Глава 3

Занятие третье. О параметрическом оценивании, о методах известных и не очень и о том, что такое бутстрэп

Мы привыкли рассматривать параметрическую модель $X_i \sim F_\theta$, где F — какое-то заданное семейство функций распределения, причем обычно довольно узкое — например, все нормальные распределения.

Будем обозначать $X_1, ..., X_n$ выборку, то есть сами случайные величины, а $x_1, ..., x_n$ — реализацию, то есть принятые ими в эксперименте значения.

3.1 Оценки и их свойства

Оценкой мы называем измеримую функцию $\widehat{\theta}(X_1, ..., X_n)$, которая приближает параметр θ или функцию $g(\theta)$ от него.

3.1.1 Свойства оценок

Важными свойствами оценок для нас были:

- Несмещенность $\mathbf{E}_\theta \widehat{\theta}(X_1, ..., X_n) = \theta$ при всех $\theta \in \Theta$.
- Состоятельность: $\widehat{\theta}(X_1, ..., X_n) \xrightarrow{P_\theta} \theta$ при всех $\theta \in \Theta$.
- Асимптотическая нормальность: $\sqrt{n}(\widehat{\theta}(X_1, ..., X_n) - \theta) \xrightarrow{d} Z \sim \mathcal{N}(0, \sigma^2(\theta))$. Величину $\sigma^2(\theta)$ называют асимптотической дисперсией. Вместо $\sqrt{n}$ иногда рассматривают другие скорости сходимости.

Если наша цель была оценить не θ, а какую-то функцию $g(\theta)$, то в правую часть 1)-3) мы будем ставить $g(\theta)$.

Свойство несмещенности нам важно, если мы повторяем опыт, тогда мы можем гарантировать себе, что у нас не будет систематического занижения или завышения оценки. Свойство состоятельности необходимо, если мы имеем возможность наращивать количество испытаний, свойство асимптотической нормальности уточняет состоятельность. Непосредственно нормальность предельного распределения дает достаточно распространенный класс предельных распределений и некоторые удобные свойства, в частности функциональную инвариантность, о которой мы сейчас поговорим.

3.1.2 Функциональная инвариантность

Отметим несколько важных для нас замечаний, связанных с так называемой функциональной инвариантностью оценок. Если оценка $\widehat{\theta}$ обладает каким-то свойством как оценка $g(\theta)$, то обладает ли $h(\widehat{\theta})$ тем же свойством как оценка $h(g(\theta))$.

- Если оценка $\widehat{\theta}(X_1, ..., X_n)$ несмещенная для функции $g(\theta)$, то $h(\widehat{\theta}(X_1, ..., X_n))$ не обязана быть несмещенной оценкой для $h(g(\theta))$. Более того, если, скажем, h строго выпукла вверх или вниз, а $\widehat{\theta}$ не константа, то это заведомо не так.
- Если оценка $\widehat{\theta}(X_1, ..., X_n)$ состоятельная для функции $g(\theta)$, а h — непрерывная функция, то $h(\widehat{\theta}(X_1, ..., X_n))$ состоятельная оценка $h(g(\theta))$.
- Если оценка $\widehat{\theta}(X_1, ..., X_n)$ асимптотически нормальная для функции $g(\theta)$ с асимптотической дисперсией $\sigma^2(\theta)$, а h — дифференцируемая функция, то $h(\widehat{\theta}(X_1, ..., X_n))$ будет асимптотически нормальной оценкой для $h(g(\theta))$ с дисперсией $(\sigma(\theta)h'(g(\theta)))^2$.

Последнее утверждение называют Delta method (в русскоязычной литературе также встречается название "Лемма об асимптотической нормальности"). Этот факт мы сформулировали для одномерной функции h, аналогичный факт верен и в случае векторной функции:

Теорема 1. *Если векторная оценка $\widehat{\theta}(X_1, ..., X_n)$ асимптотически нормальная для функции $g(\theta)$, $g : \Theta \to \mathbb{R}^k$ с асимптотической ковариацией $\Sigma(\theta)$, то есть*

$$\sqrt{n}\left(\widehat{\theta}(X_1, ..., X_n) - g(\theta)\right) \xrightarrow{d} Z \sim \mathcal{N}(\vec{0}, \Sigma(\theta)),\ n \to \infty.$$

Пусть отображение $h : \mathbb{R}^k \to \mathbb{R}^m$ дифференцируемо в точке $g(\theta)$ и имеет матрицу Якоби

$$J(t_1, ..., t_m) = \left(\frac{\partial h_i(t_1, ..., t_m)}{\partial t_j},\ i \le k,\ j \le m\right),$$

где $h(t_1, ..., t_m) = (h_1(t_1, ..., t_m), ..., t_k(t_1, ..., t_m))$. Тогда

$$\sqrt{n}\left(h(\widehat{\theta}(X_1, ..., X_n)) - h(g(\theta))\right) \xrightarrow{d} Z \sim \mathcal{N}(\vec{0}, J(g(\theta))\Sigma(\theta)J^{\mathrm{t}}(g(\theta))),\ n \to \infty,$$

то есть асимптотическая нормальность сохраняется при действии дифференцируемых отображений.

В векторной форме функциональная инвариантность верна и для состоятельности. В частности, отсюда вытекает и инвариантность около базовых операций, например, суммирования, поскольку $x + y$ — дифференцируемая функция двух переменных. Для асимптотически нормальных оценок это также верно, но с оговоркой — требуется не просто асимптотическая нормальность каждого из слагаемых, но нормальность векторной оценки, включающей оба слагаемых.

Три упомянутых свойства оценок далеко не единственные: нас может интересовать эффективность или оптимальность оценок (то есть минимальность их дисперсии в классе несмещенных оценок), робастность (то есть устойчивость к выбросам в выборке) и другие свойства.

3.2 Методы построения оценок

Два базовых метода для построения оценок, которые вы рассматривали в курсе статистики— метод моментов и метод максимального правдоподобия. Напомним как они устроены, обсудим их качества и добавим еще один метод, называемый методом спэйсингов.

3.2.1 Метод моментов

Метод моментов базируется на том, что в силу ЗБЧ и ЦПТ выборочные средние $\overline{X^k} = \frac{X_1^k + ... + X_n^k}{n}$— хорошие оценки для $\mathbf{E}_\theta X_1^k$, а именно состоятельные и при $\mathbf{E}_\theta X_1^{2k} < \infty$ асимптотически нормальные с асимптотической дисперсией $\mathbf{D}_\theta X_1^k$. С помощью функциональной инвариантности мы можем попытаться сделать из них соответствующие оценки для $\vec{\theta}$.

Для оценки параметра $\vec{\theta} = (\theta_1, ..., \theta_k)$ предлагается рассмотреть $\overline{X^1}$,...,$\overline{X^k}$ как оценки $\mu_1(\theta) = \mathbf{E}_\theta X$,...,$\mu_k(\theta) = \mathbf{E}_\theta X^k$ и применить к ним такое отображение f, которое переводит $(\mu_1(\theta), ..., \mu_k(\theta))$ в $\vec{\theta}$. Если это отображение окажется непрерывным, то $f(\overline{X}, ..., \overline{X^k})$ будет состоятельной для $\vec{\theta}$, а если дифференцируемым, то асимптотически нормальным (если, конечно $\mathbf{E}_\theta X^{2k} < \infty$).

Иначе говоря, чтобы найти оценку методом моментов (ОММ), мы должны решить систему уравнений

$$\begin{cases} \mu_1(\theta_1, ..., \theta_k) = \overline{X^1}, \\ \mu_2(\theta_1, ..., \theta_k) = \overline{X^2}, \\ \qquad ... \\ \mu_k(\theta_1, ..., \theta_k) = \overline{X^k}. \end{cases}$$

Если система оказалась несовместной, то мы можем убрать часть уравнений, заменив их на соотношения на следующие моменты.

К сожалению, метод моментов не вполне удачен в плане асимптотической дисперсии, она зачастую бывает весьма большой, особенно в случае многомерных параметров. Привязка непосредственно к моментам ограничивает возможности метода. Зачастую используют модификацию, основанную на так называемых *пробных функциях*. В этом случае рассматриваются моменты $\overline{g_i(X)}$, $i = 1, ..., k$ и рассматриваются уравнения

$$\mathbf{E}_{\theta_1,...,\theta_k} g_i(X_1) = \overline{g_i(X)}, \ i = 1, ..., k.$$

3.2.2 Метод максимального правдоподобия

Этот метод исходит из простого соображения — мы должны искать такое θ, при котором появление нашей выборки особенно вероятно. Это равносильно максимизации совместного распределения (в случае дискретной выборки) или совместной плотности (в случае абсолютно-непрерывной плотности). Таким образом, метод максимального правдоподобия предписывает искать θ, такие что

$$L(x_1, ..., x_n; \theta) = f_\theta(x_1)...f_\theta(x_n) \to \max,$$

где $f_\theta(x)$ — плотность X_i или вероятность $\mathbf{P}_\theta(X = x)$. Практическую зачастую удобнее искать аргумент максимума $\ln L(x_1, ..., x_n; \theta)$.

В некоторых случаях ОМП может быть неединственной или ее не будет совсем. Приведем некоторые случаи, когда у ОМП возникают проблемы:

1. Если распределение X_i устроено так, что при разных параметрах θ величины принимают значения из разных множеств, то есть носитель распределения зависит от параметра.

   Вопрос 1. Что будет ОМП в случае $X_i \sim R[\theta, \theta + 1]$?

2. Если плотность распределения X_i не гладко зависит от параметра.

   Вопрос 2. Что будет ОМП в случае X_i с плотностью $\exp(-|x - \theta|)/2$?

   Вопрос 3. Что будет с ОМП, если X_i бернуллиевские с параметром θ при $\theta \in \mathbb{Q} \cap (0, 1)$ и $1 - \theta$ иначе?

3. Если модель содержит большое число параметров.

   Вопрос 4. Что будет с ОМП, если $(X_{2i}, X_{2i-1}) \sim \mathcal{N}(\theta_i, \sigma^2)$. Будет ли ОМП для σ^2 состоятельной?

4. Если множество изменения параметра не является замкутым, а плотность может неограниченно расти при определенных значениях аргумента и приближении параметра к границе области изменения.
   Такого рода ситуация может привести к тому, что ОМП будет отсутствовать, поскольку глобальный максимум равен бесконечности. Подробно такой пример будет рассмотрен в следующем параграфе. Однако, ситуация может оказаться еще более коварной — скажем, максимум правдоподобия существует, и ОМП даже сходится к какому-то значению, однако вовсе не совпадающему со значением

параметра. Суть проблемы в том, что оценка максимального правдоподобия рассчитывает на то, что правдоподобие достигает максимального значения засчет того что основная масса наблюдений при таком значении параметра имеет достаточно большую плотность. Может, однако, оказаться что максимум достигается за счет огромных значений плотности в одном из элементов реализации, хотя во всех остальных точках значения плотности достаточно маленькие. Пример такого рода случая рассмотрен в задаче ниже.

Мы указали столько минусов, что может возникнуть вопрос: "Зачем вообще нужна такая оценка?" Однако, не все так плохо — зависимость носителя от параметра чаще всего не так пагубно влияет на оценку (так в примере из первого пункта ОМП не единственна, но все оценки обладают неплохими свойствами, в частности состоятельны и имеют достаточно маленькую дисперсию), отсутствие гладкости в соответствующем примере с распределением Лапласа также не приводит к качественным проблемам, хотя оценка опять же перестает быть единственной. Пример со схемой Бернулли очевидно достаточно искусственный, а в примере из пункта 3) мы работаем с $n+1$ параметром при $2n$ наблюдениях (к слову, если перейти к $X_{2i} - X_{2i-1}$ и рассмотреть соответствующую модель, то ОМП сразу же станет отличной оценкой). Наконец в ситуации последнего пункта можно рассматривать локальные максимумы, некоторые из которых будут близки к искомому параметру.

Зато в случае достаточно гладко зависящих от θ распределений, замкнутого множества изменения параметра и независимости носителя от параметра ОМП оказываются совершенно выдающимися оценками — состоятельными, асимптотически нормальными, да еще и с наилучшей возможной асимптотической дисперсией среди оценок с непрерывной асимптотической дисперсией.

Вопрос 5. Рассмотрим $\widehat{\theta}_n = \overline{X}(1 + I_{|\overline{X}|>n^{2/3}})/2$, $X_i \sim \mathcal{N}(\theta, 1)$. Показать, что эта оценка асимптотически нормальна и ее асимптотическая дисперсия равна дисперсии ОМП при $\theta \neq 0$ и меньше ее при $\theta = 0$.

Более конкретно, для сильно регулярных моделей (условия сильной регулярности могут иметь различный вид, одна из версий приведена в конце файла) ОМП удовлетворяет соотношению

$$\sqrt{n}(\widehat{\theta}(X_1, ..., X_n) - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right),$$

где $I(\theta) = \mathbf{E}\left(\frac{\partial}{\partial\theta} \ln f_\theta(X_1)\right)^2$ — информация Фишера.

Обе рассмотренных оценки являются эквивариантными, то есть если $\widehat{\theta}$ — ОМП или ОММ для θ, то $f(\widehat{\theta})$ — ОМП или ОММ для $f(\theta)$. Это удобное свойство, которое, к сожалению, плохо сочетается с несмещенностью.

Задача 1. Построить алгоритм, численно вычисляющий по однопараметрической функции плотности ОММ и ОМП.
Для решения уравнений пригодится функция uniroot, для интегрирования — функция integrate (лучше интегрировать не по всей прямой, а по конечному отрезку), а для минимизации — nlm или optimize. Функция nlm подходит и для векторного случая, также минимизацию можно осуществлять optim, а вот uniroot в многомерном случае заменяется на polyroot из библиотеки rootsolve)

Реализованная оценка ОМП с помощью optim находится в пакете stats4, эта функция носит название mle, ее аргументом является логарифмическая функция правдоподобия с обратным знаком.

Как и обещали в конце раздела приводим вам пример однопараметрического распределения, на котором ОМП работает достаточно неудачно, в частности, несостоятельна. Причина в том, что глобальный максимум достигается засчет взрывного роста плотности в окрестности одного из наблюдений при определенном значении параметра.

Задача 2. Моделировать выборку с плотностью

$$f_\theta(x) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) + \frac{1}{2\sqrt{2\pi}\sigma(\theta)} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2(\theta)}\right),\ \sigma^2(\theta) = \exp\left(-\frac{1}{\theta^2}\right).$$

Это распределение представляет смесь двух нормальных выборок — величина равновероятно выбирается либо из $\mathcal{N}(0, 1)$, либо из $\mathcal{N}(\theta, \sigma^2(\theta))$. Построить график логарифма правдоподобия и посмотрите чему оно равно при параметре, равном минимальной по модулю точке выборки.

3.2.3 Метод спейсингов

Метод спейсингов (method of maximal spacing) предлагает для оценивания параметра θ семейства распределений с плотностями, сосредоточенными на отрезке $[a, b]$, рассмотреть

$$S(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} D_i(\theta),$$

где $D_i(\theta) = \mathbf{P}_\theta(X \in [x_{(i)}, x_{(i+1)}]) = F_\theta(x_{(i+1)}) - F_\theta(x_{(i)})$, $x_{(0)} = a$, $x_{(n+1)} = b$, где $x_{(i)}$ — вариационный ряд. Тогда θ, максимизирующее S, называют оценкой методом спейсингов.

Логика довольно проста — при правильном θ $Y_i = F_\theta(X_{(i)})$ есть вариационный ряд равномерного распределения. Тогда $Y_i - Y_{i-1}$ одинаково распределены. Но $\max_{y_1+...+y_{n+1}}(y_1...y_{n+1})$ достигается при $y_i = 1/(n+1)$. Следовательно, максимизация S будет связана с выравниванием $D_i(\theta)$, что соответствует искомому θ.

Эта оценка состоятельна, а в случае регулярных оценок асимптотически эффективна. В регулярных моделях она достаточно близка к ОМП, но при этом зачастую избегает ее проблем в нерегулярных. Конечно, она неприятна в вычислительном плане, но численно ее поиск не представляет таких уж проблем.

Вопрос 6. Какая оценка методом спейсингов для равномерного распределения $R[\theta_1, \theta_2]$?

Задача 3. Пусть $F_{\theta_1,\theta_2}(x) = 1 - e^{-(x-\theta_1)^{\theta_2}}$, $x > \theta_1$. Построить численно оценку методом спейсингов. Сравнить ее с ОМП.

3.3 EM-алгоритм или что делать, когда аналитически к ОМП не подобраться

Этот раздел обязателен только тем, кто собирается сдавать тему на сложном уровне.

3.3.1 Отсутствие ОМП для смеси нормальных распределений

Вычисление ОМП может быть достаточно затруднительным или вообще невозможным. Рассмотрим, например, следующую задачу:

Пример 1. Пусть X_i являются смесью двух нормальных распределений, то есть каждое наблюдение с вероятностью p получается из распределения $\mathcal{N}(\mu_1, \sigma_1^2)$, а иначе из $\mathcal{N}(\mu_2, \sigma_2^2)$. Тогда правдоподобие имеет вид

$$L(x_1, ..., x_n; p, \mu_1, \mu_2, \sigma_1, \sigma_2) = f_{p,\mu_1,\mu_2,\sigma_1,\sigma_2}(x_1)...f_{p,\mu_1,\mu_2,\sigma_1,\sigma_2}(x_1) = \prod_{i=1}^{n} \left(\frac{p}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \frac{1-p}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \right).$$

Эта функция выглядит крайне громоздкой и максимизация ее выглядит сложной задачей. В действительности все еще хуже, если положить $p = 1/2$, $\mu_1 = x_1$, $\mu_2 = 0$, $\sigma_2 = 1$ и устремить $\sigma_1 \to 0$, то правдоподобие можно сделать сколь угодно большим.

Вопрос 7. Докажите этот факт.

Однако, даже если $\sigma_1 = 1$, $\sigma_2 = 1$, $p = 1/2$ и правдоподобие ограничено, найти его максимум будет затруднительно.

3.3.2 EM-алгоритм

Как уже говорилось выше, в этом случае спасают локальные максимумы. Предложим метод, который может довольно эффективно искать локальные максимумы в такого рода сложных задачах, называемый EM-алгоритмом (по двум шагам алгоритма — Expectation и Maximization).

Пусть мы хотим максимизировать правдоподобие $L(x_1, ..., x_n; \vec{\theta})$. Предположим, что наша задача значительно упростилась бы, если бы вместе с выборкой $x_1, ..., x_n$ мы знали некоторые скрытые наблюдения $y_1, ..., y_m$. Так для задачи из прошлого примера мы легко решили бы задачу, если бы мы знали $y_1, ..., y_n \in \{1, 2\}$, отвечающие за то к какому из нормальных распределений относится каждый элемент. Тогда EM-алгоритм предлагает следующую программу действий:

1. Выберем начальное значение параметров $\vec{\theta}^0 = (\theta_1^0, ..., \theta_k^0)$ в нашей модели.
   На $j+1$-ом шаге проделаем следующее:
2. Expectation. Вычислим
$$J(\vec{\theta}|\vec{\theta}^j) = \mathbf{E}_{\vec{\theta}^j} \left(\ln \frac{L(X_1, ..., X_n, Y_1, ..., Y_m; \vec{\theta})}{L(X_1, ..., X_n, Y_1, ..., Y_m; \vec{\theta}^j)} \middle| X_1 = x_1, ..., X_n = x_n \right).$$
3. Maximization. Выберем $\vec{\theta}_{j+1}$ так, что $J(\vec{\theta}^{j+1}|\vec{\theta}^j)$ максимально.

Можно доказать, что на каждом шаге функция правдоподобия не уменьшается.

Бывает удобно слегка переформулировать процедуру:

1. Вычислим распределение $y_1, ...y_m$ в дискретном случае
$$\mathbf{P}_{\vec{\theta}^j}(Y_1 = y_1, ..., Y_m = y_m | X_1 = x_1, ..., X_n = x_n) = \frac{L(x_1, ..., x_n, y_1, ..., y_m; \vec{\theta}^j)}{\mathbf{E}_{\vec{\theta}^j} L(x_1, ..., x_n, Y_1, ..., Y_m; \vec{\theta}^j)}$$
   или плотность $Y_1, ..., Y_m$ в непрерывном случае
$$f_{\vec{\theta}^j}(y_1, ..., y_m | X_1 = x_1, ..., X_n = x_n) = \frac{L(x_1, ..., x_n, y_1, ..., y_m; \vec{\theta}^j)}{\mathbf{E}_{\vec{\theta}^j} L(x_1, ..., x_n, Y_1, ..., Y_m; \vec{\theta}^j)}.$$
   Такое распределение называется апостериорным для $Y_1, ..., Y_m$ при условии $X_1, ..., X_n$.
2. Максимизируем по θ
$$J(\vec{\theta}|\vec{\theta}^j) = \sum_{y_1, ..., y_m} \mathbf{P}_{\vec{\theta}^j}(Y_1 = y_1, ..., Y_m = y_m | X_1 = x_1, ..., X_n = x_n) \ln L(x_1, ..., x_n, y_1, ..., y_m; \vec{\theta})$$
   в дискретном или
$$J(\vec{\theta}|\vec{\theta}^j) = \int_{\mathbb{R}^m} f_{\vec{\theta}^j}(y_1, ..., y_m | X_1 = x_1, ..., X_n = x_n) \ln L(x_1, ..., x_n, y_1, ..., y_m; \vec{\theta}) dy_1 ... dy_m$$
   в абсолютно-непрерывном. Иначе говоря,
$$J(\vec{\theta}|\vec{\theta}^j) = \mathbf{E} \ln L(x_1, ..., x_n, Y_1, ..., Y_n; \theta),$$
   взятое по апостериорному распределению $\widetilde{\mathbf{P}}(A) = \mathbf{P}_{\vec{\theta}^j}((Y_1, ..., Y_n) \in A | X_1 = x_1, ..., X_n = x_n)$.

Во второй форме видно, чем EM-алгоритм облегчает нашу задачу. Максимизация правдоподобия соответствует шагам 1'), 2'), где на первом шаге используется не известное $\vec{\theta}^j$, а неизвестное $\vec{\theta}$, по которому и ведется максимизация. Мы же упрощаем задачу, подставляя на первом шаге текущую оценку параметра $\vec{\theta}$ и улучшая ее последовательными итерациями.

Улучшая полученные оценки $\vec{\theta}^j$, мы постепенно будем приближаться к точке локального (!) максимума правдоподобия, либо, если такой точки нет, уходить в бесконечность. При этом даже в условиях прошлого примера мы вполне можем получить какую-то оценку, поскольку можем попасть в один из локальных экстремумов. С другой стороны, мы совершенно не гарантируем, что этот экстремум будет близок к оцениваемому параметру, да и то, какой именно получится экстремум, зависит от начальной оценки для наших параметров.

3.3.3 EM-алгоритм для смеси нормальных распределений

Давайте применим полученный метод к уже упоминавшейся задаче о смеси нормальных:
Пусть $X_1, ..., X_n$ — наши величины, полученные из $\mathcal{N}(\mu_1, 1)$ или $\mathcal{N}(\mu_2, 1)$ с вероятностью $p = 0.5$, $\vec{\theta} = (\mu_1, \mu_2)$, $Y_1, ..., Y_n$ — величины, отвечающие за то, из какого распределения получены величины X_i. Тогда

$$L(x_1, ..., x_n, y_1, ..., y_n; \mu_1, \mu_2) = \frac{1}{2^n} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_{y_i})^2}.$$

Воспользуемся формулой

$$J(\vec{\theta}|\vec{\theta^j}) = \mathbf{E}_{\vec{\theta}} \ln L(x_1, ..., x_n, Y_1, ..., Y_m) = -n \ln 2 - \frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \mathbf{E}_{\widetilde{\mathbf{P}}}(x_i - \mu_{Y_i})^2,$$

где математическое ожидание, как уже упоминалось, берется по мере $\widetilde{\mathbf{P}}(A)$, заданной распределением Y при условии X, то есть апостериорным распределением Y при условии X. Давайте найдем его по формуле Байеса

$$p_i^j = \widetilde{\mathbf{P}}(Y_i = 1) = \mathbf{P}_{\vec{\theta^j}}(Y_i = 1 | X_1 = x_1, ..., X_n = x_n) = \frac{\mathbf{P}(Y_i = 1) f_{X_1}(x_1) ... f_{X_{i-1}}(x_{i-1}) f_{X_i}(x_i|Y_i = 1) f_{X_{i+1}}(x_{i+1}) ... f_{X_n}(x_n)}{\prod_{j \neq i} f_{X_i}(x_i) \left(\mathbf{P}_{\vec{\theta^j}}(Y_i = 1) f_{X_i}(x_i|Y_i = 1) + \mathbf{P}_{\vec{\theta^j}}(Y_i = 0) f_{X_i}(x_i|Y_i = 0) \right)} = \frac{\exp(-\frac{1}{2}(x_i - \mu_1^j)^2)}{\exp(-\frac{1}{2}(x_i - \mu_1^j)^2) + \exp(-\frac{1}{2}(x_i - \mu_0^j)^2)}. \tag{3.1}$$

Значит

$$\mathbf{E}_{\widetilde{\mathbf{P}}}(x - \mu_{Y_i})^2 = (x - \mu_1)^2 \widetilde{\mathbf{P}}(Y_i = 1) + (x - \mu_2)^2 \widetilde{\mathbf{P}}(Y_i = 2).$$

Отсюда максимизация $J(\vec{\theta}|\vec{\theta^j})$ сводится к минимизации

$$\sum_{i=1}^{n} (p_i^j (x_i - \mu_1)^2 + (1 - p_i^j)(x_i - \mu_2)^2).$$

Дифференцируя, получаем

$$\mu_1^{j+1} = \frac{\sum_{i=1}^{n} p_i x_i}{\sum_{i=1}^{n} p_i}, \quad \mu_2^{j+1} = \frac{\sum_{i=1}^{n} (1 - p_i) x_i}{\sum_{i=1}^{n} (1 - p_i)}. \tag{3.2}$$

Итак, наш алгоритм готов:

1. Находим p_i^j на основе μ_i^j по формуле (3.1).
2. Находим μ_i^{j+1} по формуле (3.2).
3. Переходим на следующий шаг и повторяем процедуру.

Задача 4. Протестировать метод для $\mu_1 = 0$, $\mu_2 = 1$.

EM-алгоритм для нормальных данных реализован в R в пакетах EMCluster или mclust, в Python в модуле sklearn в пакете mixture. Примеры программ можно найти в разделе практикума.

3.3.4 EM-алгоритм для неполных данных

EM-алгоритм также крайне удобен, если у нас неполные данные.

Пусть $Z_1, ..., Z_n$ из полиномиального распределения. Напомню, что полиномиальное распределение соответствует эксперименту, в котором проводится m испытаний, в каждом из которых n возможных итогов с вероятностями τ_1,...,τ_n, $Z_1, ..., Z_n$ при этом будет означать количество исходов каждого из видов.

Представим себе, что $\beta \in \mathbb{R}^+$, $m \in \mathbb{N}$, $\tau_i \in [0,1]$: $\tau_1 + ... + \tau_n = 1$, X_i — пуассоновские с параметром $m\beta\tau_i$, а $(Z_1, ..., Z_n)$ — полиномиальный вектор с вероятностями $\tau_1, ..., \tau_n$ и числом испытаний m. При этом наблюдение Z_1 отсутствует среди наших наблюдений, параметры β, τ_i нам неизвестны. Мы хотим оценить β, τ_i на основе данных $(X_1, ..., X_n)$, $(Z_2, ..., Z_n)$. Правдоподобие при известном z_1 имело бы вид

$$L(x_1, ..., x_n, z_1, z_2, ..., z_n; \beta, \tau_i) = \frac{(z_1 + ... + z_n)!}{z_1!...z_n!}\tau_1^{z_1}...\tau_n^{z_n}\prod_{i=1}^{n} e^{-x_i}\frac{(m\beta\tau_i)^{x_i}}{x_i!}e^{-m\beta\tau_i}$$

При неизвестном z_1 задача максимизации сложна. Возьмем в качестве Y_1 неизвестную величину Z_1. Теперь мы можем применить EM-алгоритм (опять же во второй форме — вычисляя апостериорное распределение z_1 при известных параметрах $\vec{\theta^j}$).

Задача 5. Применить EM-алгоритм к этой задаче и реализовать его на R или Python, полагая $n = 10$, $\tau_i = 1/10$,$\beta = 0.25$, $m = 60$. Сгенерировать полиномиальную выборку с равными вероятностями исходов в R можно с помощью функции sample: sample(1:n, m, replace=TRUE). К слову, можно аналогично генерировать и неравновероятную полиномиальную выборку, указывая среди параметров prob = p, где p — вектор вероятностей. В Python аналогичный трюк можно проделать, используя random.choice из numpy, где параметр p, опять-таки, отвечает вероятностям исходов.

3.4 Доверительные беседы о доверительных интервалах и множествах

3.4.1 Доверительные интервалы

Будем рассматривать одномерные параметры $\theta \in \mathbb{R}$.

Напомним, что доверительным интервалом называют пару статистик $\widehat{\theta}_1$, $\widehat{\theta}_2$, таких, что при всех θ выполнено равенство

$$\mathbf{P}_\theta(\theta \in (\widehat{\theta}_1, \widehat{\theta}_2)) = 1 - \alpha$$

где $1 - \alpha$ — заданное число, называемое уровнем доверия. Можно требовать взамен выполнение неравенства $\geq$ или сходимости $\to$ при $n \to \infty$ (в последнем случае интервал называется асимптотическим).

Основным методом для их построения является метод центральной функции:

- выбираем некоторую статистику T, например, достаточную;
- находим некоторую функцию $g(T(X_1, ..., X_n); \theta)$, монотонную по переменной θ и имеющую распределение, независящее от параметра;
- находим отрезок [a,b], в котором $g(T(X_1, ..., X_n); \theta)$ лежит с вероятностью $1 - \alpha$;
- обращаем функцию g по второй переменной и находим $\widehat{\theta}_1 = g^{-1}(T(x_1, ..., x_n); \cdot)(a)$, $\widehat{\theta}_2 = g^{-1}(T(x_1, ..., x_n); \cdot)(b)$.

Полученные статистики и есть границы интервала. Для нахождения указанных a, b используют так называемые квантили, α-квантилью распределения F называют $F^{-1}(\alpha)$.

3.4.2 Асимптотические доверительные интервалы

Построение точных доверительных интервалов достаточно трудоемко, а вот асимптотические доверительные интервалы можно строить на основе любой асимптотически нормальной оценки. Так если $\widehat{\theta}$ — ОМП для θ в сильно регулярной модели, то ее асимптотическая дисперсия есть $1/I(\theta)$ — величина, обратная к информации Фишера, которую можно состоятельно оценить величиной $I(\widehat{\theta})$. Из асимптотической нормальности ОМП

$$\sqrt{n}\frac{\widehat{\theta} - \theta}{1/\sqrt{I(\theta)}} \xrightarrow{d} Z \sim \mathcal{N}(0, 1).$$

Но, в силу состоятельности $I(\widehat{\theta})$ как оценки $I(\theta)$

$$\sqrt{n}\frac{\widehat{\theta}-\theta}{1/\sqrt{I(\widehat{\theta})}} = \sqrt{n}\frac{\widehat{\theta}-\theta}{1/\sqrt{I(\theta)}}\sqrt{\frac{I(\widehat{\theta})}{I(\theta)}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

Тогда

$$P\left(\sqrt{n}\frac{\widehat{\theta}-\theta}{1/\sqrt{I(\widehat{\theta})}} \leq x\right) \to \Phi(x),$$

откуда мы можем построить доверительный интервал вида

$$\left(\widehat{\theta} - \frac{z_\beta}{\sqrt{n}\sqrt{I(\widehat{\theta})}}, \widehat{\theta} - \frac{z_\gamma}{\sqrt{n}\sqrt{I(\widehat{\theta})}}\right),$$

$\beta - \gamma = 1 - \alpha$.

Может показаться, что для оценки методом моментов доверительный интервал строится сложнее. Но здесь нам на помощь приходит дельта-метод. Асимптотическую дисперсию оценки $\overline{X^k}$ как оценки $\mu_k(\theta)$ мы знаем — это $\mathbf{D}_\theta X_k = \mu_{2k}(\theta) - \mu_k^2(\theta)$. Следовательно, если ОММ для θ есть $f(\overline{X^k})$, то ее асимптотическая дисперсия $\sigma^2(\theta) = (\mu_{2k}(\theta) - \mu_k^2(\theta))(f'(\mu_k(\theta)))^2$, то есть доверительный интервал будет иметь вид

$$\left(f(\overline{X^k}) - \frac{z_\beta \sigma(f(\overline{X^k}))}{\sqrt{n}}, f(\overline{X^k}) - \frac{z_\gamma \sigma(f(\overline{X^k}))}{\sqrt{n}}\right)$$

3.4.3 Доверительные множества или эллипс против прямоугольника

Этот раздел обязателен только тем, кто собирается сдавать тему на сложном уровне.
Для векторного параметра $\theta = (\theta_1, ..., \theta_k)$ нам уже не очень помогают доверительные интервалы для каждой θ_i, поскольку знание того, что каждая из θ_1,...,θ_k лежит в своем интервалом с заданной вероятностью не дает нам возможность выписать вероятность того, что все параметры одновременно попадут в соответствующий параллелипипед. Можно было бы, конечно, выделить по каждой оси $1 - \alpha/k$ интервал, тогда вероятность того, что хоть одна из θ_i не попадет в свой интервал, будет не больше α. Но этот параллелипипед будет слишком большим и в ряде случаев может быть значительно улучшен. Поэтому рассматривают доверительное множество $A(X_1, ..., X_n) \subset \mathbb{R}^k$, которое накроет мой параметр с вероятностью $1 - \alpha$, то есть

$$\mathbf{P}(\vec{\theta} \in A(X_1, ..., X_n)) = 1 - \alpha.$$

Для многомерной нормальной выборки $(X_1, ..., X_k) \sim \mathcal{N}(\vec{\mu}, \Sigma)$, мы можем утверждать, что $(\vec{X} - \vec{\mu})^{\mathrm{T}}\Sigma^{-1}(\vec{X} - \vec{\mu})$ будет вектором с распределением χ^2_k, поскольку $A^{-1}(\vec{X} - \vec{\mu}) \sim \mathcal{N}(0, E)$, где $A^{\mathrm{T}}A = \Sigma$. Но тогда мы можем сказать, что выборка попадает в эллипсоид с центром $\vec{\mu}$

$$(\vec{X} - \vec{\mu})^{\mathrm{T}}\Sigma^{-1}(\vec{X} - \vec{\mu}) \leq x$$

с вероятностью $F_{\chi^2_k}(x)$.

Вопрос 8. Показать, что это действительно так.

Таким образом, мы можем строить асимптотические доверительные множества на основе векторных асимптотически нормальных оценок. Для ОМП в роли асимптотической дисперсии $\Sigma(\theta)$ будет выступать $I^{-1}(\theta)$, где $I(\theta)$ — информационная матрица Фишера с элементами

$$cov\left(\frac{\partial}{\partial\theta_i}\ln f_\theta(X), \frac{\partial}{\partial\theta_j}\ln f_\theta(X)\right).$$

$$\sqrt{n}(\widehat{\theta} - \vec{\theta}) \xrightarrow{d} \mathcal{N}(0, \Sigma(\theta)),$$

откуда имеем доверительный эллипсоид для $\vec{\theta}$, заданный соотношением

$$(\widehat{\theta} - \vec{\theta})^{\mathrm{T}} I(\widehat{\theta})(\widehat{\theta} - \vec{\theta}) \leq y_{1-\alpha},$$

где y — квантиль χ^2_k.

Информационная матрица может быть записана как $-E_\theta G$, где G — матрица Гессе для логарифмической функции правдоподобия. Поэтому вычисление оценки $\Sigma(\widehat{\theta})$ может быть сделано с помощью подсчета гессиана при нахождении ОМП. При поиске ОМП с помощью nlm, укажем параметр hessian = TRUE. У выданной функцией nlm величины out параметр out$hessian будет содержать искомую оценку для матрицы Фишера.

Аналогично строится доверительное множество на основе ОММ, только для подсчета матрицы ковариации придется использовать многомерный Дельта-метод и ЦПТ для векторов. Более конкретно, если $\Sigma_1(\theta)$ — матрица ковариации вектора $(X, X^2, ..., X^k)$, g — отображение, такое что $g(\mu_1(\theta), ..., \mu_k(\theta)) = \theta$, а $J(\theta)$ — его матрица Якоби в точке $\mu_1(\theta), ..., \mu_k(\theta)$, то для $\widehat{\theta} = g(\overline{X}, \overline{X^2}, ..., \overline{X^k})$ выполнено соотношение

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} Z \sim \mathcal{N}(0, J^{\mathrm{T}}(\widehat{\theta})\Sigma_1(\widehat{\theta})J(\widehat{\theta})).$$

Соответственно, доверительный эллипсоид имеет вид

$$(\widehat{\theta} - \theta)^{\mathrm{T}} \Sigma^{-1}(\widehat{\theta})(\widehat{\theta} - \theta) \leq y_{1-\alpha}/n,$$

где $\Sigma = J\Sigma_1 J^{\mathrm{T}}$.

Задача 6. Построить асимптотический доверительный эллипс на основе ОМП для $X_i \sim \Gamma(\theta_1, \theta_2)$ на выборке размера 50 при $\theta_1 = 2$, $\theta_2 = 2$. Обратную матрицу к A можно построить с помощью функции solve(A).

Доверительные эллипсы нормального распределения удобно рисовать с помощью функции ellipse пакета mixtools. Так функция ellipse(mu,Sigma,alpha = .05, npoints = 50, newplot = TRUE, type = "l") строит доверительный эллипс уровня alpha для нормального $\mathcal{N}(mu, Sigma)$ распределения.

3.5 Бутстрэп или как вытаскивать себя из болота за тесемки сапогов

Построив ОМП, ОММ или оценку методом спейсингов, мы никак не гарантируем несмещенности. Но если мы часто повторяем эксперимент, нам может быть важно снизить смещение. Поэтому мы бы хотели научиться исправлять или снижать смещенность оценки.

3.5.1 Идея бутстрэпа

Идея метода бутстрэпа основывается на двух соображениях

- Если мы знаем функцию распределения какой-то выборки, то с помощью метода Монте-Карло мы можем приближенно найти математическое ожидание функций от наших величин, генерируя выборки из нашего распределения.
- Если у нас есть "хорошая оценка" $\widehat{\theta}$ неизвестного параметра θ, а семейство распределений непрерывно, то $F_{\widehat{\theta}}$ близко к F_θ

Пример 2. Итак, пусть оценка $\widehat{\theta}$ достаточно близка к θ, но обладает небольшим смещением

$$a(\theta) = \mathbf{E}_\theta \widehat{\theta}(X_1, ..., X_n) - \theta.$$

Пусть она на нашей реализации выборки приняла значение $\widehat{\theta}(x_1, ..., x_n)$. Тогда мы можем взять множество выборок $Y_{i,1}$,...,$Y_{i,n}$, $i = 1, ..., m$ из $F_{\widehat{\theta}(x_1,...,x_n)}$ и подсчитать по ним

$$\widetilde{\theta}(Y) = \sum_{i=1}^{m} \widehat{\theta}(Y_{i,1}, ..., Y_{i,n}).$$

Тогда смещение $\widetilde{\theta}(Y) - \widehat{\theta}(X_1, ..., X_n)$ при больших m близко к $a(\widehat{\theta})$. В свою очередь мы можем ожидать, что $a(\widehat{\theta})$ близко к $a(\theta)$. Таким образом, мы можем взять оценку

$$\widehat{\theta}(x_1, ..., x_n) - a(\widehat{\theta}) = 2\widehat{\theta}(x_1, ..., x_n) - \widehat{\theta}(Y)$$

и ожидать, что ее смещение значительно меньше чем было.

Метод бутстрэпа называется в честь ремешков на обуви в связи с идиомой, означающей "вытянуть себя из трясина за ремешки на обуви" (мы в таком случае вспоминаем косичку барона Мюнхгаузена). Ведь этот метод позволяет нам улучшить оценку с помощью самой этой оценки.

Задача 7. С помощью бутстрэппинга исправить смещение оценки S^2 для выборки размера 20 из $\mathcal{N}(0, 1)$.

Использовать метод бутстрэпа можно не только для оценивания смещения, но и, например, для оценки квадратичного смещения. Так мы могли бы взять $\frac{1}{m}\sum_{i=1}^{m}(\widehat{\theta}(Y_{i,1}, ..., Y_{i,n}) - \widehat{\theta}(X_1, ..., X_n))^2$ в качестве оценки дисперсии $\widehat{\theta}(X_1, ..., X_n)$.

Задача 8. Пусть известно $\overline{X}$ из некоторой выборки. Как с его помощью бутстрэппингом оценить дисперсию нашего распределения? Применить метод для $R[0, 1]$.

Этот метод имеет несколько неоспоримых плюсов — он прост в использовании и не требует вычислений, применим даже к весьма громоздким моделям. С другой стороны, мы не можем явным образом оценить его погрешность, а в случае, если оценка $\widehat{\theta}$ значимо промахнулась мимо θ, рискуем неправильно изменить оценку.

Вопрос 9. Предположим, что мы оценили среднее в модели $R[0, \theta]$ с помощью $\overline{X}$. Теперь мы берем новую выборку из $R[0, 2\overline{X}]$ и оцениваем с помощью ее среднего θ. Какую дисперсию будет иметь эта новая оценка?

3.5.2 Доверительные интервалы с помощью бутстрэпа

Бутстрэп дает свои варианты и для построения доверительного множества. Такого рода методов несколько, но мы рассмотрим наиболее простые. Рассмотрим оценку $\widehat{\theta}$ для θ. Будем брать выборки из $F_{\widehat{\theta}}$ и строить на основе них оценки $\widehat{\theta}_1$,...,$\widehat{\theta}_m$.

- Percentile интервал предлагает взять в качестве интервала для θ диапазон $(\widehat{\theta}_{([\gamma m])}, \widehat{\theta}_{([\beta m])})$ (то есть $[\gamma m]$-ое и $[\beta m]$-ое по возрастанию значения $\widehat{\theta}_i$), где $\beta - \gamma = 1 - \alpha$.
- Pivotal интервал.
  Рассмотрим $\Delta_i = \widehat{\theta}_i - \widehat{\theta}$. Мы ожидаем, что это выборка из величин, близких к $\Delta = \widehat{\theta} - \theta$. Если бы мы знали ф.р. F_Δ, то мы смогли бы построить интервал

  $$\left(F_\Delta^{-1}(\gamma), F_\Delta^{-1}(\beta)\right)$$

  для Δ, где $\beta - \gamma = 1 - \alpha$. Значит мы должны оценить $F_\Delta^{-1}(\gamma)$ и $F_\Delta^{-1}(\beta)$. Заметим, что $F_\Delta(x)$ оценивается величиной

  $$\frac{1}{m}\sum_{i=1}^{m} I_{\Delta_i \le x}.$$

  Обратная функция к такой функции в точке y — это $[ym]$-я по возрастанию точка из Δ_i. Вывод — упорядочим Δ_i и выберем те из них Δ_-, Δ_+, которые стоят на местах $[\gamma m]$ и $[\beta m]$ по возрастанию.

Тогда

$$(\widehat{\theta} - \Delta_+, \widehat{\theta} - \Delta_-)$$

и будет нашим интервалом.

Задача 9. Сравнить на выборках размера 50 для а) $\mathcal{N}(\theta, 1)$, б) $R[0, \theta]$ доверительные интервалы на основе ОММ, ОМП, бутстрэпа с помощью $\overline{X}$.

Оба интервала легко обобщаются на случай многомерного параметра: мы можем выбрать эллипс, содержащий $(1-\alpha)\cdot 100\%$ $\widehat{\theta}^*$ или $2\widehat{\theta} - \widehat{\theta}^*$.

В R удобно строить такие выборочные доверительные эллипсы с помощью data.ellipse(x, y, levels=0.9), где x, y задают массивы точек, a levels — уровень.

3.5.3 Замечания о надежности метода

Несмотря на простую форму, бутстрэп имеет под собой математические основы — такого рода оценки действительно сходятся к оцениваемому параметру для состоятельных оценок $\widehat{\theta}$. Однако, метод имеет скорее прикладное значение — он достаточно прост в использовании даже в сложных моделях.

Бутстрэп опирается на предположение, которое принято формулировать "The population is to the sample as the sample is to the bootstrap samples." Иначе говоря, при генерировании выборок из $F_{\widehat{\theta}}$ они будут отличаться от $\widehat{\theta}$ примерно также как $\widehat{\theta}$ от θ. Если это утверждение не выполнено, то метод может испортить оценку $\widehat{\theta}$.

Percentile интервал опирается на то, что оценка $\widehat{\theta}$ не имеет смещения и ее распределение достаточно симметрично. Из-за этого этот простой метод оказывается достаточно ненадежным — истинное значение параметра будет вылетать из интервала значительно чаще, чем должно (возможно даже никогда в него не попадать). Использовать этот метод стоит только если вы построили график $\widehat{\theta}^* - \widehat{\theta}$ и он оказался симметричным около 0. Pivotal значительно менее придирчив к данным. В следующий раз мы поговорим о более тонких вариантах этого метода.

Задача 10. Построить Percentile и Pivotal интервал для θ, $X_i \sim R[0, \theta]$, на основе смещенной оценки $\max(X_i)$ по 50 наблюдениям. Повторить эксперимент 100 раз. Как часто доверительный интервал накрывал истинное значение параметра?

Для ленивых эти и другие примеры работы метода разобраны в файлах BootstrapConf.R и BootstrapConf.py

3.6 Ответы на вопросы

Ответ 1. Что будет ОМП в случае $X_i \sim R[\theta, \theta+1]$?
Если $X_i \sim R[\theta, \theta+1]$, то функция правдоподобия будет иметь вид

$$L(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} \frac{1}{\theta + 1 - \theta} I_{x_i \in [\theta, \theta+1]} = I_{x_i \in [\theta, \theta+1], i \leq n}.$$

Эта функция равняется 1, если $\theta \leq \min(x_i) \leq \max(x_i) \leq \theta + 1$ и 0 иначе. Поэтому ОМП будет целый отрезок $[\max(x_i) - 1, \min(x_i)]$. Это, впрочем, не мешает всем этим ОМП быть состоятельными, поскольку этот отрезок стягивается с ростом n в точку θ.

Ответ 2. Что будет ОМП в случае X_i с плотностью $\exp(-|x-\theta|)/2$?
В этом случае функция правдоподобия имеет вид

$$L(x_1, ..., x_n; \theta) = 2^{-n} \exp\left(-\sum_{i=1}^{n} |x_i - \theta|\right)$$

и его максимизация равносильна минимизации $\sum_{i=1}^n |x_i - \theta|$. Эта функция линейна на каждом отрезке $[x_i, x_{i+1}]$, $i = 1, ..., n-1$. Давайте предположим, что a — (n+1)/2-й по возрастанию из x_i, если выборка имеет нечетный размер. Тогда при $\theta < a$ прямые имеют отрицательный наклон, а при $\theta > a$ положительный и минимум будет достигаться при $\theta = a$. А если выборка имеет четный размер и a, b — $n/2$ и $n/2+1$ из наблюдений по возрастанию, то при $\theta \in [a, b]$ график будет горизонтальный и ОМП будет весь отрезок $[a, b]$.

Ответ 3. Что будет с ОМП, если X_i бернуллиевские с параметром θ при $\theta \in \mathbb{Q} \cap (0, 1)$ и $1 - \theta$ иначе? В этом случае правдоподобие имеет вид

$$L(x_1, ..., x_n; \theta) = \theta^{x_1 + ... + x_n}(1-\theta)^{n - x_1 - ... - x_n},\ \theta \in \mathbb{Q},\ L(x_1, ..., x_n; \theta) = (1-\theta)^{x_1 + ... + x_n}\theta^{n - x_1 - ... - x_n},\ \theta \notin \mathbb{Q}.$$

Тогда ОМП будет $\overline{x}$, если $\overline{x} \in \mathbb{Q}$ и $1 - \overline{x}$ иначе. Но $\overline{x}$ рациональна по определению, поэтому оценка всегда будет $\overline{x}$. При иррациональных θ она не состоятельна.

Ответ 4. Что будет с ОМП, если $(X_{2i}, X_{2i-1}) \sim \mathcal{N}(\theta_i, \sigma^2)$. Будет ли ОМП для σ^2 состоятельной? Логарифм правдоподобия для этой задачи имеет вид

$$\ln L(x_1, ..., x_{2n}; \theta_1, ..., \theta_n, \sigma) = -2n \ln \sigma - n \ln(2\pi) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^n (x_{2i-1} - \theta_i)^2 - \sum_{i=1}^n (x_{2i} - \theta_i)^2\right).$$

ОМП $\widehat{\theta}_i$ для θ_i получаются равными $(x_{2i-1} + x_{2i})/2$, а для σ определяется из условия равенства нулю производной

$$-\frac{2n}{\sigma} + \frac{1}{\sigma^3}\left(\sum_{i=1}^n (x_{2i-1} - \widehat{\theta}_i)^2 - \sum_{i=1}^n (x_{2i} - \widehat{\theta}_i)^2\right).$$

Поскольку

$$(x_{2i-1} - \widehat{\theta}_i)^2 = (x_{2i} - \widehat{\theta}_i)^2 = (x_{2i} - x_{2i-1})^2/4,$$

то

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n (x_{2i} - x_{2i-1})^2}{4n}$$

Величины под знаком квадрата имеют $\mathcal{N}(0, 2\sigma^2)$ распределение, откуда

$$\widehat{\sigma}^2 \xrightarrow{d} \sigma^2/2,\ n \to \infty.$$

Следовательно, оценка не состоятельна.

Ответ 5. Рассмотрим $\widehat{\theta}_n = \overline{X}(1 + I_{|\overline{X}| > n^{2/3}})/2$, $X_i \sim \mathcal{N}(\theta, 1)$. Показать, что эта оценка асимптотически нормальна и ее асимптотическая дисперсия равна дисперсии ОМП при $\theta \neq 0$ и меньше ее при $\theta = 0$. При $\theta = 0$ величина $\sqrt{n}\overline{X} I_{\overline{X} > n^{2/3}}$ сходится к 0 по вероятности, поскольку

$$\frac{\overline{X}}{n^{2/3}} \xrightarrow{P} 0$$

из ЦПТ. Следовательно, при $\theta = 0$

$$\sqrt{n}(\widehat{\theta}_n - \theta) = \frac{1}{2}\sqrt{n}\overline{X} + \frac{1}{2}\sqrt{n}\left(\overline{X} I_{\overline{X} > n^{2/3}}\right) \xrightarrow{d} Z \sim \mathcal{N}\left(0, \frac{1}{4}\right).$$

При $\theta \neq 0$ величина $\sqrt{n}\overline{X} I_{\overline{X} \leq n^{2/3}}$ сходится к 0 по вероятности, откуда

$$\sqrt{n}(\widehat{\theta}_n - \theta) = \sqrt{n}(\overline{X} - \theta) - \frac{1}{2}\sqrt{n}\left(\overline{X} I_{\overline{X} \leq n^{2/3}}\right) \xrightarrow{d} Z \sim \mathcal{N}(0, 1).$$

Таким образом, у нашей оценки асимптотическая дисперсия будет $\frac{1}{4}$ при $\theta = 0$ и 1 иначе. Между тем, ОМП $\overline{X}$ имеет асимптотическую дисперсию 1. Приведенная оценка имеет ту же дисперсию при $\theta \neq 0$ и лучшую при $\theta = 0$. Это не противоречит асимптотической эффективности ОМП, поскольку асимптотическая дисперсия в данном случае разрывна.

Ответ 6. Какая оценка методом спейсингов для равномерного распределения $R[\theta_1, \theta_2]$?
Для нахождения оценки мы должны максимизировать

$$S(x_1, ..., x_n; \theta_1, \theta_2) = \frac{(x_1 - \theta_1)(x_2 - x_1)...(x_n - x_{n-1})(\theta_2 - x_n)}{(\theta_2 - \theta_1)^{n+1}},$$

где $\theta_1 \le x_1, \theta_2 \ge x_n$, $x_1 < x_2 ... < x_n$ — упорядоченная по возрастанию выборка. Для максимизации будем рассматривать функцию S без $(x_2 - x_1)...(x_n - x_{n-1})$ (назовем это выражение $\widetilde{S}$). Продифференцируем функцию $\widetilde{S}$ по θ_1

$$\frac{\partial}{\partial \theta_1}\widetilde{S} = \frac{-(\theta_2 - x_n)}{(\theta_2 - \theta_1)^{n+1}} + \frac{(n+1)(x_1 - \theta_1)(\theta_2 - x_n)}{(\theta_2 - \theta_1)^{n+2}} = 0,$$

откуда получаем два возможных варианты: $\theta_2 = x_n$, $\theta_2 - \theta_1 = (n+1)(x_1 - \theta_1)$. Аналогично два корня будет у уравнения, полученного из производной по θ_2 — $\theta_1 = x_1$, $\theta_2 - \theta_1 = (n+1)(\theta_2 - x_n)$. При $x_n = \theta_2$ или $x_1 = \theta_1$ функция S принимает нулевое значение, поэтому единственный кандидат на максимум — $\theta_2, \theta_1 : x_1 - \theta_1 = x_2 - \theta_2 = (\theta_2 - \theta_1)/(n+1)$. При этом

$$\theta_2 - \theta_1 = \theta_2 - x_n + x_n - x_1 + x_1 - \theta_1 = \frac{2(\theta_2 - \theta_1)}{n+1} + x_n - x_1,$$

откуда $\widehat{\theta_2} - \widehat{\theta_1} = (n+1)(x_n - x_1)/(n-1)$, $\widehat{\theta_1} = x_1 - (x_n - x_1)/(n-1)$, $\widehat{\theta_2} = x_n + (x_n - x_1)/(n-1)$. Убедимся, что данная критическая точка является максимумом. Можно сделать это с помощью матрицы вторых производных, а можно сослаться на то, что функция неотрицательна, стремится к нулю на бесконечности и ноль на границе нашей области, откуда единственная критическая точка обязана быть максимумом.

Ответ 7. Пусть X_i являются смесью двух нормальных распределений, то есть каждое наблюдение с вероятностью $1/2$ получается из распределения $\mathcal{N}(\mu_1, \sigma_1^2)$, а иначе из $\mathcal{N}(0, 1)$. Показать, что если положить $\mu_1 = x_1$ и устремить $\sigma_1 \to 0$, то правдоподобие можно сделать сколь угодно большим.
Вопрос довольно прозрачен и каждому лучше постараться ответить на него самостоятельно. Один из множителей в рассматриваемой ситуации растет к бесконечности, а остальные отделены от нуля. Убедиться в этом можно, оставив в первой скобке первое слагаемое, а в остальных второе.

Ответ 8. Показать для многомерной нормальной выборки $(X_1, ..., X_k) \sim \mathcal{N}(\vec{\mu}, \Sigma)$, что она попадает в эллипсоид с центром $\vec{\mu}$

$$(\vec{X} - \vec{\mu})^{\mathrm{T}}\Sigma^{-1}(\vec{X} - \vec{\mu}) \le x$$

с вероятностью $F_{\chi_k^2}(x)$.
Поскольку Σ положительно определена, то $\Sigma = S^{\mathrm{T}}S$ для некоторой S. При этом вектор $S^{-1}(\vec{X} - \vec{\mu})$ имеет распределение

$$\mathcal{N}\left(\vec{0}, S^{-1}\Sigma(S^{-1})^{\mathrm{T}}\right) = \mathcal{N}\left(\vec{0}, E\right).$$

Остается заметить, что тогда

$$(S^{-1}(\vec{X} - \vec{\mu}))^{\mathrm{T}}(S^{-1}(\vec{X} - \vec{\mu})) = (\vec{X} - \vec{\mu})^{\mathrm{T}}\Sigma^{-1}(\vec{X} - \vec{\mu}) \sim \chi_k^2.$$

Ответ 9. Предположим, что мы оценили среднее в модели $R[0, \theta]$ с помощью $\overline{X}$. Теперь мы берем новую выборку из $R[0, 2\overline{X}]$ и оцениваем с помощью ее среднего θ. Какую дисперсию будет иметь эта новая оценка?
Величины $Y_i \sim R[0, 2\overline{X}]$ могут быть представлены как $Y_i = 2\overline{X}R_i$, где $R_i \sim R[0, 1]$ независимы от $\overline{X}$. Тогда

$$\mathbf{E}Y_i^2 = 4\mathbf{E}\overline{X}^2\mathbf{E}R_i^2 = 4(\mathbf{D}\overline{X} + (\mathbf{E}\overline{X})^2)(\mathbf{D}R_i + (\mathbf{E}R_i)^2) = 4\left(\frac{\theta^2}{12} + \frac{\theta^2}{4}\right)\left(\frac{1}{12} + \frac{1}{4}\right) = \frac{4}{9}\theta^2.$$

При этом $\mathbf{E}Y_i = 2\mathbf{E}\overline{X}\mathbf{E}R_i = \theta/2$. Следовательно, дисперсия этой оценки $\frac{7}{36}\theta^2$, то есть больше той дисперсии $\frac{1}{12}\theta^2$, которая была бы у $\overline{X}$. Это логично, дополнительная рандомизация не уменьшает дисперсию.