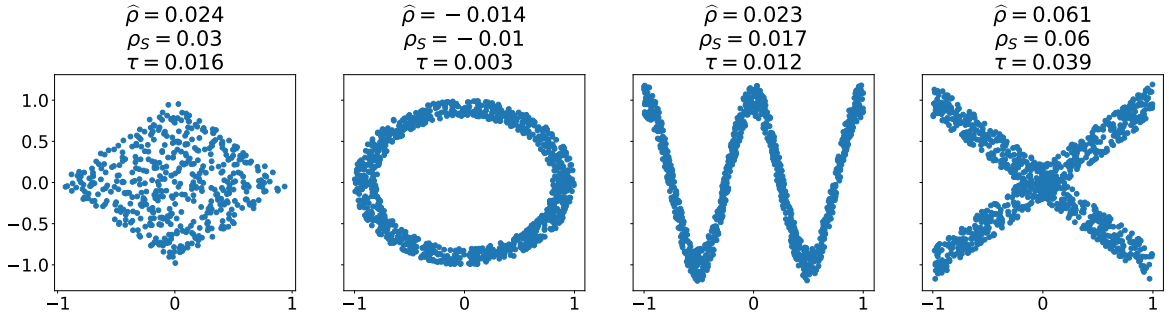
Теорема 11.3.

При верности H_0 выполнено $\mathsf{E}\tau = 0$, $\mathsf{D}\tau = \frac{2(2n+5)}{9n(n-1)}$, и есть сходимость

$$\frac{\tau}{\sqrt{\mathsf{D}\tau}} \xrightarrow{d} \mathcal{N}(0,1).$$

11.2 Критерий χ^2 и таблицы сопряжённости

Одним из недостатков коэффициентов корреляции выше является их немощность при симметричности данных. На картинках ниже имеется явная зависимость между двумя признаками, однако коэффициенты на это никак не реагирует и выдают околонулевые значения.



Данную проблему можно решить и визуальным анализом, однако мы обсудим другой, более формальный способ, который является ещё одним аналогом критерия хи-квадрат. Вообще говоря, он предназначен для категориальных признаков, но по аналогии с обычным критерием хи-квадрат его можно применять и в общем случае.

Пусть признак X принимает m значений $x_1, \ldots, x_m$, а признак Y — k значений $y_1, \ldots, y_k$. Имеется выборка (X_i, Y_i) из n наблюдений, для которой обозначим за ν_{ij} количество пар таких, что в них признак X равен x_i, а Y равен y_j, то есть

$$\nu_{ij} = \sum_{l=1}^{n} I(X_l = x_i, Y_l = y_j).$$

Полученные значения обычно записывают в таблицу, которую называют *таблицей сопряжённости*. Введём обозначения $p_{ij} = \mathsf{P}(X_1 = x_i, Y_1 = y_j)$, $p_{i\bullet} = \mathsf{P}(X_1 = x_i) = \sum_j p_{ij}$ и аналогично $p_{\bullet j} = \mathsf{P}(Y_1 = y_j) = \sum_i p_{ij}$. Нулевая гипотеза о независимости X и Y равносильна равенствам

$$p_{ij} = p_{i\bullet} p_{\bullet j}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, k,$$

то есть имеет место принадлежность некоторой подповерхности Θ_0 в исходном пространстве параметров Θ. Легко посчитать её размерность — $m + k - 2$. Действительно, параметры $p_{1\bullet}, \ldots p_{(m-1)\bullet}$, $p_{\bullet 1}, \ldots p_{\bullet(k-1)}$ берутся произвольными (в рамках разумного), числа $p_{m\bullet} = 1 - p_{1\bullet} - \ldots - p_{(m-1)\bullet}$ и $p_{\bullet k} = 1 - p_{\bullet 1} - \ldots - p_{\bullet(k-1)}$ определяются по ним однозначно, откуда из формул выше получаем все p_{ij}.

Гипотезу такого вида можно проверить с помощью параметрического критерия χ^2 из раздела 10.4, для чего нужно найти ОМП при верности нулевой гипотезы. Функция правдоподобия будет равна

$$f_{\mathbf{p}}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{m} \prod_{j=1}^{k} p_{ij}^{\nu_{ij}} = \prod_{i=1}^{m} \prod_{j=1}^{k} (p_{i\bullet} p_{\bullet j})^{\nu_{ij}} = \prod_{i=1}^{m} p_{i\bullet}^{\nu_{i\bullet}} \cdot \prod_{j=1}^{k} p_{\bullet j}^{\nu_{\bullet j}},$$

где $\nu_{i\bullet} = \sum_j \nu_{ij}$, $\nu_{\bullet j} = \sum_i \nu_{ij}$. Максимизируя каждый множитель по отдельности, получаем оценки $\widehat{p}_{i\bullet} = \nu_{i\bullet}/n$, $\widehat{p}_{\bullet j} = \nu_{\bullet j}/n$. Таким образом, по теореме 10.3 при независимости признаков имеется сходимость

$$\chi^2(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(\nu_{ij} - n\widehat{p}_{i\bullet}\widehat{p}_{\bullet j})^2}{n\widehat{p}_{i\bullet}\widehat{p}_{\bullet j}} = n \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(\nu_{ij} - \nu_{i\bullet}\nu_{\bullet j}/n)^2}{\nu_{i\bullet}\nu_{\bullet j}} \xrightarrow{d} \chi^2_{(m-1)(k-1)}, \tag{12}$$

где $(m-1)(k-1) = mk - 1 - (m + k - 2)$ — коразмерность Θ_0 в Θ.

Пример 11.2. В 1892 году в работе Фрэнсиса Гальтона «Finger Prints» изучалась наследственность типов отпечатков пальца: в виде дуг, петлей и завитков. Для этого бралась выборка из 105 пар братьев и сестёр и сравнивался тип отпечатков пальцев в каждой паре — получился набор из 9 чисел, записанных в таблицу сопряжённости: по горизонтали берётся тип отпечатков у первого ребёнка из пары, по вертикали — второго.

	Дуги	Петли	Завитки
Дуги	5	12	2
Петли	4	42	15
Завитки	1	14	10

Если предположить, что наследственность роли не играет, то зависимости быть не должно. Проверим эту гипотезу на уровне значимости 0.05 с помощью критерия χ^2. Статистику (12) можно посчитать и напрямую, но удобнее занести данные в Python и воспользоваться функцией `scipy.stats.chi2_contingency`.

```
fraternal_obs = [
    [5, 12, 2],
    [4, 42, 15],
    [1, 14, 10]
]
sps.chi2_contingency(fraternal_obs).pvalue
```

```
0.024719148645087168
```

Таким образом, гипотеза о независимости данных признаков у братьев и сестёр отклоняется на взятом уровне значимости. ■

Задачи

Задача 11.1. Докажите теорему 11.1.

Указание. Найдите условное распределение $\widehat{\rho}$ при фиксированной выборке $\mathbf{X}$ и убедитесь, что оно одинаково при любом условии. В частности, отсюда будет следовать, что утверждение теоремы справедливо даже в случае ненормальности одной из выборок.

12 Критерии однородности

Вообще говоря, критерии однородности проверяют нулевую гипотезу о том, правда ли, что две данные выборки пришли нам из одного и того же распределения. Впрочем, в этой главе будут рассмотрены и другие двухвыборочные критерии, которые проверяют равенство средних, дисперсий или медиан. Их объединяет умение проверять наличие эффекта: верно ли, что влияние некоторых факторов меняет распределение, или они не вносят существенного вклада в поведение наблюдаемых величин? Обычно нас будут интересовать следующие две ситуации:

Ind Выборки $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ независимы;

Rel Выборки $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_n)$ связаны, то есть $(X_1, Y_1), \ldots, (X_n, Y_n)$ — независимые случайные векторы.

Примером случая **Ind** могут служить выборки, которые описывают поведение некоторых метрик при A/B-тестировании, когда испытуемые делятся на две *независимые* группы: контрольную и экспериментальную, причём над последней проводят некоторые преобразования (дают новый препарат, открывают доступ к новым возможностям приложения и т.д.), и критерий должен показать, имеет ли место быть влияние сего преобразования. Иллюстрацией случая **Rel** служат показатели *одних и тех же* испытуемых, с которыми случилась некоторая метаморфоза между замерами: это может быть температура до и после принятия лекарства или успеваемость класса с течением времени. Также встречаются выборки, которые формально не связаны и могут иметь разный размер, но имеют зависимости. Например, результаты социологических опросов: не всегда спрашивают одних и тех же людей, но при этом среди них могут попадаться одинаковые.

Самым простым способом проведения теста на равенство средних является критерий Вальда, который ранее упоминался в разделе 7.1. Напомним, что в его основе лежит асимптотически нормальная оценка некоторого параметра, который мы хотим проверить на равенство каком-то числу. В качестве такого параметра можно взять разность средних $\delta = \mu_1 - \mu_2$, где $\mu_1 = \mathsf{E}X_1$ и $\mu_2 = \mathsf{E}Y_1$, то есть проверяется гипотеза

$$H_0 \colon \delta = \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1 \colon \delta \neq 0.$$

В частности,

- В случае **Ind** в качестве оценки параметра δ можно взять $\widehat{\delta} = \widehat{\mu}_1 - \widehat{\mu}_2$, где, предсказуемо, $\widehat{\mu}_1 = \overline{\mathbf{X}}$ и $\widehat{\mu}_2 = \overline{\mathbf{Y}}$. В силу независимости выборок дисперсия такой оценки равна
$$\mathsf{D}\widehat{\delta} = \mathsf{D}\widehat{\mu}_1 + \mathsf{D}\widehat{\mu}_2 = \frac{\mathsf{D}X_1}{n} + \frac{\mathsf{D}Y_1}{m},$$
причём $\mathsf{D}X_1$ и $\mathsf{D}Y_1$ оцениваются с помощью несмещённых выборочных дисперсий $S^2(\mathbf{X})$ и $S^2(Y)$ соответственно. Таким образом, имеется сходимость
$$W(\mathbf{X}, \mathbf{Y}) = \frac{\overline{\mathbf{X}} - \overline{\mathbf{Y}}}{\sqrt{\frac{S^2(\mathbf{X})}{n} + \frac{S^2(\mathbf{Y})}{m}}} \xrightarrow{d} \mathcal{N}(0,1)$$
и асимптотический критерий на уровне значимости α
$$R_\alpha = \left\{(\mathbf{x}, \mathbf{y}) \colon \, |W(\mathbf{x}, \mathbf{y})| \geqslant z_{1-\alpha/2}\right\}.$$
Напомним также о возможности модификации данного теста для односторонней альтернативы. Например, если H_0 проверяется против $H_2 \colon \delta > 0$ (то есть рассматривается возможность только лишь увеличения среднего), критерий на уровне значимости α будет иметь вид $R'_\alpha = \{(\mathbf{x}, \mathbf{y}) \colon W(\mathbf{x}, \mathbf{y}) \geqslant z_{1-\alpha}\}$.

- В случае **Rel** логично рассмотреть величины $D_i = X_i - Y_i$: для i-ого испытуемого она показывает изменение в наблюдаемой метрике. При верности гипотезы H_0 имеем $\mathsf{E}D_i = \mathsf{E}X_i - \mathsf{E}Y_i = \mu_1 - \mu_2 = 0$, поэтому в качестве оценки δ можно взять $\overline{\mathbf{D}}$. Дисперсия такой оценки по независимости испытуемых равна $n^{-1}\mathsf{D}D_1$, которую можно оценить несмещённой выборочной дисперсией $S^2(\mathbf{D})$. Итого, получается следующий асимптотический критерий уровня значимости α для проверки H_0 против H_1:

$$R_\alpha = \left\{ \sqrt{n} \cdot \left| \frac{\overline{\mathbf{D}}}{S^2(\mathbf{D})} \right| \geqslant z_{1-\alpha/2} \right\}.$$

Такой способ крайне прост, однако обладает лишь асимптотическими свойствами, и возможно в конкретной ситуации будет иметь крайне большие вероятности ошибок I и II рода. Далее в этой главе мы опишем более мощные критерии.

12.1 Тесты для нормальных выборок

Статистики критериев Вальда, которые были рассмотрены выше, можно немного видоизменить, уточнив их распределение, если сделать допущение о нормальности наших данных. Будем работать в парадигме **Ind**, а также добавим, что $(X_1, \ldots, X_n)$ есть выборка из $\mathcal{N}(\mu_1, \sigma^2)$, а $(Y_1, \ldots, Y_m)$ — выборка из $\mathcal{N}(\mu_2, \sigma^2)$, то есть параметры сдвига и масштаба нам неизвестны, но мы знаем, что данные распределены нормально и с одинаковой дисперсией. Проверим гипотезу

$$H_0\colon \mu_1 = \mu_2 \quad \text{versus} \quad H_1\colon \mu_1 \neq \mu_2.$$

При верности основной гипотезы $\overline{\mathbf{X}} - \overline{\mathbf{Y}} \sim \mathcal{N}(0, \sigma^2(1/n + 1/m))$, то есть

$$\sqrt{\frac{nm}{n+m}} \cdot \frac{\overline{\mathbf{X}} - \overline{\mathbf{Y}}}{\sigma} \sim \mathcal{N}(0,1). \tag{13}$$

Воспользуемся идеей раздела 6.2, где похожую случайную величину мы делили на корень из оценки дисперсии, чтобы неизвестная σ сократилась, а полученное отношение имело распределение Стьюдента. Выборочные дисперсии $s^2(\mathbf{X})$ и $s^2(\mathbf{Y})$ независимы как функции от независимых выборок, поэтому $ns^2(\mathbf{X})/\sigma^2$ и $ms^2(\mathbf{Y})/\sigma^2$, как величины с распределением хи-квадрат, в сумме дают

$$\frac{ns^2(\mathbf{X}) + ms^2(\mathbf{Y})}{\sigma^2} \sim \chi^2_{n+m-2}. \tag{14}$$

Выборочные среднее и дисперсия для разных выборок, очевидно, независимы, как и для одной выборки, что было показано в примере 5.6. Таким образом, величины из (13) и (14) независимы, поэтому

$$T(\mathbf{X}, \mathbf{Y}) = \sqrt{\frac{nm}{n+m}} \cdot \frac{\overline{\mathbf{X}} - \overline{\mathbf{Y}}}{S} \sim T_{n+m-2}, \text{ где } S = \sqrt{\frac{ns^2(\mathbf{X}) + ms^2(\mathbf{Y})}{n+m-2}}. \tag{15}$$

t-критерий, основанный на сей статистике, будет иметь вид

$$R_\alpha = \left\{ (\mathbf{x}, \mathbf{y})\colon |T(\mathbf{x}, \mathbf{y})| \geqslant T_{n+m-2,1-\alpha/2} \right\}.$$

Как можно заметить, в ходе рассуждений существенно используется равенство дисперсий, что не всегда получается знать априори. Первый вариант решения проблемы заключается в проверке гипотезы равенства дисперсий. Пусть независимы выборки $\mathbf{X}$ и $\mathbf{Y}$ пришли из распределения $\mathcal{N}(\mu_1, \sigma_1^2)$ и $\mathcal{N}(\mu_2, \sigma_2^2)$ соответственно, проверим гипотезу $H_0\colon \sigma_1 = \sigma_2$. Воспользуемся независимыми статистиками $ns^2(\mathbf{X})/\sigma_1^2 \sim \chi^2_{n-1}$ и $ms^2(\mathbf{Y})/\sigma_2^2 \sim \chi^2_{m-1}$. При верности нулевой гипотезы дисперсии равны, поэтому если мы поделим одно на другое, то

неизвестная σ сократится. Полученное распределение имеет специальное название, в честь которого назван и сам критерий.

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 $\zeta \sim F_{a,b}$

Итого, получаем

$$\frac{\frac{n}{n-1} \cdot s^2(\mathbf{X})}{\frac{m}{m-1} \cdot s^2(\mathbf{Y})} = \frac{S^2(\mathbf{X})}{S^2(\mathbf{Y})} \sim F_{n-1,m-1},$$

где S^2 — несмещённая оценка дисперсии. На основании этой статистики можно построить так называемый *F-критерий Фишера*, который имеет вид

$$R_\alpha = \left\{ (\mathbf{x}, \mathbf{y}) \colon \frac{S^2(\mathbf{x})}{S^2(\mathbf{y})} \in (f_{\alpha/2}, f_{1-\alpha/2}) \right\},$$

где f_p — p-квантиль распределения $F_{n-1,m-1}$.

Если же сей критерий отверг нулевую гипотезу, то можно воспользоваться *критерием Аспина-Уэлча*, который чуть менее мощный, чем t-критерий выше. Теперь уже выборки $\mathbf{X}$ и $\mathbf{Y}$ приходят из распределения $\mathcal{N}(\mu_1, \sigma_1^2)$ и $\mathcal{N}(\mu_2, \sigma_2^2)$ соответственно, то есть дисперсии могут различаться. Во главе критерия стоит статистика

$$W(\mathbf{X}, \mathbf{Y}) = \frac{\overline{\mathbf{X}} - \overline{\mathbf{Y}}}{\sqrt{\frac{S^2(\mathbf{X})}{n} + \frac{S^2(\mathbf{Y})}{m}}},$$

которая встречалась нам ранее в критерии Вальда. Если $H_0 \colon \mu_1 = \mu_2$ верна, то эта статистика приблизительно распределена как T_K, где

$$K = \left(\frac{S^2(\mathbf{X})}{n} + \frac{S^2(\mathbf{Y})}{m} \right)^2 \cdot \left(\frac{S^4(\mathbf{X})}{n^2(n-1)} + \frac{S^4(\mathbf{Y})}{m^2(m-1)} \right)^{-1}. \tag{16}$$

Откуда появляется такое причудливое число степеней свобод? Причина в следующем: по аналогии с t-критерием нам хотелось бы представить статистику $W(\mathbf{X}, \mathbf{Y})$ как отношение $\mathcal{N}(0,1)$ к $\sqrt{\chi^2_K/K}$ для некоторого K. Если нормировать числитель, получится

$$W(\mathbf{X}, \mathbf{Y}) = \underbrace{\frac{\overline{\mathbf{X}} - \overline{\mathbf{Y}}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}}_{\sim \mathcal{N}(0,1)} \Bigg/ \sqrt{\underbrace{\frac{S^2(\mathbf{X})/n + S^2(\mathbf{Y})/m}{\sigma_1^2/n + \sigma_2^2/m}}_{\approx \chi^2_K/K?}}.$$

В связи с таким представлением нужно подобрать параметр K таким образом, чтобы выделенная статистика максимально сильно походила на χ^2_K/K. Так как с ростом степеней свободы хи-квадрат становится всё больше похожим на нормальное, то достаточно приравнять их среднее и дисперсию, что уже даст достаточно точное приближение. Их средние и так совпадают, так как $S^2(\mathbf{X})$ и $S^2(\mathbf{Y})$ несмещённо оценивают σ_1^2 и σ_2^2 соответственно. Приравнивание дисперсий и решение полученного уравнения дадут нам выражение K через параметры σ_1^2 и σ_2^2, что при замене на их состоятельные оценки приводит к статистике (16).

12.2 Предположение нормальности/независимости и метод бакетов

Изложенное выше может показаться совершенно бесполезным, так как довольно редко встречаются выборки, распределённые непременно нормально. Однако, несмотря на сей факт, данное предположение часто допускают и всё равно используют критерий Стьюдента и иже с ним. Отчего же так?

Если коротко, то всему виной ЦПТ: для большого объёма данных в силу предельных теорем выборочные характеристики распределены почти что нормально, отчего многие результаты выше остаются в силе. Тем более распределение Стьюдента, в соответствии с коим распределена статистика t-критерия, с ростом n приближается к нормальному, и критерий вырождается в обычный критерий Вальда, который справедлив в куда большем числе случаев.

Впрочем, и для малых n стьюдентовское приближение порой бывает состоятельным. Если говорить опять неформально, то распределение Стьюдента, в отличие от нормального, обладает «тяжёлыми хвостами», что вполне характерно для суммы случайных величин с распределением, отличным от нормального.

Чуть более строго, присмотримся внимательно к статистике (15). Чтобы она имела распределение Стьюдента, необходимы три вещи: 1) распределение числителя нормально; 2) распределение содержимого корня в знаменателе есть хи-квадрат; 3) числитель и знаменатель независимы. В асимптотическом плане первые два пункта гарантирует ЦПТ, ведь выборочные среднее и дисперсия асимптотически нормальны, а хи-квадрат с ростом степеней свобод само по себе почти что нормально. Третий пункт даже асимптотически верен не всегда, но в широком наборе случае и он имеет место (см. задачу 1.3). Таким образом, даже в отсутствии честной нормальности данных использование t-критерия бывает предпочтительнее обычного критерия Вальда.

Куда более опасным и незаметным

12.3 Модернизации критериев согласия

Видоизменив критерии согласия, озвученные в параграфе 9, можно получить аналогичные критерии, проверяющие гипотезу о равенстве двух распределений против общей альтернативы в случае **Ind**. Здесь мы приведём краткую сводку.

Критерий Колмогорова-Смирнова в качестве статистики рассматривает наибольшее отклонение у двух эмпирических распределений:

$$D_{n,m}(\mathbf{X},\mathbf{Y}) = \sup_{z\in\mathbb{R}} |\widehat{F}_n(z) - \widehat{G}_m(z)|,$$

где $\widehat{F}_n$ и $\widehat{G}_m$ — эмпирические функции распределения, построенные по независимым выборкам $\mathbf{X} = (X_1,\dots,X_n)$ и $\mathbf{Y} = (Y_1,\dots,Y_m)$ соответственно. Как и в теореме Колмогорова, можно найти предельное распределение данной статистики, определив тем самым асимптотический критерий.

Теорема 12.1.

Пусть F и G — непрерывные распределения, из которых пришли выборки $\mathbf{X}$ и $\mathbf{Y}$ соответственно. Тогда

$$\sqrt{\frac{nm}{n+m}}\cdot D_{n,m}(\mathbf{X},\mathbf{Y}) \xrightarrow{d} K, \quad n,m\to\infty,$$

где K — распределение Колмогорова, определяемое функцией распределения (6).

Таким образом, в условиях теоремы имеется критерий уровня значимости α вида

$$R_\alpha = \left\{ (\mathbf{x}, \mathbf{y})\colon \sqrt{\frac{nm}{n+m}} \cdot D_{n,m}(\mathbf{x}, \mathbf{y}) > k_{1-\alpha} \right\},$$

который реализован в функции `scipy.stats.ks_2samp`.

Критерий Розенблатта же является модернизацией критерия Крамера-фон Мизеса-Смирнова, в котором рассматривается L_2-норма разности распределений. Только если в оригинальном критерии интеграл брался по теоретическому распределению, то в этот раз берётся эмпирическое совместное распределение: каждому наблюдению в двух выборках приписывается вес $\frac{1}{n+m}$. Формально такую функцию распределения можно записать так:

$$\widehat{H}_{n,m}(x) = \frac{n}{n+m}\widehat{F}_n(x) + \frac{m}{n+m}\widehat{G}_m(x).$$

Статистика критерия имеет вид

$$\omega^2_{n,m}(\mathbf{X}, \mathbf{Y}) = \int_{\mathbb{R}} \left(\widehat{F}_n(x) - \widehat{G}_m(x)\right)^2 d\widehat{H}_{n,m}(x).$$

При верности нулевой гипотезы $H_0\colon F = G$ и $n, m \to \infty$ статистика $\frac{nm}{n+m}\omega^2_{n,m}$ стремится к распределению F_1 из раздела 9.2, к которому стремится и статистика критерия Крамера-фон Мизеса-Смирнова. В классических библиотеках данный критерий не реализован, поэтому полезной будет следующая альтернативная формула для статистики критерия:

$$\omega^2_{n,m}(\mathbf{X}, \mathbf{Y}) = \frac{1}{nm}\left(\frac{1}{6} + \frac{1}{m}\sum_{i=1}^{n}(R_i - i)^2 + \frac{1}{n}\sum_{j=1}^{m}(S_j - j)^2\right) - \frac{2}{3}, \tag{17}$$

где R_i и S_j — ранги наблюдений $X_{(i)}$ и $Y_{(j)}$ в объединённом вариационном ряду.

Критерий Андерсона-Дарлинга устроен несколько по-другому: он берёт не отклонение эмпирических распределений между собой, а отклонение каждого из них от совместного эмпирического распределения. Это позволяет обобщить задачу на случай нескольких выборок. Конкретнее, рассмотрим k независимых выборок $\mathbf{X}_1, \ldots, \mathbf{X}_k$ размера $n_1, \ldots, n_k$ соответственно. Пусть $\widehat{F}_i$ — эмпирическая функция распределения, построенная по $\mathbf{X}_i$, а $\widehat{H}$ — по совокупности всех выборок. Введём статистику

$$\Omega^2(\mathbf{X}_1, \ldots, \mathbf{X}_k) = \sum_{i=1}^{k} n_i \int_A \frac{(\widehat{F}_i(x) - \widehat{H}(x))^2}{\widehat{H}(x)(1 - \widehat{H}(x))}\, d\widehat{H}(x),$$

где $A = \{x\colon H(x) < 1\}$, чтобы интеграл был конечным. При верности гипотез, что все распределения выборок одинаковы, с ростом n_i распределение статистики Ω^2 стремится к некоторому фиксированному распределению. Не будем опять же вдаваться в подробности его устройства, ведь для проверки гипотезы имеется удобная функция `scipy.stats.anderson_ksamp`.

Критерий однородности χ^2 можно получить как частный случай параметрического критерия χ^2 или критерия отношения правдоподобий, остановимся на первом варианте.

Пусть всё также имеется k независимых выборок $\mathbf{X}_1, \ldots, \mathbf{X}_k$ из категориальной модели, то есть наблюдения принадлежат некоторому дискретному множеству $\{1, \ldots, m\}$. В общем случае для каждой выборки вероятность выпадения конкретного признака своя, а именно

$p_{ij} = \mathsf{P}(X_{i1} = j)$. Вектор параметров модели $\{p_{ij}\}_{i=1,\dots,k}^{j=1,\dots,m}$ полностью описывается $m - k$ числами p_{ij}, $i = 1, \dots, k-1$, $j = 1, \dots, m$, так как величины p_{kj} однозначно восстанавливаются по остальным вероятностям. Таким образом, общее пространство параметров имеет размерность $m(k-1)$.

Нулевая гипотеза H_0 же состоит в том, что на самом деле $p_{1j} = p_{2j} = \dots = p_{kj}$ для всех $j = 1, \dots, m$. В таком случае множество параметров имеет размерность $k - 1$, так как $p_{11}, \dots, p_{1(k-1)}$ можно взять произвольными, а остальные p_{ij} выражаются через них. Таким образом, статистика (11) параметрического критерия χ^2 будет сходится к распределению хи-квадрат с $\dim(\Theta) - \dim(\Theta_0) = (mk - m) - (k - 1) = (m - 1)(k - 1)$ степенями свободы.

ОМП для параметров модели при верности H_0 находится довольно просто: функция правдоподобия в таком случае будет равна

$$f(\mathbf{x}_1, \dots, \mathbf{x}_k) = \prod_{i=1}^{k} \prod_{j=1}^{m} p_{ij}^{\nu_{ij}} = \prod_{j=1}^{m} p_{1j}^{\nu_{\bullet j}},$$

где ν_{ij} — количество признаков j в реализации выборки $\mathbf{x}_i$, $\nu_{\bullet j} = \sum_i \nu_{ij}$. Максимизируя полученную функцию, получаем оценки $\widehat{p}_{ij} = \nu_{\bullet j}/n$, где $n = \sum_j \nu_{\bullet j}$ — общее количество элементов в выборках. Итого, статистика примет вид

$$\chi^2(\mathbf{X}_1, \dots, \mathbf{X}_k) = \sum_{i=1}^{k} \sum_{m=1}^{j} \frac{(\nu_{ij} - \nu_{i\bullet}\widehat{p}_{ij})^2}{\nu_{i\bullet}\widehat{p}_{ij}} = \sum_{i=1}^{k} \sum_{m=1}^{j} \frac{(\nu_{ij} - \nu_{i\bullet}\nu_{\bullet j}/n)^2}{\nu_{i\bullet}\nu_{\bullet j}/n},$$

где $\nu_{i\bullet} = \sum_j \nu_{ij}$ — количество наблюдений в i-ой выборке. Заметим, что и статистика, и её предельный закон совпадает с аналогичной статистикой (12) для проверки коррелированности двух признаков: здесь роль второго признака играет номер выборки, из которого пришло наблюдение. Дополнительное удобство ещё и в том, что для проверки можно использовать ту же самую библиотечную функцию.

Пример 12.1. Весной 2024 года на курсе мат. логики Даниил Владимирович по объективным причинам задержал выдачу последнего домашнего задания и выдал его в канун экзаменов. Многим студентам показалось, что это негативно скажется на его решаемости, а стало быть и на итоговых результатах семестра. Проверим это статистически.

Возьмём табличку с ведомостью, нас будет интересовать распределение итоговых баллов (целое число от 0 до 4) за 2024 и 2023 года (ранние данные брать не будем, потому что тогда оценка была от 0 до 3). Запишем данные в таблицу сопряжённости:

Год \ Кол-во баллов	0	1	2	3	4
2023	32	63	64	25	18
2024	17	63	49	31	13

Проверим на уровне значимости 0.05 гипотезу о том, что распределение не поменялось. Вызовем соответствующую функцию и посмотрим на pvalue.

```
arr = np.array([
    [32, 63, 64, 25, 18],
    [17, 63, 49, 31, 13]
])
sps.chi2_contingency(arr).pvalue
```

```
0.21264668380037094
```

Фактический уровень значимости оказался недостаточно малым, поэтому можно сделать вывод о том, что нет статистически значимых отличий между распределениями баллов разных годов. ■

Задачи

Задача 12.1. Докажите состоятельность критерия Колмогорова-Смирнова для общей альтернативы $H_1 \colon F \neq G$.

Задача 12.2. Докажите формулу (17).

13 Множественная проверка гипотез

В большинстве ситуаций выше гипотезы проверялись по одной, хотя зачастую на практике встаёт необходимость проверять множество гипотез. Вот лишь некоторые примеры:

- Проверка гипотез об изменении метрик при A/B-тестировании, коих обычно несколько десятков или сотен;
- Применение к одной и той же выборке критериев согласия с целью нахождения семейства распределений, которому принадлежит истинное;
- Вспомогательные проверки, которые нужны для применения более мощных критериев при более ограничительных условиях (например, проверка нормальности, независимости, равенства дисперсий и т.д.);
- etc.

Каждый критерий имеет некоторый уровень значимости α, что означает вероятность $\leqslant \alpha$ ошибки I рода для одной конкретной гипотезы. Однако вероятность хотя бы одного отвержения верной гипотезы куда выше. Данный нюанс иллюстрирует следующий

Пример 13.1. В разделе 12.1 обсуждался способ проверки однородности двух независимых выборок при условии их нормальности с помощью применения двух критериев: сначала проверялось равенство дисперсий посредством F-критерия, и если оно не отвергается, то применяется t-тест для случая равенства дисперсий. Для начала реализуем F-критерий, который, к сожалению, не имеется в scipy:

```
def f_test(x, y):
    n, m = len(x), len(y)
    statistic = np.var(x, ddof=1) / np.var(y, ddof=1)
    pvalue = 1 - 2 * abs(sps.f(n - 1, m - 1).cdf(statistic) - 0.5)
    return (statistic, pvalue)
```

Каждый критерий будет проверятся на уровне значимости $\alpha = 0.05$. С помощью симуляции одинаково распределённых выборок оценим ошибку I рода, которая заключается в отвержении хотя одной из проверяемых гипотез.

```
n_iter = 100000
n, m = 30, 30
x = sps.norm.rvs(size=(n_iter, n))
y = sps.norm.rvs(size=(n_iter, m))
reject_cnt = 0
alpha = 0.05
for i in range(n_iter):
    pvalue1 = f_test(x[i], y[i])[1]
    pvalue2 = sps.ttest_ind(x[i], y[i])[1]
    reject_cnt += max((pvalue1 < alpha), (pvalue2 < alpha))
print(reject_cnt / n_iter)
```

```
0.10286
```

Ожидаемо, ошибка I рода составила примерно 0.1, что в два раза больше того, что мы хотели. ■

В случае проверки сотен или даже тысяч гипотез ситуация становится ещё более плачевной. Цель сей главы — ознакомиться со способами проверять множество гипотез так, чтобы при любом раскладе (то есть при любом наборе истинных и ложных гипотез) вероятность хотя бы одного отвержения верной гипотезы была не больше установленного числа α.

13.1 Контроль FWER и нисходящие процедуры

Для начала формально поставим задачу. Пусть имеется m выборок $\mathbf{X} = \{X_i^{(j)}\}$, где $1 \leqslant j \leqslant m$, $1 \leqslant i \leqslant n_j$, которые, вообще говоря, могут быть зависимыми и даже совпадать. На проверку поставлено m гипотез вида

$$H_j \colon \mathsf{P}_j \in \mathcal{P}_j \quad \text{versus} \quad H'_j \colon \mathsf{P}_j \notin \mathcal{P}_j,$$

среди которых выделим множество верных гипотез $M_0 = \{j \colon \mathsf{P}_j \in \mathcal{P}_j\}$, где $\mathcal{P}_j \subset \mathcal{P}$.

Нам необходимо построить m критериев $S_1, \ldots, S_m$, где, как обычно, $\mathbf{X}^{(j)} \in S_j$ равносильно отвержению гипотезы H_j. Введём следующие обозначения для количества гипотез для различных категорий.

	Верные	Ложные	Всего
Принятые	U	T	$m - R$
Отвергнутые	V	S	R
Всего	m_0	$m - m_0$	m

Определение. *Групповой вероятностью ошибки I рода* или FWER (от англ. family-wise error rate) называют величину

$$\mathrm{FWER} = \mathsf{P}(V > 0), \quad \mathsf{P} \in \mathcal{P}.$$

Контроль FWER на уровне α означает, что $\mathrm{FWER} \leqslant \alpha$ для любого $\mathsf{P} \in \mathcal{P}$, а стало быть, для любого набора верных гипотез M_0.

Пусть $\alpha_1, \ldots, \alpha_m$ — уровни значимости критериев $S_1, \ldots, S_m$ проверки гипотез $H_1, \ldots, H_m$ соответственно. Мы хотим их выбрать таким образом, чтобы $\mathrm{FWER} \leqslant \alpha$.

Метод Бонферрони заключается в топорном уменьшении всех уровней значимости в m раз: $\alpha_1 = \ldots = \alpha_m = \alpha/m$. Такое и вправду сработает:

$$\mathrm{FWER} = \mathsf{P}(V > 0) = \mathsf{P}\left(\bigcup_{j \in M_0} \{j\text{-ая гипотеза отвергается}\}\right) = \mathsf{P}\left(\bigcup_{j \in M_0} \{\mathbf{X}^{(j)} \in S_j\}\right) \leqslant$$
$$\leqslant \sum_{j \in M_0} \mathsf{P}\left(\mathbf{X}^{(j)} \in S_j\right) \leqslant m_0 \cdot \frac{\alpha}{m} \leqslant \alpha.$$

Замечание. Число m в знаменателе можно заменить на любую другую оценку сверху на число верных гипотез m_0, если таковая у нас имеется. Например, если мы применяем критерии для проверки несовместных гипотез (то есть не могут быть одновременно верными две и более гипотезы), то уровни значимости можно оставить прежними.

Данный способ хоть и прост, но чрезмерно наивен и не используется на практике в силу своей немощности: при большом m он очень сильно уменьшает пороги α_j, из-за чего отклонение от основной гипотезы должно быть чрезвычайно велико, чтобы быть обнаруженным, что бьёт по мощности.

Метод Шидака чуть более мощен, хотя он и требует независимости выборок $\mathbf{X}^{(j)}$. Он устанавливает $\alpha_1 = \ldots = \alpha_m = 1 - (1-\alpha)^{1/m}$. Как можно догадаться, он выражает FWER через вероятность не объединения, а пересечения:

$$\text{FWER} = 1 - \mathsf{P}(V=0) = 1 - \mathsf{P}\left(\bigcap_{j\in M_0} \{j\text{-ая гипотеза не отвергается}\}\right) =$$

$$= 1 - \mathsf{P}\left(\bigcap_{j\in M_0} \{\mathbf{X}^{(j)} \notin S_j\}\right) = 1 - \prod_{j\in M_0} \left(\mathbf{X}^{(j)} \notin S_j\right) \leqslant 1 - \left[(1-\alpha)^{1/m}\right]^m = \alpha.$$

Впрочем, при больших m он ведёт себя почти так же, как и печально известный метод Бонферрони, поэтому данный способ тоже не шибко полезен.

Другой подход заключается в последовательном рассмотрении p-value критериев $p_1, \ldots, p_m$ в порядке очерёдности:

$$p_{(1)} \leqslant \ldots \leqslant p_{(m)}.$$

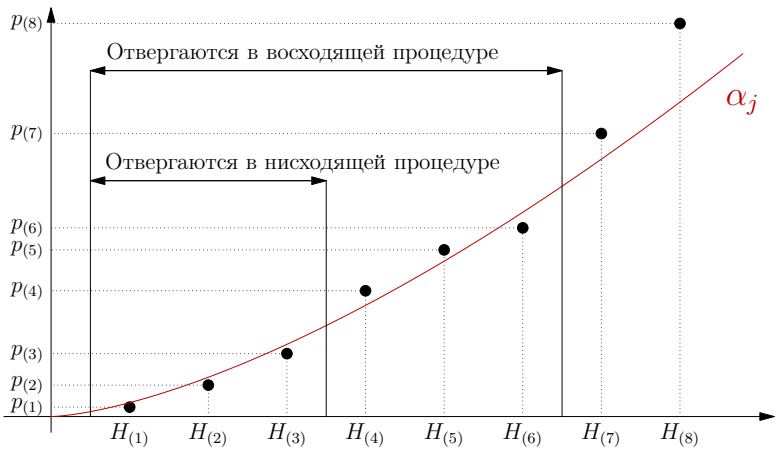
Для удобства обозначим соответствующие упорядоченным p-value гипотезы $H_{(1)}, \ldots, H_{(m)}$. Определим *нисходящую процедуру* для уровней значимости $\alpha_1, \ldots, \alpha_m$ следующим образом:

- Если $p_{(1)} > \alpha_1$, то принимаем все гипотезы $H_{(1)}, \ldots, H_{(m)}$ и останавливаемся, иначе отвергаем $H_{(1)}$ и продолжаем;
- Если $p_{(2)} > \alpha_2$, то принимаем все гипотезы $H_{(2)}, \ldots, H_{(m)}$ и останавливаемся, иначе отвергаем $H_{(2)}$ и продолжаем;
- И так далее.

Проще говоря, мы ищем минимальное такое j, при котором $p_{(j)} > \alpha_j$, и отвергаем все те гипотезы, номера которых в вариационном ряду меньше, чем j. Похоже, но наоборот, работает *восходящая процедура*:

- Если $p_{(m)} \leqslant \alpha_m$, то отвергаем все гипотезы $H_{(1)}, \ldots, H_{(m)}$ и останавливаемся, иначе принимаем $H_{(m)}$ и продолжаем;
- Если $p_{(m-1)} \leqslant \alpha_{m-1}$, то отвергаем все гипотезы $H_{(1)}, \ldots, H_{(m-1)}$ и останавливаемся, иначе принимаем $H_{(m-1)}$ и продолжаем;
- И так далее

Здесь уже мы ищем максимальное такое j, при котором $p_{(j)} \leqslant \alpha_j$, и отвергаем все те гипотезы, номера которых в вариационном ряду не больше j. Принцип работы процедур можно изобразить на рисунке:



Как можно видеть, нисходящая процедура более консервативна и начинает отклонять гипотезы с самых бесперспективных, по одной. Восходящая же не боится отвергнуть сразу охапку гипотез. О примерах восходящих процедур и их

роли в множественной проверке гипотез мы поговорим в следующем разделе, а пока остановимся на нисходящих.

Метод Холма подразумевает использование нисходящей процедуры при

$$\alpha_j = \frac{\alpha}{m - j + 1},$$

то есть $\alpha_1 = \alpha/m, \alpha_2 = \alpha/(m-1), \ldots, \alpha_m = \alpha$. Данный метод мощнее, чем метод Бонферрони, так как уровни значимости больше, чем просто α/m, что позволяет лучше отвергать гипотезы. Более того, при неимении дополнительной информации о наших выборках (например, их независимости) данный метод не улучшаем в плане мощности. Но что же насчёт вероятности ошибки I рода?

Теорема 13.1.

При заданных α_j метод Холма контролирует FWER на уровне значимости α.

Доказательство. Пусть во время процедуры какие-то верные гипотезы пришлось отвергнуть. Рассмотрим среди таковых гипотезу с наименьшим номером j в вариационном ряду, то есть все предыдущие $j-1$ гипотезы были ложными. Всего ложных гипотез $m - m_0$ штук, поэтому $j - 1 \leqslant m - m_0$, что даёт оценку $m_0 \leqslant m - j + 1$. Это значит, что уровень значимости, на котором проверялась j-ая гипотеза, был равен $\alpha/(m-j+1) \leqslant \alpha/m_0$. С учётом того, что её отвергли, p-value для её проверки был $\leqslant \alpha/m_0$. Подобное умозаключение позволяет свести доказательство к оцениванию вероятности объединения по всем верным гипотезам:

$$\text{FWER} = \mathsf{P}(V > 0) \leqslant \mathsf{P}\left(\bigcup_{l \in M_0} \{p_l \leqslant \alpha/m_0\}\right) \leqslant \sum_{l \in M_0} \mathsf{P}(p_l \leqslant \alpha/m_0) \leqslant m_0 \cdot \frac{\alpha}{m_0} = \alpha.$$

□

Если же известна информация о независимости выборок $\mathbf{X}^{(j)}$, $1 \leqslant j \leqslant m$, то процедуру можно немного усилить в плане мощности, получив **метод Шидака-Холма**. Как можно догадаться из названия, он является комбинацией методов Холма и Шидака: для него уровень значимости определяется как

$$\alpha_j = 1 - (1 - \alpha)^{\frac{1}{m-j+1}}.$$

Аналогично методу Холма доказывается, что нисходящая процедура с данными уровнями значимости будет контролировать FWER на заданном уровне. Опять же, при больших m отличий от обычного метода Холма мало, однако интерес представляет то, что данный метод наиболее мощный для независимых выборок.

13.2 Контроль FDR и восходящие процедуры

Не всегда необходимо так строго контролировать ошибку I рода, как это делают процедуры выше. В них мы боимся сделать хотя бы одно неверное отвержение, что сильно бьёт по способности процедур уметь отвергать гипотезы. Иногда именно такое умение и нужно, например, для множественной проверки важности признаков в линейной регрессии из примера 14.5, когда страх отвергнуть гипотезу о бесполезности признака может стоить нам выкидыванием действительно нужных. Компромисса можно достичь, контролируя следующую, более слабую величину, нежели FWER:

Определение. *Ожидаемой долей ложных отклонений гипотез* или FDR (от англ. false discovery rate) называют величину

$$\mathrm{FDR} = \mathsf{E}\frac{V}{\max\{R, 1\}}.$$

Контроль FDR на уровне α означает, что $\mathrm{FDR} \leqslant \alpha$ для любого $\mathsf{P} \in \mathcal{P}$.

Максимум в знаменателе взят для того, чтобы не делить на нуль в случае принятия всех гипотез. Данная характеристика действительно является более слабой по сравнению с FWER, что показывает следующее

Утверждение 13.1. *$FDR \leqslant FWER$.*

Доказательство.

$$\mathrm{FDR} = \mathsf{E}\frac{V}{\max\{R, 1\}} = \mathsf{E}\left[\frac{V}{\max\{R, 1\}} \cdot I(V > 0)\right] \leqslant \mathsf{E}I(V > 0) = \mathsf{P}(V > 0) = \mathrm{FWER}.$$

□

Таким образом, контролирование FWER на заданном уровне значимости даёт нам контроль и FDR, что, вообще говоря, может не выполняться в обратную сторону, поэтому процедуры, контролирующие FDR, могут совершать больше ошибок I рода.

Выше нам встречалась так называемая восходящая процедура, которая способна легко отвергать гипотезы. Рассмотрим некоторые её примеры.

Метод Бенджамини-Иекутиели применим в общем случае, когда о зависимостях между выборками неизвестно (в таком случае метод не улучшаем). Он основан на восходящей процедуре с

$$\alpha_j = \frac{\alpha i}{m} \cdot \left(\sum_{l=1}^{m} \frac{1}{l}\right)^{-1}.$$

Для независимых выборок можно применим более мощный **метод Бенджамини-Хохберга**, для которого уже

$$\alpha_j = \frac{\alpha i}{m}.$$

Можно показать, что выполняется следующее утверждение:

Теорема 13.2.

Два приведённых метода контролируют FDR на уровне α.

Доказательство этого факта можно найти в [1].

Пример 13.2. Смоделируем набор из $m = 1000$ пар выборок $(\mathbf{X}^{(j)}, \mathbf{Y}^{(j)})$, которые независимы в совокупности, среди которых первые $m_0 = 600$ будут выбраны из одного распределения $\mathcal{N}(0, 1)$, а остальные $m - m_0$ пар будут содержать неоднородные выборки: $\mathbf{X}^{(j)}$ будет также взято из $\mathcal{N}(0, 1)$, а вот $\mathbf{Y}^{(j)}$ уже из $\mathcal{N}(1, 1)$. Мы хотим проверить m гипотез о однородности пар выборок:

```
m, m_0 = 1000, 600
n = 30
x = sps.norm.rvs(size=(m, n))
```

```
y = np.concatenate([
    sps.norm.rvs(size=(m_0, n)),
    sps.norm(loc=1).rvs(size=(m - m_0, n))
])
is_true = np.array([True] * m_0 + [False] * (m - m_0))
pvalues = sps.ttest_ind(x, y, axis=1).pvalue
```

Для удобства напишем функцию, которая будет применять к p-value выбранный метод и выводить статистику по принятым или отвергнутым гипотезам:

```
def multiple_ttest_results(pvalues, method, alpha=0.05):
    is_rejected = method(pvalues, alpha)
    U = (~is_rejected & is_true).sum()
    T = (~is_rejected & ~is_true).sum()
    V = (is_rejected & is_true).sum()
    S = (is_rejected & ~is_true).sum()
    f =\
"""          True False
Accepted {0:>4} {1:>5}
Rejected {2:>4} {3:>5}"""
    print(f.format(U, T, V, S))
```

Сначала посмотрим, какие результаты получаются в случае прямолинейной проверки:

```
def straightforward(pvalues, alpha=0.05):
    return pvalues < alpha

multiple_ttest_results(pvalues, straightforward)
```

```
        True False
Accepted 571    18
Rejected  29   382
```

Ошибок I рода достаточно много, примерно одна двадцатая от всех истинных гипотез. Попробуем исправить ситуацию с помощью поправки Шидака, как-никак выборки у нас независимы:

```
def sidak(pvalues, alpha=0.05):
    m = len(pvalues)
    return pvalues < 1 - (1 - alpha) ** (1 / m)

multiple_ttest_results(pvalues, sidak)
```

```
        True False
Accepted 600   271
Rejected   0   129
```

Число ошибок I рода уменьшилось аж до нуля, так как вероятность увидеть здесь ненулевое значение довольно мала, не говоря уже о прежних конских цифрах. Однако мощность явно просела: мы стали принимать больше половины ложных гипотез, что никуда не годится. Можно улучшить ситуацию с помощью метода Шидака-Холма:

```
def sidak_holm(pvalues, alpha=0.05):
    m = len(pvalues)
    threshold = m
    sorted_pvalues = np.sort(pvalues)
    for i in range(m):
        if sorted_pvalues[i] > 1 - (1 - alpha) ** (1 / (m - i)):
            threshold = i
            break
    return np.argsort(np.argsort(pvalues)) < threshold

multiple_ttest_results(pvalues, sidak_holm)
```

```
        True False
Accepted 600   262
Rejected   0   138
```

Стало не особо лучше, зато всё ещё успешно контролируется FWER. Наконец испробуем метод Бенджамини-Хохберга:

```
def benjamini_hochberg(pvalues, alpha=0.05):
    m = len(pvalues)
    threshold = m
    sorted_pvalues = np.sort(pvalues)
    for i in range(m - 1, -1, -1):
        if sorted_pvalues[i] <= alpha * (i + 1) / m:
            threshold = i
            break
    return np.argsort(np.argsort(pvalues)) <= threshold

multiple_ttest_results(pvalues, benjamini_hochberg)
```

```
        True False
Accepted 590    36
Rejected  10   364
```

Теперь допускаются ошибки I рода (хоть и не в таких количествах, как в топорном способе), но мощность куда лучше, чем в предыдущих двух случаев. ■

Часть III
Прочие модели и методы в статистике

14 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 признаков 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,

$$\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{18}$$

где $Z = (Z_{ij})$ — матрица «объект-признак», $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)^T$ — столбец из неизвестных параметров. Логично допустить, что выполнены следующие ограничения на случайный вектор $\boldsymbol{\varepsilon}$:

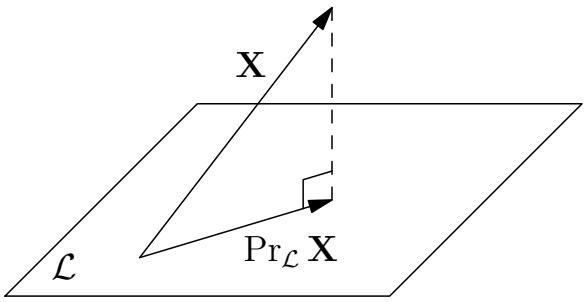
L1 Для всех i выполнено $\mathsf{E}\varepsilon_i = 0$ (в среднем ошибки нет);

L2 Дисперсия ε_i одинакова и равна неизвестному параметру σ^2, причём ε_i попарно нескоррелированы, то есть $\mathsf{D}\boldsymbol{\varepsilon} = \sigma^2 E_n$ (наблюдения друг на друга не влияют).

Итак, наша задача — по вектору наблюдений $\mathbf{X}$ оценить вектор $\boldsymbol{\theta}$ (чтобы иметь возможность находить целевую величину по другим признакам) и дисперсию σ^2 (чтобы оценить нашу уверенность в оценке и уметь строить доверительные интервалы для неё).

Для дальнейшего удобства также будет важно сделать допущение

L3 Столбцы $\mathbf{z}_1, \dots, \mathbf{z}_k$ матрицы Z *линейно независимы.*



Это позволяет интерпретировать задачу с позиций линейной алгебры: истинный вектор $l = Z\boldsymbol{\theta}$ 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 $\mathbf{X}$ может в общем случае и не лежать в $\mathcal{L}$ (см. рис.). Отсюда логично в качестве «приближения» вектора $\mathbf{X}$ выбрать его проекцию $\Pr_{\mathcal{L}} \mathbf{X}$ на это подпространство, так как она доставляет минимум расстояния между $\mathbf{X}$ и векторами из $\mathcal{L}$.

Осталось лишь найти оценку вектору параметров $\widehat{\boldsymbol{\theta}}$, отвечающую проекции $\Pr_{\mathcal{L}} \mathbf{X} = Z\widehat{\boldsymbol{\theta}}$. Так как $\Pr_{\mathcal{L}} \mathbf{X}$ – ортогональная проекция, то вектор $\boldsymbol{\delta} = \mathbf{X} - \Pr_{\mathcal{L}} \mathbf{X}$ лежит в $\mathcal{L}^{\perp}$, а значит,

он ортогонален любому вектору из $\mathcal{L}$, в частности, векторам $\mathbf{z}_1, \ldots, \mathbf{z}_k$. Следовательно, вектор $Z^T\boldsymbol{\delta}$, состоящий из скалярных произведений $\boldsymbol{\delta}$ с $\mathbf{z}_j$, – нулевой, то есть

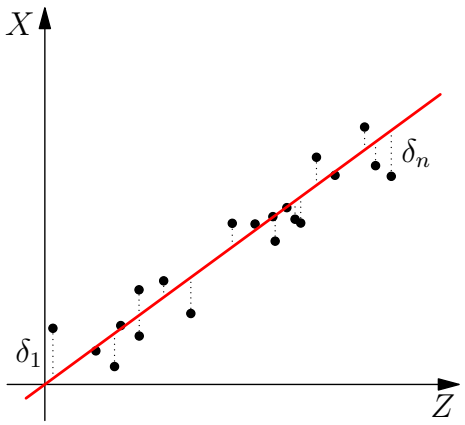
$$Z^T(\mathbf{X} - \Pr_{\mathcal{L}}\mathbf{X}) = 0 \Longrightarrow Z^T\mathbf{X} = (Z^TZ)\widehat{\boldsymbol{\theta}}.$$

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

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

В частности, оператор проектирования на $\mathcal{L}$ соответствует матрице $A = Z(Z^TZ)^{-1}Z^T$.

Определение. Полученная оценка называется *оценкой по методу наименьших квадратов.*



$$\sum_{i=1}^{n}\delta_i^2 \longrightarrow \min$$

Такое название метод получил благодаря иному способу получения сей оценки: не через линейную алгебру, а посредством анализа. Как было сказано выше, проекция доставляет минимум расстояния от вектора до пространства. Значит, при искомой оценке достигается экстремум суммы квадратов координат разности $F(\boldsymbol{\theta}) = \|Z\boldsymbol{\theta} - \mathbf{X}\|^2$. Точку экстремума же можно найти обычным дифференцированием:

$$d_{\boldsymbol{\theta}}F = d_{\boldsymbol{\theta}}\left[(Z\boldsymbol{\theta} - \mathbf{X})^T(Z\boldsymbol{\theta} - \mathbf{X})\right] = (Z\boldsymbol{\theta} - \mathbf{X})^TZd\boldsymbol{\theta} + d\boldsymbol{\theta}^TZ^T(Z\boldsymbol{\theta} - \mathbf{X}) =$$
$$\langle 2Z^T(Z\boldsymbol{\theta} - \mathbf{X}), d\boldsymbol{\theta}\rangle = 0 \Longrightarrow Z^TZ\widehat{\boldsymbol{\theta}} = Z^T\mathbf{X}, \quad \widehat{\boldsymbol{\theta}} = (Z^TZ)^{-1}Z^T\mathbf{X}.$$

14.1 Свойства МНК-оценки

Сразу выделим полезные свойства полученной оценки.

Утверждение 14.1. *Оценка по методу наименьших квадратов имеет матожидание, равное $\boldsymbol{\theta}$, и её ковариационная матрица равна $\sigma^2(Z^TZ)^{-1}$.*

Доказательство. По линейности матожидания:

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

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

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

Теперь мы можем получить требуемое:

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

□

Таким образом, полученная оценка $\widehat{\boldsymbol{\theta}}$ является несмещённой оценкой вектора $\boldsymbol{\theta}$. При выполнении допущений выше можно сказать, что такая оценка будет наилучшей в некотором хорошем классе оценок.

Теорема 14.1 (Гаусс, Марков).

В линейной регрессионной модели при выполнении условий **L1-L3** оценка $\widehat{\boldsymbol{\theta}}$ является эффективной в классе несмещённых линейных оценок $\boldsymbol{\theta}$.

Доказательство. Пусть $\boldsymbol{\theta}^*(\mathbf{X}) = C\mathbf{X}$ — некоторая несмещённая линейная оценка $\boldsymbol{\theta}$, то есть $C \in \mathbb{R}^{k\times n}$, и

$$\forall \boldsymbol{\theta} \in \mathbb{R}^k \colon \boldsymbol{\theta} = \mathsf{E}_{\boldsymbol{\theta},\sigma^2}\boldsymbol{\theta}^* = \mathsf{E}_{\boldsymbol{\theta},\sigma^2}(C\mathbf{X}) = CZ\boldsymbol{\theta} \Longrightarrow CZ = E_k.$$

Найдём ковариационную матрицу сей оценки:

$$\mathsf{D}_{\boldsymbol{\theta},\sigma^2}\boldsymbol{\theta}^* = \mathsf{D}_{\boldsymbol{\theta},\sigma^2}(C\mathbf{X}) = C \cdot (\mathsf{D}_{\boldsymbol{\theta},\sigma^2}\boldsymbol{\varepsilon}) \cdot C^T = \sigma^2 CC^T.$$

Таким образом, по определению эффективной многомерной оценки нам надо показать, что матрица $CC^T - (Z^TZ)^{-1}$ положительно полуопределена. Для этого домножим $(Z^TZ)^{-1}$ слева и справа на единичную матрицу $E_k = CZ$ и вынесем общие множители за скобку:

$$CC^T - (Z^TZ)^{-1} = CC^T - CZ(Z^TZ)^{-1}(CZ)^T = C\left(E_n - Z(Z^TZ)^{-1}Z^T\right)C^T.$$

Вспоминаем, что $Z(Z^TZ)^{-1}Z^T$ есть оператор проектирования на $\mathcal{L}$, поэтому $B = E_n - Z(Z^TZ)^{-1}Z^T$ соответствует проектору на $\mathcal{L}^\perp$, и её собственные значения равны 0 и 1. Следовательно, B есть матрица некоторой положительно полуопределённой квадратичной формы. Следовательно, матрица выше положительна полуопределена. □

Что же насчёт σ^2? Так как она характеризует меру «разброса» вокруг пространства $\mathcal{L}$, то её можно приблизить квадратом длины расстояния до $\mathcal{L}$, а именно $RSS = \left\|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\right\|^2$. Положим

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

Константа около квадрата нормы выбрана так, чтобы выполнялось

Утверждение 14.2. *$\widehat{\sigma^2}$ является несмещённой оценкой параметра σ^2.*

Доказательство. Так как $Z\widehat{\boldsymbol{\theta}} = \Pr_{\mathcal{L}}\mathbf{X}$, то $\mathbf{X} - Z\widehat{\boldsymbol{\theta}} = \Pr_{\mathcal{L}^\perp}\mathbf{X} = B\mathbf{X}$, где матрица ортогонального проектирования на $\mathcal{L}^\perp$ может быть представлена как

$$B = E_n - A = E_n - Z(Z^TZ)^{-1}Z^T,$$

и поэтому

$$\mathsf{E}_{\boldsymbol{\theta},\sigma^2}\left\|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\right\|^2 = \mathsf{E}_{\boldsymbol{\theta},\sigma^2}\|B\mathbf{X}\|^2.$$

Заметим, что $Z\boldsymbol{\theta} \in \mathcal{L}$ и $\mathbf{X} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, поэтому $\Pr_{\mathcal{L}^\perp}\mathbf{X} = \Pr_{\mathcal{L}^\perp}\boldsymbol{\varepsilon}$, и

$$\mathsf{E}_{\boldsymbol{\theta},\sigma^2}\|B\mathbf{X}\|^2 = \mathsf{E}_{\boldsymbol{\theta},\sigma^2}\|B\boldsymbol{\varepsilon}\|^2 = \mathsf{E}_{\boldsymbol{\theta},\sigma^2}\boldsymbol{\varepsilon}^TB^TB\boldsymbol{\varepsilon} \circledcirc$$

Провернём классический трюк: представим число под знаком матожидания как след одноэлементной матрицы. Это позволит нам воспользоваться свойством следа матрица о циклической перестановке:

$$\circledcirc \operatorname{tr}\mathsf{E}_{\boldsymbol{\theta},\sigma^2}\boldsymbol{\varepsilon}^TB^TB\boldsymbol{\varepsilon} = \operatorname{tr}\mathsf{E}_{\boldsymbol{\theta},\sigma^2}B^TB\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \circledcirc$$

Матрица $B = E_n - Z(Z^TZ)^{-1}Z^T$ симметрична и отвечает проектору, поэтому $B^TB = B^2 = B$, значит, по линейности матожидания

$$\textcircled{=}\ \operatorname{tr} B \cdot \mathsf{E}_{\boldsymbol{\theta},\sigma^2}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T = \operatorname{tr} B \cdot \mathsf{D}_{\boldsymbol{\theta},\sigma^2}\boldsymbol{\varepsilon} = \operatorname{tr} B \cdot \sigma^2 E_n = \sigma^2 \operatorname{tr} B = \sigma^2 \operatorname{tr}(E_n - Z(Z^TZ)^{-1}Z^T) =$$
$$= \sigma^2(n - \operatorname{tr}\left[(Z^TZ)^{-1}Z^TZ\right]) = \sigma^2(n - \operatorname{tr} E_k) = \sigma^2(n-k).$$

Таким образом, $\widehat{\sigma^2} = RSS/(n-k)$ является несмещённой оценкой σ^2, что и требовалось. □

Пример 14.1. Пусть 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. 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

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

Теперь у нас есть всё, чтобы посчитать оценку:

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

$$\widehat{\boldsymbol{\theta}} = (Z^TZ)^{-1}Z^T\mathbf{X} = \begin{pmatrix} 5/6 & -1/6 \\ -1/6 & 5/6 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \end{pmatrix} \begin{pmatrix} X_a \\ X_b \\ X_{ab}/2 \end{pmatrix} =$$

$$= \begin{pmatrix} 5/6 & -1/6 & 1/3 \\ -1/6 & 5/6 & 1/3 \end{pmatrix} \begin{pmatrix} X_a \\ X_b \\ X_{ab}/2 \end{pmatrix} \Longrightarrow$$

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

■

14.2 Взвешенный МНК

Подход из примера выше ещё называют *взвешенным МНК*: если ε_i имеют разную дисперсию, а точнее ковариационная матрица $\boldsymbol{\varepsilon}$ равна $\sigma^2 D$, где D — диагональна, то можно домножить равенство (18) на $D^{-1/2}$, что сведёт нашу задачу к обычной линейной регрессии. Обобщим эту идею.

Рассмотрим более общую модель, где шум может не только иметь разный разброс, но и быть зависимым между собой. Для этого заменим условие **L2** на

L2' Ковариационная матрица вектора $\boldsymbol{\varepsilon}$ равна $\mathsf{D}\boldsymbol{\varepsilon} = \sigma^2 V$, где V известна и положительна определена, а σ^2 — неизвестный параметр.

Так как V — п.о., то существует п.о. матрица $V^{1/2}$ такая, что $V^{1/2} \cdot V^{1/2} = V$. Теперь достаточно домножить уравнение (18) на $V^{-1/2}$, получив классическую модель линейной регрессии

$$\widetilde{\mathbf{X}} = \widetilde{Z}\boldsymbol{\theta} + \widetilde{\boldsymbol{\varepsilon}},$$

где $\widetilde{\mathbf{X}} = V^{-1/2}\mathbf{X}$, $\widetilde{Z} = V^{-1/2}Z$ и $\widetilde{\boldsymbol{\varepsilon}} = V^{-1/2}\boldsymbol{\varepsilon}$. Случайная составляющая $\widetilde{\boldsymbol{\varepsilon}}$ всё ещё имеет нулевое матожидание и матрицу ковариаций

$$\mathsf{D}\widetilde{\boldsymbol{\varepsilon}} = \mathsf{D}\left[V^{-1/2}\boldsymbol{\varepsilon}\right] = V^{-1/2} \cdot \mathsf{D}\boldsymbol{\varepsilon} \cdot V^{-1/2} = V^{-1/2} \cdot \sigma^2 V \cdot V^{-1/2} = \sigma^2 E_n.$$

Вкупе с тем фактом, что при домножении на невырожденную матрицу ранг Z не поменялся, в новой модели выполняются условия **L1**-**L3**. Подставляя матрицы $\widetilde{\mathbf{X}}$ и $\widetilde{Z}$ в определении оценки по МНК, получаем оценку

$$\widehat{\boldsymbol{\theta}} = \left(\widetilde{Z}^T\widetilde{Z}\right)^{-1}\widetilde{Z}^T\widetilde{\mathbf{X}} = \left(Z^TV^{-1}Z\right)^{-1}Z^TV^{-1}\mathbf{X}. \tag{19}$$

Пример 14.2 (*критерий Шапиро-Уилка revisited*)**.** Напомним, что перед нами стояла задача проверки гипотезы о принадлежности семейству распределений с параметрами сдвига и масштаба, которая обсуждалась в разделе 10.1. Здесь мы поймём, как можно получить формулу (9), подогнав регрессию под похожий на QQ-plot график.

Рассмотрим вариационный ряд $\mathbf{X} = (X_{(1)}, \ldots, X_{(n)})$, построенный по выборке из распределения $F_0\left(\frac{x-\mu}{\sigma}\right)$. Нормируем выборку: $Y_{(i)} = (X_{(i)} - \mu)/\sigma$, или, что эквивалентно, $\mathbf{X} = \mu\mathbf{1} + \sigma\mathbf{Y}$, где $\mathbf{1} = (1, \ldots, 1)$ — вектор из n единиц. Распределение $\mathbf{Y} = (Y_{(1)}, \ldots, Y_{(n)})$ не зависит от параметров и однозначно определяется распределением F_0, обозначим её вектор средних и матрицу ковариаций за $\mathbf{m}$ и V соответственно. Вынося из $\mathbf{Y}$ вектор средних, получим следующее равенство:

$$\mathbf{X} = \mu\mathbf{1} + \sigma\mathbf{m} + \sigma\boldsymbol{\varepsilon},$$

где $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{m}$, и как несложно понять, $\mathsf{E}\boldsymbol{\varepsilon} = \mathbf{0}$, $\mathsf{D}\boldsymbol{\varepsilon} = V$. Таким образом, получена обобщённая модель линейной регрессии, которая отчасти похожа на QQ-plot: в ней по матожиданию i-ой порядковой статистики выборки из F_0 нужно с точностью до сдвига предсказать значение i-ой по порядку величины.

В терминах написанного выше ясно, что матрица признаков Z состоит из столбцов $\mathbf{1}$ и $\mathbf{m}$. С этим знанием посчитаем наилучшую оценку μ и σ:

$$Z^TV^{-1}Z = \begin{pmatrix} \mathbf{1}^TV^{-1}\mathbf{1} & \mathbf{1}^TV^{-1}\mathbf{m} \\ \mathbf{m}^TV^{-1}\mathbf{1} & \mathbf{m}^TV^{-1}\mathbf{m} \end{pmatrix},$$

$$\left(Z^TV^{-1}Z\right)^{-1} = \frac{1}{\mathbf{1}^TV^{-1}\mathbf{1}\mathbf{m}^TV^{-1}\mathbf{m} - (\mathbf{1}^TV^{-1}\mathbf{m})^2}\begin{pmatrix} \mathbf{m}^TV^{-1}\mathbf{m} & -\mathbf{m}^TV^{-1}\mathbf{1} \\ -\mathbf{1}^TV^{-1}\mathbf{m} & \mathbf{1}^TV^{-1}\mathbf{1} \end{pmatrix},$$

$$\widehat{\mu} = \frac{\mathbf{m}^TV^{-1}(\mathbf{m}\mathbf{1}^T - \mathbf{1}\mathbf{m}^T)V^{-1}\mathbf{X}}{\mathbf{1}^TV^{-1}\mathbf{1}\mathbf{m}^TV^{-1}\mathbf{m} - (\mathbf{1}^TV^{-1}\mathbf{m})^2}, \quad \widehat{\sigma} = \frac{\mathbf{1}^TV^{-1}(\mathbf{1}\mathbf{m}^T - \mathbf{m}\mathbf{1}^T)V^{-1}\mathbf{X}}{\mathbf{1}^TV^{-1}\mathbf{1}\mathbf{m}^TV^{-1}\mathbf{m} - (\mathbf{1}^TV^{-1}\mathbf{m})^2}$$

Оценка среднего может оказаться полезной, зачастую она обладает хорошими свойствами или устойчива к выбросам, так как выражается через порядковые статистики. Однако нас интересует оценка σ: этот параметр встречается в том числе и перед $\boldsymbol{\varepsilon}$, то есть он ещё отвечает за меру разброса около прямой, поэтому, причесав его оценку, можно получить критерий «хорошести» приближения.

Для дальнейшего удобства сделаем допущение, что распределение F_0 симметрично. Из него следует, что число $\mathbf{1}^TV^{-1}\mathbf{m}$ (а значит и $\mathbf{m}^TV^{-1}\mathbf{1}$) равно нулю, что значительно упростит формулы выше:

$$\widehat{\mu} = \frac{\mathbf{1}^TV^{-1}\mathbf{X}}{\mathbf{1}^TV^{-1}\mathbf{1}}, \quad \widehat{\sigma} = \frac{\mathbf{m}^TV^{-1}\mathbf{X}}{\mathbf{m}^TV^{-1}\mathbf{m}}.$$

Обычно берут не саму оценку $\widehat{\sigma}$, а её нормированную версию, то есть такую линейную комби-

нацию компонент $\mathbf{X}$, у которой норма вектора коэффициентов равна единице

$$b = \mathbf{a}^T\mathbf{X}, \quad \text{где} \quad \mathbf{a} = \frac{\mathbf{m}^T V^{-1}}{(\mathbf{m}^T V^{-2}\mathbf{m})^{1/2}}$$

Заметим, что распределение $\widehat{\sigma}$ (а значит и b) не зависит от параметра сдвига $\mathbf{X}$, так как к наблюдению можно добавить число c и в силу симметричности получить следующее:

$$\widehat{\sigma} = \frac{\overbrace{c\mathbf{m}^T V^{-1}\mathbf{1}}^{=0} + \mathbf{m}^T V^{-1}(\mathbf{X} - c\mathbf{1})}{\mathbf{m}^T V^{-1}\mathbf{m}} = \frac{\mathbf{m}^T V^{-1}(\mathbf{X} - c\mathbf{1})}{\mathbf{m}^T V^{-1}\mathbf{m}}.$$

Чтобы избавиться от зависимости от параметра масштаба, возведём оценку b в квадрат и поделим её на выборочную дисперсию с множителем n, получив инвариантную относительно сдвига и сжатия *статистику критерия Шапиро-Уилка*:

$$W = \frac{b^2}{ns^2(\mathbf{X})} = \frac{\left(\sum a_i X_{(i)}\right)^2}{\sum \left(X_i - \overline{\mathbf{X}}\right)^2}.$$

■

14.3 Гауссовская линейная модель

Всякие физики да химики из своих внутренних побуждений часто полагают ошибки нормальными, поэтому в дальнейшем под $\boldsymbol{\varepsilon}$ 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 (бремя доказательства которого ложится на читателя, см. задачу 14.3):

Теорема 14.2.

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

Знание о полноте и достаточности статистики, как мы знаем, нередко помогает при оценивании: если получится несмещённо оценить неизвестный параметр с помощью полной достаточной статистики, то по теореме Лемана-Шеффе такая оценка будет оптимальной.

Пример 14.3. 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. Будем считать, что наша модель гауссовская, то есть все ошибки измерения распределены нормально и имеют дисперсию σ^2. 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

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

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:

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

$$= \begin{pmatrix} 1/n_1 & 0 & 0 \\ 0 & 1/n_2 & 0 \\ 0 & 0 & 1/n_3 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix} \mathbf{X} = \begin{pmatrix} \frac{1}{n_1} \sum X_i^\alpha \\ \frac{1}{n_2} \sum X_i^\beta \\ \frac{1}{n_3} \sum X_i^\gamma \end{pmatrix}$$

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.

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

■

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

Теорема 14.3 (об ортогональном разложении).

Пусть $\mathbf{X} \sim \mathcal{N}(\mathbf{a}, \sigma^2 E_n)$ — гауссовский вектор, а $L_1 \oplus \ldots \oplus L_r$ — ортогональное разложение $\mathbb{R}^n$. Тогда $\Pr_{L_1} \mathbf{X}, \ldots, \Pr_{L_r} \mathbf{X}$ независимы и нормально распределены, $\mathsf{E} \Pr_{L_i} \mathbf{X} = \Pr_{L_i} \mathbf{a}$ и для всех $i \in \{1, \ldots, r\}$

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

Доказательство. Равенство $\mathsf{E} \Pr_{L_i} \mathbf{X} = \Pr_{L_i} \mathbf{a}$ очевидно в силу линейности матожидания, а проекция есть линейное преобразование. Посему для простоты перейдём к рассмотрению вектора $\mathbf{Y} = (\mathbf{X} - \mathbf{a})/\sigma \sim \mathcal{N}(\mathbf{0}, E_n)$: теперь достаточно показать, что проекции $\mathbf{Y}$ независимы и $\| \Pr_{L_i} \mathbf{Y} \|^2 \sim \chi^2_{\dim L_i}$.

Пусть $\mathfrak{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)$ — о/н базис, согласованный с разложением $L_1 \oplus \ldots \oplus L_r$, то есть $\mathfrak{E} = (\mathfrak{E}_1, \ldots, \mathfrak{E}_r)$, где $\mathfrak{E}_i$ — базис в L_i, а S — матрица перехода от этого базиса к стандартному. Так как матрица перехода от одного о/н базиса к другому — ортогональна, то вектор $S\mathbf{Y}$ будет иметь распределение $\mathcal{N}(\mathbf{0}, S \cdot E_n \cdot S^T) = \mathcal{N}(\mathbf{0}, E_n)$, поэтому координаты $(S\mathbf{Y})_j = \mathbf{e}_j^T \mathbf{Y}$ в новом базисе независимы в совокупности и имеют стандартное нормальное

распределение.

Но заметим, что в новом базисе проекция на какое-либо L_i есть просто «откидывание» всех базисных векторов, кроме $\mathfrak{E}_i$:

$$\Pr_{L_i} \mathbf{Y} = \sum_{\mathbf{e}\in\mathfrak{E}_i} (\mathbf{e}^T Y)\cdot \mathbf{e}$$

Таким образом, проекции независимы как функции от попарно не пересекающихся наборов независимых в совокупности координат вектора $S\mathbf{Y}$. Более того, квадрат длины проекции как сумма квадратов координат с распределением $\mathcal{N}(0,1)$ имеет распределение $\chi^2_{\dim L_i}$, что и требовалось. □

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

$$\frac{1}{\sigma^2}\|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\|^2 \sim \chi^2_{\dim \mathcal{L}^\perp} = \chi^2_{n-k}. \tag{20}$$

Что же касается $\widehat{\boldsymbol{\theta}}$, то она, как линейное преобразование гауссовского вектора $\mathbf{X}$, имеет распределение $\mathcal{N}(\boldsymbol{\theta}, \sigma^2(Z^TZ)^{-1})$ (параметры мы нашли ранее в утверждении 14.1). Следовательно, $\widehat{\theta}_i \sim \mathcal{N}(\theta_i, \sigma^2\left[(Z^TZ)^{-1}\right]_{ii})$, или, что эквивалентно,

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

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 (20). 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 $Z\widehat{\boldsymbol{\theta}}$ и $\mathbf{X} - Z\widehat{\boldsymbol{\theta}}$ будут независимыми, а значит,

$$\widehat{\boldsymbol{\theta}} = (Z^TZ)^{-1}Z^T\cdot Z\widehat{\boldsymbol{\theta}} \quad \text{и} \quad \widehat{\sigma^2} = \|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\|^2/(n-k)$$

также независимы как функции от независимых случайных векторов. Поэтому распределение

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

не зависит от неизвестных параметров. В разделе 6.2 мы уже знакомились с распределением этой величины — это распределение Стьюдента. Таким образом,

$$\frac{\widehat{\theta}_i - \theta_i}{\sqrt{\widehat{\sigma^2}\left[(Z^TZ)^{-1}\right]_{ii}}} \sim T_{n-k}. \tag{21}$$

Полученные распределения приведённых статистик позволяют искать доверительные интервалы для θ_i и σ^2. Из соотношения (20) имеем доверительный интервал для σ^2:

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

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

где $\chi^2_{n-k,p}$ – p-квантиль распределения χ^2_{n-k}. В то же время из (21) имеем следующие ДИ для θ_i:

$$\mathsf{P}_{\boldsymbol{\theta},\sigma^2}\left(\widehat{\theta}_i - T_{n-k,(1+\gamma)/2}\sqrt{\widehat{\sigma^2}\left[(Z^TZ)^{-1}\right]_{ii}} < \theta_i < \widehat{\theta}_i + T_{n-k,(1+\gamma)/2}\sqrt{\widehat{\sigma^2}\left[(Z^TZ)^{-1}\right]_{ii}}\right) = \gamma,$$

где $T_{n-k,p}$ – p-квантиль распределения T_{n-k}. Здесь мы воспользовались симметричностью распределения Стьюдента, благодаря которой выполнено $T_{n-k,(1+\gamma)/2} = -T_{n-k,(1-\gamma)/2}$.

Впрочем, из сказанного выше легко понять, как построить доверительный интервал для произвольной линейной комбинации неизвестных параметров, что особенно полезно в контексте примера 14.3. Пусть необходимо оценить функцию $\tau(\boldsymbol{\theta}) = \mathbf{c}^T\boldsymbol{\theta}$, где $\mathbf{c} \in \mathbb{R}^k$ — вектор из коэффициентов, с которыми мы берём те или иные компоненты вектора $\boldsymbol{\theta}$. По линейности матожидания несмещённой оценкой $\tau(\boldsymbol{\theta})$ (а значит и оптимальной) будет оценка $\mathbf{c}^T\widehat{\boldsymbol{\theta}} \sim \mathcal{N}(\mathbf{c}^T\boldsymbol{\theta}, \sigma^2\mathbf{c}^T(Z^TZ)^{-1}\mathbf{c})$ (в случае $\tau(\boldsymbol{\theta}) = \theta_i$ вектор $\mathbf{c}$ будет содержать лишь одну единицу на i-ой позиции посреди нулей, поэтому дисперсия оценки будет равной $\sigma^2\left[(Z^TZ)^{-1}\right]_{ii}$, как было получено выше). Таким образом,

$$\frac{\mathbf{c}^T\widehat{\boldsymbol{\theta}} - \mathbf{c}^T\boldsymbol{\theta}}{\sqrt{\sigma^2\mathbf{c}^T(Z^TZ)^{-1}\mathbf{c}}} \sim \mathcal{N}(0,1) \Longrightarrow \frac{\mathbf{c}^T\widehat{\boldsymbol{\theta}} - \mathbf{c}^T\boldsymbol{\theta}}{\sqrt{\widehat{\sigma^2}\mathbf{c}^T(Z^TZ)^{-1}\mathbf{c}}} \sim T_{n-k},$$

$$\mathsf{P}_{\boldsymbol{\theta},\sigma^2}\left(\mathbf{c}^T\widehat{\boldsymbol{\theta}} - T_{n-k,(1+\gamma)/2}\sqrt{\widehat{\sigma^2}\mathbf{c}^T(Z^TZ)^{-1}\mathbf{c}} < \mathbf{c}^T\boldsymbol{\theta} < \mathbf{c}^T\widehat{\boldsymbol{\theta}} + T_{n-k,(1+\gamma)/2}\sqrt{\widehat{\sigma^2}\mathbf{c}^T(Z^TZ)^{-1}\mathbf{c}}\right) = \gamma,$$

14.4 Проверка линейных гипотез

В модели гауссовской линейной регрессии часто возникают некоторые предположения касательно роли и взаимосвязи компонент вектора параметров $\boldsymbol{\theta}$. Обычно их можно записать как систему линейных уравнений, которой должны удовлетворять параметры. Таким образом, наша гипотеза будет состоять в предположении, что $\boldsymbol{\theta}$ лежит в некоторой гиперплоскости, то есть

$$H_0\colon T\boldsymbol{\theta} = \boldsymbol{\tau},$$

где $T \in \mathbb{R}^{m\times k}$, $\boldsymbol{\tau} \in \mathbb{R}^m$ – известные величины, причём будем допускать, что $\operatorname{rk} T = m \leqslant k$. Отсюда собственно и название гипотезы: мы накладываем некоторые линейные ограничения на параметр $\boldsymbol{\theta}$.

Из утверждения 14.1 мы знаем матожидание и ковариационную матрицу у $\widehat{\boldsymbol{\theta}}$, а значит, можем найти её и у $T\widehat{\boldsymbol{\theta}}$:

$$\mathsf{E}_{\boldsymbol{\theta},\sigma^2}T\widehat{\boldsymbol{\theta}} = T\mathsf{E}_{\boldsymbol{\theta},\sigma^2}\widehat{\boldsymbol{\theta}} = T\boldsymbol{\theta},$$

$$\mathsf{D}_{\boldsymbol{\theta},\sigma^2}T\widehat{\boldsymbol{\theta}} = T\left[\mathsf{D}_{\boldsymbol{\theta},\sigma^2}\widehat{\boldsymbol{\theta}}\right]T^T = \sigma^2\underbrace{T(Z^TZ)^{-1}T^T}_{=B} = \sigma^2 B.$$

$T\widehat{\boldsymbol{\theta}}$, как линейное преобразование над нормально распределённым $\widehat{\boldsymbol{\theta}}$, само нормально распределено, то есть

$$T\widehat{\boldsymbol{\theta}} \sim \mathcal{N}(T\boldsymbol{\theta}, \sigma^2 B).$$

В силу максимальности ранга T матрица B будет невырожденной, а значит — положительно определённой, поэтому у неё существует $\sqrt{B}$, откуда

$$\frac{1}{\sigma}\sqrt{B}^{-1}(T\widehat{\boldsymbol{\theta}} - T\boldsymbol{\theta}) \sim \mathcal{N}(0, E_m) \Longrightarrow$$

$$\left\|\frac{1}{\sigma}\sqrt{B}^{-1}(T\widehat{\boldsymbol{\theta}} - T\boldsymbol{\theta})\right\|^2 = \frac{1}{\sigma^2}(T\widehat{\boldsymbol{\theta}} - T\boldsymbol{\theta})^TB^{-1}(T\widehat{\boldsymbol{\theta}} - T\boldsymbol{\theta}) \sim \chi^2_m.$$

При верности гипотезы $T\boldsymbol{\theta} = \boldsymbol{\tau}$, а значит, при подстановке этого тождества в выражение выше получаем статистику, зависящую от $\widehat{\boldsymbol{\theta}}$,

$$\frac{1}{\sigma^2}(T\widehat{\boldsymbol{\theta}} - \boldsymbol{\tau})^T B^{-1}(T\widehat{\boldsymbol{\theta}} - \boldsymbol{\tau}) \sim \chi^2_m.$$

Вспомним, что у нас в запасе есть независящая от $\widehat{\boldsymbol{\theta}}$ статистика

$$\frac{1}{\sigma^2}\|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\|^2 \sim \chi^2_{n-k}.$$

Поделив одно на другое, мы избавимся от неизвестной σ^2, да ещё и получим «хорошее» распределение Фишера. При верности H_0 имеем

$$\frac{(T\widehat{\boldsymbol{\theta}} - \boldsymbol{\tau})^T B^{-1}(T\widehat{\boldsymbol{\theta}} - \boldsymbol{\tau})}{\|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\|^2} \cdot \frac{n-k}{m} \sim F_{m,n-k}.$$

Итоговый критерий записывается так:

$$R = \left\{ \frac{(T\widehat{\boldsymbol{\theta}} - \tau)^T B^{-1}(T\widehat{\boldsymbol{\theta}} - \tau)}{\|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\|^2} \cdot \frac{n-k}{m} > f_{1-\alpha} \right\},$$

где f_p – p-квантиль распределения Фишера со степенями свободы m и $n-k$.

Пример 14.4 (*t-test revisited*)**.** 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 $\boldsymbol{\theta}$. Таким образом:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \\ Y_1 \\ \vdots \\ Y_m \end{pmatrix} = Z\boldsymbol{\theta} + \boldsymbol{\varepsilon} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \boldsymbol{\varepsilon},$$

где $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 E_{n+m})$. Стало быть, H_0 есть линейная гипотеза:

$$H_0 \colon T\boldsymbol{\theta} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = 0 = \tau.$$

Найдём величины, участвующие в критерии выше:

$$Z^T Z = \begin{pmatrix} n & 0 \\ 0 & m \end{pmatrix}, \quad \widehat{\boldsymbol{\theta}} = \begin{pmatrix} \overline{\mathbf{X}} \\ \overline{\mathbf{Y}} \end{pmatrix}, \quad B = T(Z^T Z)^{-1} T^T = \frac{1}{n} + \frac{1}{m},$$

$$(T\widehat{\boldsymbol{\theta}} - \tau)^T B^{-1}(T\widehat{\boldsymbol{\theta}} - \tau) = \frac{nm(\overline{\mathbf{X}} - \overline{\mathbf{Y}})^2}{n+m}, \quad \|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\|^2 = \sum_{i=1}^{n}(X_i - \overline{\mathbf{X}})^2 + \sum_{j=1}^{m}(Y_j - \overline{\mathbf{Y}})^2.$$

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 $n+m-2$ (сравните с t-критерием Стьюдента из раздела 12.1, правда похоже?). Процедуру выше можно

обобщить на случай нескольких выборок, что позволяет проверять сложную гипотезу о том, что у независимых групп одинаковые средние (см. задачу 14.6). ■

Пример 14.5. Может возникнуть ситуация, когда среди признаков определённо имеются лишние, которые не дают существенного вклада в нахождении целевой величины. Было бы крайне полезно проверять признаки на их полезность, так как выкидывание бесполезных признаков позволяет понизить размерность задачи. *Гипотезу о значимости признаков* (без потери общности, первых m признаков) можно формализовать как

$$H_0\colon \theta_1 = \ldots = \theta_m = 0.$$

Гипотезу можно переписать как $H_0\colon T\boldsymbol{\theta} = \mathbf{0}$, где

$$T = \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 & \ldots & 0 \end{pmatrix}.$$

■

Задачи

Задача 14.1. Докажите, что линейный переход от одного пространства признаков в другое (то есть когда в качестве матрицы «объект-признак» берётся не Z, а ZS, где $S \in GL(k, \mathbb{R})$) концептуально не меняет оценку МНК (и заодно объясните, что это значит).

Задача 14.2. Пусть помимо k вещественных признаков имеется категориальный признак, принимающий d различных значений. Для j-ого из них рассматривается своя линейная зависимость с n_j объектами:

$$\mathbf{X}^{(j)} = Z^{(j)}\boldsymbol{\theta} + \boldsymbol{\varepsilon}^{(j)},$$

где $\mathbf{X}^{(j)} \in \mathbb{R}^{n_j}$, $Z^{(j)} \in \mathbb{R}^{n_j \times k}$. Сведите задачу к одной линейной регрессии и покажите, что оценка МНК не поменяется. Чем это отличается от тупого добавления $d-1$ onehot признаков?

Задача 14.3. Докажите теорему 14.2, а именно покажите, что статистика $(\widehat{\boldsymbol{\theta}}, \|\mathbf{X} - Z\widehat{\boldsymbol{\theta}}\|^2)$ является **(а)** достаточной; **(б)** полной в модели гауссовской линейной регрессии.

Указание. **(а)** Теорема Пифагора: существует; **(б)** Чтобы применить достаточное условие полноты для экспоненциального семейства, попробуйте рассмотреть другую статистику, которая «содержит столько же информации», сколько и исходная.

Задача 14.4. Постройте доверительный интервал уровня доверия γ для целевой величины, отвечающего некоторому набору признаков $\mathbf{z}_0$ (то есть $\mathbf{z}_0^T\boldsymbol{\theta} + \varepsilon$, где $\varepsilon \sim \mathcal{N}(0, \sigma^2)$) для модели гауссовской линейной регрессии.

Задача 14.5. Покажите, что оценка (19) в обобщённой модели линейной регрессии является наилучшей в среднеквадратичном подходе в классе линейных несмещённых оценок при выполнении условий **L1**, **L2'** и **L3**.

Задача 14.6 (*one-way ANOVA*)**.** Пусть имеется k выборок $\mathbf{X}_1, \ldots, \mathbf{X}_k$, причём i-ая из них $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})$ состоит из n_i наблюдений, которые распределены как $\mathcal{N}(\mu_i, \sigma^2)$ (дисперсии, как и ранее, считаются равными, хоть и неизвестными). Сведя задачу к линейной регрессии, предложите критерий для проверки гипотезы $H_0\colon \mu_1 = \mu_2 = \ldots = \mu_k$ против общей альтернативы.

15 Байесовский подход

15.1 Мотивация

Здесь мы собираемся не просто изучить новый подход в статистике, а постичь целую философию, новую парадигму мышления, противопоставляемая тому пониманию статистики, которому нас учили ранее.

Обычный статистический вывод построен на *частотном* (или *фреквентистском*) *подходе* (от англ. frequentism). Ранее мы всегда предполагали, что модель (или параметры, которые её описывают) в реальности определена и фиксирована — просто мы её не знаем. Единственный способ уменьшить неопределённость — потреблять всё больше и больше наблюдений, которые в силу различных предельных законов должны с ростом выборки более точно описывать модель. Нам это позволяет сделать *принцип статистической устойчивости частот*: с самых азов мы связывали вероятность с долей экспериментов, в которых событие осуществилось, отчего важную роль играет повторяемость эксперимента. В то же время неизвестные параметры детерминированные, а значит, к ним нельзя применить вероятностные суждения. Мы не могли ранее сказать что-то типа «гипотеза верна с вероятностью 0.73» или «параметр больше 0 с вероятностью 0.42», так как всё это фиксированные константы.

В *байесовском подходе* вероятность интерпретируется как субъективная оценка, степень уверенности которой может быть основана на нашем опыте или вере, она не обязана подкрепляться частотностью наблюдаемого события. Посему любая величина интерпретируется как случайная, что позволяет нам использовать статистический аппарат для извлечения выводов о неизвестных нам характеристиках. Наши предположения о неизвестном параметре 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 $\mathcal{B}(\Theta)$).

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 σ-алгебр, а мера $\widetilde{\mathsf{P}}$ задаётся обобщённой плотностью $\rho_{\theta,X}(t, \mathbf{x}) = q(t) \cdot \rho(\mathbf{x}|t)$. Обобщённая плотность $\rho(\mathbf{x}|t)$, 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. Условное распределение параметра θ при условии выборки $X_1, \ldots, X_n$, чья плотность (напоминаем) может быть высчитана по формуле

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

называется *апостериорной плотностью* параметра θ.

В этом ещё одно преимущество байесовского подхода: вместо обычной точечной оценки или доверительного интервала мы получаем целое распределение, из которого можно смастерить точечную оценку любыми удобными способами. Например, можно взять среднее значение по плотности, что есть попросту

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

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

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

Польза полученной оценки проявляется в свете следующего подхода в сравнении оценок.

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

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

где Q — априорное распределение на множестве параметров Θ.

Теорема 15.1.

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

Другой, более простой вариант — взять моду полученного распределения, то есть самое правдоподобное значение параметра (такую оценку иногда называют «Байесом для бедных»). Удобство такого метода заключается в ненадобности считать интеграл в знаменателе апостериорной плотности: он не зависит от параметра, поэтому максимизация $\rho(t|\mathbf{x})$ по θ равносильна максимизации $q(t)\rho(\mathbf{x}|t)$, что даёт оценку

$$\theta^*(\mathbf{X}) = \arg\max_{t\in\Theta} q(t)\rho(\mathbf{X}|t) = \arg\max_{t\in\Theta} \ln\left[q(t)\rho(\mathbf{X}|t)\right] = \arg\max_{t\in\Theta} \left[\ln\rho(\mathbf{X}|t) + \ln q(t)\right].$$

Заметим, что полученная оценка очень напоминает ОМП, только на этот раз под знак $\arg\max$ добавлен *регуляризатор* $\ln q(t)$, которой накладывает дополнительные ограничения на параметр. Подобные регуляризационные свойства позволяют использовать байесовский подход при любых размерах выборки, в отличие от прежних методов, результаты которых актуальны лишь для достаточно большого набора данных.

Пример 15.1. Применим байесовский подход к модели гауссовской линейной регрессии $\mathbf{X} = Z\boldsymbol{\theta} + \varepsilon$, где $\varepsilon \sim \mathcal{N}(0, \sigma^2 E)$ (см. главу 14). Возьмём в качестве априорного распределения на вектор параметров $\boldsymbol{\theta}$ гауссовский вектор с независимыми компонентами $\mathcal{N}(0, \gamma^2 E_k)$. В нём формализовано наше желание быть вектору $\boldsymbol{\theta}$ не очень большим по норме: вероятностная масса у нормального распределения сконцентрирована как раз у начала координат, а насколько сильно — определяется параметром γ^2.

Более подробно, априорное распределение задаётся плотностью

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

Значит, апостериорному распределению соответствует

$$\rho(\mathbf{t}|\mathbf{x}) \sim \exp\left(-\frac{1}{2\gamma^2}\mathbf{t}^T\mathbf{t} - \frac{1}{2\sigma^2}(\mathbf{x}-Z\mathbf{t})^T(\mathbf{x}-Z\mathbf{t})\right).$$

Если мы попытаемся максимизировать эту плотность (то есть найти байесовскую оценку «для бедных»), то получим следующую задачу оптимизации:

$$\|\mathbf{X}-Z\mathbf{t}\|^2 + \frac{\sigma^2}{\gamma^2}\|\mathbf{t}\|^2 \to \min_{\mathbf{t}}.$$

Такой способ называется Ridge regression. Его преимущество в том, что мы «штрафуем» вектор $\boldsymbol{\theta}$ за излишне большие координаты, что позволяет получать более стабильные решения с меньшей дисперсией. 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 $\rho(\mathbf{t}|\mathbf{x})$. С одной стороны,

$$\rho(\mathbf{t}|\mathbf{x}) \sim \exp\left[\frac{1}{\sigma^2}\mathbf{x}^TZ\mathbf{t} - \frac{1}{2}\mathbf{t}^T\left(\frac{1}{\gamma^2}E + \frac{1}{\sigma^2}Z^TZ\right)\mathbf{t}\right].$$

С другой,

$$\rho(\mathbf{t}|\mathbf{x}) \sim \exp\left(-\frac{1}{2}(\mathbf{t}-a)^T\Sigma^{-1}(\mathbf{t}-a)\right) \sim \exp\left(a^T\Sigma^{-1}\mathbf{t} - \frac{1}{2}\mathbf{t}^T\Sigma^{-1}\mathbf{t}\right).$$

Получается, что

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

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

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

Получили практически решение задачи обычной линейной регрессии, но теперь к матрице Z^TZ 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^TZ на $\frac{\sigma^2}{\gamma^2}$ вправо, отчего определитель, как произведение собственных чисел, отдаляется от нуля.

Теперь возьмём следующее априорное распределение: компоненты вектора $\boldsymbol{\theta}$ независимы, $\theta_i \sim \text{Laplace}(\lambda)$. У данного распределения более тяжёлые хвосты, поэтому такая модель лучше объясняет данные с выбросами. Имеем априорную плотность

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

Следовательно, апостериорная плотность имеет вид

$$\rho(\mathbf{t}|\mathbf{x}) \sim \exp\left(-\lambda\|\mathbf{t}\|_1 - \frac{1}{2\sigma^2}(\mathbf{x}-Z\mathbf{t})^T(\mathbf{x}-Z\mathbf{t})\right).$$

Аналитически получить формулу для байесовской оценки проблематично, зато можно приблизить моду данного распределения с помощью градиентного спуска. Соответствующую задачу оптимизации можно сформулировать так:

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

Здесь имеется похожая регуляризация, что и в случае выше, 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»). ■

15.2 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 лишь некоторые способы.

15.2.1 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* $\{\mathsf{P}_\theta \colon \theta \in \Theta\}$.

Для примера рассмотрим выборку $X_1, \ldots, X_n$ из распределения $\mathrm{Bern}(\theta)$. Её совместная плотность равна

$$\rho(\mathbf{x}|t) = t^{\sum x_i}(1-t)^{n-\sum x_i}.$$

Чтобы получить апостериорную плотность, надо домножить $\rho(\mathbf{x}|t)$ на априорную плотность и потом нормировать это произведение, деля на некоторый интеграл. Таким образом, $\rho(t|\mathbf{x})$ пропорциональна $q(t)\rho(\mathbf{x}|t)$, а стало быть, надо подобрать семейство для $q(t)$ таким образом, чтобы домножение на $\rho(\mathbf{x}|t)$ 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 $\mathrm{Beta}(\alpha, \beta)$, 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

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

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

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

Пример 15.2. Найдём сопряжённые семейства для ещё некоторых известных семейств распределений.

- Пусть $X_1, \ldots, X_n$ — выборка из распределения $\mathrm{U}(0, \theta)$. Имеем совместную плотность $\rho(\mathbf{x}|t) = t^{-n} I(0 < x_1, \ldots, x_n < t)$. 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

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

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

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

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

- Пусть $X_1, \ldots, X_n$ — выборка из распределения $\mathrm{Pois}(\theta)$. Имеем совместную плотность

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

  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

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

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

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

- Пусть $X_1, \ldots, X_n$ — выборка из распределения $\mathcal{N}(\theta, 1)$. Имеем совместную плотность

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

  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

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

Следовательно, для реализации выборки $\mathbf{x}$ апостериорным распределением является $\mathcal{N}\left(\widehat{a}(\mathbf{x}),\widehat{\sigma^2}(\mathbf{x})\right)$. 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:

$$\begin{cases}\dfrac{\widehat{a}}{\widehat{\sigma^2}}=\sum x_i+\dfrac{a}{\sigma^2},\\[2mm] \dfrac{1}{\widehat{\sigma^2}}=n+\dfrac{1}{\sigma^2}.\end{cases}\Longrightarrow \widehat{a}(\mathbf{x})=\frac{\sum x_i+\frac{a}{\sigma^2}}{n+\frac{1}{\sigma^2}}.$$

Отсюда находим, что

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

- Пусть $X_1,\ldots,X_n$ — выборка из распределения $\mathcal{N}(0,\theta)$. Имеем совместную плотность

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

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

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

и апостериорным распределением является Inv-Gamma $\left(\alpha+n/2,\lambda+\frac{1}{2}\sum x_i^2\right)$. Давайте поймём, как выглядит матожидание у $\xi\sim$ Inv-Gamma(a,b):

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

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

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

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

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

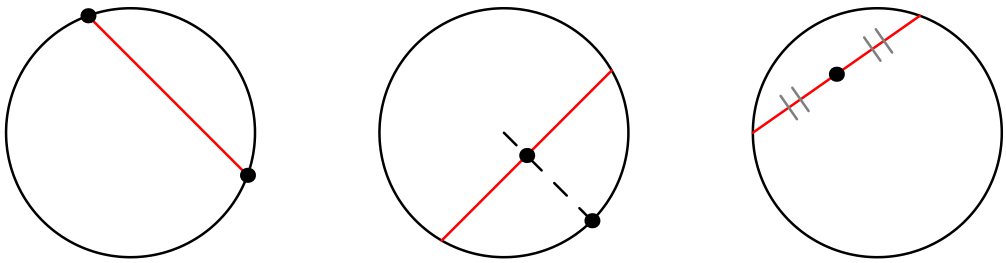
■

15.2.2 Распределение Джеффриса

В случае неимения каких-либо априорных знаний о параметре возникает логичное желание задать на Θ равномерное распределение. Да, с неограниченным носителем так не выйдет (ну, или почти не выйдет, см. пример 15.5), 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, что иллюстрирует следующий известный

Пример 15.3 (*парадокс Бертрана*)**.** Пусть наше семейство распределений параметризовано случайной хордой единичной окружности, и перед нами стоит задача задать на них какое-то априорное распределение. Работать непосредственно с хордами неудобно, легче задать их какими-то другими численными параметрами, на которые, в свою очередь, можно ввести равномерное распределение. Однако так можно сделать множеством способов (см. картинки):

1. Определить хорду через два конца, которые будут иметь равномерное распределение на окружности (левый рисунок);
2. Определить хорду через радиус окружности и точку на нём, через которую пройдёт хорда, перпендикулярно радиусу. Радиус проведём к точке, равномерно взятой на окружности, а точка на радиусе будет взята с равномерным распределением на нём (выделен пунктиром на рисунке по центру);
3. Определить хорду через её середину, которой зададим равномерное распределение на единичном круге (правый рисунок).



Во всех случаях хорда определяется через нечто, имеющее «естественное» равномерное распределение, однако, как несложно убедиться, во всех трёх случаях итоговое распределение на хордах получится своим. Выходит, задавая равномерное распределение, мы всё-таки вносим какую-то информацию о параметре, чего хотелось бы избежать. ■

Это является мотивацией к идее, что априорная плотность должна быть устойчива к замене переменной. Иными словами, априорное распределение нужно выбирать таким образом, чтобы при любой параметризации распределение на исходных параметрах было одним и тем же.

Рассмотрим одномерный случай. Напомним, что если к случайной величине ξ с плотностью $\rho_\xi(x)$ применяется диффеоморфизм φ, то плотность пересчитывается как

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

И тут на сцене появляется информация Фишера. Зададимся вопросом: как поменяется информация Фишера $I_{\mathbf{X}}(\theta)$, если отныне параметром бы будем считать не θ, а некоторую

$\sigma = \varphi(\theta)$? Ответ неожиданный и приятный:

$$I_{\mathbf{X}}(\sigma) = \mathsf{E}_\theta \left(\frac{\partial \ln \rho_\theta(\mathbf{X})}{\partial \varphi(\theta)} \right)^2 = \mathsf{E}_\theta \left(\frac{\partial \ln \rho_\theta(\mathbf{X})}{\partial \theta} \cdot \frac{\partial \theta}{\partial \varphi(\theta)} \right)^2 = \left(\frac{\partial \theta}{\partial \varphi(\theta)} \right)^2 \mathsf{E}_\theta \left(\frac{\partial \ln \rho_\theta(\mathbf{X})}{\partial \theta} \right)^2 =$$
$$= \frac{1}{\left(\frac{\partial \varphi(\theta)}{\partial \theta} \right)^2} \cdot I_{\mathbf{X}}(\theta) = \frac{1}{\varphi'(\theta)^2} \cdot I_{\mathbf{X}}(\theta) = \frac{1}{\varphi'(\varphi^{-1}(\sigma))^2} \cdot I_{\mathbf{X}}(\varphi^{-1}(\sigma)).$$

Вот те раз! Прямо как в формуле плотности при замене переменной появляется производная в знаменателе, правда на этот раз в квадрате. Значит, если мы возьмём в качестве априорного распределения

$$q(t) \sim \sqrt{I_{\mathbf{X}}(t)},$$

то при смене параметризации информация Фишера поменяется так же, как и плотность, следовательно, распределение на θ всегда будет одним и тем же, даже если изначально в качестве параметра была взята $\sigma = \varphi(\theta)$.

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

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

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

Иногда бывает так, что $\sqrt{I_{\mathbf{X}}(\theta)} \notin L_1(\Theta)$ 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 с $\int_\Theta q(t)dt = \infty$ называют *несобственными* (англ. improper prior).

Пример 15.5. Посмотрим, как ведут себя такие распределения и насколько адекватными получаются из них оценки.

- Пусть $X_1, \ldots, X_n$ — выборка из распределения $\mathrm{Pois}(\theta)$. Найдём информацию Фишера для распределения Пуассона:

$$\rho_t(x) = \frac{t^x e^{-t}}{x!}; \quad \ln \rho_t(x) = x \ln t - t - \ln x!;$$

$$\frac{\partial}{\partial t}\ln\rho_t(x) = \frac{x}{t} - 1; \quad i(t) = \mathsf{D}_t\left(\frac{X_1}{t} - 1\right) = \frac{1}{t^2}\mathsf{D}_t X_1 = \frac{1}{t}.$$

Получается, что распределение Джеффриса имеет плотность

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

что не интегрируемо на $(0; +\infty)$. Но при этом

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

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

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

- Пусть теперь $X_1, \ldots, X_n$ – выборка из распределения $\mathcal{N}(\theta, 1)$. Информацию Фишера позаимствуем из задачи 3.6: $i(t) = 1/\sigma^2 = 1$, то есть распределение Джеффриса будет равномерным распределением на $\mathbb{R}$ (следовательно, несобственным). В таком случае

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

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

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

■

Как можно видеть, несмотря на то что затея с несобственными распределениями кажется неадекватной, она даёт нам довольно неплохие оценки. В последнем случае полученная байесовская оценка и вовсе совпала с оценкой максимального правдоподобия: в отличие от предыдущих оценок в ней нет никаких дополнительных сдвигов и добавок, который регуляризируют нашу оценку, исходя из наших априорных знаний. Это наталкивает на мысль, что полученное распределение является *неинформативным*, то есть не вносящим никаких дополнительных знаний о параметре.

Аналогичными свойствами обладает несобственное распределение $q(t) = \frac{1}{t(1-t)}$ для выборки из $\mathrm{Bern}(\theta)$, которое формально можно толковать как $\mathrm{Beta}(0, 0)$. Как мы убедились ранее, для априорного бета-распределения при пересчёте плотности первый параметр увеличивается на число единиц в выборке, а второй — на число нулей. Поэтому нули в качестве параметров априорного распределения можно понимать как отсутствие какой-либо информации об известных результатах — до проведения эксперимента мы не знаем ни про выпавшие единицы, ни про выпавшие нули.

15.3 Связь с минимаксными оценками

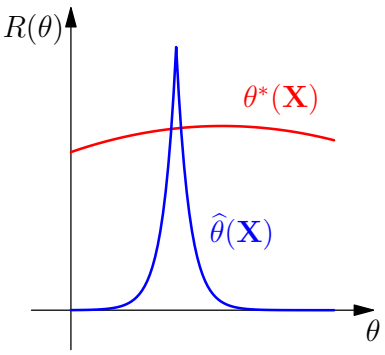
В контексте байесовского подхода полезно рассмотреть иной способ сравнения оценок, который тесно связан с текущим. Он, как и ранее, ранжирует функции риска с помощью одного числа. В байесовском подходе это была L_1-норма по некоторой вероятностному (или несобственному) распределению. Сейчас, дабы не утруждать себя выбором априорного распределения, предлагается взять L_∞-норму.

Определение. Пусть $\widehat{\theta}$ и θ^* — оценки параметра θ. Говорят, что оценка $\widehat{\theta}$ *лучше оценки* θ^* *в минимаксном подходе*, если

$$\sup_{\theta\in\Theta} R(\widehat{\theta},\theta) \leqslant \sup_{\theta\in\Theta} R(\theta^*,\theta).$$

Оценка $\widehat{\theta}$ называется *минимаксной*, если она лучше любой другой оценки в минимаксном подходе, то есть

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



Может показаться, что такая метрика не совсем адекватна. Она штрафует оценки, которые могут сильно ошибаться при некоторых значениях параметра, даже если их в каком-то смысле не очень много, а при других значениях она показывает себя отлично. Например, на рисунке оценка $\widehat{\theta}(\mathbf{X})$ будет лучше в байесовском подходе, если взять равномерное распределение на Θ, потому что в среднем она ошибается незначительно. Но в минимаксном выиграет $\theta^*(\mathbf{X})$, хотя всюду функция риска довольно велика. Впрочем, это не значит, что данный подход совсем не годен, уместность его использования зависит от конкретной ситуации. Байесовский подход на то и байесовский, что он минимизирует взвешенную ошибку, основанную на нашем представлении о том, как часто встречается то или иное значение параметра в природе. Минимаксный подход же удобен, когда никакой информации о параметре нет, и нам хотелось бы перестраховаться на самый неблагоприятный случай.

Несмотря на такое фундаментальное различие между данными подходами, между ними имеется тесная связь, позволяющая из байесовости оценки получать минимаксность. Ключевым наблюдением здесь будет тот факт, что для вероятностных мер $\|\cdot\|_\infty \geqslant \|\cdot\|_1$, отсюда появляется возможность оценить снизу произвольную оценку по минимаксной метрике.

Теорема 15.2.

Пусть $\widehat{\theta}(\mathbf{X})$ — наилучшая оценка в байесовском подходе относительно априорного вероятностного распределения Q и функции потерь R, причём $R(\widehat{\theta},\theta) \equiv \text{const}$. Тогда она является и минимаксной относительно этой функции потерь.

Доказательство. Пусть $\theta^*(\mathbf{X})$ — произвольная оценка. Её минимаксную норму можно оценить снизу следующим образом:

$$\sup_{\theta\in\Theta} R(\theta^*,\theta) \geqslant \int_\Theta R(\theta^*,\theta)\,\mathsf{Q}(d\theta) \geqslant \int_\Theta R(\widehat{\theta},\theta)\,\mathsf{Q}(d\theta),$$

где последнее неравенство следует из оптимальности оценки $\widehat{\theta}$ в байесовском подходе. Но так как её функция риска постоянна, то последнее выражение в точности равно $\sup_{\theta\in\Theta} R(\widehat{\theta},\theta)$, откуда следует минимаксность оценки $\widehat{\theta}$. □

Таким образом, нам достаточно найти «хорошее» априорное распределение, которое выравнивает функцию риска, тогда байесовская оценка автоматически будет минимаксной.

Пример 15.6. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из $\mathrm{Bern}(\theta)$, найдём в этой модели минимаксную оценку при квадратичной функции потерь. Ранее мы уже поняли, что байесовская оценка легко считается при априорном распределении $\mathrm{Beta}(\alpha, \beta)$: в таком случае оптимальная оценка равна $\widehat{\theta}(\mathbf{X}) = \frac{\alpha + \sum X_i}{\alpha + \beta + n}$

Попробуем подобрать параметры α и β так, чтобы $R(\widehat{\theta}, \theta) = \mathsf{E}_\theta(\widehat{\theta} - \theta)^2 \equiv \mathrm{const}$.

$$\mathsf{E}_\theta(\widehat{\theta} - \theta)^2 = \mathsf{D}_\theta\widehat{\theta} + (\mathsf{E}_\theta\widehat{\theta} - \theta)^2 = \frac{\mathsf{D}_\theta \sum X_i}{(\alpha + \beta + n)^2} + \left(\frac{\alpha + n\theta}{\alpha + \beta + n} - \theta\right)^2 =$$

$$= \frac{n\theta(1-\theta) + (\alpha(1-\theta) - \beta\theta)^2}{(\alpha + \beta + n)^2} \sim ((\alpha+\beta)^2 - n)\theta^2 + (n - 2\alpha(\alpha+\beta))\theta + \alpha^2 \equiv \mathrm{const} \Longrightarrow$$

$$\begin{cases} \alpha + \beta = \sqrt{n}, \\ \alpha(\alpha + \beta) = \dfrac{n}{2} \end{cases}; \quad \alpha = \beta = \frac{\sqrt{n}}{2}.$$

Таким образом, по теореме 15.2 оценка

$$\widehat{\theta}(\mathbf{X}) = \frac{\sqrt{n}/2 + \sum X_i}{\sqrt{n} + n} = \frac{\overline{\mathbf{X}} + \frac{1}{2\sqrt{n}}}{1 + \frac{1}{\sqrt{n}}}$$

является минимаксной при квадратичной функции потерь. ■

Полученная выше оценка называется *оценкой Ходжеса-Лемана*. Несложно заметить, что она является выпуклой комбинацией обычного среднего и $1/2$, и поэтому она смещена.

Задачи

Задача 15.1. Покажите, что знание о параметре может обновляться постепенно по мере поступления новых наблюдений. Более формально, пусть наблюдения $X_1, \ldots, X_n$ подаются по одному, и старая апостериорная плотность становится априорной, из которой посредством нового наблюдения X_i получается новая апостериорная плотность. Докажите, что в итоге получится та же плотность, как если бы она была получена по всей выборке сразу.

Задача 15.2. Докажите, что байесовская оценка является функцией от достаточной статистики.

Задача 15.3. Докажите теорему 15.1.

Задача 15.4. Найдите семейство распределений, сопряжённое $\{\mathcal{N}(0, 1/\theta) \colon \theta > 0\}$.

Задача 15.5. В примере 15.5 для выборки $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$, неинформативное распределение дало нам обычное выборочное среднее в качестве байесовской оценки. Докажите, что если взять честное вероятностное распределение в качестве Q, то такой байесовской оценки получиться не может.

Задача 15.6. В модели масштаба $X_1, \ldots, X_n \sim \mathcal{N}(0, \theta^2)$, $\theta > 0$, придумайте неинформативное распределение для параметра масштаба θ. Аргументируйте ваш выбор.

16 Робастность

Выделим в отдельный параграф важное свойство оценок, упоминаемое в обсуждении метода выборочных квантилей. Мотивация его такова: в реальном мире данные не обязаны подстраиваться под нашу идеальную модель. Вероятнее всего наблюдения подчиняются несколько иному распределению, которое немного отклоняется от наших предположений. Нам бы хотелось, чтобы получаемые статистические результаты не сильно от этого страдали.

Цель этого параграфа — понять, как можно измерить устойчивость оценки при «шевелении» истинного распределения и какие статистики лучше всего справляются с такого рода проблемами. Статистики с подобными свойствами называют *робастными*. Неформально робастные оценки — это такие оценки, которые устойчивы при небольшом отклонении от предположений выбранной модели. Попробуем дать математически точное, формальное описание этого термина.

Пусть $\mathcal{P}$ — множество всех распределений на пространстве $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ (обычно мы будем рассматривать одномерный случай $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$). Начнём с того, а как вообще понимать, когда распределения близки, а когда — не очень? Для этого можно явно задать метрику на $\mathcal{P}$, но проще вести дело с окрестностями: для каждого распределения $\mathsf{P} \in \mathcal{P}$ можно ввести его ε-окрестность $B_\varepsilon(\mathsf{P}) \subset \mathcal{P}$, содержимое которой мы и будем понимать как распределения, «близкие» к P. Один из подходов — брать окрестности, порождающие топологию слабой сходимости. Например, в одномерном случае зачастую рассматривают *расстояние Леви*, чьи окрестности на языке функций распределения имеют вид

$$\mathcal{L}_\varepsilon(F_0) = \{F \in \mathcal{P} \,|\, \forall x \in \mathbb{R}\colon F_0(x-\varepsilon) - \varepsilon \leqslant F(x) \leqslant F_0(x+\varepsilon) + \varepsilon\}.$$

Отклонение от распределения можно понимать и по-другому. Распространённый случай — загрязнение выборки, когда небольшая доля наблюдений, которые называют *выбросами*, приходит не из истинного распределения P_0, а произвольного мусорного Q. Если считать, что вероятность появления выброса равна ε, то итоговое распределение будет равно $\mathsf{P}' = (1-\varepsilon)\mathsf{P}_0 + \varepsilon\mathsf{Q}$. Таким образом, окрестность загрязнения можно определить как

$$\mathcal{C}_\varepsilon(\mathsf{P}_0) = \{\mathsf{P} \in \mathcal{P} \,|\, \exists \mathsf{Q} \in \mathcal{P}\colon \mathsf{P} = (1-\varepsilon)\mathsf{P}_0 + \varepsilon\mathsf{Q}\}.$$

Впрочем, в контексте сходимости распределений данное расстояние не совсем интуитивное. Например, если в качестве P_0 взять дискретное распределение на прямой, то для любого $\varepsilon < 1$ окрестность $\mathcal{C}_\varepsilon(\mathsf{P}_0)$ не будет содержать непрерывные распределения, хотя P_0 можно легко представить как слабый предел непрерывных распределений.

Пусть G — некоторый функционал на множестве $\mathcal{P}$. Мы хотим понять, насколько plug-in оценка $\widehat{\theta} = G(\widehat{\mathsf{P}}_n)$ устойчива при оценивании параметра $G(\mathsf{P}_0)$. Полезной может оказаться следующая характеристика, которая показывает, насколько сильно может в теории поменяться значение функционала, если в пределах разумного пошевелить распределение:

$$b_{\mathsf{P}_0}(\varepsilon) = \sup_{\mathsf{P} \in B_\varepsilon(\mathsf{P}_0)} |G(\mathsf{P}) - G(\mathsf{P}_0)|,$$

где $B_\varepsilon(\mathsf{P}_0)$ — окрестность в каком-то из смыслов выше (будем указывать тип окрестности в верхнем индексе, например, $b^{\mathcal{C}}_{\mathsf{P}_0}(\varepsilon)$, если это неясно из контекста).

Отметим, что для $\varepsilon = 1$ загрязнённое распределение будет целиком состоять из мусора, и тогда величина $b^{\mathcal{C}}_{\mathsf{P}_0}(\varepsilon)$ будет показывать максимальное отклонение от искомой статистики $G(\mathsf{P}_0)$ — хуже уже не будет.

Самое базовое требование от робастной plug-in оценки — непрерывность функционала, её породившего.

Определение. Оценку $G(\widehat{\mathsf{P}})$ называют *качественно робастной* при $\mathsf{P} = \mathsf{P}_0$, если функционал G непрерывен в точке P_0 относительно слабой сходимости, то есть $b^{\mathcal{L}}_{\mathsf{P}_0}(\varepsilon) \to 0$ при $\varepsilon \to 0$.

Иногда вместо слабой сходимости указывают равномерную сходимость функций распределения, что, впрочем, то же самое в случае непрерывного распределения P_0.

Пример 16.1. Функционал среднего $G(\mathsf{P}) = \int x\,\mathsf{P}(dx)$ не обладает качественной робастностью. Действительно, для $\mathsf{P}_\varepsilon = (1-\varepsilon)\mathsf{P}_0 + \varepsilon\delta_x$, где δ_x — распределение, сконцентрированное в точке x, имеем слабую сходимость $\mathsf{P}_\varepsilon \xrightarrow{d} \mathsf{P}_0$, при этом для каждого конкретного ε величину $G(\mathsf{P}_\varepsilon)$ можно сделать сколь угодно большой при подходящем выборе x.

С другой стороны, функционал медианы $\mu(\mathsf{P})$ таким недугом не обладает, потому что если у распределения P_0 медиана определена однозначно, то из слабой сходимости следует сходимость медиан. ■

Такое определение робастности уже позволяет отсеивать бесперспективные варианты по типу интегральных функционалов, однако не отвечает на вопрос, насколько сильное отклонение мы можем себе позволить. Хотелось бы количественно оценить доступный «запас прочности» у оценки, что позволяет сделать следующее

Определение. *Пороговым значением (точкой)* (англ. breakdown point) функционала $G(\widehat{\mathsf{P}})$ в точке P_0 называют величину

$$\varepsilon^* = \sup\{\varepsilon \in [0;1]\colon b^{\mathcal{C}}_{\mathsf{P}_0}(\varepsilon) < b^{\mathcal{C}}_{\mathsf{P}_0}(1)\}.$$

Оценку $G(\widehat{\mathsf{P}})$ называют *количественно робастной* при $\mathsf{P} = \mathsf{P}_0$, если пороговое значение в этой точке больше 0.

Как было отмечено выше, $\varepsilon = 1$ соответствует полному хаосу и наибольшему отклонению функционала, таким образом, пороговое значение показывает, какая доля выборки может быть загрязнена, чтобы получить результаты чуть лучше, чем полностью случайные.

Пример 16.2. Из рассуждений примера 16.1 следует, что для функционала среднего $G(\mathsf{P}) = \int x\,\mathsf{P}(dx)$ и любого $\varepsilon > 0$ выполнено $b_{\mathsf{P}}(\varepsilon) = \infty$ — даже если общая доля выбросов незначительна, можно сделать среднее сколь угодно большим, устремляя размер выброса к бесконечности. Таким образом, пороговое значение для среднего равно 0, и оценка не робастна.

Разберёмся с медианой. Если $\varepsilon > 1/2$, то очевидно $b(\varepsilon) = \infty$: можно взять сколь угодно большой $x \in \mathbb{R}$, тогда для $\mathsf{P}_\varepsilon = (1-\varepsilon)\mathsf{P}_0 + \varepsilon\delta_x \in b_{\mathsf{P}_0}(\varepsilon)$ справедлива оценка $F_{\mathsf{P}_\varepsilon}(x-0) = (1-\varepsilon)F_{\mathsf{P}_0}(x-0) < 1/2$, значит, $\mu(\mathsf{P}_\varepsilon) \geqslant x \to \infty$. Теперь пусть $\varepsilon < 1/2$, и $F_\varepsilon = (1-\varepsilon)F_0 + \varepsilon H$ для некоторой функции распределения H. Оценим значение $F_\varepsilon^{-1}(1/2)$ (как обычно принято, $F^{-1}(x) = \inf\{y\colon F(y) \geqslant x\}$). Заметим, что для точки $x < F^{-1}\left(\frac{1-2\varepsilon}{2(1-\varepsilon)}\right)$ верна оценка

$$F_\varepsilon(x) = (1-\varepsilon)F_0(x) + \varepsilon H(x) < (1-\varepsilon)\cdot\frac{1-2\varepsilon}{2(1-\varepsilon)} + \varepsilon\cdot 1 = \frac{1}{2},$$

поэтому $\mu(F_\varepsilon) \geqslant F^{-1}\left(\frac{1-2\varepsilon}{2(1-\varepsilon)}\right)$. Аналогично показывается, что $\mu(F_\varepsilon) \leqslant F^{-1}\left(\frac{1}{2(1-\varepsilon)}\right)$, причём несложно убедиться, что подходящим выбором H можно добиться значения

медианы, сколько угодно близкого к полученным границам. Итого,

$$b_{F_0}(\varepsilon) = \max\left\{\mu(F_0) - F^{-1}\left(\frac{1-2\varepsilon}{2(1-\varepsilon)}\right), F^{-1}\left(\frac{1}{2(1-\varepsilon)}\right) - \mu(F_0)\right\} < \infty \quad \text{при } \varepsilon < 1/2,$$

и пороговое значение равно $\varepsilon^*(\mu) = 1/2$. Аналогично показывается, что функционал $G_p(F) = F^{-1}(p)$ — p-квантиль распределения — имеет пороговое значение $\min(p, 1-p)$. ■

16.1 Функция влияния

Пороговое значение раскрывает функционал с глобальной точки зрения, показывая, насколько сильно можно испортить выборку. Но к сожалению, оно не конкретизирует поведение plug-in оценки при меньшем загрязнении выборки: полезно было бы оценить меру ухудшения оценки, когда доля выбросов достаточно мала. В этом нам поможет следующая важная величина, которая имеет множество приложений в непараметрической статистике.

Определение. *Производная по Гато* функционала G в точке P по направлению Q определяется как

$$L_{\mathsf{P}}(\mathsf{Q}) = \lim_{\varepsilon \to 0} \frac{G((1-\varepsilon)\mathsf{P} + \varepsilon\mathsf{Q}) - G(\mathsf{P})}{\varepsilon}.$$

Функцией влияния функционала G в точке P называют функцию $I_{\mathsf{P}} \colon x \mapsto L_{\mathsf{P}}(\delta_x)$.

Таким образом, функция влияния, оправдывая своё название, показывает асимптотическое изменение оценки при появлении в выборке большого размера выброса x.

16.2 Симметричные распределения

17 Бутстреп

Как мы могли убедиться ранее, порой не всегда обычная точечная оценка $T_n(\mathbf{X}) = T_n(X_1, \ldots, X_n)$ даст нам полноценную информацию о значении неизвестного параметра. Немаловажную роль также играет значение дисперсии нашей статистики $\mathsf{D}T_n(\mathbf{X})$, так как её высокое значение может свидетельствовать о нашей неуверенности в результате. Однако не всегда эту дисперсию можно легко посчитать аналитически или даже приблизить.

Во-первых, что очевидно, значение дисперсии может зависеть от значения неизвестного параметра, который мы собственно и хотим оценить. Во-вторых, в большинстве случаев мы вовсе не знаем, как параметризовать семейство: не всегда у нас имеется хоть какая-нибудь априорная информация о неизвестном распределении, которому подчиняется выборка. Иногда нам может повести, и оценка $T_n(\mathbf{X})$ окажется асимптотически нормальной с известной асимптотической дисперсией:

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

откуда можно сделать вывод, что дисперсию $\mathsf{D}_\theta T_n(\mathbf{X})$ якобы можно приблизить значением σ^2/n. Но сходимость по распределению — вещь ненадёжная, требующая достаточного числа данных, и даже для относительно больших выборок приближение нормальным распределением может оказаться несостоятельным, а оценки дисперсии — тем более.

17.1 Принцип работы

На помощь приходит новый, по истине волшебный инструмент, предложенный Брэдли Эфроном в [4], и имя ему — *бутстреп* (bootstrap). Он сочетает в себе две идеи:

- Раз уж исходное распределение выборки мы не знаем, давайте заменим его на эмпирическое, то есть будем искать не дисперсию оригинальной статистики

  $$\mathsf{D}_{\mathsf{P}} T_n = \mathsf{D}T_n(X_1, \ldots, X_n), \quad X_1, \ldots, X_n \sim \mathsf{P} \in \mathcal{P},$$

  а дисперсию статистики от выборки из эмпирического распределения

  $$\mathsf{D}_{\widehat{\mathsf{P}}} T_n = \mathsf{D}T_n(X_1^*, \ldots, X_n^*), \quad X_1^*, \ldots, X_n^* \sim \widehat{\mathsf{P}},$$

  где, напомним, эмпирическое распределение $\widehat{\mathsf{P}}$ задаётся функцией распределения

  $$\widehat{F_n}(x) = \sum_{i=1}^{n} \frac{I(x \leqslant X_i)}{n}.$$

- Оценим дисперсию $\mathsf{D}_{\widehat{\mathsf{P}}} T_n$ с помощью *метода Монте-Карло*, который в общем случае выглядит так: если нам надо оценить $\mathsf{E}g(\mathbf{X})$, где $X \sim F$, то можно смоделировать B случайных величин $X_1, \ldots, X_B \sim F$, чьё среднее будет стремиться к оцениваемому числу по закону больших чисел:

  $$\overline{h(\mathbf{X})} = \frac{1}{B} \sum_{i=1}^{B} h(X_i) \xrightarrow{\mathsf{P}} \mathsf{E}h(\mathbf{X}).$$

  Очень похоже на метод моментов за тем лишь исключением, что на этот раз выборка не приходит к нам сверху, а генерируется нами самостоятельно. В частности, дисперсия оценивается выборочным аналогом:

  $$\frac{1}{B} \sum_{i=1}^{B} (X_i - \overline{X})^2 = \frac{1}{B} \sum_{i=1}^{B} X_i^2 - \left(\frac{1}{B} \sum_{i=1}^{B} X_i \right)^2 \xrightarrow{\mathsf{P}} \mathsf{E}X_1^2 - (\mathsf{E}X_1)^2 = \mathsf{D}X_1.$$

Эти два простых шага удачно дополняют друг друга, и в общем случае нельзя избавиться от одного из них:

- Без первого шага нам пришлось бы семплировать из неизвестного распределения P вместо $\widehat{\mathsf{P}}$, что чаще всего не представляется возможным.
- Без второго шага нам пришлось бы аналитически вычислять $\mathsf{D}_{\widehat{\mathsf{P}}} T_n$, которая, вообще говоря, равна монструозной сумме по n индексам, i-ый из которых соответствует i-ому наблюдению X_i^*, который может равновероятно принимать значения из $(X_1, \ldots, X_n)$:

$$\mathsf{D}_{\widehat{\mathsf{P}}} T_n(X_1^*, \ldots, X_n^*) =$$

$$= \int T_n^2(x_1, \ldots, x_n)\, d\widehat{F}_n(x_1) \ldots d\widehat{F}_n(x_n) - \left(\int T_n(x_1, \ldots, x_n)\, d\widehat{F}_n(x_1) \ldots d\widehat{F}_n(x_n) \right)^2 =$$

$$= \frac{1}{n^n} \sum_{i_1=1}^{n} \ldots \sum_{i_n=1}^{n} T^2(X_{i_1}, \ldots, X_{i_n}) - \left(\frac{1}{n^n} \sum_{i_1=1}^{n} \ldots \sum_{i_n=1}^{n} T(X_{i_1}, \ldots, X_{i_n}) \right)^2,$$

  что само по себе вызывает отвращение, не говоря уже о необходимости считать порядка n^n слагаемых.

Бутстреп избавляет нас от этих проблем: Монте-Карло позволяет не вычислять какие-либо интегралы, приближая их многократным семплированием, а замена P на $\widehat{\mathsf{P}}$ даёт возможность это семплирование осуществить. Действительно, как нам вообще смоделировать случайную величину X с распределением $\widehat{F}_n$? Так как сие распределение дискретно и придаёт каждому элементу выборки $X_1, \ldots, X_n$ вероятностную массу $1/n$, то достаточно просто равновероятно выбрать какой-нибудь элемент выборки, то есть если случайный индекс j берётся равномерно из множества $\{1, \ldots, n\}$, то величина X_j будет иметь искомое распределение.

Итого, алгоритм нахождения оценки дисперсии $v_{\text{boot}}(T_n)$ таков:

- Для каждого $i = 1, \ldots, B$, где B достаточно велико:
  - Выбираем случайные индексы $j_1, \ldots, j_n$ равновероятно и независимо из множества $\{1, \ldots, n\}$;
  - Получаем выборку $X^{*(i)} = (X_1^{*(i)}, \ldots, X_n^{*(i)})$, где $X_k^{*(i)} = X_{j_k} \sim \widehat{\mathsf{P}}$.
- По сгенерированным выборкам считаем статистики $T_{n,i}^* = T_n\big(X^{*(i)}\big)$;
- В качестве оценки дисперсии берём

$$v_{\text{boot}}(T_n) = \frac{1}{B} \sum_{i=1}^{B} \left(T_{n,i}^* - \overline{T_n^*} \right)^2.$$

Данная процедура на самом деле предлагает нам нечто большее: на втором шаге мы получаем целую выборку $\mathbf{T}_n^* = (T_{n,1}^*, \ldots, T_{n,B}^*)$ из распределения, которое в силу предложения о близости F и $\widehat{F}_n$ само близко к истинному распределению $T_n(\mathbf{X})$. Отсюда мы можем получить оценки для куда более широкого класса характеристик, например, квантилей, коэффициентов асимметрии/эксцесса и т.д.

Впрочем не всегда бутстреповское распределение так уж хорошо приближает настоящее. Стоит помнить, что эмпирическое распределение $\widehat{F}$ сильно отличается от истинного F своей дискретностью и конечностью носителя, поэтому некоторые оценки могут вести себя слишком предсказуемо при сэмплировании из $\widehat{F}$ и не давать никакой информации о её настоящем распределении.

Пример 17.1. Пусть $X_1, \ldots, X_n \sim F$ — данная нам свыше выборка, для которой нам хотелось бы оценить бутстрепом распределение максимума $X_{(n)}$. В отличие от, например, среднего, которое в каком-то смысле «смешивает» свои компоненты между собой, максимум выборки из дискретного распределения $\widehat{F}_n$ всегда равен какой-то точке из его носителя. Более того, распределение бутстреповского максимума, грубо говоря, не зависит от самой выборки: легко показать, что для фиксированного i выполнено $\mathsf{P}(X^*_{(n)} = X_{(n-i)}) \to e^{-i} - e^{-i-1}$. ■

17.2 Бутстрепные доверительные интервалы

Знание распределения оценки (в том числе асимптотического) помогало нам ранее строить доверительные интервалы для оцениваемых параметров. Сейчас, когда у нас появился мощный инструмент, дающий большее понимание в поведении оценок, мы можем ввести новые способы в построении доверительных интервалов, основанных на бутстреповской выборке оценок. Мы не будем углубляться в их состоятельность и корректность, больше деталей про них можно прочитать в [**?**, tibshirani1993introduction]

Задача 17.1. Для выборки $(X_1, \ldots, X_n)$, значения в которой будем предполагать различными, существует всего n^n возможных реализаций бутстреповской выборки. Однако почти все разумные статистики не используют порядок в выборке. Сколько же тогда реализаций будет, если порядок не учитывать?

Задача 17.2. Для некоторых статистик бутстреповскую дисперсию можно посчитать аналитически, не используя метод Монте-Карло (такие оценки ещё называют *идеальными бутстреповскими*). Посчитайте идеальную бутстреповскую оценку дисперсии выборочного среднего $\overline{g(\mathbf{X})}$.

Список литературы

[1] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[2] Jacob Cohen. The earth is round (p< .05). *American psychologist*, 49(12):997, 1994.

[3] Ralph B D'agostino, Albert Belanger, and Ralph B D'Agostino Jr. A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44(4):316–321, 1990.

[4] B Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7: 1–26, 1979.

[5] Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

[6] Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

[7] А. А. Боровков. *Математическая статистика*. Лань, 2010.

[8] А. И. Козбарь. *Прикладная математическая статистика. Для инженеров и научных работников*. М.: ФИЗМАТЛИТ, 2006.

[9] Гаральд Крамер. *Математические методы статистики*. Регулярная и хаотическая динамика, 2003.

[10] М. Б. Лагутин. *Наглядная математическая статистика: учеб. пособие. 5-е изд.(эл.)*. Москва: БИНОМ. Лаборатория знаний, 2015.

[11] А. Н. Ширяев. *Вероятность*. МЦНМО, 2021.