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

Вспоминаем умные слова

В рамках этого занятия мы довольно много будем повторять стандартный курс статистики. Мы привыкли рассматривать именно параметрическую модель $X_i \sim F_\theta$, где F — какое-то заданное семейство функций распределения, причем обычно довольно узкое — например, все нормальные распределения.

Рассматривая модель $X_i \sim F_\theta$, где $\theta \in \Theta$, мы называли оценкой функцию $T(X_1, ..., X_n)$, которая приближает параметр θ.
При этом важными свойствами для нас были:
1) Несмещенность $E_\theta T(X_1, ..., X_n) = \theta$,
2) Состоятельность: $T(X_1, ..., X_n) \xrightarrow{P_\theta} \theta$,
3) Асимптотическая нормальность: $\sqrt{n}(T(X_1, ..., X_n) - \theta) \xrightarrow{d} Z \sim \mathcal{N}(0, \sigma^2(\theta))$. 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.

Отметим несколько важных для нас замечаний:
1) Если оценка $T(X_1, ..., X_n)$ несмещенная для функции $g(\theta)$, то $h(T(X_1, ..., X_n))$ не обязана быть несмещенной оценкой для $h(g(\theta))$. Более того, если, скажем, h строго выпукла вверх или вниз и нелинейна, то это заведомо не так.
2) Если оценка $T(X_1, ..., X_n)$ состоятельная для функции $g(\theta)$, а h — непрерывная функция, то $h(T(X_1, ..., X_n))$ состоятельная оценка $h(g(\theta))$
3) Если оценка $T(X_1, ..., X_n)$ асимптотически нормальная для функции $g(\theta)$ с асимптотической дисперсией $\sigma^2(\theta)$, а h — дифференцируемая функция, то $h(T(X_1, ..., X_n))$ 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 f, аналогичный факт верен и в случае векторной функции, асимптотическая дисперсия при этом заменяется асимптотической матрицей ковариации $\Sigma(\theta)$, которая после применения f переходит в $J\Sigma J^t$, где J — матрица Якоби f в точке $g(\theta)$.

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.

О двух старых и одном новом методе

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 $E_\theta X_1^k$, а именно состоятельные и при $E_\theta X_1^{2k} < \infty$ асимптотически нормальные с асимптотической дисперсией $D_\theta X_1^k$.

Для оценки параметра $\vec{\theta} = (\theta_1, ..., \theta_k)$ предлагает рассмотреть $\overline{X^1}$,...,$\overline{X^k}$ как оценки $\mu_1(\theta) = E_\theta X$,...,$\mu_k(\theta) = E_\theta X^k$ 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 гладким, то асимптотически нормальным (если, конечно $E_\theta X^{2k} < \infty$).

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

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

К сожалению, метод моментов не вполне удачен в плане асимптотической дисперсии, она зачастую бывает весьма большой, особенно в случае многомерных параметров.

Метод максимального правдоподобия 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 $P_\theta(X = x)$. 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. Наиболее типичные случаи:
1) Если распределение X_i устроено так, что при разных параметрах θ величины принимают значения из разных множеств.
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 ОМП, если X_i бернуллиевские с параметром θ при $\theta \in \mathbb{Q} \cap (0, 1)$ и $1 - \theta$ иначе.
3) Если множество изменения параметра не является замкнутым
Такого рода ситуация может привести к тому, что правдоподобие и вовсе будет отсутствовать и будет рассмотрен в следующем параграфе.

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

Вопрос 4. Рассмотрим $\hat{\theta}_n = \overline{X}(1+I_{|\overline{X}|>n^{2/3}})/2$, $X_i \sim \mathcal{N}(\theta, 1)$. 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

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

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

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

Упражнение 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.

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 $D_i(\theta) = P_\theta(X \in [x_{(i)}, x_{(i+1)}]) = F_\theta(x_{(i+1)}) - F_\theta(x_{(i)})$, $x_{(0)} = a$, $x_{(n+1)} = b$, 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.
Вопрос 5. Какая оценка методом спейсингов для равномерного распределения $R[\theta_1, \theta_2]$?
Упражнение 2. $F_{\theta_1,\theta_2}(x) = 1 - e^{-(x-\theta_1)^{\theta_2}}$, $x > \theta_1$. Построить численно оценку методом спейсингов. Сравнить ее с ОМП.

EM-алгоритм или что делать, когда аналитически не подобраться

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

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

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 в наша 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, то все бы значительно упростилось.
Тогда EM-алгоритм предлагает следующую программу действий:
0) Выберем начальное значение параметров $\vec{\theta}^0 = (\theta_1^0, ..., \theta_k^0)$ в нашей модели.
На $j + 1$-ом шаге проделаем следующее:
1) Expectation. Вычислим

$$J(\vec{\theta}|\vec{\theta}^j) = E_{\vec{\theta}^j} \left(\ln \frac{L(X_1, ..., X_n, Y_1, ..., Y_m; \vec{\theta})}{L(X_1, ..., X_n, Y_1, ..., Y_m; \vec{\theta}^j)} \middle| X_1 = x_1, ..., X_n = x_n \right).$$

2) 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

$$P_{\vec{\theta}^j}(Y_1 = y_1, ..., Y_m = y_m | X_1 = x_1, ..., X_n = x_n) = \frac{L(x_1, ..., x_n, y_1, ..., y_m; \vec{\theta}^j)}{\mathbf{E}_{\vec{\theta}^j} L(x_1, ..., x_n, Y_1, ..., Y_m; \vec{\theta}^j)}$$

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

$$J(\vec{\theta}|\vec{\theta}^j) = \sum_{y_1, ..., y_m} P_{\vec{\theta}^j}(Y_1 = y_1, ..., Y_m = y_m | X_1 = x_1, ..., X_n = x_n) \ln L(x_1, ..., x_n, y_1, ..., y_m; \vec{\theta})$$

в дискретном или

$$J(\vec{\theta}|\vec{\theta}^j) = \int_{\mathbb{R}^m} f_{\vec{\theta}^j}(y_1, ..., y_m | X_1 = x_1, ..., X_n = x_n) \ln L(x_1, ..., x_n, y_1, ..., y_m; \vec{\theta}) dy_1 ... dy_m.$$

Иначе говоря,

$$J(\vec{\theta}|\vec{\theta}^j) = E \ln L(x_1, ..., x_n, Y_1, ..., Y_n; \theta),$$

взятое по апостериорному распределению $\tilde{P}(A) = P_{\vec{\theta}^j}((Y_1, ..., Y_n) \in A | X_1 = x_1, ..., X_n = x_n)$.

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 примера 1 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.

Давайте применим полученный метод к уже упоминавшейся задаче о смеси нормальных:
Пример 2. $X_1, ..., X_n$ — 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

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

где математическое ожидание, как уже упоминалось, берется по мере $\tilde{P}(A)$. При этом

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

Апостериорное распределение Y_i найдем по формуле Байеса

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

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

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

Дифференцируя, получаем

$$\mu_1^{j+1} = \frac{\sum_{i=1}^n p_i x_i}{\sum_{i=1} n p_i}, \mu_2^{j+1} = \frac{\sum_{i=1}^n p_i x_i}{\sum_{i=1} n p_i}$$

Упражнение 3. Протестировать метод для $\mu_1 = 0$, $\mu_2 = 1$.

EM-алгоритм особенно часто используется в двух случаях: если плотность представима в виде смеси (в этом случае y_i берутся как номер распределения, к которому относится x_i) или если у нас неполные данные. О первой задаче мы поговорили, приведем пример второй.

Пример 3. 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. 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. 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 в форме 1'), 2') — вычислить апостериорное распределение z_1 при известных параметрах $\vec{\theta}^j$,

Упражнение 4. Применить EM-алгоритм к этой задаче и реализовать его на R, полагая $n = 10$, $\tau_i = 1/10$,$\beta = 0.25$, $m = 60$. Сгенерировать полиномиальную выборку с равными вероятностями исходов 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.

Бутстрап или как прикрывать недостатки оценок

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.

Идея метода бутстрапа заключается в том, что если оценка $\widehat{\theta}(X_1, ..., X_n)$ близка к настоящему параметру θ, а семейство распределений непрерывно зависит от параметра, то распределение $F_{\widehat{\theta}}$ будет похоже на F_θ. Таким образом, мы можем взять выборку Y_1,...,Y_n из $F_{\widehat{\theta}}$, и подсчитать по ней $\widehat{\theta}(Y_1, ..., Y_n)$. Тогда смещение $\widehat{\theta}(Y_1, ..., Y_n) - \widehat{\theta}(X_1, ..., X_n)$ будет близко к смещению $\widehat{\theta}(X_1, ..., X_n) - \theta$. Таким образом, метод бутстрапа предписывает взять оценку $2\widehat{\theta}(X_1, ..., X_n) - \widehat{\theta}(Y_1, ..., Y_n)$.

Метод бутстрапа 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) Также и этот метод позволяет нам улучшить оценку с помощью самой этой оценки.

Упражнение 5. 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 $(\hat\theta(Y_1,...,Y_n) - \hat\theta(X_1,...,X_n))^2$ в качестве оценки дисперсии $\hat\theta(X_1,...,X_n)$.

Упражнение 6. Пусть известно $\overline{X}$ из выборки. Как с его помощью бутстрэппингом оценить дисперсию известного 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 $\hat\theta$ значимо промахнулась мимо θ, рискуем неправильно изменить оценку.

Вопрос 7. 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?

Доверительные беседы о доверительных интервалах

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

Напомним, что доверительным интервалом называют пару статистик $\hat\theta_1$, $\hat\theta_2$, таких, что при всех θ выполнено равенство

$$P_\theta(\theta \in (\hat\theta_1, \hat\theta_2)) = 1-\alpha$$

где $1-\alpha$ — заданное число, называемое уровнем доверия. Можно требовать взамен выполнение неравенства $\geq$.

Построение точных доверительных интервалов достаточно трудоемко, а вот асимптотические доверительные интервалы (т.е. те, где равенство заменено на равенство в пределе) можно строить на основе любой асимптотически нормальной оценки. Так если $\hat\theta$ — 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 $I(\hat\theta)$. Из асимптотической нормальности ОМП

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

Но, в силу состоятельности $I(\hat\theta)$ как оценки $I(\theta)$

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

Тогда

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

откуда мы можем построить доверительный интервал вида

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

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

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 $D_\theta X_k = \mu_{2k}(\theta) - \mu_k^2(\theta)$. Следовательно, если ОММ для θ есть $f(\overline{X^k})$, то ее асимптотическая дисперсия $\sigma^2(\theta) = (\mu_{2k}(\theta) - \mu_k^2(\theta))(f'(\mu_k(\theta)))^2$, т.е. доверительный интервал будет иметь вид

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

Свои варианты для построения доверительного множества дает и бутстрэп. Такого рода методов несколько, но мы рассмотрим наиболее простой — pivotal интервал.
Для этого рассмотрим оценку $\hat{\theta}$ для θ. Будем брать выборки из $F_{\hat{\theta}}$ и строить на основе них оценки $\hat{\theta}_1$,...,$\hat{\theta}_m$. Рассмотрим $\Delta_i = \hat{\theta}_i - \hat{\theta}$. Мы ожидаем, что это выборка из величин, близких к $\hat{\theta} - \theta$. Упорядочим Δ_i и выберем те из них Δ_-, Δ_+, которые стоят на местах $[\gamma m]$ и $[\beta m]$ по возрастанию. Тогда

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

и будет нашим интервалом.
Упражнение 7. Сравнить на выборках размера 50 для а) $\mathcal{N}(\theta, 1)$, б) $R[0, \theta]$ доверительные интервалы на основе ОММ, ОМП, бутстрэпа с помощью $\overline{X}$.

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 вероятностей накрытия каждой из θ_1,...,θ_k своим интервалом 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 $1 - \alpha$.

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

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

с вероятностью $F_{\chi^2_k}(x)$.

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

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

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

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

$$(\hat{\theta} - \vec{\theta})^t I(\hat{\theta})(\hat{\theta} - \vec{\theta}) \leq y_{1-\alpha},$$

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 $\Sigma(\hat{\theta})$ 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 $\hat{\theta} = g(\overline{X}, \overline{X^2}, ..., \overline{X^k})$ выполнено соотношение

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

Соответственно, доверительный эллипсоид имеет вид

$$(\hat{\theta} - \theta)^t \Sigma^{-1}(\hat{\theta})(\hat{\theta} - \theta) \leq y_{1-\alpha}/n,$$

где $\Sigma = J\Sigma_1 J^t$.

Упражнение 8. * 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 $\mathcal{N}(mu, Sigma^2)$ распределения.

Ответы на вопросы

Вопрос 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

Вопрос 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 $[a, b]$.

Вопрос 3. 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.

Вопрос 4. При $\theta = 0$ $\sqrt{n}\overline{X} I_{\overline{X} > n^{2/3}}$ сходится к 0 по вероятности, поскольку

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

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

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

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

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

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.

Вопрос 5. 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 $\theta_1 \leq x_1$, $\theta_2 \geq x_n$, $x_1 < x_2 ... < x_n$ — упорядоченная по возрастанию выборка. Для максимизации будем рассматривать функцию S без $(x_2 - x_1)...(x_n - x_{n-1})$ (назовем это выражение $\tilde{S}$). Продифференцируем функцию $\tilde{S}$ по θ_1

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

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 $\hat{\theta}_2 - \hat{\theta}_1 = (n+1)(x_n - x_1)/(n-1)$, $\hat{\theta}_1 = x_1 - (x_n - x_1)/(n-1)$, $\hat{\theta}_2 = x_n + (x_n - x_1)/(n-1)$. 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.

Вопрос 6. 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.

Вопрос 7. 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

$$EY_i^2 = 4E\overline{X}^2 ER_i^2 = 4(D\overline{X} + (E\overline{X})^2)(DR_i + (ER_i)^2) = 4\left(\frac{\theta^2}{12} + \frac{\theta^2}{4}\right)\left(\frac{1}{12} + \frac{1}{4}\right) = \frac{4}{9}\theta^2.$$

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