МЕХАНИКО-
МАТЕМАТИЧЕСКИЙ
ФАКУЛЬТЕТ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in
ЛЕКЦИИ УЧЕНЫХ МГУ

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ШАБАНОВ
ДМИТРИЙ АЛЕКСАНДРОВИЧ

МЕХМАТ МГУ

Содержание

1. Лекция 1

Введение

Предмет изучения математической статистики

Разберемся с тем, что изучает математическая статистика, какие задачи ставятся и какие цели мы хотим достигать.
Начнем с того, что является предметом изучения **теории вероятностей**. Кратко его можно охарактеризовать как математический анализ случайных явлений. Если говорить неформально, то у нас есть природа некоторого явления (с математической точки зрения это значит, что мы знаем распределение), и мы хотим выяснить как ведут себя характеристики, наблюдаемые в эксперименте.

Математическая статистика - это часть теории вероятностей, в которой изучаются обратные задачи, то есть нам известны экспериментальные данные, и требуется вынести суждение о природе случайного явления.

Пример

Рассмотрим классический пример, который показывает разницу в постановках задач математической статистики и теории вероятностей.

Пусть в городе имеется всего N жителей, среди них есть M заболевших $COVID-19$.

Задача теории вероятностей
Какова вероятность того, что в случайной выборке из n человек будет ровно m заболевших?
В этой задаче M известно (без него задачу решить не получится). Ответом в этой задаче является гипергеометрическое распределение.

В задаче математической статистики все наоборот. Мы провели наблюдение и выяснили, что среди n жителей ровно m заболевших.
Задача математической статистики
Среди случайно выбранных n жителей оказалось ровно m заболевших. Как можно оценить общее число заболевших в городе?

Здесь M уже выступает в роли неизвестного параметра.

Примеры статистических задач

Рассмотрим модель простой линейной регрессии:

$$y_i = \theta \cdot x_i + \varepsilon_i, \quad i = 1, ..., n.$$

Здесь y_i, x_i - известные величины, θ - неизвестный параметр, который мы хотели бы оценить, ε_i - случайные неизвестные ошибки измерений. Это простая модель, более общую мы рассмотрим позже.

В качестве **примера** рассмотрим движение объекта по прямой. У нас есть прибор, которым мы измеряем движение объекта, но мы не знаем его скорость. Мы предполагаем, что прибор неточный, поэтому есть некоторые ошибки измерений. В итоге мы получаем $\theta \cdot x_i$ - настоящее положение объекта, с некоторой ошибкой ε_i.

- x_i - моменты измерения положения объекта,
- y_i - измеренное положение объекта в момент времени x_i,
- ε_i - случайная ошибка измерений,
- θ - неизвестная скорость объекта.

Можно выделить следующие постановки статистических задач:

1) **Точечное оценивание**.
Здесь мы хотим отыскать такую функцию $\hat{\theta} = f(x_1, ..., x_n, y_1, ..., y_n)$, что в некотором смысле приближает неизвестную нам θ.

2) **Интервальное оценивание**.
Здесь нам не так важно знать точное значение θ. Мы хотим получить это значение с некоторой точностью, например, указав такое $\varepsilon > 0$, что $|\hat{\theta} - \theta| < \varepsilon$ с большой вероятностью, например, с вероятностью 0,95.

3) **Проверка статистических гипотез**
Здесь требуется проверить (подтвердить или опровергнуть) на основе данных $x_1, ..., x_n, y_1, ..., y_n$ некоторое суждение вида $\theta \geqslant \theta_0$ или $\theta = \theta_0$. Например, нас интересует, произойдет ли повышение уровня воды в реке выше какой-либо критической отметки, причем неважно насколько выше.

4) **Проверка однородности**
Рассмотрим две группы данных $X_1, ..., X_n$ и $Y_1, ..., Y_m$, полученных в результате проведения случайного эксперимента в разных условиях, например, измерения разными приборами.

Необходимо выяснить, верно ли, что условия не влияют на результат? То есть верно ли, что $X_i \stackrel{d}{=} Y_j$ (данные одинаково распределены).

5) **Проверка независимости**
Пусть результат эксперимента имеет два параметра (фактора) (A, B), принимающие конечное число значений, $a_1, ..., a_k, b_1, ..., b_m$. После проведения серии независимых экспериментов получены количественные данные n_{ij}, $i = 1, ..., k$, $j = 1, ..., m$, по сочетаниям значений факторов.
Верно ли, что факторы A и B независимы?

Математическая статистика - это теория принятия оптимальных статистических решений. Цель - найти оптимальные решения, основанные на статистических данных.

Вероятностно-статистическая модель

Определение 1.1. Пусть имеется наблюдение $\mathbf{X}$ (результат проведенного эксперимента), которое мы считаем случайным элементом (случайным вектором и т.д.). Множество всех возможных значений $\mathbf{X}$ называется **выборочным пространством** и обозначается $\mathcal{X}$. Мы считаем, что $\mathbf{X}$ получен как результат случайного выбора из $\mathcal{X}$ с неизвестным распределением P.

Замечание
Пусть P и неизвестно, зачастую нам известен класс распределений $\mathcal{P}$, которому принадлежит P. Например, мы пронаблюдали некоторые данные, и мы предполагаем, что они взяты из нормального распределения, но нам неизвестны параметры этого распределения.
Если у нас есть такой класс распределений, то мы можем образовать тройку.

Определение 1.2. Тройка $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mathcal{P})$, где $\mathcal{X}$ - выборочное пространство, $\mathcal{B}_{\mathcal{X}}$ - σ-алгебра на нем, а $\mathcal{P}$ - семейство распределений (вероятностных мер) на $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, называется **вероятностно-статистической моделью**.

Как правило $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$

Выборка и эмпирическое распределение

Определение 1.3. Если наблюдение $\mathbf{X} = (X_1, ..., X_n)$ есть набор независимых одинаково распределенных случайных величин, то $\mathbf{X}$ называется **выборкой** размера n из некоторого распределения.

Предположим, что у нас есть выборка $\mathbf{X} = (X_1, ..., X_n)$ размера n из неизвестного распределения P_X на $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Пусть $B \in \mathcal{B}(\mathbb{R})$. Как можно восстановить $\mathrm{P}_X(B)$?

Определение 1.4. Для любого $B \in (\mathbb{R})$ положим $\mathrm{P}_n^*(B) = \dfrac{\mu(B)}{n}$, где $\mu(B)$ - это число элементов выборки, попавших в множество B, то есть

$$\mathrm{P}_n^*(B) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{X_i \in B\}.$$

Распределением P_n^*, как функция от $B \in \mathcal{B}(\mathbb{R})$, называется **эмпирическим распределением**, построенным при выборке $X_1, ..., X_n$. Это дискретное распределение, оно имеет практически равномерное распределение на элементах $X_1, ..., X_n$.

Эмпирическая функция распределения

Насколько эмпирическое распределение близко к настоящему?

Утверждение
Пусть $(X_n, n \in \mathbb{N})$ - набор н.о.р.с.в. (независимых одинаково распределенных величин) с распределением P_X. Тогда для любого $B \in \mathcal{B}(\mathbb{R})$ выполнено

$$\mathrm{P}_n^*(B) \xrightarrow{\text{п.н.}} \mathrm{P}_X(B), \quad n \to \infty,$$

где п.н. - почти наверное.
Доказательство.
Заметим, что $\mathrm{P}_n^*(B) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{X_i \in B\}$ - сумма н.о.р.с.в. Согласно УЗБЧ (усиленному закону больших чисел) имеем

$$\mathrm{P}_n^*(B) \xrightarrow{\text{п.н.}} \mathrm{EI}\{X_1 \in B\} = \mathrm{P}(X_1 \in B) = \mathrm{P}_X(B).$$

То есть наша сумма сходится почти наверное к своему математическому ожиданию $\mathrm{EI}\{X_i \in B\}$. Это то же самое, что вероятность того, что X_1 попал в B, это и есть $P_X(B)$. ∎

Замечание
К сожалению, данная сходимость является точечной, то есть при большом n мы можем хорошо аппроксимировать $\mathrm{P}_X(B)$ для конкретного B, но аппроксимация

настоящего распределения для всех множеств одновременно невозможна. Однако если рассмотреть "хорошие" множества B, то равномерная аппроксимация получается.

А можно ли аппроксимировать все множества одновременно? Ответ на этот вопрос строго отрицательный, даже для самых простых распределений.
Например, предположим, что у нас есть выборка из равномерного распределения на интервале $(0,1)$, то есть $X_1, ..., X_n \sim U(0,1)$. Мы хотим выяснить чему равно выражение $\sup\limits_{B \in \mathcal{B}(\mathbb{R})} |P_n^*(B) - P_X(B)|$. Это случайная величина, к тому же не совсем понятно, насколько она измерима, так как мы берем супремум по всей борелевской σ-алгебре. На самом деле, это выражение почти всегда равно 1, так как эмпирическое распределение - это дискретное распределение, сосредоточенное в числах $X_1, ..., X_n$. Если мы возьмем множество $B = [0,1] \setminus \{X_1, ..., X_n\}$, то $\mathrm{P}_n^* = 0$, так как там нет элементов выборки, соответственно его вероятность с точки зрения эмпирического распределения нулевая, а настоящее распределение $\mathrm{P}_X(B) = 1$, так как мы выкинули из отрезка конечное число точек, следовательно мера B совпадает с мерой отрезка. Получается, что постановка задачи такого вида бессмысленна, то есть мы не можем приближать одновременно все множества.

Чтобы рассмотреть нечто похожее, необходимо взять супремум не по борелевской σ-алгебре, а по множеству из некоторого класса.

Мы помним из теории вероятностей, что само распределение как вероятностная мера однозначно определяется своей функцией распределения. Соответственно, у эмпирического распределения тоже есть функция распределения.

Эмпирическая функция распределения

Определение 1.5. Пусть $\mathbf{X} = (X_1, ..., X_n)$ есть выборка размера n из неизвестного распределения P_X. Тогда величина

$$F_n^*(x) = \mathrm{P}_n^*((-\infty, x])) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{X_i \leqslant x\}$$

называется **эмпирической функцией распределения** выборки $X_1, ..., X_n$.

Попробуем нарисовать ее график. Эмпирическое распределение является дискретным, соответственно мы имеем дело с функцией распределения дискретного распределения.
Для удобства предположим, что выборка упорядочена, то есть $X_1 < X_2 < ... < X_n$. Тогда наша эмпирическая функция распределения выглядит следующим образом:

- До точки X_1 элементов выборки нет, поэтому значение $F_n^*(x)$ равно нулю,

- Когда мы достигли точки X_1, появляется одна точка, поэтому мы получаем скачок в значение $\frac{1}{n}$ (сумма становится равной 1), и так далее,
- Если элементы выборки совпали, то скачки могут быть на большее значение,
- Синим цветом на графике изобразим настоящую функцию распределения.

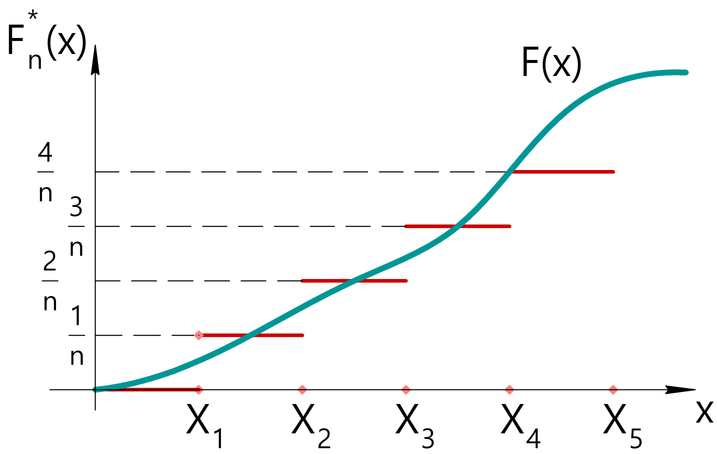

Рис. 1

Возникает вопрос: как хорошо $F_n^*(x)$ приближает настоящую функцию $F_n(x)$? Мы знаем, что в каждой точке есть сходимость (из утверждения, доказанного выше). Оказывается, что для функции распределения ситуация проще, чем для всего эмпирического распределения, а именно, мы покажем, что $F_n^*(x)$ равномерно приближает $F(x)$ на всей прямой.

Теорема (Гливенко-Кантелли)

Теорема 1.1. (Гливенко-Кантелли)
Пусть $(X_n, n \in \mathbb{N})$ - н.о.р.с.в. с функцией распределения $F(x)$. Тогда

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

Доказательство.
Будем считать, что случайные величины $(X_n, n \in \mathbb{N})$ заданы на вероятностном пространстве $(\Omega, \mathcal{F}, \mathrm{P})$. Проверим сначала, что величина

$$D_n(\omega) = \sup_{x\in\mathbb{R}} |F_n^*(x,\omega) - F(x)|$$

является случайной величиной. Заметим, что для каждого ω (элементарный исход) функция $|F_n^*(x,\omega)-F(x)|$ непрерывна справа (это следует из свойств функций распределения) и не может достигать своего максимума на бесконечностях, так как обе функции на $+\infty$ стремятся к 1, а на $-\infty$ к 0. Стало быть,

$$D_n(\omega) = \sup_{x\in\mathbb{Q}} |F_n^*(x,\omega) - F(x)|$$

является супремумом счетного числа ограниченных случайных величин. Значит, D_n - это тоже случайная величина.

Перейдем к доказательству того, что последовательность сходится к нулю п.н. Зафиксируем произвольным образом $N \in \mathbb{N}$ и положим

$$x_{k,N} = min\left\{x \in \mathbb{R} : F(x) \geqslant \frac{k}{N}\right\},$$

$k = 1, ..., N-1$ и положим $x_{0,N} = -\infty$, $x_{N,N} = +\infty$ (разрежем нашу прямую на отрезки по точкам $x_{k,N}$).
Возьмем произвольный x. Если $x \in [x_{k,N}, x_{k+1,N})$, то попробуем оценить сверху выражение $F_n^*(x) - F(x)$. Воспользуемся свойствами функций распределения, а именно: неубыванием и непрерывностью справа. Тогда:

$$F_n^*(x) - F(x) \leqslant F_n^*(x_{k+1,N} - 0) - F(x_{k,N}) =$$

$$= F_n^*(x_{k+1,N} - 0) - F(x_{k+1,N} - 0) + F(x_{k+1,N} - 0) - F(x_{k,N}) \leqslant$$

$$\leqslant F_n^*(x_{k+1,N} - 0) - F(x_{k+1,N} - 0) + \frac{1}{N},$$

где $F_n^*(x_{k+1,N} - 0)$ - значение предела слева функции F_n^* в точке $x_{k+1,N}$. Значение функции $F(x)$ берется в точке $x_{k,N}$, так как $F(x) \geqslant F(x_{k,N})$.
$F(x_{k+1,N} - 0) - F(x_{k,N}) \leqslant \frac{1}{N}$, так как в точке $x_{k,N}$ значение функции хотя-бы $\frac{k}{N}$, а в точке $x_{k+1,N}$ значение хотя-бы $\frac{k+1}{N}$, но это наименьшая точка с таким свойством, поэтому предел слева будет меньше либо равен $\frac{k+1}{N}$.

Аналогично получаем оценку снизу:

$$F_n^*(x) - F(x) \geqslant F_n^*(x_{k,N}) - F(x_{k+1,N} - 0) =$$

$$= F_n^*(x_{k,N}) - F(x_{k,N}) + F(x_{k,N}) - F(x_{k+1,N} - 0) \geqslant$$

$$\geqslant F_n^*(x_{k,N}) - F(x_{k,N}) - \frac{1}{N}.$$

Попробуем, с учетом полученного, оценить разность $|F_n^*(x) - F(x)|$ для всех $x \in \mathbb{R}$:

$$|F_n^*(x) - F(x)| \leqslant$$

$$\leqslant \max_{\substack{1 \leqslant k \leqslant N-1 \\ 1 \leqslant l \leqslant N-1}} (|F_n^*(x_{k,N}) - F(x_{k,N})|, |F_n^*(x_{l,N} - 0) - F(x_{l,N} - 0)|) + \frac{1}{N} .$$

Вспомним, что $F_n^*(y) = \frac{1}{n}\sum\limits_{i=1}^{n} \mathrm{I}\{X_i \leqslant y\}$, а $F_n^*(y-0)$ - предел слева функции распределения, который либо на единицу меньше (оказались в точке X_i), либо равен значению (по непрерывности). Мы посчитаем количество точек, которые строго меньше чем y, то есть $F_n^*(y-0) = \frac{1}{n}\sum\limits_{i=1}^{n} \mathrm{I}\{X_i < y\}$. Согласно УЗБЧ для каждого $y \in \mathbb{R}$ выполнено $F_n^*(y) \xrightarrow{\text{п.н.}} F(y)$ и $F_n^*(y-0) \xrightarrow{\text{п.н.}} F(y-0)$. Обозначим:

$$\Omega_N = \{\omega : \forall k, l,\ F_n^*(x_{k,N}, \omega) \longrightarrow F(x_{k,N}), F_n^*(x_{l,N} - 0, \omega) \longrightarrow F(x_{l,N} - 0)\} .$$

Тогда $\mathrm{P}(\Omega_N) = 1$ для любого N. В то же время для любого $\omega \in \Omega_N$ выполнено

$$\overline{\lim_{n\to\infty}} D_n(\omega) \leqslant \frac{1}{N} .$$

Обозначим $\Omega' = \bigcap\limits_{N=1}^{\infty} \Omega_N$. Тогда $\mathrm{P}(\Omega') = 1$ и для любого $\omega \in \Omega'$

$$\exists \overline{\lim_{n\to\infty}} D_n(\omega) = 0 ,$$

что и доказывает искомую сходимость $D_n \xrightarrow{\text{п.н.}} 0$ при $n \longrightarrow +\infty$. ■

Замечание
Теорема Гливенко-Кантелли показывает, что основная задача математической статистики (восстановление неизвестного распределения) вполне обоснована. Например, мы можем равномерно по всем точкам на прямой восстановить функцию распределения.

Глядя на эту теорему возникает вопрос о скорости сходимости D_n к нулю: сколько наблюдений необходимо взять, чтобы можно было сказать, что наша функция распределения эмпирическая? Ответ на этот вопрос очень нетривиален, и получим мы его ближе к концу курса.

Параметрическая модель

Обсудим теперь в каких вероятностных пространствах мы будем работать.

Определение 1.6. Вероятностно-статистическая модель $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mathcal{P})$ называется **параметрической**, если семейство $\mathcal{P}$ параметризовано, то есть $\mathcal{P} = \{\mathrm{P}_\theta : \theta \in \Theta\}$. При этом $\mathrm{P}_{\theta_1} \neq \mathrm{P}_{\theta_2}$, если $\theta_1 \neq \theta_2$.

Примеры

- Биномиальное распределение: $\mathrm{Bin}(1, \theta)$, $\theta \in (0, 1)$.
- Нормальное распределение: $\mathcal{N}(a, \sigma^2)$, $\theta = (a, \sigma^2)$, $a \in \mathbb{R}$, $\sigma > 0$.
- Экспоненциальное распределение: $\mathrm{Exp}(\theta)$, $\theta > 0$.

Преимущество параметрических моделей в том, что если наше распределение параметризовано, то вместо того, чтобы восстанавливать все распределение, мы можем попытаться восстановить конкретное значение параметра, что гораздо проще.

Наблюдение $\mathbf{X}$ мы считаем результатом случайного выбора элемента из $\mathcal{X}$. Формально, можно понимать $\mathbf{X}$ как тождественное отображение из $\mathcal{X}$ в $\mathcal{X}$:

$$\mathbf{X}(x) = x \quad \forall x \in \mathcal{X}\ .$$

Тогда для любого $\theta \in \Theta$ случайный вектор $\mathbf{X}$ определен на вероятностном пространстве $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mathrm{P}_\theta)$ и принимает значения в $\mathcal{X}$.

Замечание
Тогда P_θ будет и распределением $\mathbf{X}$: для любого $B \in \mathcal{B}_{\mathcal{X}}$

$$\mathrm{P}_\theta(\mathbf{X} \in B) = \mathrm{P}_\theta(B).$$

Если $\mathbf{X}= (X_1, ..., X_n)$ - это выборка, то естественно оценивать и параметризовывать распределение P_θ только одного элемента выборки. Как тогда построить формальную модель?

Случай конечной выборки

Предположим, что мы построили вероятносто-статистическую модель $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mathcal{P})$ для одного элемента выборки. Тогда рассмотрим тройку $(\mathcal{X}^n, \mathcal{B}^n_{\mathcal{X}}, \mathcal{P}^n)$, где

- $\mathcal{X}^n = \underbrace{\mathcal{X} \times ... \times \mathcal{X}}_{n \text{ штук}}$;
- $\mathcal{B}_{\mathcal{X}}^n = \underbrace{\mathcal{B}_{\mathcal{X}} \otimes ... \otimes \mathcal{B}_{\mathcal{X}}}_{n \text{ штук}}$ - σ-алгебра, порожденная прямоугольниками.
  Формально
  $$\mathcal{B}_{\mathcal{X}}^n = \sigma(B_1 \times ... \times B_N : B_i \in \mathcal{B}_{\mathcal{X}}).$$
- $\mathcal{P}^n = \{\mathrm{P}_\theta^n, \theta \in \Theta\}$, где $\mathrm{P}_\theta^n = \mathrm{P}_\theta \otimes ... \otimes \mathrm{P}_\theta$ - вероятностная мера на $(\mathcal{X}^n, \mathcal{B}_{\mathcal{X}}^n)$, которая на прямоугольниках задается по правилу
  $$\mathrm{P}_\theta^n(B_1 \times ... \times B_n) = \prod_{i=1}^n \mathrm{P}_\theta(B_i).$$

Лемма
Вероятностная мера P_θ^n существует и единственна с подобным свойством.

Замечание
Говорят, что вероятностное пространство $(\mathcal{X}^n, \mathcal{B}_{\mathcal{X}}^n, \mathcal{P}^n)$ является прямым произведением вероятностных мер $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mathcal{P})$.

Как моделировать выборку? Мы хотим, чтобы случайные величины были независимы и одинаково распределены, и хотим иметь одно и то же распределение P_θ. Для каждого $i = 1, ..., n$ определим отображение $X_i : \mathcal{X}^n \longrightarrow \mathcal{X}$ по правилу
$$X_i(x_1, ..., x_n) = x_i.$$
Тогда для любых $B_1, ..., B_n \in \mathcal{B}_{\mathcal{X}}$ выполнено
$$\mathrm{P}_\theta^n(X_1 \in B_1, ..., X_n \in B_n) = \mathrm{P}_\theta^n(B_1 \times ... \times B_n) = \prod_{i=1}^n \mathrm{P}_\theta(B_i) = \prod_{i=1}^n \mathrm{P}_\theta^n(X_i \in B_i)$$
Значит, $X_1, ..., X_n$ - независимы и одинаково распределены с одним и тем же распределением $\mathrm{P}_\theta(B_i)$.

Случай бесконечной выборки

В асимптотических вопросах возникает потребность в выборке $\{X_n, n \in \mathbb{N}\}$ неограниченного размера. В этом случае рассматриваем тройку $(\mathcal{X}^\infty, \mathcal{B}_{\mathcal{X}}^\infty, \mathcal{P}^\infty)$, где

- $\mathcal{X}^\infty = \{(x_1, x_2, ...) : x_i \in \mathcal{X}\}$ - пространство последовательностей из $\mathcal{X}$;
- $\mathcal{B}_{\mathcal{X}}^\infty$ - цилиндрическая σ-алгебра на $\mathcal{X}^\infty$. Пусть $B_n \in \mathcal{B}_{\mathcal{X}}^n$, тогда определим цилиндр с основанием B_n;
  $$\mathcal{F}_n(B_n) = \{x \in \mathcal{X}^\infty, x = (x_1, x_2, ...) : (x_1, ..., x_n) \in B_n\}$$
  Тогда $\mathcal{B}_{\mathcal{X}}^\infty$ - σ-алгебра, порожденная всеми цилиндрами:
  $$\mathcal{B}_{\mathcal{X}}^\infty = \sigma(\mathcal{F}_n(B_n) : n \in \mathbb{N}, B_n \in \mathcal{B}_{\mathcal{X}}^n).$$

- $\mathcal{P}^\infty = \{\mathrm{P}_\theta^\infty, \theta \in \Theta\}$, где P_θ^∞ - вероятностная мера на $(\mathcal{X}^\infty, \mathcal{B}_\mathcal{X}^\infty)$, которая на цилиндрах задается по правилу

$$\mathrm{P}_\theta^\infty(\mathcal{F}_n(\mathcal{B}_n)) = \mathrm{P}_\theta^n(B_n).$$

Теорема 1.2. (Колмогоров)
Если $(\mathcal{X}, \mathcal{B}_\mathcal{X}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, то вероятностная мера P_θ^∞ существует и единственная с подобным свойством.

Как моделировать выборку? Для каждого $i \in \mathbb{N}$ определим отображение $X_i : \mathcal{X}^\infty \longrightarrow \mathcal{X}$ по правилу (такому же, как в n-мерном пространстве)

$$X_i(x_1, x_2, ...) = x_i$$

Тогда для любых $B_1, ..., B_n \in \mathcal{B}_\mathcal{X}$ выполнено

$$\mathrm{P}_\theta^\infty(X_1 \in B_1, ..., X_n \in B_n) = \mathrm{P}_\theta^\infty(\mathcal{F}_n(B_1 \times ... \times B_n)) =$$

$$= \mathrm{P}_\theta^n(B_1 \times ... \times B_n) = \prod_{i=1}^n \mathrm{P}_\theta(B_i) = \prod_{i=1}^n \mathrm{P}_\theta^\infty(X_i \in B_i).$$

Значит, $(X_n, n \in \mathbb{N})$ - независимы и одинаково распределены с распределением P_θ.

Замечание
В дальнейшем мы будем опускать индексы n и ∞ у мер P_θ^n и P_θ^∞, обозначая их также P_θ, как и распределение одного наблюдения X_i.

В записях вида

$$\mathrm{P}_\theta(\mathbf{X} \in B), \quad \mathrm{E}_\theta f(\mathbf{X})$$

предполагается, что вероятности и математические ожидания берутся в пространстве $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \mathrm{P}_\theta)$, то есть при условии, что θ - истинное значение параметра. Отметим, что в построенной модели смена параметра не требует смены отображения $\mathbf{X}$.

В параметрической модели вопрос о нахождении истинного распределения сводится к вопросу о нахождении истинного значения параметра (что, конечно, выглядит куда проще).

2. Лекция 2

Формально, материал этой лекции относится к теории вероятностей, но в силу того, что зачастую в базовом курсе он отсутствует, мы пройдем его в нашем курсе и изучим те основные инструменты работы со сходимости, которые нам будут крайне нужны в рамках математической статистики.

Сходимости случайных векторов

Определение 2.1. Пусть $\{\xi_n, n \in \mathbb{N}\}$, ξ - случайные векторы размерности m. Тогда последовательность ξ_n сходится к ξ

- с вероятностью 1 (почти наверное, $\xi_n \xrightarrow{\text{п.н.}} \xi$), если
$$\mathrm{P}(\lim_{n \longrightarrow +\infty} \xi_n = \xi) = 1.$$
- по вероятности ($\xi_n \xrightarrow{\mathrm{P}} \xi$), если $\forall \varepsilon > 0$ выполнено
$$\mathrm{P}(\|\xi_n - \xi\|_2 \geqslant \varepsilon) \to 0 \quad \text{при} \quad n \to \infty,$$
где $\|x\|_2 = \sqrt{x_1^2 + ... + x_m^2}$ для $x \in \mathbb{R}^m$.
- по распределению ($\xi_n \xrightarrow{d} \xi$), если для любой ограниченной непрерывной функции $f : \mathbb{R}^m \longrightarrow \mathbb{R}$ выполнено
$$\mathrm{E}f(\xi_n) \longrightarrow \mathrm{E}f(\xi) \quad \text{при} \quad n \to \infty.$$

Естественно задать вопрос: чем отличаются данные сходимости от одномерных?

Упражнение (1)
Для сходимости почти наверное и по вероятности векторная сходимость эквивалентна соответствующим сходимостям компонент: если $\xi = (\xi^{(1)}, ..., \xi^{(m)})$, $\xi_n = (\xi_n^{(1)}, ..., \xi_n^{(m)})$, то

$$\xi_n \xrightarrow{\text{п.н.}} \xi \Leftrightarrow \forall i = 1, ..., m \quad \xi_n^{(i)} \xrightarrow{\text{п.н.}} \xi^{(i)},$$
$$\xi_n \xrightarrow{\mathrm{P}} \xi \Leftrightarrow \forall i = 1, ..., m \quad \xi_n^{(i)} \xrightarrow{\mathrm{P}} \xi^{(i)}.$$

Упражнение (2)
Для сходимости по распределению векторная сходимость влечет сходимость компонент: если $\xi = (\xi^{(1)}, ..., \xi^{(m)})$, $\xi_n = (\xi_n^{(1)}, ..., \xi_n^{(m)})$, то

$$\xi_n \xrightarrow{d} \xi \Leftrightarrow \forall i = 1, ..., m \quad \xi_n^{(i)} \xrightarrow{d} \xi^{(i)}.$$

Обратное утверждение, вообще говоря, неверно.

Упражнение (3, взаимоотношение видов сходимостей)

$$\xi_n \xrightarrow{\text{п.н.}} \xi \Rightarrow \xi_n \xrightarrow{\mathrm{P}} \xi \Rightarrow \xi_n \xrightarrow{d} \xi.$$

Слабая сходимость вероятностных мер

Слабая сходимость вероятностных мер пригодится нам в дальнейшем, когда мы будем говорить о доказательстве критерия Колмогорова. Пока напомним о том, что это такое и поговорим о том, как интерпретируется сходимость по распределению в терминах сходимости самих распределений.

Пусть (S, ρ) - метрическое пространство.

Определение 2.2. Борелевской σ-алгеброй, $\mathcal{B}(S)$, на (S, ρ) называется минимальная σ-алгебра, содержащая все открытые множества в S.

Определение 2.3. Пусть задано метрическое пространство (S, ρ) и последовательность $\{\mathrm{Q}_n, n \in \mathbb{N}\}$ вероятностных мер на $(S, \mathcal{B}(S))$. Будем говорить, что Q_n слабо сходится к вероятностной мере Q на $(S, \mathcal{B}(S))$, если для любой ограниченной непрерывной функции $f : S \longrightarrow \mathbb{R}$ выполнено

$$\lim_{n\to\infty} \int\limits_S f(x)\mathrm{Q}_n(dx) = \int\limits_S f(x)\mathrm{Q}(dx).$$

Обозначение: $\mathrm{Q}_n \xrightarrow{\mathrm{w}} \mathrm{Q}$.

Теорема Александрова

Утверждение Если пространство (S, ρ) сепарабельно, то $\mathcal{B}(S)$ является минимальной σ-алгеброй, содержащей все открытые шары.

Одним из главных инструментов работы со слабой сходимостью является следующая теорема Александрова, которая дает серию эквивалентных утверждений, каждое из которых эквивалентно слабой сходимости. Соответственно, чтобы доказать слабую сходимость, будет достаточно проверить одно из свойств, представленных в теореме.

Теорема 2.1. (А.Д. Александров)
Пусть $\{\mathrm{Q}_n, n \in \mathbb{N}\}$ и Q - вероятностные меры на метрическом пространстве (S, ρ). Тогда следующие утверждения эквивалентны:

1) $\mathrm{Q}_n \xrightarrow{\mathrm{w}} \mathrm{Q}$,

2) $\overline{\lim\limits_{n\to\infty}} \mathrm{Q}_n(F) \leqslant \mathrm{Q}(F)$ для любого замкнутого множества $F \subset S$,

3) $\underline{\lim\limits_{n\to\infty}} \mathrm{Q}_n(G) \geqslant \mathrm{Q}(G)$ для любого открытого множества $G \subset S$,

4) Для любого борелевского множества $B \in \mathcal{B}(S)$ такого, что $\mathrm{Q}(\partial B) = 0$, выполнено $\mathrm{Q}_n(B) \longrightarrow \mathrm{Q}(B)$ при $n \to \infty$.

Пояснение
Пусть есть некоторое множество $B \subset S$. Рассмотрим его замыкание $[B]$ в S (для этого необходимо дополнить множество B всеми предельными точками). Тогда граница $\partial B = [B] \cap [\overline{B}]$, где $\overline{B}$ - дополнение к множеству B.

Обозначение для свойства 4: $\mathrm{Q}_n \Rightarrow \mathrm{Q}$ (сходимость в основном).

Доказательство теоремы Александрова можно найти в курсе теории вероятностей.

Взаимоотношение видов сходимостей

Замечание
Пусть P_η обозначает распределение случайного вектора η. Тогда

$$\xi_n \xrightarrow{d} \xi \Leftrightarrow \mathrm{P}_{\xi_n} \xrightarrow{\mathrm{w}} \mathrm{P}_\xi \Leftrightarrow \text{ (т. Александрова) } \mathrm{P}_{\xi_n} \Rightarrow \mathrm{P}_\xi.$$

Отсюда можно вывести факт, который объясняет, зачем вообще нужна сходимость по распределению. Рассмотрим теорему, которая говорит о том, чему эквивалентна сходимость по распределению для случайных величин.

Теорема 2.2. Если $\{\xi_n, n \in \mathbb{N}\}$, ξ - случайные величины, то $\xi_n \xrightarrow{d} \xi$ тогда и только тогда, когда $F_{\xi_n}(x) \longrightarrow F_\xi(x)$ для всех $x \in \mathbb{C}(F)$, где $\mathbb{C}(\mathbb{F})$ - множество точек непрерывности функции распределения $F_\xi(x)$.

Для случайных векторов мы можем тоже предложить подобное утверждение.

Предельные теоремы

Вспомним предельные теоремы, и где вообще могут возникать сходимости.

Утверждение (УЗБЧ для случайных векторов)
Пусть $\{X_n, n \in \mathbb{N}\}$ - последовательность независимых одинаково распределенных случайных векторов из $\mathbb{R}^m$, $\mathrm{E}X_1$ конечно. Тогда

$$\frac{X_1 + ... + X_n}{n} \xrightarrow{\text{п.н.}} \mathrm{E}X_1 \ .$$

Доказательство.
Следует из одномерного УЗБЧ и эквивалентности векторной сходимости п.н. сходимости п.н. всех компонент. ∎

Теорема 2.3. (многомерная центральная предельная теорема)
Пусть $\{X_n, n \in \mathbb{N}\}$ - независимые одинаково распределенные случайные векторы из $\mathbb{R}^m$, $\mathrm{E}X_n = a$, $\mathrm{D}X_n = \Sigma$ (матрица ковариации). Обозначим $S_n = X_1 + ... + X_n$. Тогда

$$\sqrt{n}\left(\frac{S_n}{n} - a\right) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Теорема о наследовании сходимости

Ключевым утверждением, позволяющим работать со сходимостями, является теорема о наследовании сходимости.

Теорема 2.4. (**о наследовании сходимости**)
Пусть $\{\xi_n, n \in \mathbb{N}\}$, ξ - случайные векторы размерности m. Пусть $h(x) : \mathbb{R}^m \longrightarrow \mathbb{R}^k$ - функция, непрерывная почти всюду относительно распределения ξ (т.е. $\exists B \in \mathcal{B}(\mathbb{R}^m)$ такое, что h непрерывна на B и $\mathrm{P}(\xi \in B) = 1$). Тогда

1) $\xi_n \xrightarrow{\text{п.н.}} \xi \Rightarrow h(\xi_n) \xrightarrow{\text{п.н.}} h(\xi)$,

2) $\xi_n \xrightarrow{\mathrm{P}} \xi \Rightarrow h(\xi_n) \xrightarrow{\mathrm{P}} h(\xi)$,

3) $\xi_n \xrightarrow{d} \xi \Rightarrow h(\xi_n) \xrightarrow{d} h(\xi)$.

Доказательство первого и второго пункта довольно простые. В третьем пункте придется воспользоваться теоремой Александрова.

Если бы h была просто непрерывной функцией, то доказательство стало бы намного проще.
Покажем это для третьего пункта: необходимо показать, что $h(\xi_n) \xrightarrow{d} h(\xi)$, где $h(\xi_n) \in \mathbb{R}^k$. Возьмем ограниченную непрерывную функцию $f : \mathbb{R}^k \longrightarrow \mathbb{R}$. Необходимо удостовериться в том, что $\mathrm{E}f(h(\xi_n)) \longrightarrow \mathrm{E}f(h(\xi))$. Так как функция $f \circ h : \mathbb{R}^m \longrightarrow \mathbb{R}$ тоже ограничена и непрерывна, так как композиция непрерывных функций - непрерывная функция, а внешняя функция ограничена, то $\mathrm{E}f(h(\xi_n)) \longrightarrow \mathrm{E}f(h(\xi))$ верно в силу того, что $\xi_n \xrightarrow{d} \xi$. Перейдем к доказательству теоремы.
Доказательство.
1.

$$\mathrm{P}(\lim_{n\to\infty} h(\xi_n) = h(\xi)) \geqslant \mathrm{P}(\lim_{n\to\infty} h(\xi_n) = h(\xi), \xi \in B) \underset{\text{на } B}{\overset{\text{т.к. } h \text{ непр.}}{\geqslant}} \mathrm{P}(\lim_{n\to\infty} \xi_n = \xi, \xi \in B) = 1$$

так как оба события имеют полную вероятность.

2. Пусть $h(\xi_n) \overset{\mathrm{P}}{\not\longrightarrow} h(\xi)$. Тогда $\exists \varepsilon_0, \delta_0 > 0$ и подпоследовательность ξ_{n_k} такая, что

$$\mathrm{P}(\|h(\xi_{n_k}) - h(\xi)\|_2 \geqslant \varepsilon_0) \geqslant \delta_0 \quad \forall k.$$

Но $\xi_{n_k} \xrightarrow{\mathrm{P}} \xi$, значит, существует подпоследовательность, сходящаяся почти наверное: $\xi_{n_{k_s}} \xrightarrow{\text{п.н.}} \xi$ при $s \to \infty$. Согласно 1. получаем, что $h(\xi_{n_{k_s}}) \xrightarrow{\text{п.н.}} h(\xi)$.
Значит, $h(\xi_{n_{k_s}}) \xrightarrow{\mathrm{P}} h(\xi)$ при $s \to \infty$. Противоречие с выбором подпоследовательности ξ_{n_k}.

3. Обозначим Q_n - распределение $h(\xi_n)$, Q - распределение $h(\xi)$. Хотим показать, что $\mathrm{Q}_n \xrightarrow{\mathrm{w}} \mathrm{Q}$. По теореме Александрова достаточно показать, что для любого замкнутого $F \subset \mathbb{R}^k$ выполнено

$$\overline{\lim_n} \mathrm{Q}_n(F) \leqslant \mathrm{Q}(F).$$

Имеем,

$$\overline{\lim_n} \mathrm{Q}_n(F) = \overline{\lim_n} \mathrm{P}_{\xi_n}(h^{-1}(F)) \leqslant \overline{\lim_n} \mathrm{P}_{\xi_n}([h^{-1}(F)]) \overset{\text{т.к. } \mathrm{P}_{\xi_n} \xrightarrow{\mathrm{w}} \mathrm{P}_\xi}{\leqslant} \mathrm{P}_\xi([h^{-1}(F)]).$$

Но в силу замкнутости F

$$[h^{-1}(F)] \subset \overline{B} \cup h^{-1}(F),$$

ведь если $x_n \longrightarrow x$, $x_n \in h^{-1}(F)$ и $x \in B$, то $h(x) \in F$. Учитывая, что $\mathrm{P}_\xi(\overline{B}) = 0$, получаем, что $\mathrm{P}_\xi([h^{-1}(F)]) = \mathrm{P}_\xi(h^{-1}(F)) = \mathrm{Q}(F)$. ■

Лемма Слуцкого

Второй важнейший инструмент работы со сходимостями - это лемма Слуцкого.

Теорема 2.5. (лемма Слуцкого)
Пусть $\{\xi, n \in \mathbb{N}\}$ и $\{\eta, n \in \mathbb{N}\}$ - две последовательности случайных векторов (вообще говоря, разной размерности). Если $\xi_n \xrightarrow{d} \xi$ и $\eta_n \xrightarrow{d} C$, где C - константа, тогда

$$(\xi_n, \eta_n) \xrightarrow{d} (\xi, C).$$

Мы упоминали в начале лекции, что сходимость компонент по распределению не означает сходимости векторной по распределению. Это крайне неудобно, так как теорема о наследовании сходимости требует векторной сходимости. Утверждение леммы Слуцкого состоит в том, что если одна из координат сходится по распределению к константе, то выполнена и двумерная сходимость по распределению. В то же время, теорема о наследовании сходимости позволяет брать функции от многомерных сходимостей. Тем самым, можно удачно сочетать данные два инструмента.
Доказательство.
Пусть $\xi \in \mathbb{R}^m$, $C \in \mathbb{R}^k$. Необходимо показать, что для любой ограниченной непрерывной функции $f : \mathbb{R}^{m+k} \longrightarrow \mathbb{R}$ выполнено

$$\mathrm{E}f(\xi_n, \eta_n) \longrightarrow \mathrm{E}f(\xi, C) \quad \text{при} \quad n \to \infty$$

Заметим, что достаточно проверить указанную сходимость только для равномерно непрерывных функций f. Например, для сходимости характеристических функций нам даже достаточно полагать, что $f(x) = \cos\langle x, t\rangle$ или $f(x) = \sin\langle x, t\rangle$.

Зафиксируем $\varepsilon > 0$. Положим $M = \max\limits_{x \in \mathbb{R}^{m+k}} |f(x)|$. Выберем $\delta > 0$ так, чтобы для любого $n \geqslant n_0(\varepsilon)$ были выполнены условия:

- $$2M \cdot \mathrm{P}(\|\eta_n - C\| \geqslant \delta) \leqslant \frac{\varepsilon}{2};$$
- для любых $x \in \mathbb{R}^m$ и $\|y - y'\| \leqslant 2\delta$ верно, что
$$|f(x, y) - f(x, y')| \leqslant \frac{\varepsilon}{2}\,.$$

Далее рассмотрим функции

$$_1(x) = \max_{y : \|y - C\| \leqslant \delta} f(x, y), \quad g_2(x) = \min_{y : \|y - C\| \leqslant \delta} f(x, y).$$

Это непрерывные ограниченные функции на $\mathbb{R}^m$, причем

$$|g_i(x) - f(x, C)| \leqslant \frac{\varepsilon}{2}\,.$$

Теперь разобьем искомую величину на сумму

$$\mathrm{E}f(\xi_n, \eta_n) = \mathrm{E}(f(\xi_n, \eta_n)\mathrm{I}\{\|\eta_n - C\| \leqslant \delta\}) + \mathrm{E}(f(\xi_n, \eta_n)\mathrm{I}\{\|\eta_n - C\| > \delta\}).$$

При $n \geqslant n_0(\varepsilon)$ второе слагаемое не превосходит $M \cdot \mathrm{P}(\|\eta_n - C\| \geqslant \delta) \leqslant \dfrac{\varepsilon}{4}$.
Первое же слагаемое можно оценить следующим образом:

$$\mathrm{E}(g_2(\xi_n)\mathrm{I}\{\|\eta_n - C\| \leqslant \delta\}) \leqslant \mathrm{E}(f(\xi_n, \eta_n)\mathrm{I}\{\|\eta_n - C\| \leqslant \delta\}) \leqslant \mathrm{E}(g_1(\xi_n)\mathrm{I}\{\|\eta_n - C\| \leqslant \delta\}).$$

Далее, заметим, что

$$|\mathrm{E}g_i(\xi_n) - \mathrm{E}(g_i(\xi_n)\mathrm{I}\{\|\eta_n - C\| \leqslant \delta\})| \leqslant \mathrm{E}(|g_i(\xi_n)|\mathrm{I}\{\|\eta_n - C\| > \delta\}) \leqslant$$

$$\leqslant M \cdot \mathrm{P}(\|\eta_n - C\| \geqslant \delta) \leqslant \frac{\varepsilon}{4}\,.$$

Из приведенных неравенств получаем, что при $n \geqslant n_0(\varepsilon)$

$$\mathrm{E}g_2(\xi_n) - \frac{\varepsilon}{2} \leqslant \mathrm{E}f(\xi_n, \eta_n) \leqslant \mathrm{E}g_1(\xi_n) + \frac{\varepsilon}{2}\,.$$

Перейдем к пределу при $n \longrightarrow +\infty$ и используем тот факт, что $\xi_n \xrightarrow{d} \xi$:

$$\mathrm{E}g_2(\xi) - \frac{\varepsilon}{2} \leqslant \varliminf_{n\to\infty} \mathrm{E}f(\xi_n, \eta_n) \leqslant \varlimsup_{n\to\infty} \mathrm{E}f(\xi_n, \eta_n) \leqslant \mathrm{E}g_1(\xi) + \frac{\varepsilon}{2}\,.$$

Остается вспомнить, что для любого x выполнено $|g_i(x) - f(x, C)| \leqslant \dfrac{\varepsilon}{2}$. В итоге:

$$\mathrm{E}f(\xi, C) - \varepsilon \leqslant \varliminf_{n\to\infty} \mathrm{E}f(\xi_n, \eta_n) \leqslant \varlimsup_{n\to\infty} \mathrm{E}f(\xi_n, \eta_n) \leqslant \mathrm{E}f(\xi, C) + \varepsilon\,.$$

В силу произвольности ε получаем, что существует

$$\lim_{n\to\infty} \mathrm{E}f(\xi_n, \eta_n) = \mathrm{E}f(\xi, C).$$

■

Полезно следующее следствие из леммы Слуцкого.

Следствие
Пусть $\{\xi, n \in \mathbb{N}\}$ и $\{\eta, n \in \mathbb{N}\}$ - две последовательности случайных величин. Если $\xi_n \xrightarrow{d} \xi$ и $\eta_n \xrightarrow{d} C$, где C - константа. Тогда

$$\xi_n + \eta_n \xrightarrow{d} \xi + C, \quad \xi_n \cdot \eta_n \xrightarrow{d} \xi \cdot C.$$

Доказательство.
Согласно лемме Слуцкого выполнено $(\xi_n, \eta_n) \xrightarrow{d} (\xi, C)$. Тогда применяя теорему о наследовании сходимости для непрерывных функций суммы и произведения в $\mathbb{R}^2$, получаем, что

$$\xi_n + \eta_n \xrightarrow{d} \xi + C, \quad \xi_n \cdot \eta_n \xrightarrow{d} \xi \cdot C.$$

■

Пример применения

Рассмотрим следующее применение теоремы о наследовании сходимости и леммы Слуцкого.

Пример
Пусть $\xi_n \xrightarrow{d} \xi$ - случайные величины, а $h(x)$ - функция, дифференцируемая в точке $a \in \mathbb{R}$. Найдите предел сходимости по распределению у последовательности

$$\frac{h(a + b_n\xi_n) - h(a)}{b_n},$$

где $b_n \to 0$ - произвольная последовательность положительных чисел.

Решение

Рассмотрим функцию

$$H(x) = \begin{cases} \dfrac{h(x+a) - h(a)}{x}, & x \neq 0; \\ h'(a), & x = 0. \end{cases}$$

Ясно, что H непрерывна в нуле.
По лемме Слуцкого $b_n\xi_n \xrightarrow{d} 0 \cdot \xi = 0$. Следовательно, по теореме о наследовании сходимости

$$H(b_n\xi_n) \xrightarrow{d} H(0) = h'(a).$$

Снова применяем лемму Слуцкого:

$$\xi_n \cdot H(b_n\xi_n) \xrightarrow{d} \xi \cdot h'(a).$$

Остается заметить, что

$$\xi_n \cdot H(b_n\xi_n) = \frac{h(a + b_n\xi_n) - h(a)}{b_n}.$$

Общий случай

Полезно и следующее обобщение разобранного примера на многомерный случай.

Утверждение
Пусть $\xi_n \xrightarrow{d} \xi$ - случайные векторы из $\mathbb{R}^m$. Пусть $H(x) : \mathbb{R}^m \longrightarrow \mathbb{R}^k$ - вектор-функция, дифференцируемая в точке $a \in \mathbb{R}^m$. Обозначим через

$$H'(a) = \left(\frac{\partial H_i(x)}{\partial x_j}, \quad i = 1, ..., k, \quad j = 1, ..., m \right) \Bigg|_{x=a}$$

матрицу частных производных. Тогда для любой последовательности положительных чисел $b_n \to 0$ выполнено

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

Доказательство оставляется в качестве упражнения.

Асимптотическая нормальность

В математической статистике важную роль играют асимптотически нормальные оценки. Один из основных способов их получения состоит в применении данных утверждений к (многомерной) центральной предельной теореме.

Следствие (дельта-метод)
Пусть $\{X, n \in \mathbb{N}\}$ - независимые одинаково распределенные случайные векторы из $\mathbb{R}^m$, $\mathrm{E}X_n = a$, $\mathrm{D}X_n = \Sigma$. Обозначим $S_n = X_1 + ... + X_n$. Пусть $h : \mathbb{R}^m \longrightarrow \mathbb{R}$ - дифференцируемая в точке $a \in \mathbb{R}^m$ функция. Тогда

$$\sqrt{n}\left(h\left(\frac{S_n}{n} \right) - h(a) \right) \xrightarrow{d} \langle h'(a), \mathcal{N}(0, \Sigma) \rangle.$$

Решение

Подберем параметры для предыдущего утверждения. Согласно многомерной ЦПТ

$$\sqrt{n}\left(\frac{S_n}{n} - a \right) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Тогда положим:

- $\xi_n = \sqrt{n}\left(\dfrac{S_n}{n} - a \right)$,

- $\xi \sim \mathcal{N}(0, \Sigma)$,
- $H(x) = h(x)$,
- $a = a$,
- $b_n = \dfrac{1}{\sqrt{n}}$ (положительная и стремится к 0).

Тогда

$$\frac{H(a + b_n\xi_n) - H(a)}{b_n} = \sqrt{n}\left(h\left(\frac{S_n}{n}\right) - h(a)\right),$$

$$H'(a)\xi = h'(a)^{\mathrm{T}}\xi \sim \langle h'(a), \mathcal{N}(0, \Sigma)\rangle.$$

3. Лекция 3

Статистики и оценки

Пусть $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mathcal{P})$ - вероятностно-статистическая модель. Пусть $\mathbf{X}$ - наблюдение в этой модели. Пусть $(E, \mathcal{E})$ - измеримое пространство.

Определение 3.1. Статистикой $S(\mathbf{X})$ называется измеримая функция от наблюдения $\mathbf{X}$. Напомним, что отображение $S : \mathcal{X} \longrightarrow E$ является измеримым, если для любого множества $B \in \mathcal{E}$ выполнено

$$S^{-1}(B) = \{x : S(x) \in B\} \in \mathcal{B}_{\mathcal{X}}.$$

Важный момент: если модель является параметрической, то отображение S не должно зависеть от параметра!

Пусть $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mathcal{P})$, $\mathcal{P} = \{\mathrm{P}_\theta : \theta \in \Theta\}$ - параметрическая модель. Пусть $\mathbf{X}$ - наблюдение в этой модели.

Определение 3.2. Если статистика $S(\mathbf{X})$ принимает значения в Θ, то ее можно назвать **оценкой** неизвестного параметра $\theta \in \Theta$. Можно также оценивать функции от параметра, $\tau(\theta)$. В этом случае $S(\mathbf{X})$ должна принимать значения в $\tau(\Theta)$.

Замечание
Мы будем использовать данные два термина почти как эквивалентные.

Примеры статистик и оценок

Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка, $X_i \in \mathbb{R}$.

Общая идея построения хороших статистик.
Если Q - распределение вероятностей на прямой, а $G(\mathrm{Q})$ - некоторый функционал, то $G(\mathrm{P}_n^*)$ - потенциально хорошая статистика. Здесь P_n^* - эмпирическое распределение, построенное по выборке $\mathbf{X} = (X_1, ..., X_n)$.

1) **Выборочные усреднения**.
   Пусть $g(x)$ - некоторая борелевская функция $\mathbb{R}$, а

$$G(\mathrm{Q}) = \int_{\mathbb{R}} g(x)\mathrm{Q}(dx).$$

   Тогда, если подставить в качестве аргумента эмпирическое распределение (равномерное распределение на элементах выборки), то мы по сути получим математическое ожидание функции $g(x)$ по этому распределению, то есть

$$G(\mathrm{P}_n^*) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) = \overline{g(\mathbf{X})}$$

- среднее значение функции $g(x)$ по выборке. Хорошие примеры:

- $g(x) = x$, тогда $\overline{\mathbf{X}} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ - выборочное среднее;
- $g(x) = x^k$, тогда $\overline{\mathbf{X}^k} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^k$ - выборочный момент порядка k.

2) **Функции от выборочных усреднений**.
Пусть $g_1(x), ..., g_k(x)$ - борелевские функции на $\mathbb{R}$, а $h : \mathbb{R}^k \longrightarrow \mathbb{R}$ - борелевская функция нескольких переменных, тогда можно рассмотреть

$$G(\mathrm{P}_n^*) = h\Big(\overline{g_1(\mathbf{X})}, ..., \overline{g_k(\mathbf{X})}\Big).$$

Хорошие примеры:

- $h(x, y) = x - y^2$, $g_1(x) = x^2$, $g_2(x) = x$, тогда получаем статистику $S^2 = \overline{\mathbf{X}^2} - (\overline{\mathbf{X}})^2$ - выборочную дисперсию;

Упражнение
Проверить, что

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{\mathbf{X}})^2.$$

- $M_k = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{\mathbf{X}})^k$ - выборочный центральный момент порядка k.

3) **Порядковые статистики**.
Введем обозначения (упорядочим выборку):

$$\begin{aligned}
&X_{(1)} = \min\{X_1, ..., X_n\},\\
&X_{(2)} = \min\{(X_1, ..., X_n) \setminus X_{(1)}\},\\
&\vdots\\
&X_{(n)} = \max\{X_1, ..., X_n\}.
\end{aligned}$$

Вектор $(X_{(1)}, ..., X_{(n)})$ называется вариационным рядом.

4) **Выборочные квантили.**

Определение 3.3. Пусть распределение Q имеет функцию распределения $F(x)$, а $p \in (0, 1)$. Тогда $p-$квантилью распределения Q называется

$$\zeta_p = \min\{x : F(x) \geqslant p\}.$$

Если $F(x)$ непрерывна, то для любого $p \in (0, 1)$ будет выполнено $F(\zeta_p) = p$.

Изобразим, как может выглядеть эта величина. Возьмем произвольную функцию распределения и рассмотрим три принципиальных случая.

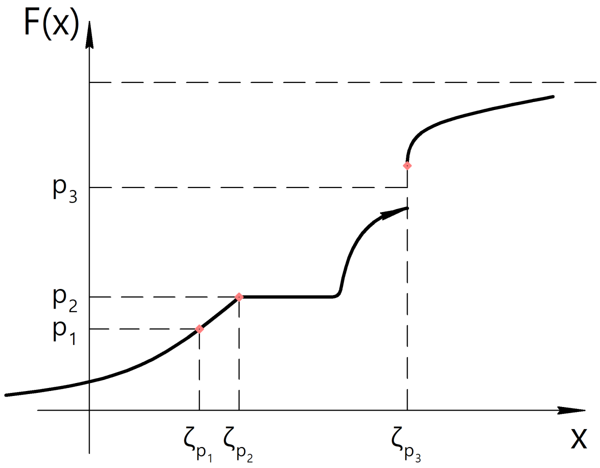

Рис. 2

Первый случай: ζ_{p_1} - квантиль честная, то есть у нас ровно одна точка, в которой достигается значение p_1.
Второй случай: если у функции распределения есть отрезок постоянства (множество точек, в которых значение ровно p_2). Тогда среди этих точек мы берем наименьшую - ζ_{p_2}.
Третий случай: в точке p_3 значение не достигается. Тогда в качестве квантиля мы берем наименьшую точку, в которой значение превышает p_3 - ζ_{p_3}.

Определение 3.4. Выборочной p-квантилью называется p-квантиль эмпирического распределения P_n^*, то есть

$$Z_{n,p} = \begin{cases} X_{([np]+1)}, & \text{если } np \notin \mathbb{Z}; \\ X_{(np)}, & \text{если } np \in \mathbb{Z}. \end{cases}$$

Напомним, что эмпирическое распределение - дискретное, сосредоточенное в точках $X_1, ..., X_n$ и принимающее значения равновероятно, то есть эмпирическая функция распределения идет "лесенкой" и принимает значения вида $\frac{k}{n}$. Соответственно, почти всегда у нас будет третий случай, когда нет точки, в которой значение нашей функции распределения в точности равно p.

Замечание: в качестве функционала выступает квантиль распределения.

5) **М-оценки**.
Пусть $\psi(x, \theta)$, $\theta \in \Theta$ - некоторый набор функций. Тогда M-оценкой называется

статистика вида

$$S(X) = arg\max_{\theta\in\Theta}\left(\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\theta)\right).$$

Примером такой оценки будет оценка максимального правдоподобия.

Свойства оценок

Теперь поймем, какие мы хотели бы видеть свойства у оценок, чтобы можно было рассматривать их в качестве хороших.

Несмещённость

Пусть $\mathbf{X}$ - наблюдение с неизвестным распределением $\mathrm{P}\in\{\mathrm{P}_\theta : \theta\in\Theta\}$, где $\theta\in\mathbb{R}^k$.

Определение 3.5. Оценка $\hat{\theta}(\mathbf{X})$ называется несмещённой оценкой параметра θ, если для любого $\theta\in\Theta$ выполняется равенство

$$\mathrm{E}_\theta\hat{\theta}(\mathbf{X}) = \theta.$$

Вспомним, что запись E_θ означает, что при взятии математического ожидания мы предполагаем, что выборка $\mathbf{X}$ взята из распределения P_θ:

$$\mathrm{E}_\theta\hat{\theta}(\mathbf{X}) = \int\limits_{\mathcal{X}}\hat{\theta}(x)\mathrm{P}_\theta(dx).$$

Пример
Пусть $\mathbf{X} = (X_1,...,X_n)$ - выборка из $\mathcal{N}(\theta,1)$. Тогда $\overline{\mathbf{X}}$ (выборочное среднее) и X_1 - несмещённые оценки θ.

Состоятельность

В асимптотических вопросах мы предполагаем, что $\mathbf{X}$=$(X_1,...,X_n)$ - это выборка неограниченного размера из неизвестного распределения $\mathrm{P}\in\{\mathrm{P}_\theta : \theta\in\Theta\}$, $\Theta\in\mathbb{R}^k$.

Определение 3.6. Оценка $\hat{\theta}_n(X_1,...,X_n)$ (точнее - последовательность оценок) называется состоятельной оценкой параметра θ, если для любого $\theta\in\Theta$

$$\hat{\theta}(X_1,...,X_n)\xrightarrow{\mathrm{P}_\theta}\theta,\quad n\to\infty.$$

Запись означает, что для любого $\theta\in\Theta$ и для любого $\varepsilon>0$

$$\mathrm{P}_\theta\Bigg(\|\hat{\theta}(X_1,...,X_n)-\theta\|\geqslant\varepsilon\Bigg)\to 0,\quad n\to\infty.$$

Пример
Пусть $\mathbf{X}(X_1,...,X_n)$ - выборка из $\mathcal{N}(\theta,1)$. Тогда по ЗБЧ оценка $\overline{\mathbf{X}}$ является состоятельной.

Сильная состоятельность

Определение 3.7. Оценка $\hat{\theta}_n(X_1, ..., X_n)$ (точнее - последовательность оценок) называется сильно состоятельной оценкой параметра θ, если для любого $\theta \in \Theta$

$$\hat{\theta}(X_1, ..., X_n) \xrightarrow{\mathrm{P}_\theta - \text{п.н.}} \theta, \quad n \to \infty.$$

Запись означает, что для любого $\theta \in \Theta$

$$\mathrm{P}_\theta\left(\lim_{n\to\infty} \hat{\theta}_n(X_1, ..., X_n) = \theta\right) = 1.$$

Пример
Пусть $\mathbf{X}(X_1, ..., X_n)$ - выборка из $\mathcal{N}(\theta, 1)$. Тогда по УЗБЧ оценка $\overline{\mathbf{X}}$ является сильно состоятельной.

Асимптотическая нормальность

Ограничимся одномерным вариантом. Пусть $\Theta \subset \mathbb{R}$.

Определение 3.8. Оценка $\hat{\theta}_n(X_1, ..., X_n)$ (точнее - последовательность оценок) называется асимптотически нормальной оценкой параметра θ, если для любого $\theta \in \Theta$

$$\sqrt{n}\left(\hat{\theta}_n(X_1, ..., X_n) - \theta\right) \xrightarrow{d_\theta} \mathcal{N}(0, \sigma^2(\theta)), \quad n \to \infty.$$

Функция $\sigma^2(\theta)$ называется асимптотической дисперсией оценки $\hat{\theta}_n$.

Запись означает, что для любого $\theta \in \Theta$ и для любого $\in \mathbb{R}$ выполнено

$$\lim_{n\to\infty} \mathrm{P}_\theta\left(\sqrt{n}\left(\hat{\theta}_n(X_1, ..., X_n) - \theta\right) \leqslant x\right) = \Phi\left(\frac{x}{\sigma(\theta)}\right),$$

где Φ - функция Лапласа (функция распределения стандартного нормального закона).

Пример
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из $\mathrm{Bin}(1, \theta)$. Тогда по ЦПТ оценка $\overline{\mathbf{X}}$ является асимптотически нормальной.

Замечания

1) Асимптотические свойства (состоятельность, сильная состоятельность и асимптотическая нормальность) имеют смысл только в случае выборки большого размера. Несмещённость - это "конечное" свойство.

2) Оценивать можно не только сам параметр θ, но и функции от него, $\tau(\theta)$. Определения свойств при этом сохраняются.

Наследование свойств

Вопрос: пусть $\hat{\theta}_n(X_1, ..., X_n)$ хорошо приближает θ, как тогда хорошо приблизить $\tau(\theta)$?

Допустим, что $\hat{\theta}_n$ является состоятельной оценкой. В качестве хорошей оценки для $\tau(\theta)$ можно взять $\tau(\hat{\theta}_n)$:

Утверждение (1)

Пусть $\hat{\theta}_n(X_1, ..., X_n)$ - сильно состоятельная (состоятельная) оценка параметра θ. Если $\tau(\theta)$ непрерывна на $\Theta \subset \mathbb{R}^k$, то $\tau(\hat{\theta}_n(X_1, ..., X_n))$ будет сильно состоятельной (состоятельной) оценкой $\tau(\theta)$.

Доказательство. Сразу следует из теоремы о наследовании сходимости. ∎

Утверждение (2)

Пусть $\hat{\theta}_n(x_1, ..., X_n)$ - асимптотически нормальная оценка параметра $\theta \in \mathbb{R}$ с асимптотической дисперсией $\sigma^2(\theta)$. Пусть $\tau(\theta)$ дифференцируема на $\Theta \subset \mathbb{R}$.

Тогда $\tau(\hat{\theta}_n(X_1, ..., X_n))$ - асимптотически нормальная оценка параметра $\tau(\theta)$ с асимптотической дисперсией $\sigma^2(\theta) \cdot (\tau'(\theta))^2$.

Доказательство.

По условию для любого $\theta \in \Theta$ выполнено

$$\sqrt{n}\left(\hat{\theta}_n(X_1, ..., X_n) - \theta\right) \xrightarrow{d_\theta} \mathcal{N}(0, \sigma^2(\theta)).$$

Применяя дельта-метод, получаем, что

$$\sqrt{n}\left(\tau(\hat{\theta}_n(X_1, ..., X_n)) - \tau(\theta)\right) \xrightarrow{d_\theta} \tau'(\theta) \cdot \mathcal{N}(0, \sigma^2(\theta)).$$

Остается заметить, что $\tau'(\theta) \cdot \mathcal{N}(0, \sigma^2(\theta)) = \mathcal{N}(0, \sigma^2(\theta) \cdot (\tau'(\theta))^2)$. ∎

Замечание

Утверждения работают и в обратную сторону. Если мы умеем хорошо оценивать $\tau(\theta)$, существует обратная функция τ^{-1} и она непрерывна или дифференцируема.

Пример

Пусть $X_1, ..., X_n$ - выборка из экспоненциального распределения $\mathrm{Exp}(\theta)$, $\theta > 0$. Найти асимптотически нормальную оценку θ.

Решение

Плотность X_i равна $p_\theta(x) = \theta \cdot e^{-\theta x}\mathrm{I}\{x > 0\}$. Отсюда

$$\mathrm{E}_\theta X_i = \frac{1}{\theta}, \quad \mathrm{D}_\theta X_i = \frac{1}{\theta^2}\ .$$

Согласно ЦПТ

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

Положим $\tau(x) = \dfrac{1}{x}$. Тогда по утверждению 2

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

или

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

Взаимоотношение свойств

Между свойствами есть следующие простые соотношения:

$$\text{Сильная состоятельность} \quad \Rightarrow \quad \text{Состоятельность;}$$

$$\text{Асимптотическая нормальность} \quad \Rightarrow \quad \text{Состоятельность;}$$

Методы построения оценок

Вопрос: какие можно предложить методы построения оценок с хорошими (прежде всего - асимптотическими) свойствами?

Принцип подстановки

Пусть для параметрического семейства $\{\mathrm{P}_\theta, \theta \in \Theta\}$ нашелся такой функционал G, что для всех $\theta \in \Theta$ выполняется равенство

$$G(\mathrm{P}_\theta) = \theta.$$

Тогда оценкой θ по методу подстановки называется

$$\theta^*(X_1, ..., X_n) = G(\mathrm{P}_n^*),$$

где P_n^* - эмпирическое распределение, построенное по выборке $\mathbf{X} = (X_1, ..., X_n)$.

Вопрос: какие функционалы можно брать?

Идея: моменты определяют параметр распределения. То есть, у нас есть параметрическое семейство. Посчитаем моменты, которые являются функциями параметра. Если эти функции однозначно определяют параметр, то мы можем взять моменты в качестве функционала.

Метод моментов

Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из неизвестного распределения $\mathrm{P} \in \{\mathrm{P}_\theta, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$. Пусть борелевские функции $g_1(x), ..., g_k(x)$ таковы, что набор значений

$$\mathrm{E}_\theta g_1(X_1), ..., \mathrm{E}_\theta g_k(X_1)$$

однозначно определяет параметр θ. Это означает, что функция $m : \Theta \to \mathbb{R}^k$, где

$$m_i(\theta) = \mathrm{E}_\theta g_i(X_1), \quad i = 1, ..., k,$$

является биекцией между Θ и $m(\Theta)$.

Определение 3.9. Оценкой θ по методу моментов с пробными функциями $g_1(x), ..., g_k(x)$ называется $\theta^*(X_1, ..., X_n)$ - решение следующей системы уравнений относительно θ:

$$\begin{cases} m_1(\theta) = \overline{g_1(\mathbf{X})}, \\ \vdots \\ m_k(\theta) = \overline{g_k(\mathbf{X})}. \end{cases}$$

Другими словами:

$$\theta^*(X_1, ..., X_n) = m^{-1}\Big(\overline{g_1(\mathbf{X})}, ..., \overline{g_k(\mathbf{X})}\Big).$$

Замечание (1)
Если множество Θ не является открытым, то вполне может оказаться, что вектор $\Big(\overline{g_1(\mathbf{X})}, ..., \overline{g_k(\mathbf{X})}\Big)$ с большой положительной вероятностью не попадет в область определения функции m^{-1}. Так что разумно считать, что θ принимает все возможные значения.

Замечание (2)
Обычно используют стандартные пробные функции

$$g_1(x) = x, ..., g_k(x) = x^k.$$

Пример
Пусть $(X_1, ..., X_n)$ - выборка из нормального распределения $\mathcal{N}(a, \sigma^2)$, $a \in \mathbb{R}$, $\sigma^2 > 0$, $\theta = (a, \sigma^2)$. Найти оценку θ по методу моментов.

Решение

Вспомним, что

$$\mathrm{E}_\theta X_1 = a, \quad \mathrm{E}_\theta X_1^2 = a^2 + \sigma^2.$$

Получаем систему

$$\begin{cases} a^* = \overline{\mathbf{X}} \\ (a^*)^2 + (\sigma^*)^2 = \overline{\mathbf{X}^2} \end{cases}$$

Решая, находим ответ:

$$a^* = \overline{\mathbf{X}}, \quad (\sigma^*)^2 = \overline{\mathbf{X}^2} - (\overline{\mathbf{X}})^2 = S^2.$$

Теорема 3.1. (состоятельность оценки метода моментов)
Пусть $\theta_n^*(X_1, ..., X_n) = m^{-1}\left(\overline{g_1(\mathbf{X})}, ..., \overline{g_k(\mathbf{X})}\right)$ - корректно определенная оценка по методу моментов. Если функция m^{-1} непрерывна на множестве $m(\Theta)$, то θ_n^* - сильно состоятельная оценка параметра θ.

Доказательство.
Зафиксируем произвольное $\theta \in \Theta \subset \mathbb{R}^k$. Тогда согласно УЗБЧ для любого $j = 1, ..., k$

$$\overline{g_j(\mathbf{X})} = \frac{1}{n}\sum_{i=1}^n g_j(X_i) \xrightarrow{\mathrm{P}_\theta-\text{п.н.}} \mathrm{E}_\theta g_j(X_1) = m_j(\theta).$$

По теореме о наследовании сходимости получаем:

$$\theta_n^*(X_1, ..., X_n) = m^{-1}\left(\overline{g_1(\mathbf{X})}, ..., \overline{g_k(\mathbf{X})}\right) \xrightarrow{\mathrm{P}_\theta-\text{п.н.}} m^{-1}(m_1(\theta), ..., m_k(\theta)) = m^{-1}(m(\theta)) = \theta$$

■

Упражнение
Если в дополнение к условия теоремы для любого $\theta \in \Theta \subset \mathbb{R}^k$ и любого $j = 1, ..., k$, конечно $\mathrm{E}_\theta g_j^2(X_1)$, а функция m^{-1} дифференцируема на $m(\Theta)$, то для каждого $j = 1, ..., k$ оценка $\theta_{j,n}^*(X_1, ..., X_n)$ параметра θ_j по методу моментов является асимптотически нормальной.

Итак, мы увидели, что метод моментов поставляет состоятельные и асимптотически нормальные оценки параметра. Однако у него есть целый ряд недостатков:

- Подбор пробных функций. Если стандартные не подошли, то возникает проблема поиска подходящих функций. Например, рассмотрим распределение Коши со сдвигом: у нас есть выборка $(X_1, ..., X_n)$, и мы взяли плотность распределения Коши со сдвигом $p_\theta(x) = \dfrac{1}{\pi(1 + (x - \theta)^2)}$, $\theta \in \mathbb{R}$. Тогда математическое ожидание $\mathrm{E}_\theta X_1$ не определено, и, соответственно, ни одна из стандартных функций не может быть использована. Можно попробовать взять

$\mathrm{E}_\theta \ln(|X_1| + 1)$. Такое математическое ожидание конечно, но метод моментов требует, чтобы мы могли посчитать эту функцию, как функцию от параметра. Это довольно сложно.

- Вторая трудность применения метода в том, что для нахождения оценки нужно знать функцию $m(\theta)$, что с практической точки зрения не представляется возможным.

- Получается, что применимость метода ограничена стандартными распределениями, для которых возможно аналитическое вычисление моментов, как функций от параметров.

На следующей лекции

Вернемся к примеру с распределением Коши со сдвигом. Забудем про идею с логарифмом. Что в таком случае можно предложить в качестве оценки параметра, если метод моментов не работает? Другая идея состоит в том, чтобы взять другой функционал. Для данного распределения θ не является математическим ожиданием, но оно является $\frac{1}{2}$-квантилью, потому что плотность симметрична относительно θ, соответственно в точке θ функция распределения будет достигать значения $\frac{1}{2}$. Можем ли мы надеяться на то, что выборочная $\frac{1}{2}$-квантиль хорошо приближает наш параметр θ? На следующей лекции мы разберем теорему, которая доказывает, что выборочные квантили хорошо приближают настоящие, соответственно их можно использовать в качестве настоящих квантилей. Кроме того, в следующий раз мы посмотрим на принципы, по которым мы можем сравнивать оценки и разберем разные подходы к сравнению оценок.

4. Лекция 4

Сегодняшняя лекция будет посвящена продолжению изучения методов получения хороших оценок, а также принципам сравнению оценок.

Выборочные квантили

На прошлой лекции мы познакомились с методом моментов. Но есть распределения, где подбор нужных пробных функций крайне затруднен.

Например, пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из распределения Коши со сдвигом $\theta \in \mathbb{R}$: плотность

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

Все стандартные моменты у такого распределения либо не определены, либо бесконечны. Применение других функций выглядит безнадежным. Можно ли что-то предложить в качестве альтернативы методу моментов?

Идея: квантили определяют параметр распределения.

В примере θ является $\frac{1}{2}$-квантилью распределения (медианой распределения).

Напомним, что если $\mathbf{X} = (X_1, ..., X_n)$ - выборка, то выборочной p-квантилью называется p-квантиль эмпирического распределения P_n^*, то есть

$$Z_{n,p} = \begin{cases} X_{([np]+1)}, & \text{если } np \notin \mathbb{Z}; \\ X_{(np)}, & \text{если } np \in \mathbb{Z}. \end{cases}$$

Первой целью лекции будет доказательство теоремы о том, что выборочная квантиль является асимптотически нормальной оценкой настоящей квантили.

Теорема 4.1. (асимптотическая нормальность выборочной квантили)
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка растущего размера из распределения с функцией распределения $F(x)$. Пусть $z_p, p \in (0,1)$ - его p-квантиль, причем F дифференцируема в точке z_p и $F'(z_p) > 0$. Тогда утверждается, что $Z_{n,p}$ будет асимптотически нормальной оценкой настоящей квантили z_p, и выполнена следующая сходимость:

$$\sqrt{n}(Z_{n,p} - z_p) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{(F'(z_p))^2}\right) \quad \text{при } n \to +\infty.$$

Доказательство.
Доказательство будет опираться на следующий факт.

Упражнение
Пусть дана последовательность биномиальных случайных величин $\xi_n \sim \mathrm{Bin}(n, p_n)$, $n \in \mathbb{N}$, причем $p_n \to p \in (0, 1)$. Тогда

$$\frac{\xi_n - \mathrm{E}\xi_n}{\sqrt{\mathrm{D}\xi_n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Утверждение легко выводится из центральной предельной теоремы в схеме серий, но может быть доказано непосредственно методом характеристических функций.

Идея состоит в том, чтобы свести интересующую нас сходимость к данному упражнению. Для этого рассмотрим событие $\{X_{(k)} \leqslant x\}$ и заметим, что оно означает, что не менее k из случайных величин $X_1, ..., X_n$ попали в промежуток $(-\infty, x]$. Следовательно, его можно представить следующим образом:

$$\{X_{(k)} \leqslant x\} = \left\{\sum_{i=1}^{n} \mathrm{I}\{X_i \leqslant x\} \geqslant k\right\},$$

то есть хотя бы k элементов выборки попали на полуинтервал $(-\infty, x]$. Здесь уже видна связь необходимого нам события с упражнением, так как это сумма независимых индикаторов, то есть схема Бернулли, только в нашем случае вероятность того, что $X_i \leqslant x$ будет как-то зависеть от n. Запишем интересующую нас вероятность:

$$\mathrm{P}(\sqrt{n}(X_{n,p} - z_p) \leqslant x) = \mathrm{P}\left(Z_{n,p} \leqslant z_p + \frac{x}{\sqrt{n}}\right) = \mathrm{P}\left(\sum_{i=1}^{n} \mathrm{I}\left\{X_i \leqslant z_p + \frac{x}{\sqrt{n}}\right\} \geqslant np\right).$$

Обозначим

$$Y_n = \sum_{i=1}^{n} \mathrm{I}\left\{X_i \leqslant z_p + \frac{x}{\sqrt{n}}\right\}.$$

Это биномиальная случайная величина со следующими характеристиками:

$$\mathrm{E}Y_n = nF\left(z_p + \frac{x}{\sqrt{n}}\right) \sim nF(z_p) = np,$$

$$\mathrm{D}Y_n = nF\left(z_p + \frac{x}{\sqrt{n}}\right)\left(1 - F\left(z_p + \frac{x}{\sqrt{n}}\right)\right) \sim np(1-p).$$

Согласно упражнению получаем, что

$$\frac{Y_n - \mathrm{E}Y_n}{\sqrt{\mathrm{D}Y_n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Далее,

$$\mathrm{P}(Y_n \geqslant np) = \mathrm{P}\left(\frac{Y_n - \mathrm{E}Y_n}{\sqrt{\mathrm{D}Y_n}} \geqslant \frac{np - \mathrm{E}Y_n}{\sqrt{\mathrm{D}Y_n}}\right).$$

Найдем асимптотику выражения $\dfrac{np - \mathrm{E}Y_n}{\sqrt{\mathrm{D}Y_n}}$. Заметим, что

$$\mathrm{E}Y_n = n \cdot F\left(z_p + \frac{x}{\sqrt{n}}\right) = n\left(F(z_p) + \frac{x}{\sqrt{n}}F'(z_p) + o(n^{-\frac{1}{2}})\right) = np + xF'(z_p)\sqrt{n} + o(\sqrt{n}).$$

Значит,

$$\frac{np - \mathrm{E}Y_n}{\sqrt{\mathrm{D}Y_n}} = \frac{-\, xF'(z_p)\sqrt{n} + o(\sqrt{n})}{\sqrt{np(1-p)(1+o(1))}} = -\ \frac{xF'(z_p)}{\sqrt{p(1-p)}} + o(1).$$

Следовательно,

$$\lim_{n\to+\infty} \mathrm{P}(\sqrt{n}(Z_{n,p} - z_p) \leqslant x) = 1 - \Phi\left(-\ \frac{xF'(z_p)}{\sqrt{p(1-p)}}\right) = \Phi\left(\frac{xF'(z_p)}{\sqrt{p(1-p)}}\right),$$

что и означает искомую сходимость:

$$\sqrt{n}(Z_{n,p} - z_p) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{(F'(z_p))^2}\right) \quad \text{при } n \to +\infty.$$

■

Выборочная медиана

Определение 4.1. Медианой распределения называется его $\frac{1}{2}$-квантиль. Выборочной медианой $\hat{\mu} = \hat{\mu}(X_1, ..., X_n)$ выборки $(X_1, ..., X_n)$ называется величина

$$\hat{\mu}(X_1, ..., X_n) = \begin{cases} X_{(k+1)}, & \text{если } n = 2k+1; \\ \dfrac{X_{(k)} + X_{(k+1)}}{2}, & \text{если } n = 2k. \end{cases}$$

Следствие
В условиях теоремы об асимптотической нормальности выборочной квантили выполнено

$$\sqrt{n}(\hat{\mu} - z_{1/2}) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4(F'(z_{1/2}))^2}\right).$$

Доказательство.
Отметим, что доказательство теоремы остается верным, если вместо $Z_{n,p}$ рассматривать любую порядковую статистику $X_{(k)}$ при $k = np + O(1)$, то есть

$$\sqrt{n}(X_{(k)} - z_p) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{(F'(z_p))^2}\right) \quad \text{при} \quad n \to +\infty.$$

Отметим, что для любого $x \in \mathbb{R}$ по подпоследовательности четных чисел $n = 2k$ выполнено

$$\lim_{k\to+\infty} \mathrm{P}(\sqrt{n}(X_{(k+1)} - z_{1/2}) \leqslant x) \leqslant$$
$$\leqslant \lim_{k\to+\infty} \mathrm{P}(\sqrt{n}(\hat{\mu} - z_{1/2}) \leqslant x) \leqslant$$
$$\leqslant \lim_{k\to+\infty} \mathrm{P}(\sqrt{n}(X_{(k)} - z_{1/2}) \leqslant x),$$

что и доказывает искомую сходимость, так как крайние пределы равны. ■

Пример
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из распределения Коши со сдвигом $\theta \in \mathbb{R}$: плотность

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

Заметим, что наше распределение подходит под условия теоремы: плотность непрерывна и положительна в точке θ. Тогда θ - это медиана распределения и, стало быть, по теореме выборочная медиана будет асимптотически нормальной (и состоятельной!) оценкой параметра:

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

Сравнение оценок

Мы уже предъявили несколько методов, которые позволяют находить оценки с хорошими свойствами. Пусть $\hat{\theta}(\mathbf{X})$, $\theta^*(\mathbf{X})$ - две хорошие оценки параметра $\theta \in \Theta$. Как понять, какая из них лучше оценивает θ? Для этого нужны способы сравнения оценок. В рамках курса мы разберем 4 подхода и будем решать задачи, связанные с нахождением наилучших оценок в этих подходах. 3 из них основаны на работе с функцией риска, 4ый подход связан со сравнением асимптотически нормальных оценок. В первую очередь поймем, какие могут быть способы определения близости оценки и параметра.

Определение 4.2. Борелевская неотрицательная функция $\rho(x, y) \geqslant 0$ называется функцией потерь. Если $\theta^*(\mathbf{X})$, то ее величиной потерь называется $\rho(\theta^*(\mathbf{X}), \theta)$.

Примеры

- $\rho(x, y) = |x - y|$;

- $\rho(x, y) = (x - y)^2$ - квадратичная функция потерь;
- если $\theta \in \mathbb{R}^d$, $d > 1$, то для неотрицательно определенной матрицы $A \in \mathrm{Mat}(d \times d)$ можно определить

$$\rho(x, y) = \langle A(x - y), x - y \rangle$$

Функция риска

Сама по себе функция потерь мало что нам дает, так как она зависит от параметра, зависит от x и является случайной. Поэтому введем понятие функции риска, которая позволяет избавиться хотя-бы от одной зависимости.

Определение 4.3. При заданной функции потерь $\rho(x, y) \geqslant$ функцией риска оценки $\theta^*(\mathbf{X})$ называется

$$R_{\theta^*}(\theta) = \mathrm{E}_\theta \rho(\theta^*(\mathbf{X}), \theta).$$

Базовое сравнение оценок, в первую очередь, основано на сравнении их функций риска. Здесь есть сразу несколько подходов, мы рассмотрим три основных: равномерный, байесовский и минимаксный.

Равномерный подход

Правило сравнения. Считаем, что оценка $\theta^*(\mathbf{X})$ лучше оценивает параметр θ, чем оценка $\hat{\theta}(\mathbf{X})$, если для любого $\theta \in \Theta$ выполнено

$$R_{\theta^*}(\theta) \leqslant R_{\hat{\theta}}(\theta),$$

причем для некоторого $\theta \in \Theta$ неравенство строгое.

Определение 4.4. Оценка $\theta^*(\mathbf{X})$ называется наилучшей в классе $\mathcal{K}$, если она лучше любой другой оценки $\hat{\theta}(\mathbf{x}) \in \mathcal{K}$.

Почему необходимо рассматривать именно классы? Оказывается, что постановка задачи сравнения во всем классе оценок без ограничений бессмысленна.

Замечание
Наилучшая оценка не всегда существует. Например, если $\mathcal{K}$ - это класс всех возможных оценок, $\theta \subset \mathbb{R}$ и $\rho(x, y) = (x - y)^2$ - это квадратичная функция потерь, то наилучшей оценки нет.

Действительно, можно для любого $\theta_0 \in \Theta$ рассмотреть оценку $\hat{\theta}_0(\mathbf{X}) \equiv \theta_0$, то ее функция риска в точке θ_0 обращается в ноль:

$$R_{\hat{\theta}_0}(\theta_0) = \mathrm{E}_{\theta_0}\left(\hat{\theta}_0(\mathbf{X}) - \theta_0\right)^2 = 0.$$

Тем самым, функция риска потенциальной наилучшей оценки равна тождественному нулю, то есть сама она равна всем константам одновременно. Противоречие.

Отсюда возникает вопрос: какие классы разумно рассматривать? Задача о поиске наилучшей оценки в равномерном подходе имеет смысл в классе оценок с одинаковым математическим ожиданием. именно в нем мы будем ее решать.

Определение 4.5. Равномерный подход с квадратичной функцией потерь называется среднеквадратичным.

Определение 4.6. Оценка $\theta^*(\mathbf{X})$ называется допустимой оценкой θ, если нет другой оценки $\hat{\theta}_0$, которая оценивает θ лучше, чем $\theta^*(\mathbf{X})$, в равномерном подходе.

Байесовский подход

Правило сравнения. Пусть Q - некоторое заданное распределение вероятностей на Θ (априорное распределение). Тогда для оценки $\theta^*(\mathbf{X})$ с функцией риска $R_{\theta^*}(\theta)$ определим

$$\rho_{\mathrm{Q}}(\theta^*) = \mathrm{E}_{\mathrm{Q}} R_{\theta^*}(\theta) = \int\limits_{\Theta} R_{\theta^*}(t)\mathrm{Q}(dt).$$

Например, если Q имеет плотность $q(t)$ по мере λ, то

$$\rho_{\mathrm{Q}}(\theta^*) = \int\limits_{\Theta} R_{\theta^*}(t)q(t)\lambda(dt).$$

Определение 4.7. Оценка $\theta^*(\mathbf{X})$ называется наилучшей оценкой θ в байесовском подходе для заданного априорного распределения Q (байесовской оценкой), если

$$\rho_{\mathrm{Q}}(\theta^*) = \inf_{\hat{\theta}(\mathbf{X})} \rho_{\mathrm{Q}}(\hat{\theta}(\mathbf{X})).$$

Байесовский подход интересен тем, что мы получаем серию оценок. Понятно, что наилучшая оценка зависит от априорного распределения Q. Эти оценки интересны тем, что они являются допустимыми.

Упражнение
Пусть оценка $\theta^*(\mathbf{X})$ является наилучшей оценкой θ в баейсовском подходе для некоторого априорного распределения Q. Тогда $\theta^*(\mathbf{X})$ - это допустимая оценка θ.

Минимаксный подход

Правило сравнения. Пусть $\theta^*(\mathbf{X})$ - оценка θ с функцией риска $R_{\theta^*}(\theta)$. Тогда определим

$$\rho_{max}(\theta^*) = \sup_{\theta\in\Theta} R_{\theta^*}(\theta).$$

Определение 4.8. Оценка $\theta^*(\mathbf{X})$ называется наилучшей оценкой θ в минимаксном подходе (минимаксной оценкой), если

$$\rho_{max}(\theta^*) = \inf_{\hat{\theta}(\mathbf{X})} \rho_{max}(\hat{\theta}),$$

то есть у $\theta^*(\mathbf{X})$ наименьшее максимальное значение функции риска.

Как мы увидим в дальнейшем, минимаксный и байесовский подходы тесно связаны.

Проиллюстрируем на чем основаны подходы. Предположим, что $\Theta = [0, 1]$, априорное распределение Q= $\cup[0, 1]$ (равномерное распределение на отрезке), и нарисуем примеры 2ух функций риска. Попытаемся сравнить оценки θ^* и $\hat{\theta}$ во всех подходах.

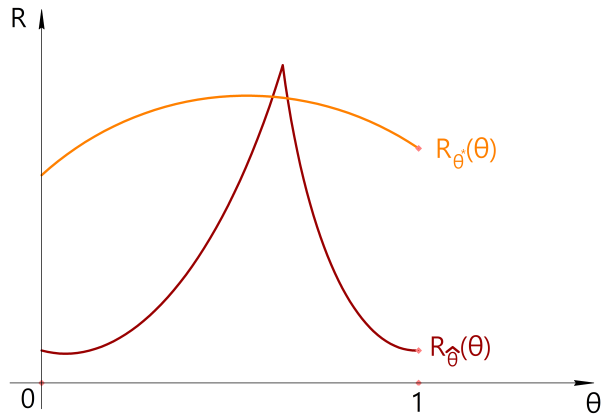

Рис. 3

Мы можем сделать следующие выводы:

- θ^* лучше в минимаксном подходе, так как у нее максимальное значение функции риска ниже;
- $\hat{\theta}$ в свою очередь лучше в байесовском подходе, так как интегральное значение - это площадь графика;
- В равномерном подходе оценки не сравнимы.

Асимптотический подход

Это принципиально другой подход к сравнению оценок. Он апеллирует к асимптотике, то есть мы имеем дело с последовательностями асимптотически нормальных оценок. Этот подход практически никак не связан с предыдущими и не использует функции риска.
Правило сравнения. Пусть $\hat{\theta}_{n,1}(X_1, ..., X_n)$ и $\hat{\theta}_{n,2}(X_1, ..., X_n)$ - две асимптотически

нормальные оценки $\theta \in \mathbb{R}$ с асимптотическими дисперсиями $\sigma_1^2(\theta)$ и $\sigma_2^2(\theta)$. Тогда оценка $\hat{\theta}_{n,1}(X_1, ..., X_n)$ лучше оценивает параметр θ, чем оценка $\hat{\theta}_{n,2}(X_1, ..., X_n)$, если для любого $\theta \in \Theta$ выполнено

$$\sigma_1^2(\theta) \leqslant \sigma_2^2(\theta),$$

причем для некоторого $\theta \in \Theta$ неравенство строгое.

Определение 4.9. Оценка $\hat{\theta}_n(X_1, ..., X_n)$ называется наилучшей в асимптотическом подходе, если она лучше любой другой оценки.

Замечание
Мы увидим, что получить совсем честное решение здесь невозможно. Какова бы ни была асимптотически нормальная оценка с положительной асимптотической дисперсией, всегда можно подобрать другую оценку, у которой асимптотическая дисперсия будет меньше хотя бы в одной точке.

Пример
Пусть $X_1, ..., X_n$ - выборка из $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. Сравните в асимптотическом подходе оценки $\overline{\mathbf{X}}$ и $\hat{\mu}$.

Решение

Обе оценки являются асимптотически нормальными. Асимптотическая нормальность $\overline{\mathbf{X}}$ следует из ЦПТ

$$\sqrt{n}(\overline{\mathbf{X} - \theta} \xrightarrow{d} \mathcal{N}(0, 1).$$

По теореме об асимптотической нормальности выборочной квантили получаем:

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

Вывод: выборочное среднее лучше.

Дальнейшая цель: научиться строить наилучшие оценки в разных подходах.

5. Лекция 5

Сегодняшняя лекция будет посвящена неравенству Рао-Крамера. Мы познакомимся с тем, что такое информация Фишера и эффективные оценки, и получим одно из частичных решений нахождения наилучшей оценки в равномерном подходе.

Доминируемые семейства

Далее в курсе мы часто будем предполагать, что имеем дело с так называемыми доминируемыми параметрическими семействами, имеющими (обобщенные) плотности.

Определение 5.1. Пусть $\{\mathrm{P}_\theta, \theta \in \Theta\}$ - параметрическое семейство распределений на выборочном пространстве $(\mathcal{X}, \mathcal{B}_\mathcal{X})$. Если для любого $\theta \in \Theta$ мера P_θ имеет плотность $p_\theta(x)$ по одной и той же (может быть, σ-конечной) мере μ на $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, то семейство $\{\mathrm{P}_\theta, \theta \in \Theta\}$ называется доминируемым относительно μ.

Напомним, что плотность понимается в смысле производной Радона-Никодима:

$$p_\theta(\theta) = \frac{d\mathrm{P}_\theta}{d\mu}, \quad \mathrm{P}_\theta(B) = \int\limits_B p_\theta(x)\mu(dx).$$

Примеры:

1) Если μ - это мера Лебега на $\mathbb{R}$, то P_θ - это абсолютно непрерывные распределения. В этом случае обобщенная плотность - это обычная плотность из курса теории вероятностей. Например, в семействе нормальных распределений $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$, плотность равна

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}.$$

2) Если μ - это считающая мера (мера, которая сопоставляет каждому множеству количество целых точек внутри него) на $\mathbb{Z}$, то P_θ - это дискретные распределения на $\mathbb{Z}$. В этом случае обобщенная плотность - это вероятность получить данное значение. Например, в семействе пуассоновских распределений $\mathrm{Pois}(\theta)$, $\theta > 0$, плотность равна

$$p_\theta(x) = \mathrm{P}_\theta(\{x\}) = \frac{\theta^x}{x!} e^{-\theta} \mathrm{I}\{x \in \mathbb{Z}_+\}.$$

Условия регулярности

Мы разберем частичное решение в среднеквадратичном подходе (равномерный подход с квадратичной функцией потерь).

Пусть $\mathbf{X}$ - наблюдение с неизвестным распределением P$\in \{\mathrm{P}_\theta, \theta \in \Theta\}$, где $\{\mathrm{P}_\theta, \theta \in \Theta\}$ - доминируемое семейство относительно меры μ с плотностью $p_\theta(x)$. Введем следующие условия регулярности:

1) $\Theta \subset \mathbb{R}$ - открытый интервал (может быть, бесконечный).
2) Множество $A = \{x : p_\theta(x) > 0\}$ не зависит от θ (носитель плотности не зависит от параметра). Это условие уже отбрасывает некоторые семейства распределений, например, под него не попадает семейство равномерное распределение на отрезке $[0, \theta]$, потому что носитель зависит от параметра.
3) Для любой статистики $S(\mathbf{X})$ с равномерно ограниченным вторым моментом, $\mathrm{E}_\theta S^2(\mathbf{X}) < M < +\infty$ для всех $\theta \in \Theta$, выполнено: для любого $\theta \in \Theta$
$$\frac{\partial}{\partial\theta}\mathrm{E}_\theta S(\mathbf{X}) = \mathrm{E}_\theta\left(S(\mathbf{X}) \cdot \frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X})\right).$$
4) Величина
$$I_{\mathbf{X}}(\theta) = \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X})\right)^2$$
положительна и конечна для всех $\theta \subset \Theta$.

Обсудим подробнее условие 3:
Оно означает, что мы можем дифференцировать под знаком интеграла. Действительно
$$\frac{\partial}{\partial\theta}\mathrm{E}_\theta S(\mathbf{X}) = \frac{\partial}{\partial\theta}\int\limits_A S(\mathbf{x})p_\theta(\mathbf{x})\mu(d\mathbf{x}) = \int\limits_A S(\mathbf{x})\frac{\partial}{\partial\theta}p_\theta(\mathbf{x})\mu(d\mathbf{x}) =$$
$$= \int\limits_A S(\mathbf{x})\frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{x}) \cdot p_\theta(\mathbf{x})\mu(d\mathbf{x}) = \mathrm{E}_\theta\left(S(\mathbf{X}) \cdot \frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X})\right)$$

Вклад и информация Фишера

Величины, участвующие в условии 4, имеют специальные названия.

Определение 5.2. Случайная величина

$$U_\theta(\mathbf{X}) = \frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X})$$

называется вкладом наблюдения $\mathbf{X}$.

Определение 5.3. Функция $I_{\mathbf{X}}(\theta)$ называется количеством информации о параметре θ, содержащимся в наблюдении $\mathbf{X}$ (информации по Фишеру).

$$I_{\mathbf{X}}(\theta) = \mathrm{E}_\theta(U_\theta(\mathbf{X}))^2.$$

Неравенство Рао-Крамера

Оказывается, что в случае регулярных семейств есть нижняя граница для дисперсии у несмещенной оценки некоторой функции от параметра. Эта граница постулируется утверждением, которое называется неравенством Рао-Крамера

Теорема 5.1. (неравенство Рао-Крамера)
Пусть выполнены условия регулярности 1-4. Если $\hat{\theta}(\mathbf{X})$ - это несмещенная оценка $\tau(\theta)$ с равномерно ограниченным вторым моментом $\mathrm{E}_\theta(\hat{\theta}(\mathbf{X}))^2 < M < +\infty$ для любого $\theta \in \Theta$, то для всех $\theta \in \Theta$ выполняется неравенство

$$\mathrm{D}_\theta\hat{\theta}(\mathbf{X}) \geqslant \frac{(\tau'(\theta))^2}{I_{\mathbf{X}}(\theta)}\ .$$

Доказательство.
Воспользуемся третьим условием регулярности, подставив в нее $S(\mathbf{X}) \equiv 1$:

$$\frac{\partial}{\partial\theta}\mathrm{E}_\theta S(\mathbf{X}) = \frac{\partial}{\partial\theta}\mathrm{E}_\theta 1 = 0, \quad \mathbf{E}_\theta(S(\mathbf{X})U_\theta(\mathbf{X})) = \mathrm{E}_\theta U_\theta(\mathbf{X}).$$

Следовательно, $\mathrm{E}_\theta U_\theta(\mathbf{X}) = 0$.

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

$$\frac{\partial}{\partial\theta}\mathrm{E}_\theta S(\mathbf{X}) = \frac{\partial}{\partial\theta}\mathrm{E}_\theta\hat{\theta}(\mathbf{X}) = \tau'(\theta),$$

$$\mathrm{E}_\theta(S(\mathbf{X})U_\theta(\mathbf{X})) = \mathrm{E}_\theta\Big(\hat{\theta}(\mathbf{X})U_\theta(\mathbf{X})\Big) = \mathrm{E}_\theta\Big((\hat{\theta}(\mathbf{X}) - \tau(\theta))U_\theta(\mathbf{X})\Big).$$

Тем самым, $\tau'(\theta) = \mathrm{E}_\theta\Big((\hat{\theta}(\mathbf{X}) - \tau(\theta))U_\theta(\mathbf{X})\Big)$. Применяя неравенство Коши-Буняковского, получаем искомое соотношение:

$$(\tau'(\theta))^2 \leqslant \mathrm{E}_\theta(\hat{\theta}(\mathbf{X}) - \tau(\theta))^2 \cdot \mathrm{E}_\theta U_\theta^2(\mathbf{X}) = \mathrm{D}_\theta\hat{\theta}(\mathbf{X}) \cdot I_{\mathbf{X}}(\theta).$$

■

Следствие 1
Если в условиях неравенства Рао-Крамера выполнено, что $\tau(\theta) = \theta$, то для всех $\theta \in \Theta$

$$\mathrm{D}_\theta \hat{\theta} \geqslant \frac{1}{I_{\mathbf{X}}(\theta)} .$$

Вопрос: чему равняется сама информация?

Утверждение
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - это выборка. Обозначим через

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

информацию Фишера одного наблюдения. Тогда

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

Следствие 2
Если в условиях неравенство Рао-Крамера $\mathbf{X} = (X_1, ..., X_n)$ - это выборка и $\tau(\theta)$ не зависит от n, то

$$\mathrm{D}_\theta \hat{\theta}(\mathbf{X}) \geqslant \frac{(\tau'(\theta))^2}{n \cdot i(\theta)} = \Omega\left(\frac{1}{n}\right).$$

Вопрос: возможно ли равенство в неравенстве Рао-Крамера?

Определение 5.4. Пусть $\hat{\theta}(\mathbf{X})$ - несмещенная оценка $\tau(\theta)$. Будем называть ее эффективной оценкой $\tau(\theta)$, если для нее достигается равенство в неравенстве Рао-Крамера, то есть для всех $\theta \in \Theta$

$$\mathrm{D}_\theta \hat{\theta}(\mathbf{X}) = \frac{(\tau'(\theta))^2}{I_{\mathbf{X}}(\theta)} .$$

Критерий эффективности

Лемма (критерий эффективности)
В условиях неравенства Рао-Крамера $\hat{\theta}(\mathbf{X})$ будет эффективной оценкой $\tau(\theta)$ тогда и только тогда, когда для любого $\theta \in \Theta$ P_θ-п.н. выполняется равенство

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

Доказательство.
Неравенство Рао-Крамера получалось применением неравенства Коши-Буняковского к равенству:

$$\mathrm{E}_\theta\Bigg((\hat{\theta}(\mathbf{X}) - \tau(\theta)) U_\theta(\mathbf{X})\Bigg) = \tau'(\theta).$$

Как известно, в неравенстве Коши-Буняковского равенство достигается тогда и только тогда, когда между случайными величинами есть линейная зависимость с вероятностью 1. Тут надо понимать, что поскольку мы фиксируем θ и применяем неравенство, то линейная зависимость между $\hat{\theta}(\mathbf{X}) - \tau(\theta)$ и $U_\theta(\mathbf{X})$ для каждого θ своя. Следовательно, $\hat{\theta}(\mathbf{X})$ будет эффективной оценкой для $\tau(\theta)$ только в том случае, если существует такая функция $c(\theta)$, что

$$\hat{\theta}(\mathbf{X}) - \tau(\theta) = c(\theta) U_\theta(\mathbf{X}).$$

Найдем $c(\theta)$. Домножим обе части последнего равенства на $U_\theta(\mathbf{X})$ и возьмем математическое ожидание:

$$\mathrm{E}_\theta\Bigg((\hat{\theta}(\mathbf{X}) - \tau(\theta)) U_\theta(\mathbf{X})\Bigg) = \mathrm{E}_\theta(c(\theta) U_\theta^2(\mathbf{X})).$$

Но левая часть есть $\tau'(\theta)$, а второй момент для вклада - информация Фишера. Тогда

$$\tau'(\theta) = c(\theta) I_{\mathbf{X}}(\theta) \Longrightarrow c(\theta) = \frac{\tau'(\theta)}{I_{\mathbf{X}}(\theta)}.$$

■

Разберем пример, который показывает, как необходимо подходить к задачам такого рода. Идея состоит в том, что необходимо вычислить вклад и посмотреть, представляется ли он в виде произведения функции от параметра на разность функции от x и функции от параметра. Если такое представление возможно, то функция от x будет эффективной оценкой того, что из нее вычитается, а информацию Фишера можно посчитать с помощью равенства, в котором наш коэффициент выражается через производную оцениваемой функции и информацию Фишера.

Пример

Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из распределения Бернулли $\mathrm{Bin}(1, \theta)$, где $\theta \in (0, 1)$. Найдите эффективную оценку θ и информацию $i(\theta)$ одного наблюдения.

Решение

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

$$p_\theta(\mathbf{X}) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} \theta^{\mathrm{I}\{X_i=1\}}(1-\theta)^{\mathrm{I}\{X_i=0\}} = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}.$$

Теперь несложно найти вклад выборки:

$$\ln p_\theta(\mathbf{X}) = \sum_{i=1}^{n} (X_i \ln\theta + (1-X_i)\ln(1-\theta)),$$

$$U_\theta(\mathbf{X}) = \frac{X_1 + ... + X_n}{\theta} - \frac{n - X_1 - ... - X_n}{1-\theta} = n\left(\frac{\overline{\mathbf{X}}}{\theta} - \frac{1-\overline{\mathbf{X}}}{1-\theta}\right) = \frac{n}{\theta(1-\theta)}(\overline{\mathbf{X}} - \theta).$$

Отсюда мы можем сделать вывод, что $\overline{\mathbf{X}}$ - эффективная оценка для $\tau(\theta) = \theta$. Значит $\tau'(\theta) = 1$ и

$$\frac{I_{\mathbf{X}}(\theta)}{\tau'(\theta)} = c^{-1}(\theta) = \frac{n}{\theta(1-\theta)}$$

Отсюда получаем информацию Фишера:

$$I_{\mathbf{X}} = \frac{n}{\theta(1-\theta)}, \quad i(\theta) = \frac{1}{\theta(1-\theta)}\ .$$

Выводы

1) Эффективная оценка для заданного семейства существует максимум для одной функции $\tau(\theta)$ (с точностью до умножения и добавления констант).

2) Требуется регулярность семейства распределений.

3) Если эффективная оценка существует, то плотность $\mathbf{X}$ имеет специальный вид:

$$p_\theta(\mathbf{X}) = h(\mathbf{X})e^{A(\theta)\hat{\theta}(\mathbf{X}) + B(\theta)},$$

где функции A, B не зависят от $\mathbf{X}$, а h - не зависит от θ.

4) Эффективная оценка $\tau(\theta)$ - это несмещенная оценка $\tau(\theta)$ с равномерно наименьшей дисперсией. Значит она - наилучшая оценка θ в среднеквадратичном подходе в классе всех несмещенных оценок $\tau(\theta)$:

$$\mathrm{E}_\theta(\hat{\theta}(\mathbf{X}) - \theta)^2 = \mathrm{D}_\theta\hat{\theta}(\mathbf{X}) + (\tau(\theta) - \theta)^2.$$

О выполнении условий регулярности

Естественно задаться вопросом о выполнении условий регулярности. Условия 1, 2 и 4 не вызывают вопросов, но для доказательства неравенства Рао-Крамера принципиально условие 3, прямая проверка которого очевидно затруднительная. Однако имеет место следующая теорема.

Теорема 5.2. Пусть для почти всех x по мере μ функция $\sqrt{p_\theta(x)}$ непрерывно дифференцируема по θ, а также информация Фишера $I_{\mathbf{X}}(\theta)$ конечна, положительна и непрерывна по θ. Тогда выполнено условие регулярности 3.

Доказательство теоремы можно найти в литературе, мы его опустим и приведем несколько замечаний.

Замечание (1)
Для конкретной оценки $\hat{\theta}(\mathbf{X})$ для выполнения неравенства Рао-Крамера достаточно конечности второго момента и возможности дифференцирования по параметру под знаком интеграла. Последнее можно проверить отдельно.

Замечание (2)
Условие равномерной ограниченности второго момента не является ограничительным. Если в условиях регулярности оценка $\hat{\theta}(\mathbf{X})$ такова, что функция $\mathrm{E}_\theta(\hat{\theta}(\mathbf{X}))^2$ непрерывна по θ, то она будет ограничена на любом отрезке внутри Θ. Сузив значение параметра до интервала внутри отрезка, мы получим неравенство Рао-Крамера на интервале. Но раз это верно для любого интервала внутри Θ, то оно будет выполнено и на всем Θ.

Информация Фишера статистик

Понятие информации Фишера можно ввести не только для самого наблюдения, но и для статистик от него, чье семейство распределений является доминируемым.

Определение 5.5. Пусть $S(\mathbf{X})$ - это некоторая статистика, имеющая плотсноть $g_\theta(s)$ по мере λ, то есть семейство ее распределений $(G_\theta, \theta \in \Theta)$ является доминируемым относительно λ. Тогда информацией Фишера статистики $S(\mathbf{X})$ называется

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

У информации Фишера есть несколько свойств, которые частично объясняют ее название.

Свойство (1)
Если распределение $S(\mathbf{X})$ не зависит от θ, то $I_s(\theta) = 0$.
Доказательство.
очевидно, так как в таком случае $\ln g_\theta(S(\mathbf{X}))$ не будет зависеть от θ и при дифференцировании по нему обратится в 0. ■

Свойство (2)
Пусть $S(\mathbf{X})$ и $T(\mathbf{X})$ - две независимые статистики для всех $\theta \in \Theta$. Обозначим $H(\mathbf{X}) = (S(\mathbf{X}), T(\mathbf{X}))$. Тогда в условиях регулярности выполнено равенство

$$I_h(\theta) = I_S(\theta) + I_T(\theta).$$

Доказательство.
Пусть статистика $S(\mathbf{X})$ имеет плотность $g_\theta(s)$ по мере λ_1, а статистика $T(\mathbf{X})$ - плотность $f_\theta(t)$ по мере λ_2. Тогда случайный вектор $H(\mathbf{X}) = (S(\mathbf{X}), T(\mathbf{X}))$ имеет совместную плотность $h_\theta(s,t) = g_\theta(s) f_\theta(t)$ по мере $\lambda_1 \times \lambda_2$. Тогда

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

Следовательно, в силу условий регулярности

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\ln h_\theta(H(\mathbf{X}))\right) = \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\ln g_\theta(S(\mathbf{X}))\right) + \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\ln f_\theta(T(\mathbf{X}))\right) = 0.$$

Но тогда в силу независимости

$$I_H(\theta) = \mathrm{D}_\theta\left(\frac{\partial}{\partial\theta}\ln h_\theta(H(\mathbf{X}))\right) = \mathrm{D}_\theta\left(\frac{\partial}{\partial\theta}\ln g_\theta(S(\mathbf{X})) + \frac{\partial}{\partial\theta}\ln f_\theta(T(\mathbf{X}))\right) =$$

$$= \mathrm{D}_\theta\left(\frac{\partial}{\partial\theta}\ln g_\theta(S(\mathbf{X}))\right) + \mathrm{D}_\theta\left(\frac{\partial}{\partial\theta}\ln f_\theta(T(\mathbf{X}))\right) = I_s(\theta) + I_T(\theta).$$

■

Свойство (3)
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - это выборка. Обозначим через

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

информацию Фишера одного наблюдения. Тогда

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

Доказательство.
Доказательство следует из свойства (2) по индукции. ■

Свойство (4)
В условиях регулярности для любой статистики $S(\mathbf{X})$ выполнено: для любого $\theta \in \Theta$

$$I_S(\theta) \leqslant I_{\mathbf{X}}(\theta).$$

Доказательство.
Доказательство этого свойства использует свойство условного математического ожидания (будет рассмотрено позже). Рекомендуется вернутся к доказательству этого свойства после изучения свойства математического ожидания.

Пусть $p_\theta(x)$ - это плотность $\mathbf{X}$ по мере μ, а $g_\theta(s)$ - плотность статистики $S(\mathbf{X})$ по мере λ. Покажем, что

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

Левая часть выражения означает, что мы берем условное математическое ожидание от $\frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X})$ по статистике $S(\mathbf{X})$.

Проверим по определению. Правая часть должна удовлетворять двум свойствам: интегральному и быть $S(\mathbf{X})$-измеримой. Правая часть $S(\mathbf{X})$-измерима, поэтому остается проверить интегральное свойство. Для любого борелевского множества B надо проверить, что

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X}) \cdot \mathrm{I}\{S(\mathbf{X}) \in B\}\right) = \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\ln g_\theta(S(\mathbf{X})) \cdot \mathrm{I}\{S(\mathbf{X}) \in B\}\right).$$

Посчитаем левую часть. Обозначим $A = \{x : p_\theta(x) > 0\}$. Тогда

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X}) \cdot \mathrm{I}\{S(\mathbf{X}) \in B\}\right) = \int\limits_A \frac{\partial \ln p_\theta(\mathbf{x})}{\partial\theta} p_\theta(\mathbf{x})\mathrm{I}\{S(\mathbf{x}) \in B\}\mu(d\mathbf{x}) =$$

$$= \int\limits_A \frac{\partial p_\theta(\mathbf{x})}{\partial\theta}\mathrm{I}\{S(\mathbf{x}) \in B\}\mu(d\mathbf{x}) = \frac{\partial}{\partial\theta}\int\limits_A p_\theta(\mathbf{x})\mathrm{I}\{S(\mathbf{x}) \in B\}\mu(d\mathbf{x}) = \frac{\partial}{\partial\theta}\mathrm{P}_\theta(S(\mathbf{X}) \in B).$$

Но с другой стороны

$$\frac{\partial}{\partial\theta}\mathrm{P}_\theta(S(\mathbf{X}) \in B) = \frac{\partial}{\partial\theta}\int\limits_B g_\theta(\mathbf{s})\lambda(d\mathbf{s}) = \int\limits_B \frac{\partial \ln g_\theta(\mathbf{s})}{\partial\theta} g_\theta(\mathbf{s})\lambda(d\mathbf{s}) =$$

$$= \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\ln g_\theta(S(\mathbf{X}))\cdot \mathrm{I}\{S(\mathbf{X})\in B\}\right).$$

Мы получили представление для условного математического ожидания. Теперь необходимо доказать неравенство для информаций. Рассмотрим величину

$$M(\theta) = \mathrm{E}_\theta\left(\frac{\partial \ln p_\theta(\mathbf{X})}{\partial\theta}\cdot\frac{\partial \ln g_\theta(S(\mathbf{X}))}{\partial\theta}\right).$$

Согласно неравенству Коши-Буняковского

$$M^2(\theta) \leqslant \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X})\right)^2\right)\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\ln g_\theta(S(\mathbf{X}))\right)^2\right) = I_{\mathbf{X}}(\theta)\cdot I_S(\theta).$$

Теперь докажем, что $M(\theta) = I_s(\theta)$. Для ого заметим, что по формуле полной вероятности

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

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

$$\mathrm{E}_\theta\left(\frac{\partial \ln p_\theta(\mathbf{X})}{\partial\theta}\cdot\frac{\partial \ln g_\theta(S(\mathbf{X}))}{\partial\theta}\right) = \mathrm{E}_\theta\left(\frac{\partial \ln g_\theta(S(\mathbf{X}))}{\partial\theta}\mathrm{E}\left(\frac{\partial \ln p_\theta(\mathbf{X})}{\partial\theta}\middle| S(\mathbf{X})\right)\right) =$$

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

Тогда

$$I_S^2(\theta) \leqslant I_{\mathbf{X}}(\theta) I_S(\theta),$$

что и дает искомое неравенство

$$I_S(\theta) \leqslant I_{\mathbf{X}}(\theta).$$

■

Свойство (5)
В условиях регулярности $I_s(\theta) = I_{\mathbf{X}}(\theta)$ для любого $\theta\in\Theta$ тогда и только тогда, когда статистика $S(\mathbf{X})$ является достаточной для семейства распределений $(\mathrm{P}, \theta\in\Theta)$.

Данное свойсвто мы докажем позднее, когда пройдем достаточные статистики.

Приведем пример того, как можно вычислять информации Фишера статистик от нашей выборки.

Пример
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из экспоненциального распределения $\mathrm{Exp}(\theta)$. Далее, пусть $X_{(1)} = \min\limits_{1\leqslant i\leqslant n} X_i$. Найти $I_{X_{(1)}}(\theta)$ и $I_{\mathbf{X}}(\theta)$.

Решение

Начнем с того, что найдем распределение $X_{(1)}$. Для этого заметим, что для $x > 0$

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

Следовательно, $X_{(1)} \sim \mathrm{Exp}(n\theta)$. Теперь найдем информацию Фишера. Начнем с $I_{\mathbf{X}}(\theta)$. Вспомним, что $I_{\mathbf{X}}(\theta) = n \cdot i(\theta)$. Тогда

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

Тогда

$$i(\theta) = \mathrm{E}_\theta\left(X_1 - \frac{1}{\theta}\right)^2 = \mathrm{D}_\theta X_1 = \frac{1}{\theta^2} \Longrightarrow I_{\mathbf{X}}(\theta) = \frac{n}{\theta^2}.$$

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

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

Тогда

$$I_{X_{(1)}} = \mathrm{E}_\theta\left(nX_{(1)} - \frac{1}{\theta}\right)^2 = n^2 \mathrm{D}_\theta X_{(1)} = \frac{1}{\theta^2}.$$

6. Лекция 6

Многомерный вариант

На прошлой лекции мы получили нижнюю оценку дисперсии несмещенной оценки функции одномерного параметра. Пусть теперь $\mathbf{X}$ - наблюдение с неизвестным распределением из доминируемого параметрического семейства $\{\mathrm{P}_{\boldsymbol{\theta}}, \theta \in \Theta\}$ с плотностью $p_{\boldsymbol{\theta}}(x)$, где $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$, $k > 1$ - многомерный параметр. Как можно определить информацию Фишера в подобной ситуации?

Введем следующие обозначения:

1) $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k)^T$ (вектор-столбец);

2) вклад наблюдения $\mathbf{X}$ - вектор-столбец частных производных логарифма плотности:
$$U_{\boldsymbol{\theta}}(\mathbf{X}) = \left(\frac{\partial}{\partial \boldsymbol{\theta}_1} \ln p_{\boldsymbol{\theta}}(\mathbf{X}), ..., \frac{\partial}{\partial \boldsymbol{\theta}_k} \ln p_{\boldsymbol{\theta}}(\mathbf{X})\right)^T;$$

3) информационная матрица размера $k \times k$:
$$I_{\mathbf{X}}(\boldsymbol{\theta}) = \left(\mathrm{E}_{\boldsymbol{\theta}}\left(\frac{\partial \ln p_\theta(\mathbf{X})}{\partial \boldsymbol{\theta}_i} \cdot \frac{\partial \ln p_\theta(\mathbf{X})}{\partial \boldsymbol{\theta}_i}\right),\ i, j = 1, ..., k\right).$$

Замечание
В условиях регулярности, как мы помним, вклад имеет нулевое среднее, поэтому информационная матрица будет его матрицей ковариации:
$$I_{\mathbf{X}}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}(U_{\boldsymbol{\theta}}(\mathbf{X}) U_{\boldsymbol{\theta}}^T(\mathbf{X})) = \mathrm{D}_{\boldsymbol{\theta}} U_{\boldsymbol{\theta}}(\mathbf{X}).$$

4) если $\tau : \Theta \to \mathbb{R}^d$ - гладкая вектор-функция, то через $\tau'(\boldsymbol{\theta})$ мы обозначим ее матрицу Якоби:
$$\tau'(\boldsymbol{\theta}) = \left(\frac{\partial \tau_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j},\ i = 1, ..., d,\ j = 1, ..., k\right) \in \mathbf{Mat}(d \times k).$$

5) если $\hat{\theta}(\mathbf{X}) \in \mathbb{R}^d$ - то несмещенная оценка $\tau(\boldsymbol{\theta})$, то ее матрица ковариаций есть
$$\mathrm{D}_{\boldsymbol{\theta}} \hat{\theta}(\mathbf{X}) = \mathrm{E}_{\boldsymbol{\theta}}\left((\hat{\theta}(\mathbf{X}) - \tau(\boldsymbol{\theta}))(\hat{\theta}(\mathbf{X}) - \tau(\boldsymbol{\theta}))^T\right) \in \mathbf{Mat}(d \times d).$$

Матричное неравенство Коши-Буняковского

Для доказательства многомерного аналога неравенства Рао-Крамера нам понадобится и многомерный аналог неравенства Коши-Буняковского.

Лемма (матричное неравенство Коши-Буняковского) Пусть $\mathbf{\Psi}$ и $\mathbf{H}$ - случайные матрицы, $\mathbf{\Psi} \in \mathbf{Mat}(n' \times m')$, $\mathbf{H} \in \mathbf{Mat}(k' \times m')$, причем матрица $\mathrm{E}(\mathbf{H}\mathbf{H}^T)$ обратима. Тогда

$$\mathrm{E}(\mathbf{\Psi}\mathbf{\Psi}^T) \geqslant \mathrm{E}(\mathbf{\Psi}\mathbf{H}^T)(\mathrm{E}\mathbf{H}\mathbf{H}^T)^{-1}\mathrm{E}(\mathbf{H}\mathbf{\Psi}^T),$$

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

$$\mathbf{Z} = \mathrm{E}(\mathbf{\Psi}\mathbf{H}^T)(\mathrm{E}\mathbf{H}\mathbf{H}^T)^{-1}.$$

Замечание
Запись $\mathbf{A} \geqslant \mathbf{B}$ означает, что матрица $\mathbf{A} - \mathbf{B}$ неотрицательно определена.
Доказательство.
Для начала вспомним, что для любой матрицы $\mathbf{A}$ матрица $\mathbf{A}\mathbf{A}^T \geqslant 0$. Тогда для любой неслучайной матрицы $\mathbf{Z} \in \mathbf{Mat}(n' \times k')$ выполнено следующее:

$$(\mathbf{\Psi} - \mathbf{Z}\mathbf{H})(\mathbf{\Psi} - \mathbf{Z}\mathbf{H})^T \geqslant 0.$$

Возьмем математическое ожидание:

$$\mathrm{E}(\mathbf{\Psi} - \mathbf{Z}\mathbf{H})(\mathbf{\Psi} - \mathbf{Z}\mathbf{H})^T \geqslant 0.$$

Теперь раскроем по линейности:

$$\mathrm{E}\mathbf{\Psi}\mathbf{\Psi}^T - \mathbf{Z}\mathrm{E}\mathbf{H}\mathbf{\Psi}^T - (\mathrm{E}\mathbf{\Psi}\mathbf{H}^T)\mathbf{Z}^T + \mathbf{Z}(\mathrm{E}\mathbf{H}\mathbf{H}^T)\mathbf{Z}^T \geqslant 0.$$

Положим $\mathbf{Z} = \mathrm{E}\mathbf{\Psi}\mathbf{H}^T(\mathrm{E}\mathbf{H}\mathbf{H}^T)^{-1}$. Тогда вышеприведенное неравенство примет нужный вид:

$$\mathrm{E}\mathbf{\Psi}\mathbf{\Psi}^T - \mathrm{E}\mathbf{\Psi}\mathbf{H}^T(\mathrm{E}\mathbf{H}\mathbf{H}^T)^{-1}\mathrm{E}\mathbf{H}\mathbf{\Psi}^T \geqslant 0.$$

Для критерия равенства остается заметить, что

$$\mathrm{E}(\mathbf{\Psi} - \mathbf{Z}\mathbf{H})(\mathbf{\Psi} - \mathbf{Z}\mathbf{H})^T = 0 \Leftrightarrow \mathbf{\Psi} = \mathbf{Z}\mathbf{H}.$$

■

Условия регулярности

Для выполнения неравенства Рао-Крамера снова будут нужны условия регулярности. Они во многом аналогичны одномерному случаю.

1) Множество значений параметра $\Theta \subset \mathbb{R}$ - открытое связное множество в $\mathbb{R}^k$.

2) Множество $A = \{x : p_{\boldsymbol{\theta}} > 0\}$ не зависит от $\boldsymbol{\theta}$

3) Для любой статистики $S(\mathbf{X})$ с равномерно ограниченным по $\boldsymbol{\theta}$ вторым моментом $\mathrm{E}_{\boldsymbol{\theta}} S^2(\mathbf{X}) < M < +\infty$ выполнено: для любого $\boldsymbol{\theta} \in \Theta$
$$\frac{\partial}{\partial \boldsymbol{\theta}_j} \mathrm{E}_{\boldsymbol{\theta}} S(\mathbf{X}) = \mathrm{E}_{\boldsymbol{\theta}} \left(S(\mathbf{X}) \cdot \frac{\partial}{\partial \boldsymbol{\theta}_j} \ln p_{\boldsymbol{\theta}}(\mathbf{X}) \right),$$
$j = 1, ..., k$.

4) Информационная матрица $I_{\mathbf{X}}(\boldsymbol{\theta})$ конечна и положительно определена для всех $\boldsymbol{\theta} \in \Theta$.

Многомерное неравенство Рао-Крамера

Теперь мы готовы сформулировать и доказать многомерное неравенство Рао-Крамера.

Теорема 6.1. (многомерное неравенство Рао-Крамера)
Пусть выполнены условия регулярности. Пусть $\hat{\theta}(\mathbf{X})$ - это несмещенная оценка $\tau(\boldsymbol{\theta}) \in \mathbb{R}^d$ с равномерно ограниченной по $\boldsymbol{\theta}$ матрицей $\mathrm{E}_{\boldsymbol{\theta}} \hat{\theta}(\mathbf{X}) \hat{\theta}(\mathbf{X})^T$. Тогда для любого $\boldsymbol{\theta} \in \Theta$ выполняется неравенство
$$\mathrm{D}_{\boldsymbol{\theta}} \hat{\theta}(\mathbf{X}) \geqslant \tau'(\boldsymbol{\theta}) I_{\mathbf{X}}^{-1}(\boldsymbol{\theta}) (\tau'(\boldsymbol{\theta}))^T,$$
где $\tau'(\boldsymbol{\theta})$ - матрица Якоби размера $(d \times k)$

Доказательство.
Доказательство повторяет рассуждения в одномерном случае. Подставляя $S(\mathbf{X}) = 1$ в условие 3, получаем, что
$$0 = \mathrm{E}_{\boldsymbol{\theta}} \left(\frac{\partial}{\partial \boldsymbol{\theta}_j} \ln p_{\boldsymbol{\theta}}(\mathbf{X}) \right),$$

$j = 1, ..., k$, то есть $\mathrm{E}_{\boldsymbol{\theta}} U_{\boldsymbol{\theta}}(\mathbf{X}) = 0$. Подставляя $S(\mathbf{X}) = \hat{\theta}_i(\mathbf{X})$, получаем

$$\frac{\partial \tau_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} = \mathrm{E}_{\boldsymbol{\theta}} \left(\hat{\theta}_i(\mathbf{X}) \cdot \frac{\partial}{\partial \boldsymbol{\theta}_j} \ln p_{\boldsymbol{\theta}}(\mathbf{X}) \right),$$

или в матричном виде

$$\tau'(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}} \hat{\theta}(\mathbf{X})(U_{\boldsymbol{\theta}}(\mathbf{X}))^T \underset{\mathrm{E}_{\boldsymbol{\theta}} U_{\boldsymbol{\theta}}(\mathbf{X})=0}{\overset{\text{т.к.}}{=}} \mathrm{E}_{\boldsymbol{\theta}}(\hat{\theta}(\mathbf{X}) - \tau(\boldsymbol{\theta}))(U_{\boldsymbol{\theta}}(\mathbf{X}))^T.$$

Обозначим

$$\boldsymbol{\Psi} = \hat{\theta}(\mathbf{X}) - \tau(\boldsymbol{\theta}) \in \mathbf{Mat}(d \times 1),$$

$$\mathbf{H} = U_{\boldsymbol{\theta}}(\mathbf{X}) \in \mathbf{Mat}(k \times 1).$$

Тогда

$$\tau'(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}} \boldsymbol{\Psi} \mathbf{H}^T.$$

Согласно матричному неравенству Коши-Буняковского, получаем искомое неравенство:

$$\mathrm{D}_{\boldsymbol{\theta}} \hat{\theta}(\mathbf{X}) = \mathrm{E} \boldsymbol{\Psi} \boldsymbol{\Psi}^T \geqslant \mathrm{E} \boldsymbol{\Psi} \mathbf{H}^T (\mathrm{E} \mathbf{H} \mathbf{H}^T)^{-1} \mathrm{E} \mathbf{H} \boldsymbol{\Psi}^T = \tau'(\boldsymbol{\Psi}) I_{\mathbf{X}}^{-1}(\boldsymbol{\Psi}) (\tau'(\boldsymbol{\Psi}))^T.$$

■

Критерий равенства

Как и в одномерном случае, можно предложить явный критерий достижения равенства в многомерном неравенстве Рао-Крамера.

Утверждение (критерий равенства)
В многомерном неравенстве Рао-Крамера равенство для всех $\boldsymbol{\theta}$ достигается тогда и только тогда, когда для всех $\boldsymbol{\theta} \in \Theta$ с вероятностью 1 выполняется равенство

$$\hat{\theta}(\mathbf{X}) - \tau(\boldsymbol{\theta}) = \tau'(\boldsymbol{\theta}) I_{\mathbf{X}}^{-1}(\boldsymbol{\theta}) U_{\boldsymbol{\theta}}(\mathbf{X}).$$

Доказательство этого утверждения оставляется читателю в качестве упражнения.

О выполнении условий регулярности

Как и в одномерном случае можно предложить достаточное условие выполнения сложного условия регулярности 3.

Теорема 6.2. Пусть для почти всех x по мере μ функция $\sqrt{p_{\boldsymbol{\theta}}(x)}$ непрерывно дифференцируема по $\boldsymbol{\theta}_j$, $j = 1, ..., k$, а также пусть информационная матрица $I_{\mathbf{X}}(\boldsymbol{\theta})$ положительно определена и непрерывна по θ. Тогда выполнено условие регулярности 3.

Функция правдоподобия

Предположим, что мы получили некоторые данные, например, вектор из нулей и единиц: 0,0,1,1,1,0,1,1,0,1. Получилось 10 чисел: 6 единиц и 4 нуля.
Попытаемся понять, какое это распределение? У нас есть 2 гипотезы: это биномиальное распределение $\mathrm{Bin}(1, \frac{1}{2})$ (схема Бернулли симметричная), или это пуассоновское распределение $\mathrm{Pois}(\frac{1}{2})$ с параметром $\frac{1}{2}$. Возникает вопрос: какое распределение взять?
В качестве критерия проверки можно предложить следующее: попробуем понять, с какой вероятностью данная нам последовательность реализовалась в каждом распределении.

- Для биномиального - $(\frac{1}{2})^{10}$.
- Для пуассоновского распределения ноль выпадает с вероятностью $e^{-\frac{1}{2}}$, единица с вероятностью $\frac{1}{2}e^{-\frac{1}{2}}$, поэтому мы получаем $(\frac{1}{2})^6 e^{-5}$.

Сравнив полученные 2 числа можно сделать вывод, что для схемы Бернулли выпадение данного набора чисел более вероятно, чем для пуассоновского распределения с параметром $\frac{1}{2}$. Подобный подход сравнения - метод максимального правдоподобия, то есть мы выбираем те распределения, для которых реализация конкретного набора данных более вероятна, чем для других.

Введем понятие функции правдоподобия.

Определение 6.1. Пусть $\mathbf{X}$ - наблюдение с неизвестным распределением
$\mathrm{P} \in \{\mathrm{P}_\theta, \theta \in \Theta\}$, где $\{\mathrm{P}_\theta, \theta \in \Theta\}$ есть доминируемое семейство с плотностью $p_\theta(\mathbf{X})$ по мере μ.
Тогда функцией правдоподобия называется случайная величина $f_\theta(\mathbf{X}) = p_\theta(\mathbf{X})$.

Замечание
Если $\mathbf{X} = (X_1, ..., X_n)$ - выборка, то плотность случайного вектора разбивается в произведение плотностей координат:

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

Оценка максимального правдоподобия

Определение 6.2. Оценкой параметра θ по методу максимального правдоподобия, или же оценкой максимального правдоподобия (ОМП), называется

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

Данное определение уже накладывает несколько ограничений: во-первых, что максимум существует и единственен, а во-вторых, что он является борелевской функцией от $\mathbf{X}$.

Философия метода: "мы живем в наиболее вероятном мире". Выпавший результат наблюдения должен был произойти с как можно большей вероятностью.

Этот метод хорошо иллюстрируется, когда мы говорим о дискретных распределениях, так как мы сравниваем вероятности получения конкретного набора данных.

В абсолютно непрерывном случае мы в качестве вероятности получить значение x используем плотность в точке x. Плотность не совсем подходит, так как может быть больше 1, но идея такова.

Пример 1
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из равномерного распределения $\mathrm{U}(0, \theta)$, $\theta > 0$. Найти оценку максимального правдоподобия параметра θ.

Решение

Когда мы говорим об оценке максимального правдоподобия, важно говорить о том, каково множество значений, то есть по какому множеству мы будем максимизировать нашу функцию плотности. В данном случае мы возьмем все множество. Выпишем функцию правдоподобия:

$$f_\theta(\mathbf{X}) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} \frac{1}{\theta}\mathrm{I}\{0 \leqslant X_i \leqslant \theta\} = \frac{1}{\theta^n}\mathrm{I}\{0 \leqslant X_{(1)} \leqslant X_{(n)} \leqslant \theta\}.$$

Теперь нам нужно максимизировать ее, как функцию от θ. Заметим, что θ^{-n} - это монотонно убывающая функция, поэтому нужно взять такое минимальное θ, что функция правдоподобия не обратится в нуль. Стало быть, $\hat{\theta}(\mathbf{X}) = X_{(n)}$.

Пример 2
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из нормального распределения $\mathcal{N}(a, \sigma^2)$, $a \in \mathbb{R}$, $\sigma > 0$. Найдите оценку максимального правдоподобия параметра $\boldsymbol{\theta} = (a, \sigma^2)$.

Решение

Опять же, начнем с того, что выпишем функцию правдоподобия:

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

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - a)^2\right\}.$$

Нам нужно максимизировать ее одновременно по a и по σ^2. Но работать с экспонентой достаточно неудобно, поэтому прологарифмируем ее:

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

Это гладкая функция, поэтому приравниваем частные производные к нулю и решаем систему (для удобства возьмем параметром σ^2):

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

Отсюда несложно получить, что решением будут оценки

$$\begin{aligned} \hat{a}(\mathbf{X}) &= \frac{1}{n}\sum_{i=1}^n X_i = \overline{\mathbf{X}}, \\ \hat{\sigma^2}(\mathbf{X}) &= \frac{1}{n}\sum_{i=1}^n \left(X_i - \frac{1}{n}\sum_{i=1}^n X_i\right)^2 = S^2. \end{aligned}$$

Заметим, что решение получилось единственным, это говорит о том, что это будет точка максимума. Это несложно проверить стандартными методами анализа. Тем самым, $\hat{\theta}(\mathbf{X}) = (\overline{\mathbf{X}}, S^2)$.

В следующий раз

На следующей лекции мы будем обсуждать асимптотические свойства оценки максимального правдоподобия. Мы увидим, что эти свойства достаточно хороши, и оценка максимального правдоподобия поставляет нам наилучшие оценки в асимптотическом подходе к сравнению оценок.

Оценка максимального правдоподобия хороша с практической точки зрения, так как ее нахождение требует решения экстремальных задач, задач оптимизации, а методы решения этих задач весьма разнообразны и продвинуты.

7. Лекция 7

На этой лекции мы разберем свойства оценки максимального правдоподобия. В прошлый раз мы ввели понятие оценки максимального правдоподобия и разобрали несколько примеров. Сегодня мы разберем, почему эта оценка весьма хороша.

Экстремальное свойство правдоподобия

У оценок максимального правдоподобия и функции правдоподобия есть хорошие асимптотические свойства (состоятельность и асимптотическая нормальность). Но они требуют некоторых условий регулярности. Будем постепенно формулировать их и доказывать свойства.

R0 Параметрическое семейство распределений $\{\mathrm{P}_\theta, \theta \in \Theta\}$ - это доминируемое семейство с плотностью $p_\theta(x)$ по мере μ и различимыми распределениями, то есть $\mathrm{P}_{\theta_0} = \mathrm{P}_{\theta_1}$ тогда и только тогда, когда $\theta_0 = \theta_1$.

R1 $\mathbf{X} = (X_1, ..., X_n)$ - выборка растущего размера из неизвестного распределения $\mathrm{P} \in \{\mathrm{P}_\theta, \theta \in \Theta\}$.

R2 $A = \{x : p_\theta > 0\}$ не зависит от θ. Здесь $p_\theta(x)$ - плотность одного элемента выборки.

Напомним, что через $f_\theta(X_1, ..., X_n)$ мы обозначаем функцию правдоподобия.

Теорема 7.1. (экстремальное свойство правдоподобия)
В условиях регулярности **R0-R2** для всех различных $\theta_1, \theta_2 \in \Theta$ выполнено

$$\lim_{n\to\infty} \mathrm{P}_{\theta_0}(f_{\theta_0}(X_1, ..., X_n) > f_{\theta_1}(X_1, ..., X_n)) = 1.$$

Смысл утверждения: при сравнении значений функций правдоподобия в истинном значении параметра и не истинном, значение в точке истинного значения будет почти всегда больше.

Доказательство.
Будем считать, что все X_i принадлежат A, то есть $p_\theta(x) > 0$. Посмотрим, при каких условиях выполняется событие $f_{\theta_0}(\mathbf{X}) > f_{\theta_1}(\mathbf{X})$. Для этого прологарифмируем и преобразуем выражение:

$$\ln \frac{f_{\theta_1}(\mathbf{X})}{f_{\theta_0}(\mathbf{X})} < 0 \Longleftrightarrow \ln \prod_{i=1}^{n} \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)} < 0 \Longleftrightarrow \frac{1}{n} \prod_{i=1}^{n} \ln \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)} < 0.$$

В левой части последнего неравенства стоит сумма независимых одинаково распределенных случайных величин, которая делится на их количество. По усиленному закону больших чисел

$$\frac{1}{n}\sum_{i=1}^{n}\frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)} < 0 \xrightarrow{\mathrm{P}_{\theta_0}-\text{п.н.}} \mathrm{E}_{\theta_0}\left(\ln\frac{p_{\theta_1}(X_1)}{p_{\theta_0}(X_1)}\right).$$

Теперь докажем, что

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

Примечание
Если добавить минус к этому математическому ожиданию, то мы получим широко известную дивергенцию Кульбака-Лейблера. По сути, мы доказываем то, что она неотрицательна и то, что она равна нулю тогда и только тогда, когда плотности равны почти всюду.

Логарифм - функция, которая выпукла вверх, поэтому воспользуемся неравенством Йенсена:

$$\mathrm{E}_{\theta_0}\left(\ln\frac{p_{\theta_1}(X_1)}{p_{\theta_0}(X_1)}\right) \leqslant \ln \mathrm{E}_{\theta_0}\left(\frac{p_{\theta_1}(X_1)}{p_{\theta_0}(X_1)}\right) = \ln\int\limits_A \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)}p_{\theta_0}(x)\mu(dx) = \ln\int\limits_A p_{\theta_1}(x)\mu(dx) = 0.$$

Мы поняли, что математическое ожидание $\leqslant 0$. Почему же неравенство, которое необходимо проверить, строгое? Предположим, что это не так, то есть

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

Но в таком случае можно воспользоваться критерием равенства для неравенства Йенсена: $\phi(\mathrm{E}\xi) = \mathrm{E}(\phi(\xi))$ тогда и только тогда, когда ϕ линейна почти всюду. но $\ln(x)$ нелинейна. Тогда получаем, что аргумент должен быть равен единице почти везде: $\mu(\{x : p_{\theta_0}(x) \neq p_{\theta_1}(x)\}) = 0$. Но это означает, что $\mathrm{P}_{\theta_0} = \mathrm{P}_{\theta_1}$, что противоречит условию **R0**. В итоге получаем, что

$$\lim_{n\to\infty}\mathrm{P}_{\theta_0}(f_{\theta_0}(\mathbf{X}) > f_{\theta_1}(\mathbf{X})) = \mathrm{P}_{\theta_0}\left(\mathrm{E}_{\theta_0}\left(\ln\frac{p_{\theta_1}(X_1)}{p_{\theta_0}(X_1)}\right) < 0\right) = 1.$$

■

Следствие
Если Θ конечно, то оценка максимального правдоподобия состоятельна.
Доказательство.
Пусть $\hat{\theta}_n(\mathbf{X})$ - это оценка максимального правдоподобия. Тогда по экстремальному свойству правдоподобия для любого $\theta_0 \in \Theta$

$$\lim_{n\to\infty}\mathrm{P}_{\theta_0}(\hat{\theta}_n(\mathbf{X}) = \theta_0) = \lim_{n\to\infty}\mathrm{P}_{\theta_0}(\forall\theta \neq \theta_0 : f_{\theta_0}(\mathbf{X}) > f_{\theta}(\mathbf{X}))) = 1.$$

■

Состоятельность решения уравнения правдоподобия

Теперь попробуем разобраться со случаем, когда множество значений параметра не является дискретным множеством. Введем еще два условия регулярности:

R3 Θ есть открытый интервал на $\mathbb{R}$.

R4 $p_\theta(x)$ непрерывно дифференцируема по θ для всех $a \in A$.

Теорема 7.2. (состоятельность решения уравнения правдоподобия)
В условиях регулярности **R0-R4** уравнение правдоподобия

$$\frac{\partial}{\partial\theta} f_\theta(X_1, ..., X_n) = 0$$

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

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

$$A_n = \{f_{\theta_0}(X_1, ..., X_n) > f_{\theta_0+\delta}(X_1, ..., X_n), f_{\theta_0}(X_1, ..., X_n) > f_{\theta_0-\delta}(X_1, ..., X_n)\}$$

Тогда согласно экстремальному свойству правдоподобия

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

Попробуем изобразить событие A_n и понять, как мы можем найти корни уравнения правдоподобия.

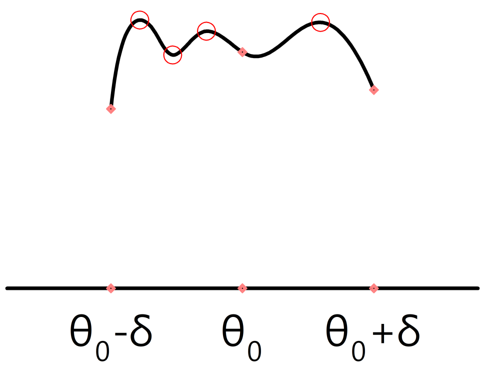

Рис. 4

Рассмотрим наше множество Θ и указанный отрезок. Мы знаем, что значение функции в середине отрезка больше, чем на краях. Функция гладкая, поэтому на отрезке возникают локальные максимумы, то есть точки, в которых производная обнуляется, соответственно, внутри отрезка мы обязательно находим одно из решений уравнения правдоподобия. Если это событие выполнено, то мы получаем, что решение существует, причем это решение будет в некоторой небольшой окрестности точки θ_0.

Допустим, что на этом отрезке есть несколько корней (не обязательно конечное число). Пусть $\tilde{\theta}(\mathbf{X})$ - ближайший к θ_0 корень уравнения правдоподобия. Он корректно определен, так как в силу непрерывности производной множество корней замкнуто. Тогда для любого $\varepsilon > 0$ будет выполнено

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

Действительно, зафиксируем ε и заметим, что рассуждения выше верны и для $\delta = \varepsilon$. Следовательно, с вероятностью, стремящейся к 1, на отрезке $[\theta_0 - \varepsilon, \theta_0 + \varepsilon]$ будет корень. Но $\tilde{\theta}(\mathbf{X})$ - ближайший к θ_0 корень. Тогда он тоже лежит на этом отрезке с вероятностью, стремящейся к 1. ■

Мы доказали теорему. Но она, на самом деле, задает больше вопросов, чем дает ответов.

1) Корней уравнения правдоподобия может быть несколько. В доказательстве мы выбираем ближайший из них к истинному значению. Но как его выбрать, если истинное значение нам неизвестно?

2) Даже если мы его найдем, то он зависит от истинного значения, то есть вообще не является оценкой!

3) Почему $\tilde{\theta}(\mathbf{X})$ есть точка максимума? Мы только сказали, что это корень уравнения, он может оказаться точкой минимума или же точкой перегиба.

4) Корень существует не всегда, а только с большой вероятностью.

Впрочем, если уравнение правдоподобия всегда имеет только один корень, то все становится на свои места. Первый, второй (корень один, поэтому он не зависит от параметра) и четвертый вопросы получают ответы. И остается только третий.

Следствие
Если в условиях **R0-R4** для всех $\mathbf{X} = (X_1, ..., X_n)$ уравнение правдоподобия имеет единственный корень $\hat{\theta}(\mathbf{X})$, то с вероятностью стремящейся к 1, $\hat{\theta}(\mathbf{X})$ будет оценкой максимального правдоподобия и, значит, оценка максимального правдоподобия будет состоятельной.
Доказательство.
По сути, доказательство повторяет предыдущие рассуждения. Пусть θ_0 - истинное

значение параметра. Снова возьмем любое $\delta > 0$ такое, что $[\theta_0 - \delta, \theta_0 + \delta] \subset \Theta$ и заметим, что

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

Однако если выполнено A_n, то внутри $[\theta_0 - \delta, \theta_0 + \delta]$ есть точка локального максимума. Как известно, в ней производная равна нулю, и, следовательно, она будет корнем уравнения правдоподобия. Но тогда эта точка есть $\hat{\theta}(\mathbf{X})$.

Осталось понять, почему это точка глобального максимума. Мы также предполагаем, что значение в точке θ_0 больше, чем на концах отрезка. Точка максимума $\hat{\theta}(\mathbf{X})$ - решение уравнения правдоподобия. Может ли быть так, что $\hat{\theta}(\mathbf{X})$ - не глобальный максимум? Допустим, что есть точка, например, вне нашего отрезка, в котором значение больше. В силу гладкости функции мы получаем, что между этими двумя точками есть точка локального минимума, где производная обнуляется. Поэтому, если наша точка не глобальный максимум, то обязательно найдется еще одно решение уравнения, что противоречит предположению. Тогда

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

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

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

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

Асимптотическая нормальность решения уравнения правдоподобия

В некоторых дополнительных условиях регулярности оценка максимального правдоподобия будет не только состоятельной, но еще и асимптотически нормальной.

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

R6 Интеграл

$$\int_A p_\theta(x)\mu(dx)$$

можно дважды дифференцировать под знаком интеграла.

R7 Для всех $\theta \in \Theta$ информация Фишера положительна и конечна

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

где $i(\theta)$ - информация Фишера одного элемента выборки.

R8 Для любого $\theta_0 \in \Theta$ существует $\delta > 0$ и функция $M(x)$ такая, что для всех $\theta \in [\theta_0 - \delta, \theta_0 + \delta]$

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

Сформулируем теорему об асимптотической нормальности.

Теорема 7.3. В условиях регулярности **R0-R8** любая состоятельная последовательность $(\hat\theta(\mathbf{X}), n \in \mathbb{N})$ корней уравнения правдоподобия удовлетворяет свойству асимптотической нормальности: для всех $\theta_0 \in \Theta$

$$\sqrt{n}(\hat\theta(X_1, ..., X_n) - \theta_0) \xrightarrow{d_{\theta_0}} \mathcal{N}\left(0, \frac{1}{i(\theta_0)}\right).$$

Доказательство.
Обозначим через $\mathcal{L}(\mathbf{X}, \theta) = \ln f_\theta(\mathbf{X})$ логарифмическую функцию правдоподобия, и введем обозначение $\mathcal{L}^{(n)}(\mathbf{X}, \theta)$ для n-й частной производной $\mathcal{L}(\mathbf{X}, \theta)$ по θ.

Далее, пусть θ_0 есть истинное значение параметра, то есть $\hat\theta(\mathbf{X})$ сходится к θ_0 по вероятности P_{θ_0}. Разложим $\mathcal{L}'(\mathbf{X}, \theta)$ в ряд Тейлора в точке θ_0:

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

где $\tilde\theta$ находится между θ и θ_0. Далее подставим $\theta = \hat\theta_n(\mathbf{X})$:

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

где $\tilde\theta_n = \tilde\theta_n(\mathbf{X}, \theta_0)$ находится между $\hat\theta_n(\mathbf{X})$ и θ_0. Теперь вспомним, что $\hat\theta_n(\mathbf{X})$ есть решения уравнения правдоподобия. Тогда левая часть равенства $\mathcal{L}'(\mathbf{X}, \hat\theta_n(\mathbf{X})) = 0$. Теперь выразим $\sqrt{n}(\hat\theta_n(\mathbf{X}) - \theta_0)$ из данного равенства:

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

$$= \frac{\frac{1}{-\sqrt{n}}\mathcal{L}'(\mathbf{X}, \theta_0)}{\frac{1}{n}\mathcal{L}''(\mathbf{X}, \theta_0) + \frac{1}{2n}\mathcal{L}'''(\mathbf{X}, \tilde\theta_n)(\hat\theta_n(\mathbf{X}) - \theta_0)}.$$

Теперь рассмотрим это выражение по частям.

- Начнем с числителя. Заметим, что

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

Как мы уже знаем, возможность дифференцирования под знаком интеграла показывает, что $\mathrm{E}_{\theta_0}U_{\theta_0}(X_1) = 0$. Тогда по центральной предельной теореме

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

- Теперь рассмотрим первое слагаемое в знаменателе. По усиленному закону больших чисел

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

Докажем, что

$$-\mathrm{E}_{\theta_0}\left(\frac{\partial^2}{\partial\theta^2}\ln p_\theta(X_1)\right) = i(\theta).$$

Заметим, что

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

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

Теперь возьмем математическое ожидание и воспользуемся тем, что мы можем два раза дифференцировать под знаком интеграла:

$$\mathrm{E}_{\theta_0}\left(\frac{\partial^2}{\partial\theta^2}\ln p_\theta(X_1)\right) = \int\limits_A \frac{\partial^2}{\partial\theta^2}\ln p_\theta(x)p_\theta(x)\mu(dx) =$$

$$= -\int\limits_A\left(\frac{\partial}{\partial\theta}\ln p_\theta(x)\right)^2 p_\theta(x)\mu(dx) + \int\limits_A \frac{\partial^2}{\partial\theta^2}p_\theta(x)\mu(dx) =$$

$$= -i(\theta) + \frac{\partial^2}{\partial\theta^2}\int\limits_A p_\theta(x)\mu(dx) = -i(\theta).$$

Отсюда получаем, что

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

- Теперь перейдем ко второму слагаемому в знаменателе и покажем, что он стремится к нулю по вероятности. Заметим, что по условию $\hat{\theta}_n(\mathbf{X})$ сходится к θ_0 по вероятности. Из этого можно сделать вывод, что $\tilde{\theta}_n$ тоже стремится к θ_0 по вероятности. Далее, по **R8**

$$\left|\frac{1}{n}\mathcal{L}'''(\mathbf{X}, \tilde{\theta}_n)\right| \leqslant \frac{1}{n}\sum_{i=1}^{n} M(X_i) \xrightarrow{\mathrm{P}_{\theta_0}} \mathrm{E}_{\theta_0} M(X_1).$$

Тогда получаем произведение ограниченной по вероятности случайной величины на сходящуюся к нулю по вероятности. Произведение сходится к нулю по вероятности.

В итоге

$$\frac{1}{2n}\mathcal{L}'''(\mathbf{X}, \tilde{\theta_n})(\hat{\theta}_n(\mathbf{X}) - \theta_0) \xrightarrow{\mathrm{P}_{\theta_0}} 0.$$

Собирая все вместе и применяя лемму Слуцкого, получаем, что

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

■

Сама теорема напрямую не говорит про оценку максимального правдоподобия, она снова апеллирует к некоторому решению уравнения правдоподобия. Сформулируем следствие, которое даст нам асимптотическую нормальность оценки максимального правдоподобия.

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

Теорема Бахадура

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

Если в условиях регулярности, похожих на **R0-R8** (в частности, условия для третьей производной $\ln p_\theta(x)$ заменяются на условия для второй), то имеет место следующий асимптотический аналог неравенства Рао-Крамера.

Теорема 7.4. (Бахадур)
Если в некоторых условиях регулярности оценка $\hat{\theta}_n(\mathbf{X})$ является асимптотически нормальной оценкой параметра θ с асимптотической дисперсией $\sigma^2(\theta)$, то
$\sigma^2(\theta) \geqslant i^{-1}(\theta)$ почти всюду по мере Лебега.

Теорему Бахадура мы докажем на следующей лекции, а сейчас обсудим ее результат.

Рассмотрим пример, показывающий, что неравенство действительно может не выполняться в отдельных точках.

Пример
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из нормального распределения $\mathcal{N}(\theta, 1)$. Введем следующую оценку параметра θ :

$$\hat{\theta}_n(\mathbf{X}) = \begin{cases} \overline{\mathbf{X}}, & \text{если } |\overline{\mathbf{X}}| \geqslant n^{-\frac{1}{4}}; \\ \beta\overline{\mathbf{X}}, & \text{если } |\overline{\mathbf{X}}| < n^{-\frac{1}{4}}, \end{cases}$$

где $\beta \in (0, 1)$ фиксированное число. Найдите асимптотическую дисперсию $\sigma^2(\theta)$ и сравните ее с обратной информацией Фишера $i^{-1}(\theta)$ одного элемента.

Решение

Как известно, усиленный закон больших чисел имеет скорость сходимости порядка $O(n^{-\frac{1}{2}})$. Следовательно, если $\theta \neq 0$, то

$$\lim_{n\to\infty} \mathrm{P}_\theta(\hat{\theta}_n(\mathbf{X}) = \overline{\mathbf{X}}) = 1,$$

Теперь рассмотрим распределение $\sqrt{n}(\hat{\theta}_n(\mathbf{X}) - \theta)$. Для любого $x \in \mathbb{R}$ имеем:

$$\begin{aligned} &\mathrm{P}_\theta(\sqrt{n}(\hat{\theta}_n(\mathbf{X}) - \theta) \leqslant x) = \\ &= \mathrm{P}_\theta\bigg(\sqrt{n}(\hat{\theta}_n(\mathbf{X}) - \theta) \leqslant x,\ \hat{\theta}_n(\mathbf{X}) = \overline{\mathbf{X}}\bigg) + \mathrm{P}_\theta\bigg(\sqrt{n}(\hat{\theta}_n(\mathbf{X}) - \theta) \leqslant x,\ \hat{\theta}_n(\mathbf{X}) = \beta\overline{\mathbf{X}}\bigg) = \\ &\qquad = \mathrm{P}_\theta(\sqrt{n}(\overline{\mathbf{X}} - \theta) \leqslant x) + o(1). \end{aligned}$$

Мы знаем, что $\mathrm{P}_\theta(\sqrt{n}(\overline{\mathbf{X}} - \theta) \leqslant x) = \Phi(x)$ - функция распределения $\mathcal{N}(0, 1)$, откуда получаем, что

$$\sqrt{n}(\hat{\theta}_n(\mathbf{X}) - \theta) \xrightarrow{d_\theta} \mathcal{N}(0, 1).$$

Пусть теперь $\theta = 0$. В таком случае, наоборот,

$$\lim_{n\to\infty} \mathrm{P}_\theta(\hat{\theta}_n(\mathbf{X}) = \beta\overline{\mathbf{X}}) = 1.$$

Тогда

$$\mathrm{P}_\theta(\sqrt{n}(\hat{\theta}_n(\mathbf{X})-\theta)\leqslant x)=$$
$$=\mathrm{P}_\theta\left(\sqrt{n}\overline{\mathbf{X}}\leqslant x,\ \hat{\theta}_n(\mathbf{X})=\overline{\mathbf{X}}\right)+\mathrm{P}_\theta\left(\sqrt{n}\overline{\mathbf{X}}\leqslant\frac{x}{\beta},\ \hat{\theta}_n(\mathbf{X})=\beta\overline{\mathbf{X}}\right)=$$
$$=\mathrm{P}_\theta\left(\sqrt{n}\overline{\mathbf{X}}\leqslant\frac{x}{\beta}\right)+o(1).$$

Следовательно,

$$\lim_{n\to\infty}\mathrm{P}_\theta(\sqrt{n}(\hat{\theta}_n(\mathbf{X})-\theta)\leqslant x)=\Phi\left(\frac{x}{\beta}\right)\Longrightarrow\sqrt{n}(\hat{\theta}_n(\mathbf{X})-\theta)\xrightarrow{d_\theta}\mathcal{N}(0,\beta^2).$$

В итоге, $\sqrt{n}(\hat{\theta}_n(\mathbf{X})-\theta)\xrightarrow{d_\theta}\mathcal{N}(0,\sigma^2(\theta))$, где $\sigma^2(\theta)=1$ для всех $\theta\neq 0$ и $\sigma^2(0)=\beta^2<1$. Однако информация Фишера всегда равна 1:

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

Эффективность оценки максимального правдоподобия

Следствие
В условиях регулярности **R0-R8** и условиях предыдущего следствия оценка максимального правдоподобия является наилучшей оценкой θ в асимптотическом подходе в классе асимптотически нормальных оценок с непрерывной асимптотической дисперсией.

Определение 7.1. Пусть $\hat{\theta}_n(\mathbf{X})$ - асимптотически нормальная оценка параметра θ с асимптотической дисперсией $\sigma^2(\theta)$. Если $\sigma^2(\theta)=i^{-1}(\theta)$, то оценка $\hat{\theta}_n(\mathbf{X})$ называется асимптотически эффективной оценкой θ.

Оказывается, что оценка максимального правдоподобия не только асимптотически эффективна, но и просто эффективна.

Теорема 7.5. (эффективность оценки максимального правдоподобия)
Если в условиях неравенства Рао-Крамера $\hat{\theta}(\mathbf{X})$ является эффективной оценкой θ, то $\hat{\theta}(\mathbf{X})$ есть оценка максимального правдоподобия.

Доказательство.
Воспользуемся критерием эффективности оценки при $\tau(\theta)=\theta$: для любого $\theta\in\Theta$ выполнено

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

Так как информация Фишера положительна, то $\mathcal{L}'(\theta, \mathbf{X})$ имеет тот же знак, что и $\hat{\theta}(\mathbf{X}) - \theta$: при $\theta < \hat{\theta}(\mathbf{X})$ выполнено $\mathcal{L}'(\theta, \mathbf{X}) > 0$ и наоборот. Тогда получаем, что $\hat{\theta}(\mathbf{X})$ - это единственная точка максимума на $p_\theta(\mathbf{X})$ (как функция от θ), то есть $\hat{\theta}(\mathbf{X})$ есть оценка максимального правдоподобия. ■

8. Лекция 8

Теорема Бахадура

В прошлый раз мы узнали, что в условиях регулярности оценка максимального правдоподобия является асимптотически нормальной с асимптотической дисперсией $\dfrac{1}{i(\theta)}$.

Вопрос: может ли асимптотическая дисперсия быть меньше?

Оказывается, принципиально это невозможно. Данный результат носит название теоремы Бахадура. Для его точной формулировки нам снова понадобятся условия регулярности.

R1 Параметрическое семейство распределений $\{\mathrm{P}_\theta, \theta \in \Theta\}$ - это доминируемое семейство с плотностью $p_\theta(x)$ по мере μ и различимыми распределениями, то есть $\mathrm{P}_{\theta_0} = \mathrm{P}_{\theta_1}$ тогда и только тогда, когда $\theta_0 = \theta_1$.

R2 $\mathbf{X} = (X_1, ..., X_n)$ - выборка растущего размера из неизвестного распределения $\mathrm{P} \in \{\mathrm{P}_\theta, \theta \in \Theta\}$.

R3 Θ есть открытый интервал на $\mathbb{R}$.

R4 Функция $l(x,\theta) = \ln p_\theta(x)$ дважды непрерывно дифференцируема по θ для всех $x \in \mathcal{X}$.

R5 Информация Фишера

$$i(\theta) = \mathrm{E}_\theta(l'(X_1,\theta))^2$$

положительна и конечна для всех $\theta \in \Theta$.

R6 Для всех $\theta \in \Theta$ выполнены равенства

$$\mathrm{E}_\theta l'(X_1,\theta) = 0, \quad -\mathrm{E}_\theta l''(X_1,\theta) = i(\theta),$$

R7 Для любого $\theta_0 \in \Theta$ существуют такие $\delta = \delta(\theta_0) > 0$ и функция $M(x) \geqslant 0$, что для всех $\theta \in [\theta_0 - \delta, \theta_0 + \delta]$ выполнено

$$|l''(x,\theta)| \leqslant M(x) \text{ и } \mathrm{E}_{\theta_0} M(X_1) < +\infty.$$

Теорема 8.1. (Бахадур)
Если в условиях регулярности **R1-R7** оценка $\tau_n(\mathbf{X})$ является асимптотически нормальной оценкой параметра θ с асимптотической дисперсией $\sigma^2(\theta)$, то $\sigma^2(\theta) \geqslant \dfrac{1}{i(\theta)}$ почти всюду по мере Лебега.

Вспомогательная теорема

Сразу приступить к ее доказательству мы не можем, нам понадобится одна вспомогательная теорема, которая представляет самостоятельный интерес.

Теорема 8.2. Пусть в условиях регулярности **R1-R7** для последовательности оценок $\{\tau_n(\mathbf{X}), n \in \mathbb{N}\}$ параметра θ выполнено свойство асимптотической нормальности: для всех $\theta \in \Theta$

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

Если для некоторого $\theta_0 \in \Theta$ и $\theta_n = \theta_0 + n^{-\frac{1}{2}}$ выполнено неравенство

$$\varliminf_{n\to\infty} \mathrm{P}_{\theta_n}(\tau_n(\mathbf{X}) \leqslant \theta_n) \leqslant \frac{1}{2}, \tag{1}$$

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

Доказательство.
Для доказательства Теоремы 8.2 необходимо сперва доказать несколько вспомогательных утверждений и ввести несколько обозначений. Начнем с того, что введем логарифмическую функцию правдоподобия:

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

Далее, введем следующую случайную величину:

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

Покажем, что для случайной величины T_n выполнено одно интересное свойство.

Лемма (1)
При фиксированном θ_0 последовательность случайных величин T_n сходится по распределению к стандартному нормальному распределению:

$$T_n \xrightarrow{d_{\theta_0}} \mathcal{N}(0, 1).$$

Доказательство. Воспользуемся дважды непрерывной дифференцируемостью $l(x, \theta)$ и разложим $L(\mathbf{X}, \theta_n)$ в ряд Тейлора в окрестности θ_0:

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

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

где $\tilde{\theta}_n \in (\theta_0, \theta_n)$ - некоторая промежуточная точка. Далее, мы хотим понять, куда сходятся слагаемые, так как если перенести из правой части равенства $\mathcal{L}(\mathbf{X}, \theta_0)$ в левую, то мы получим выражение $\mathcal{L}(\mathbf{X}, \theta_n) - \mathcal{L}(\mathbf{X}, \theta_0)$, которое стоит в скобках в выражении для T_n. Введем следующую случайную величину:

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

Докажем, что эта последовательность стремится к нулю почти наверное:

$$\xi_n \xrightarrow{\mathrm{P}_{\theta_0}-\text{п.н.}} 0.$$

Для этого введем следующую функцию: для $\varepsilon \in (0, \delta(\theta_0))$ ($\delta(\theta_0)$ - величина из условия **R7**)

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

Далее, положим $m(\varepsilon) = \mathrm{E}_{\theta_0} A(X_1, \varepsilon)$. Заметим, что $A(x, \varepsilon) \downarrow 0$ при $\varepsilon \to 0+$, так как чем меньше ε, тем меньше множество, по которому мы берем sup. Далее, воспользуемся свойством регулярности **R7**:

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

Тем самым, можно применить теорему Лебега о мажорируемой сходимости. Из нее следует, что $m(\varepsilon) \to 0$ при $\varepsilon \to 0$.

Далее, заметим, что если взять $\varepsilon^2 > \frac{1}{n}$, то

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

Однако по усиленному закону больших чисел

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

Из вышеуказанного следует, что для любого $\varepsilon > 0$

$$\varlimsup_{n\to\infty} |\xi_n| \leqslant \frac{1}{2}m(\varepsilon) \quad \mathrm{P}_{\theta_0} - \text{п.н.}$$

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

Теперь преобразуем T_n следующим образом:

$$T_n = \frac{1}{\sqrt{i(\theta_0)}}\left(\mathcal{L}(\mathbf{X},\theta_0) + \frac{1}{\sqrt{n}}\mathcal{L}'(\mathbf{X},\theta_0) + \frac{1}{2n}\mathcal{L}''(\mathbf{X},\tilde{\theta}_n) - \mathcal{L}(\mathbf{X},\theta_0) + \frac{i(\theta_0)}{2}\right) =$$

$$= \frac{1}{\sqrt{i(\theta_0)}}\left(\frac{1}{\sqrt{n}}\mathcal{L}'(\mathbf{X},\theta_0) + \frac{1}{2n}\mathcal{L}''(\mathbf{X},\theta_0) + \xi_n + \frac{i(\theta_0)}{2}\right).$$

Осталось найти предел по распределению данного выражения.

- Начнем с $n^{-\frac{1}{2}}\mathcal{L}'(\mathbf{X},\theta_0)$. Заметим, что из условия регулярности **R6** и центральной предельной теоремы следует, что

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

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

$$\frac{1}{2n}\mathcal{L}''(\mathbf{X},\theta_0) = \frac{1}{2n}\sum_{i=1}^{n} l''(X_i,\theta_0) \xrightarrow{\mathrm{P}_{\theta_0}-\text{п.н.}} \frac{1}{2}\mathrm{E}_{\theta_0} l''(X_i,\theta_0) = -\frac{i(\theta_0)}{2}.$$

- Из доказанного выше следует, что

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

В итоге, по лемме Слуцкого получаем, что

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

Лемма (1) доказана. ■

Простым следствием леммы (1) является следующее утверждение.

Лемма (2)
Для любых $y, \alpha > 0$ выполнено

$$\mathrm{E}_{\theta_0}(e^{\alpha T_n} \cdot \mathrm{I}\{T_n < y\}) \to \mathrm{E}(e^{\alpha\xi} \cdot \mathrm{I}\{\xi < y\})$$

при $n \to +\infty$, где $\xi \sim \mathcal{N}(0,1)$.

Доказательство.
Рассмотрим функцию $f(x) = e^{\alpha\ min(x,y)}$. Она является непрерывной и ограниченной на $\mathbb{R}$. Тогда в силу леммы 1

$$\mathrm{E}_{\theta_0} f(T_n) \to \mathrm{E} f(\xi).$$

С другой стороны,

$$f(x) - e^{\alpha x} \cdot \mathrm{I}\{x < y\} = e^{\alpha y} \cdot \mathrm{I}\{x \geqslant y\}.$$

Стало быть,

$$E_{\theta_0}(e^{\alpha T_n} \cdot \mathrm{I}\{T_n < y\}) = \mathrm{E}_{\theta_0} f(T_n) - \mathrm{E}_{\theta_0}(e^{\alpha y} \cdot \mathrm{I}\{T_n \geqslant y\}) = \mathrm{E}_{\theta_0} f(T_n) - e^{\alpha y} \mathrm{P}_{\theta_0}(T_n \geqslant y) \to$$

$$\to \mathrm{E} f(\xi) - e^{\alpha y} \mathrm{P}(\xi \geqslant y) = \mathrm{E} f(\xi) - \mathrm{E}(e^{\alpha y} \cdot \mathrm{I}\{\xi \geqslant y\}) = \mathrm{E}(e^{\alpha \xi} \cdot \mathrm{I}\{\xi < y\}).$$

Лемма (2) доказана. ∎

Лемма (3)
Пусть $\Phi(x)$ - функция распределения стандартной нормальной случайной величины. Тогда для всех $y \in \mathbb{R}$ выполнено

$$\lim_{n\to\infty} \mathrm{P}_{\theta_n}(T_n \geqslant y) = 1 - \Phi(y - \sqrt{i(\theta_0)}.$$

Доказательство.
Напомним, что T_n есть функция от выборки: $T_n = T_n(\mathbf{X})$. Рассмотрим вероятность

$$\mathrm{P}_{\theta_n}(T_n < y) = \int\limits_{\mathcal{X}^n} p_{\theta_n}(x_1, ..., x_n) \mathrm{I}\{T_n(x_1, ..., x_n) < y\} \mu(dx_1, ..., dx_n) =$$

$$= \int\limits_{\mathcal{X}^n} e^{\mathcal{L}(\mathbf{x}, \theta_n)} \mathrm{I}\{T_n(\mathbf{x}) < y\} \mu(d\mathbf{x}).$$

Вспомним, что из определения статистики T_n следует, что

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

Подставим это выражение в интеграл. Тогда по лемме (2)

$$\mathrm{P}_{\theta_n}(T_n < y) = e^{-\frac{i(\theta_0)}{2}} \int\limits_{\mathcal{X}^n} e^{\mathcal{L}(\mathbf{x}, \theta_0) + T_n(\mathbf{x})\sqrt{i(\theta_0)}} \mathrm{I}\{T_n(\mathbf{x}) < y\} \mu(d\mathbf{x}) =$$

$$= e^{-\frac{i(\theta_0)}{2}} \mathrm{E}_{\theta_0} \left(e^{T_n(\mathbf{X})\sqrt{i(\theta_0)}} \mathrm{I}\{T_n(\mathbf{x}) < y\} \right) \to e^{-\frac{i(\theta_0)}{2}} \mathrm{E} \left(e^{\sqrt{i(\theta_0)} \cdot \xi} \mathrm{I}\{\xi < y\} \right),$$

где $\xi \sim \mathcal{N}(0,1)$. Осталось заметить, что предел равен искомому выражению:

$$e^{-\frac{i(\theta_0)}{2}} \mathrm{E} \left(e^{\sqrt{i(\theta_0)} \cdot \xi} \mathrm{I}\{\xi < y\} \right) = e^{-\frac{i(\theta_0)}{2}} \int\limits_{-\infty}^{y} e^{\sqrt{i(\theta_0)} \cdot z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz =$$

$$= \int\limits_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\sqrt{i(\theta_0)})^2}{2}} dz = \Phi(y - \sqrt{i(\theta_0)}).$$

Лемма (3) доказана. ■

Мы близки к завершению доказательства теоремы 8.2. Зафиксируем $y > \sqrt{i(\theta_0)}$ и введем два события:

$$D_n = \{\tau_n \geqslant \theta_n\} \text{ и } S_n = \{T_n \geqslant y\}.$$

Согласно лемме 3 получаем, что

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

Теперь заметим, что по условию (1) теоремы 8.2:

$$\frac{1}{2} \geqslant \varliminf_{n\to\infty} \mathrm{P}_{\theta_n}(\tau_n(\mathbf{X}) \leqslant \theta_n) = 1 - \varlimsup_{n\to\infty} \mathrm{P}_{\theta_n}(D_n).$$

Следовательно, $\varlimsup\limits_{n\to\infty} \mathrm{P}_{\theta_n}(D_n) \geqslant \frac{1}{2}$. Значит, существует некоторая подпоследовательность индексов $\{n_k, k \in \mathbb{N}\}$ такая, что

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

Это означает, что при достаточно больших k выполнено

$$\mathrm{P}_{\theta_{n_k}}(D_{n_k}) > \mathrm{P}_{\theta_{n_k}}(S_{n_k}).$$

Следующая лемма говорит. что данное соотношение верно и для θ_0.

Лемма (4)
Для всех достаточно больших k выполнено

$$\mathrm{P}_{\theta_0}(D_{n_k}) > \mathrm{P}_{\theta_0}(S_{n_k}).$$

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

$$S_n = \{T_n(\mathbf{X}) \geqslant y\} = \left\{\frac{1}{\sqrt{i(\theta_0)}}\left(\mathcal{L}(\mathbf{X}, \theta_n) - \mathcal{L}(\mathbf{X}, \theta_0) + \frac{i(\theta_0)}{2}\right) \geqslant y\right\} =$$

$$= \{\mathcal{L}(\mathbf{X}, \theta_n) - \mathcal{L}(\mathbf{X}, \theta_0) \geqslant y'\} = \{p_{\theta_n}(\mathbf{X}) \geqslant \lambda p_{\theta_0}(\mathbf{X})\},$$

где λ - некоторое положительное число, $y' = y\sqrt{i(\theta_0)} - \dfrac{i(\theta_0)}{2}$.

Тогда для всех $\mathbf{x}$ будет выполнено неравенство:

$$(p_{\theta_n}(\mathbf{x}) - \lambda p_{\theta_0}(\mathbf{x}))\mathrm{I}\{\mathbf{x} \in S_n]\} \geqslant (p_{\theta_n}(\mathbf{x}) - \lambda p_{\theta_0}(\mathbf{x}))\mathrm{I}\{\mathbf{x} \in D_n\}.$$

Действительно,

- либо $\mathbf{x} \in S_n$, тогда левая часть равна $p_{\theta_n}(\mathbf{x}) - \lambda p_{\theta_0}(\mathbf{x}) \geqslant 0$, а индикатор $\mathrm{I}\{\mathbf{x} \in D_n\} \leqslant 1$;
- либо $\mathbf{x} \notin S_n$, тогда левая часть равна нулю, правая неположительна, так как $p_{\theta_n}(\mathbf{x}) - \lambda p_{\theta_0}(\mathbf{x}) < 0$.

Проинтегрируем данное неравенство по мере μ и получим:

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

Простой перегруппировкой мы получим желаемое:

$$\lambda(\mathrm{P}_{\theta_0}(D_n) - \mathrm{P}_{\theta_0}(S_n)) \geqslant \mathrm{P}_{\theta_n}(D_n) - \mathrm{P}_{\theta_n}(S_n) > 0.$$

Лемма (4) доказана. ■

Теперь найдем пределы $\mathrm{P}_{\theta_0}(D_n)$ и $\mathrm{P}_{\theta_0}(S_n)$ с ростом n. Согласно лемме (1) мы уже знаем, что

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

Далее, по условию теоремы

$$\lim_{n\to\infty} \mathrm{P}_{\theta_0}(D_n) = \lim_{n\to\infty} \mathrm{P}_{\theta_0}\left(\tau_n(\mathbf{X}) \geqslant \theta_0 + \frac{1}{\sqrt{n}}\right) = \lim_{n\to\infty} \mathrm{P}_{\theta_0}(\sqrt{n}(\tau_n(\mathbf{X}) - \theta_0) \geqslant 1) =$$

$$= \lim_{n\to\infty} \mathrm{P}_{\theta_0}\left(\frac{\sqrt{n}(\tau_n(\mathbf{X}) - \theta_0)}{\sigma(\theta_0)} \geqslant \frac{1}{\sigma(\theta_0)}\right) = 1 - \Phi\left(\frac{1}{\sigma(\theta_0)}\right).$$

Следовательно, из леммы (4) мы получаем, что для любого $y > \sqrt{i(\theta_0)}$ выполняется неравенство

$$\frac{1}{\sigma(\theta_0)} \leqslant y \Longleftrightarrow \sigma^2(\theta_0) \geqslant \frac{1}{y^2}\ .$$

Отсюда и вытекает искомое неравенство $\sigma^2(\theta_0) \geqslant \dfrac{1}{i(\theta_0)}$.

Теорема 8.2 доказана. ■

Доказательство теоремы Бахадура

Приступим к доказательству теоремы Бахадура.
Доказательство.
В силу теоремы 2 нам достаточно проверить, что неравенство (1) выполнено почти всюду по мере Лебега на Θ.

Рассмотрим следующую функцию: для всех $\theta \in \Theta$ положим

$$h_n(\theta) = \left| \mathrm{P}_\theta(\tau_n(\mathbf{X}) \leqslant \theta) - \frac{1}{2} \right| = \left| \mathrm{P}_\theta(\sqrt{n}(\tau_n(\mathbf{X}) - \theta) \leqslant 0) - \frac{1}{2} \right|,$$

и положим $h_n(\theta) = 0$ для $\theta \in \mathbb{R} \setminus \Theta$. Тогда $h_n(\theta) \in [0, \frac{1}{2}]$, кроме того из условия асимптотической нормальности мы получаем, что для всех $\theta \in \mathbb{R}$ выполнено

$$\lim_{n \to +\infty} h_n(\theta) = 0.$$

Теперь введем вторую функцию $g_n(\theta) = h_n(\theta + n^{-\frac{1}{2}})$, а также рассмотрим случайную величину $\xi \sim \mathcal{N}(0,1)$. Посчитаем математическое ожидание $g_n(\xi)$.

$$\mathrm{E}g_n(\xi) = \int\limits_{-\infty}^{+\infty} g_n(x) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \int\limits_{-\infty}^{+\infty} h_n\left(x + \frac{1}{\sqrt{n}}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

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

$$\mathrm{E}g_n(\xi) = \int\limits_{-\infty}^{+\infty} h_n(y) \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(y^2 - \frac{2y}{\sqrt{n}} + \frac{1}{n} \right) \right\} dy =$$

$$= \int\limits_{-\infty}^{+\infty} h_n(y) \exp\left\{ \frac{y}{\sqrt{n}} - \frac{1}{2n} \right\} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \mathrm{E}\left(h_n(\xi) \exp\left\{ \frac{\xi}{\sqrt{n}} - \frac{1}{2n} \right\} \right).$$

В силу свойств функций h_n выполнено

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

По теореме Лебега о мажорируемой сходимости получаем, что $\mathrm{E}g_n(\xi) \to 0$. Так как $g_n(x) \in [0, \frac{1}{2}]$, о это означает, что $g_n(\xi)$ сходится к нулю по вероятности. Но тогда можно выбрать подпоследовательность индексов $\{n_k, k \in \mathbb{N}\}$ такую, что $g_{n_k} \to 0$ п.н.

Остается заметить, что если $g_{n_k} \to 0$, то для такого θ будет выполнено условие (1):

$$\lim_{k\to+\infty} P_{\theta_{n_k}}(\tau_{n_k} < \theta_{n_k}) = \frac{1}{2} .$$

В силу доказанного $g_{n_k} \to 0$ почти всюду по мере Q, где Q - стандартное нормальное распределение. Но это эквивалентно тому, что $g_{n_k}(\theta) \to 0$ почти всюду по мере Лебега на $\mathbb{R}$. ■

9. Лекция 9

Эта лекция будет посвящена разделу теории вероятностей, который называется условное математическое ожидание. С помощью условного математического ожидания мы планируем разобраться с теми подходами к сравнению оценок, которые мы еще не разобрали, а именно с равномерным, байесовским и минимаксным.

Условное математическое ожидание

Вопрос: в теории вероятностей есть понятие условной вероятности. Пусть ξ и η - две случайные величины. Как можно определить $\mathrm{E}(\xi|\eta = y)$ и $\mathrm{E}(\xi|\eta)$?

Рассмотрим самый простой случай - простые случйные величины. Как известно, для них

$$\mathrm{E}\xi = \sum_{i=1}^{n} x_i \mathrm{P}(\xi = x_i).$$

Естественно ввести $\mathrm{E}(\xi|\eta = y)$ следующим образом:

$$\mathrm{E}(\xi|\eta = y) = \sum_{i=1}^{n} x_i \mathrm{P}(\xi = x_i|\eta = y).$$

Вопрос: как обобщить данное определение на случай непростой случайной величины ξ?

Домножим на $\mathrm{P}(\eta = y)$ правую часть равенства. Тогда

$$\sum_{i=1}^{n} x_i \mathrm{P}(\xi = x_i, \eta = y) = \mathrm{E}\left(\sum_{i=1}^{n} x_i \cdot \mathrm{I}\{\xi = x_i\}\mathrm{I}\{\eta = y\}\right) = \mathrm{E}(\xi \cdot \mathrm{I}\}\eta = y\}).$$

С другой стороны, если положить $\mathrm{E}(\xi|\eta = y) = \phi(y)$, то получим равенство

$$\phi(y)\mathrm{P}(\eta = y) = \mathrm{E}(\phi(y)\mathrm{I}\{\eta = y\}) = \mathrm{E}(\phi(\eta)\mathrm{I}\{\eta = y\}).$$

Тем самым, разумно сказать, что $\mathrm{E}(\xi|\eta = y)$ - это такая функция $\phi(y)$, что для любого y выполнено равенство

$$\mathrm{E}(\xi \cdot \mathrm{I}\{\eta = y\}) = \mathrm{E}(\phi(\eta)\mathrm{I}\{\eta = y\}).$$

Вопрос: как обобщить данное определение на случай непростой случайной величины η?

Пусть $(\Omega, \mathcal{F}, \mathrm{P})$ - вероятностное пространство, ξ - случайная величина на нем, а $\mathfrak{C} \subset \mathcal{F}$ под-σ-алгебра в $\mathcal{F}$.

Определение 9.1. Напомним, что σ-алгеброй, порожденной случайной величиной ξ называется

$$\mathcal{F}_\xi = \{\xi^{-1}(B) - \{\xi \in B\} : B \in \mathcal{B}(\mathcal{R})\}.$$

Случайная величина ξ называется $\mathfrak{C}$-измеримой, если порожденная ею σ-алгебра входит в $\mathfrak{C}$:

$$\mathcal{F}_\xi \subset \mathfrak{C}.$$

Определение 9.2. Условным математическим ожиданием ξ относительно σ-алгебры $\mathfrak{C}$ называется случайная величина $\mathrm{E}(\xi|\mathfrak{C})$, удовлетворяющая следующим свойствам:

1) $\mathrm{E}(\xi|\mathfrak{C})$ является $\mathfrak{C}$-измеримой случайной величиной (свойство измеримости);

2) для любого $A \in \mathfrak{C}$ выполняется равенство $\mathrm{E}(\xi \cdot I_A) = \mathrm{E}(\mathrm{E}(\xi|\mathfrak{C})I_A)$ или

$$\int\limits_A \xi d\mathrm{P} = \int\limits_A \mathrm{E}(\xi|\mathfrak{C}) d\mathrm{P} \quad \text{(интегральное свойство).}$$

Вопрос: почему $\mathrm{E}(\xi|\mathfrak{C}) \neq \xi$ в общем случае?
Ответ: ξ не обязательно является $\mathfrak{C}$-измеримой! Соответственно, мы пытаемся подобрать такую $\mathfrak{C}$-измеримую случайную величину, для которой интегрирование по любому событию из $\mathfrak{C}$ дает то же самое, что и для ξ.

Существование УМО

Определение 9.3. Пусть $(\Omega, \mathcal{F}, \mathrm{P})$ - вероятностное пространство.
Функция $\nu : \mathcal{F} \to \mathbb{R}$ называется (конечным) зарядом (или мерой со знаком), если ν - счетно-аддитивна на $\mathcal{F}$, то есть

$$\nu\left(\bigsqcup_{n=1}^{\infty} A_n\right) = \sum_{i=1}^{\infty} \nu(A_n),$$

где ряд в правой части сходится абсолютно, и

$$\sup_{a \in \mathcal{F}} |\nu(A)| < +\infty.$$

Заряд ν является абсолютно непрерывным относительно P, если из того, что $\mathrm{P}(A) = 0$ следует, что $\nu(A) = 0$.

Теорема 9.1. (Радон-Никодим)
Пусть $(\Omega, \mathcal{F}, \mathrm{P})$ - вероятностное пространство, а ν - заряд, абсолютно непрерывный относительно P. Тогда существует единственная (с точностью до равенства п.н.)

случайная величина $\eta \in L^1(\Omega, \mathcal{F}, \mathrm{P})$ (с конечным математическим ожиданием) такая, что для любого $A \in \mathcal{F}$ выполнено

$$\nu(A) = \int\limits_A \eta d\mathrm{P} = \mathrm{E}(\eta \cdot I_A).$$

Замечание
В этом случае $\eta = \dfrac{d\nu}{d\mathrm{P}}$ называется производной Радона-Никодима.

Лемма (о существовании УМО)
Если $\mathrm{E}|\xi| < \infty$, то для любой σ-алгебры $\mathfrak{C} \subset \mathcal{F}$ условное математическое ожидание $\mathrm{E}(\xi|\mathfrak{C})$ существует и единственно с точностью до равенства почти наверное.
Доказательство.
Рассмотрим вероятностное пространство $(\Omega, \mathfrak{C}, \mathrm{P})$. Для любого $A \in \mathfrak{C}$ положим

$$\nu(A) = \int\limits_A \xi d\mathrm{P} = \mathrm{E}(\xi \cdot I_A).$$

Это заряд на $(\Omega, \mathfrak{C}, \mathrm{P})$, абсолютно непрерывный относительно P. По теореме Радона-Никодима существует единственная (п.н.) случайная величина $\eta \in L^1(\Omega, \mathfrak{C}, \mathrm{P})$ такая, что для любого $A \in \mathfrak{C}$

$$\nu(A) = \int\limits_A \eta d\mathrm{P} = \mathrm{E}(\eta \cdot I_A).$$

Случайная величина η является $\mathfrak{C}$-измеримой и удовлетворяет интегральному свойству. Значит, по определению $\eta = \mathrm{E}(\xi|\mathfrak{C})$.

Единственность следует из единственности производной Радона-Никодима в теореме. ■

Замечание
Условное математическое ожидание определятся с точностью до равенства п.н. Иногда бывает важным осуществлять правильный подбор его "варианта".

Мы разобрались со существованием. Оставшуюся часть лекции будем обсуждать свойства УМО.

Дискретные σ-алгебры

Начнем с явной формулы, по которой можно будет вычислить УМО в ситуации, когда $\mathfrak{C}$ - дискретная σ-алгебра.

Напоминание

Дискретная σ-алгебра $\mathfrak{C}$ - это σ-алгебра, которая порождена некоторым (не более чем счетным) разбиением $\{D_n, n \in \mathbb{N}\}$ пространства Ω. То есть, в такую σ-алгебру входят сами D_n и всевозможные их объединения.

Лемма

Если σ-алгебра порождена разбиением $\{D_n, n \in \mathbb{N}\}$ с условием $\mathrm{P}(D_n) > 0$ для всех $n \in \mathbb{N}$, то для любой случайной величины ξ с условием $\mathrm{E}|\xi| < +\infty$ выполнено

$$\mathrm{E}(\xi|\mathfrak{C})(\omega) = \sum_{i=1}^{\infty} \frac{\mathrm{E}(\xi \cdot \mathrm{I}_{D_n})}{\mathrm{P}(D_n)} \cdot \mathrm{I}_{D_n}(\omega).$$

Попробуем осознать некоторый смысл УМО. Пусть $\Omega = [0,1]$ разрезан на 4 части D_i, $i = 1,2,3,4$. Случайная величина ξ на Ω - функция на отрезке. Что же такое условное математическое ожидание? Из последней формулы мы поняли, что в данной ситуации условным математическим ожиданием ξ относительно σ-алгебры, порожденной таким разбиением, будет случайная величина, которая является константой на каждом из множеств D_i.

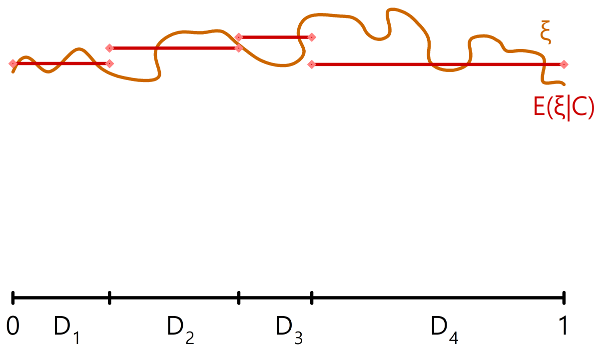

Рис. 5

По сути, условное математическое ожидание - это усреднение нашей случайной величины по множествам из σ-алгебры. В данном случае это происходит явным образом, так как множества простые (не пересекаются, из счетное количество). В общей ситуации все не так явно.

Перейдем теперь к доказательству леммы.

Доказательство.
План доказательства прост. В правой части нам предлагают кандидата в УМО, в силу единственности надо проверить для него два свойства: свойство измеримости и интегральное свойство.

Правая часть есть линейная комбинация несовместных индикаторов событий $D_n \in \mathfrak{C}$, так что это $\mathfrak{C}$-измеримая случайная величина.

Проверим интегральное свойство. Если $A \in \mathfrak{C}$, то оно представляет собой лишь дизъюнктное объединение элементов D_n. В связи с этим достаточно рассмотреть ситуацию $A = D_m$ для некоторого $m \in \mathbb{N}$. Имеем:

$$\mathrm{E}\left(\left(\sum_{n=1}^{\infty} \frac{\mathrm{E}(\xi \cdot \mathrm{I}_{D_n})}{\mathrm{P}(D_n)} \cdot \mathrm{I}_{D_n}\right)\mathrm{I}_{D_m}\right) \overset{(*)}{=} \mathrm{E}\left(\frac{\mathrm{E}(\xi \cdot \mathrm{I}_{D_m})}{\mathrm{P}(D_m)} \cdot \mathrm{I}_{D_m}\right) = \frac{\mathrm{E}(\xi \cdot \mathrm{I}_{D_m})}{\mathrm{P}(D_m)} \cdot \mathrm{P}(D_m) = \mathrm{E}(\xi \cdot \mathrm{I}_{D_m}),$$

$(*)$ - из-за несовместности событий D_n остается только слагаемое при $n = m$. ■

Свойства УМО

В общем случае таких хороших формул, как в прошлой лемме, нет, и учиться считать УМО в общем случае мы будем на следующей лекции. Сейчас разберем свойства УМО.

Всюду далее предполагаем конечность необходимых математических ожиданий у рассматриваемых случайных величин, чтобы УМО были корректно определены и существовали.

Свойство (1)
Если случайная величина ξ является $\mathfrak{C}$-измеримой, то $\mathrm{E}(\xi|\mathfrak{C}) = \xi$.
Доказательство.
Очевидно, что для ξ выполняются свойство измеримости и интегральное свойство. ■

Свойство (2, линейность)
Для любых случайных величин ξ, η и σ-алгебры $\mathfrak{C}$ выполнено, что для любых $a, b \in \mathbb{R}$:

$$\mathrm{E}(a\xi + b\eta|\mathfrak{C}) = a\mathrm{E}(\xi\mathfrak{C}) + b\mathrm{E}(\eta|\mathfrak{C}).$$

Доказательство.
Для начала заметим, что и $\mathrm{E}(\xi|\mathfrak{C})$ и $\mathrm{E}(\eta|\mathfrak{C})$ являются $\mathfrak{C}$-измеримыми случайными величинами. Тогда $a\mathrm{E}(\xi|\mathfrak{C}) + b\mathrm{E}(\eta|\mathfrak{C})$ - это тоже $\mathfrak{C}$-измеримая случайная величина.

Осталось проверить интегральное свойство. В силу линейности математического ожидания для любого $A \in \mathfrak{C}$

$$\mathrm{E}((a\xi + b\eta) \cdot \mathrm{I}_A) = a\mathrm{E}(\xi \cdot \mathrm{I}_A) + b\mathrm{E}(\eta \cdot \mathrm{I}_A).$$

Согласно интегральному свойству

$$a\mathrm{E}(\xi \cdot \mathrm{I}_A) + b\mathrm{E}(\eta \cdot \mathrm{I}_A) = a\mathrm{E}(\mathrm{E}(\xi|\mathfrak{C})I_A) + b\mathrm{E}(\mathrm{E}(\eta|\mathfrak{C})I_A)$$

Свернем сумму назад:

$$a\mathrm{E}(\mathrm{E}(\xi|\mathfrak{C})I_A) + b\mathrm{E}(\mathrm{E}(\eta|\mathfrak{C})I_A) = \mathrm{E}((a\mathrm{E}(\xi|\mathfrak{C}) + b\mathrm{E}(\eta|\mathfrak{C})) \cdot \mathrm{I}_A).$$

Тем самым, получаем желаемое. ∎

Свойство (3, формула полной вероятности)

$$\mathrm{E}(\mathrm{E}(\xi|\mathfrak{C})) = \mathrm{E}\xi.$$

Доказательство.
Воспользуемся интегральным свойством, положив $\mathfrak{C} \ni A = \Omega$:

$$\mathrm{E}\xi = \mathrm{E}(\xi \cdot \mathrm{I}_A) = \mathrm{E}(\mathrm{E}(\xi|\mathfrak{C}) \cdot \mathrm{I}_A)\mathrm{E}(\mathrm{E}\xi|\mathfrak{C})).$$

∎

Свойство (4)
Если ξ независима с $\mathfrak{C}$, то $\mathrm{E}(\xi|\mathfrak{C}) = \mathrm{E}\xi$.
Доказательство.
Напомним, что ξ независима с $\mathfrak{C}$ тогда и только тогда, когда для любого $A \in \mathfrak{C}$ случайная величина ξ независима с I_A. Заметим, что $\mathrm{E}\xi$ - это константа. Тогда $\mathrm{E}\xi$ является $\mathfrak{C}$-измеримой. Проверим интегральное свойство: для любого $A \in \mathfrak{C}$

$$\mathrm{E}(\xi \cdot \mathrm{I}_A) = |\text{независимость}| = \mathrm{E}\xi \cdot \mathrm{E}(\mathrm{I}_A) = \mathrm{E}(\mathrm{E}\xi \cdot \mathrm{I}_A).$$

∎

Свойство (5, сохранение отношения порядка)
Если $\xi \leqslant \eta$ п.н., то

$$\mathrm{E}(\xi|\mathfrak{C}) \leqslant \mathrm{E}(\eta|\mathfrak{C}) \text{ п.н.}$$

Доказательство.
Рассмотрим случайную величину $\delta = \mathrm{E}(\xi|\mathfrak{C}) - \mathrm{E}(\eta|\mathfrak{C})$. Она является $\mathfrak{C}$-измеримой

случайной величиной. Теперь рассмотрим $\mathrm{E}(\delta \cdot \mathrm{I}\{\delta > 0\})$. Заметим, что $\delta \cdot \mathrm{I}\{\delta > 0\} \geqslant 0$. Далее, рассмотрим событие $B = \{\delta > 0\} \in \mathfrak{C}$. Тогда в силу выбора δ

$$\mathrm{E}(\delta \cdot \mathrm{I}\{\delta > 0\}) = \mathrm{E}(\mathrm{E}(\xi|\mathfrak{C}) - \mathrm{E}(\eta|\mathfrak{C}))\mathrm{I}_B).$$

Пользуясь линейностью математического ожидания и интегральным свойством, получаем, что

$$\mathrm{E}(\delta \cdot \mathrm{I}\{\delta > 0\}) = \mathrm{E}(\xi \cdot \mathrm{I}_B - \eta \cdot \mathrm{I}_B) \leqslant 0 \text{ так как } \xi \leqslant \eta.$$

Тогда получаем, что $\delta \cdot \mathrm{I}\{\delta < 0\} = 0$ п.н., то есть $\delta \leqslant$ п.н. ■

Свойство (6, неравенство модуля)

$$|\mathrm{E}(\xi|\mathfrak{C})| \leqslant \mathrm{E}(|\xi||\mathfrak{C}).$$

Доказательство. Следует из свойств (2) и (5). ■

Свойство (7, телескопическое)
Пусть ξ - случайная величина, а $\mathfrak{C}_1 \subset \mathfrak{C}_2$ - две σ-алгебры. Тогда

1) $\mathrm{E}(\mathrm{E}(\xi|\mathfrak{C}_1)|\mathfrak{C}_2) = \mathrm{E}(\xi|\mathfrak{C}_1)$,

2) $\mathrm{E}(\mathrm{E}(\xi|\mathfrak{C}_2)|\mathfrak{C}_1) = \mathrm{E}(\xi|\mathfrak{C}_1)$.

Доказательство.
Первый пункт почти очевиден. Заметим, что случайная величина $\mathrm{E}(\xi|\mathfrak{C}_1)$ является $\mathfrak{C}_1$ измеримой, а, значит, и $\mathfrak{C}_2$-измеримой (т.к. $\mathfrak{C}_1 \subset \mathfrak{C}_2$). Далее, применяем свойство (1).

Теперь докажем второй утверждение. Заметим, что свойство измеримости у нас уже есть. Проверим интегральное свойство. Для любого $A \in \mathfrak{C}_1 \subset \mathfrak{C}_2$:

$$\mathrm{E}(\mathrm{E}(\xi|\mathfrak{C}_1)\mathrm{I}_A) = |\text{инт. свойство для } \mathfrak{C}_1| = \mathrm{E}(\xi \cdot \mathrm{I}_A) =$$

$$= |\text{инт. свойство для } \mathfrak{C}_2| = \mathrm{E}(\mathrm{E}(\xi|\mathfrak{C}_2)\mathrm{I}_A).$$

■

Свойство (8, предельный переход под знаком УМО)
Пусть $\{\xi_n, n \in \mathbb{N}\}$, ξ - случайные величины, $\mathrm{E}|\xi| < \infty$, $\mathrm{E}|\xi_n| < \infty$. Тогда

1) Если $0 \leqslant \xi_n \uparrow \xi$ п.н., то $\mathrm{E}(\xi_n|\mathfrak{C}) \uparrow \mathrm{E}(\xi|\mathfrak{C})$ п.н.

2) Пусть $\xi_n \xrightarrow{\text{п.н.}} \xi$ и $|\xi_n| \leqslant \delta$ для всех n, причем $\mathrm{E}\delta < +\infty$. Тогда

$$\mathrm{E}(\xi_n|\mathfrak{C}) \xrightarrow{\text{п.н.}} \mathrm{E}(\xi|\mathfrak{C}).$$

Доказательство.

1) В силу того, что ξ_n монотонно возрастают, то и $\mathrm{E}(\xi_n|\mathfrak{C})$ монотонно возрастают (по свойству (5)). Тогда п.н. существует предел $\eta = \lim\limits_{n\to\infty} \mathrm{E}(\xi_n|\mathfrak{C})$. Проверим, что $\eta = \mathrm{E}(\xi|\mathfrak{C})$.

   Заметим, что η является $\mathfrak{C}$-измеримой случайной величиной, как предел п.н. $\mathfrak{C}$-измеримых случайных величин. Проверим интегральное свойство.

   Для любого $A \in \mathfrak{C}$ п.н. выполнено $\xi_n \cdot \mathrm{I}_A \uparrow \xi \cdot \mathrm{I}_A$ и $\mathrm{E}(\xi_n|\mathfrak{C}) \cdot \mathrm{I}_A \uparrow \eta \cdot \mathrm{I}_A$. Тогда по теореме о монотонной сходимости:

$$\mathrm{E}(\xi \cdot \mathrm{I}_A) = \lim_{n\to\infty} \mathrm{E}(\xi_n \cdot \mathrm{I}_A) = |\text{инт. свойство}| = \lim_{n\to\infty} \mathrm{E}(\mathrm{E}(\xi_n|\mathfrak{C})\mathrm{I}_A) = \mathrm{E}(\eta \cdot \mathrm{I}_A).$$

2) Рассмотрим $\eta_n = \sup\limits_{m\geqslant n} |\xi_m - \xi_n|$. Тогда в силу свойства (5)

$$|\mathrm{E}(\xi_n|\mathfrak{C}) - \mathrm{E}(\xi|\mathfrak{C})| = |\mathrm{E}(\xi_n - \xi|\mathfrak{C})| \leqslant \mathrm{E}(|\xi_n - \xi||\mathfrak{C}) \leqslant \mathrm{E}(\eta_n|\mathfrak{C}).$$

   Далее, $\eta_n \downarrow 0$ п.н. и $|\eta_n| \leqslant 2\delta$. Согласно пункту 1) получаем, что

$$\mathrm{E}(\eta_n|\mathfrak{C}) \xrightarrow{\text{п.н.}} \mathrm{E}(0|\mathfrak{C}) = 0.$$

■

Свойство (9)
Пусть η - это $\mathfrak{C}$-измеримая случайная величина, а математические ожидания $\mathrm{E}\xi\eta$ и $\mathrm{E}\xi$ конечны. Тогда

$$\mathrm{E}(\xi \cdot \eta|\mathfrak{C}) = \eta\mathrm{E}(\xi|\mathfrak{C}).$$

Доказательство.
Заметим, что $\eta\mathrm{E}(\xi|\mathfrak{C})$ есть $\mathfrak{C}$-измеримая случайная величина (произведение двух $\mathfrak{C}$-измеримых). Проверим интегральное свойство.

Пусть сначала $\eta = \mathrm{I}_B$ для некоторого $B\mathfrak{C}$. Тогда для любого $A \in \mathfrak{C}$

$$\mathrm{E}(\eta\xi\cdot\mathrm{I}_A) = \mathrm{E}(\xi\cdot\mathrm{I}_B\cdot\mathrm{I}_A) = \mathrm{E}(\xi\cdot\mathrm{I}_{A\cap B}) = \mathrm{E}(\mathrm{E}(\xi|\mathfrak{C})\mathrm{I}_{A\cap B}) = \mathrm{E}(\mathrm{E}(\xi|\mathfrak{C})\mathrm{I}_B\cdot\mathrm{I}_A) = \mathrm{E}(\eta\mathrm{E}(\xi|\mathfrak{C})\mathrm{I}_A).$$

Заметим, что $\mathrm{E}(\eta\xi\cdot\mathrm{I}_A)$ и $\mathrm{E}(\eta\mathrm{E}(\xi|\mathfrak{C})\mathrm{I}_A)$ - линейные функции от случайной величины η. Тогда в силу линейности математического ожидания, интегральное свойство будет выполнено и для простых случайных величин вида $\eta = \sum\limits_{k=1}^{m} c_k\mathrm{I}_{B_k}$, $c_k \in \mathbb{R}$, $B_k \in \mathfrak{C}$. Для произвольной η возьмем последовательность простых $\mathfrak{C}$ измеримых случайных величин η_n такую, что $\eta_n \xrightarrow{\text{п.н.}} \eta$ и $|\eta_n| \leqslant |\eta|$. Тогда по свойству (8) получаем, что

$$\mathrm{E}(\eta_n\xi|\mathfrak{C}) \xrightarrow{\text{п.н.}} \mathrm{E}(\eta\xi|\mathfrak{C}).$$

Однако

$$\mathrm{E}(\eta_n\xi|\mathfrak{C}) = \eta_n\mathrm{E}(\xi|\mathfrak{C}) \xrightarrow{\text{п.н.}} \eta\mathrm{E}(\xi|\mathfrak{C}).$$

Следовательно, $\mathrm{E}(\xi\eta|\mathfrak{C}) = \eta\mathrm{E}(\xi|\mathfrak{C})$. ∎

Свойство (10, неравенство Йенсена)
Пусть φ - выпуклая книзу функция. тогда

$$\mathrm{E}(\varphi(\xi)|\mathfrak{C}) \geqslant \varphi(\mathrm{E}(\xi|\mathfrak{C})).$$

Доказательство.
В силу выпуклости для любого $x \in \mathbb{R}$ существует $\lambda(x)$ такая, что для любого $y \in \mathbb{R}$ выполнено

$$\varphi(y) \geqslant \varphi(x) + \lambda(x)(y - x).$$

Положим $y = \xi$, $x = \mathrm{E}(\xi|\mathfrak{C})$ и возьмем УМО относительно $\mathfrak{C}$ от обеих частей неравенства. Пользуясь свойствами (1) и (9), получим:

$$\mathrm{E}(\varphi(\xi)|\mathfrak{C}) \geqslant \mathrm{E}\Big(\varphi(\mathrm{E}(\xi|\mathfrak{C}))|\mathfrak{C}\Big) + \mathrm{E}\Big(\lambda(\mathrm{E}(\xi|\mathfrak{C}))(\xi - \mathrm{E}(\xi|\mathfrak{C}))|\mathfrak{C}\Big) =$$
$$= |\text{выносим } \lambda(\mathrm{E}(\xi|\mathfrak{C})) \text{ как } \mathfrak{C}\text{- измеримую с.в.}| =$$
$$= \varphi(\mathrm{E}(\xi|\mathfrak{C})) + \lambda(\mathrm{E}(\xi|\mathfrak{C}))\mathrm{E}(\xi - \mathrm{E}(\xi|\mathfrak{C})|\mathfrak{C}) = \varphi(\mathrm{E}(\xi|\mathfrak{C})).$$

∎

10. Лекция 10

Условное математическое ожидание

Нам понадобятся еще несколько определений, связанных с УМО. Пусть $(\Omega, \mathcal{F}, \mathrm{P})$ - вероятностное пространство.

Определение 10.1. Пусть $A \in \mathcal{F}$ - событие. Тогда для под-σ-алгебры $\mathfrak{C} \subset \mathcal{F}$ определим

$$\mathrm{P}(A|\mathfrak{C}) := \mathrm{E}(I_A|\mathfrak{C}).$$

Если ξ и η - две случайные величины, то полагаем

$$\mathrm{E}(\xi|\eta) := \mathrm{E}(\xi|\mathcal{F}_n),$$

где $\mathcal{F}_n$ - σ-алгебра, порожденная случайной величиной η.

Замечание
Случайная величина ζ является $\mathcal{F}_n$-измеримой тогда и только тогда, когда существует такая борелевская функция $\phi(x)$, что $\zeta = \phi(\eta)$.

Определение 10.2. Пусть ξ и η - две случайные величины. Тогда величиной $\mathrm{E}(\xi|\eta = y)$ называется такая борелевская функция $\varphi(y)$, что для любого $B \in \mathcal{B}(\mathbb{R})$ выполняется равенство

$$\mathrm{E}(\xi \cdot \mathrm{I}\{\eta \in B\}) = \mathrm{E}(\varphi(\eta) \cdot \mathrm{I}\{\eta \in B\}) = \int\limits_B \varphi(y)\mathrm{P}_\eta(dy).$$

Нам необходима такая конструкция, так как мы хотим вычислять УМО ξ относительно другой случайной величины. Для этого удобно сначала вычислять подобные математические ожидания $\mathrm{E}(\xi|\eta = y)$, так как мы имеем дело не со случайными величинами, а с формально более простыми вещами, например, с борелевскими функциями.

Лемма
Если $\mathrm{E}|\xi| < +\infty$, то $\mathrm{E}(\xi|\eta = y)$ существует и единственная (с точностью до равенства P_η-п.н.).
Доказательство.
Для борелевских множеств $B \in \mathcal{B}(\mathbb{R})$ положим

$$\mathrm{Q}(B) = \mathrm{E}(\xi \cdot \mathrm{I}\{\eta \in B\}).$$

Заметим, что это заряд на вероятностном пространстве $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathrm{P}_\eta)$, абсолютно непрерывный относительно P_η. По теореме Радона-Никодима существует единственная P_η-п.н. случайная величина φ на $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathrm{P}_\eta)$ (борелевская функция!) такая, что для любого $B \in \mathcal{B}(\mathbb{R})$

$$\mathrm{Q}(B) = \int\limits_B \varphi(y) \mathrm{P}_\eta(dy).$$

По определению получаем, что $\varphi(y) = \mathrm{E}(\xi | \eta = y)$. ■

Вопрос: как связаны $\mathrm{E}(\xi|\eta)$ и $\mathrm{E}(\xi|\eta = y)$?

Утверждение

$$\mathrm{E}(\xi|\eta = y) = \varphi(y) \Longleftrightarrow \mathrm{E}(\xi|\eta) = \varphi(\eta).$$

Доказательство.

$\boxed{\Rightarrow}$ Пусть $B \in \mathcal{B}(\mathbb{R})$. Тогда

$$\mathrm{E}(\xi \cdot \mathrm{I}\{\eta \in B\}) = \int\limits_B \varphi(y) \mathrm{P}_\eta(dy) = \mathrm{E}(\varphi(\eta) \cdot \mathrm{I}\{\eta \in B\}).$$

Значит, $\varphi(\eta) = \mathrm{E}(\xi|\eta)$.

$\boxed{\Leftarrow}$ То же самое в обратном порядке:

$$\mathrm{E}(\xi \cdot \mathrm{I}\{\eta \in B\}) = \mathrm{E}(\varphi(\eta) \cdot \mathrm{I}\{\eta \in B\}) = \int\limits_B \varphi(y) \mathrm{P}_\eta(dy).$$

Следовательно, $\varphi(y) = \mathrm{E}(\xi|\eta = y)$. ■

Свойства УМО

Условное математическое ожидание $\mathrm{E}(\xi|\eta = y)$ обладает многими свойствами условного математического ожидания $\mathrm{E}(\xi|\eta)$. В частности:

- линейность,
- сохранение отношения порядка,
- предельный переход под знаком УМО,
- неравенство Йенсена.

Доказываются они точно так же, как и свойства $\mathrm{E}(\xi|\eta)$. В связи с этим детали мы опустим.

Условное распределение и условная плотность

Условные математические ожидания можно считать с помощью условных распределений. Правильное определение (регулярного) условного распределения - следующее.

Определение 10.3. Условным распределением случайной величины ξ относительно случайной величины η называется такая функция $\mathrm{P}(B, y)$, $B \in \mathcal{B}(\mathbb{R})$, $y \in \mathbb{R}$, что

1) $\mathrm{P}(\cdot, y)$ является распределением вероятностей на $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ для каждого $y \in \mathbb{R}$;

2) $\mathrm{P}(B, \cdot)$ является борелевской функцией для каждого $B \in \mathcal{B}(\mathbb{R})$;

3) Для любых $B \in \mathcal{B}(\mathbb{R})$ и $A \in \mathcal{B}(\mathbb{R})$ выполняется равенство

$$\mathrm{P}(\xi \in B, \eta \in A) = \mathrm{E}(\mathrm{P}(B, \eta) \cdot \mathrm{I}\{\eta \in A\}).$$

Теорема 10.1. Условное распределение существует.

Обозначение: $\mathrm{P}(B, y) = \mathrm{P}(\xi \in B | \eta = y)$.

Замечание
Для любого $B \in \mathcal{B}(\mathbb{R})$ можно было бы определить

$$\mathrm{P}(\xi \in B | \eta = y) := \mathrm{E}(\mathrm{I}\{\xi \in B\} | \eta = y).$$

Тогда будут выполнены свойства 2 и 3, но свойство 1 (грубо говоря, счетная аддитивность по B) не будет выполнено для всех множеств B сразу.

Определение 10.4. Условной плотностью случайной величины ξ относительно случайной величины η называется плотность условного распределения (по некоторой мере μ), то есть такая функция $f_{\xi|\eta}(x|y)$, $x, y \in \mathbb{R}$, что для любого $B \in \mathcal{B}(\mathbb{R})$ выполняется равенство:

$$\mathrm{P}(\xi \in B | \eta = y) = \int\limits_B f_{\xi|\eta}(x|y) \mu(dx).$$

Теорема о вычислении УМО

С помощью условных плотностей можно находить условные математические ожидания по той же формуле, что и обычные математические ожидания с помощью обычных плотностей.

Теорема 10.2. Если существует условная плотность $f_{\xi|\eta}(x|y)$ случайной величины ξ относительно случайной величины η, то для любой борелевской функции $g(x) : \mathbb{R} \to \mathbb{R}$ выполнено:

$$\mathrm{E}(g(\xi)|\eta = y) = \int\limits_{\mathbb{R}} g(x) f_{\xi|\eta}(x|y)\mu(dx).$$

Доказательство.
Пусть сначала $g(x) = \mathrm{I}\{x \in A\}$ для некоторого борелевского множества A. Тогда для любого y

$$E(g(\xi)|\eta = y) = \mathrm{P}(\xi \in A|\eta = y) = \int\limits_{A} f_{\xi|\eta}(x|y)\mu(dx) =$$

$$= \int\limits_{\mathbb{R}} \mathrm{I}\{x \in A\} f_{\xi|\eta}(x|y)\mu(dx) = \int\limits_{\mathbb{R}} g(x) f_{\xi|\eta}(x|y)\mu(dx).$$

Раз формула выполнена для индикаторов, то в силу линейности математического ожидания формула верна и для линейных комбинация индикаторов, то есть для простых функций $g(x)$. Для произвольной функции $g(x)$ справедливость формулы устанавливается предельным переходом от простых функций. ■

Вычисление условной плотности

Вопрос: как вычислять условную плотность?

Теорема 10.3. (достаточное условие существования условной плотности)
Пусть случайные величины ξ и η имеют совместную плотность $f_{\xi,\eta}(x, y)$ по мере $\mu \times \nu$ на $\mathbb{R}^2$. Тогда функция

$$f_{\xi|\eta}(x|y) = \begin{cases} \dfrac{f_{\xi,\eta}(x, y)}{f_\eta(y)}, & \text{если } f_\eta(y) > 0; \\ 0, & \text{иначе;} \end{cases}$$

является условной плотностью ξ по мере μ. Здесь $f_\eta(y)$- это плотность случайной величины η по мере ν.

Доказательство.
Надо проверить, что функция $P(B, y) = \int\limits_{B} f_{\xi|\eta}(x|y)\mu(dx)$ является условным распределением ξ относительно η, то есть проверить свойства 1-3 из определения.

Свойство 2 вытекает из теоремы Фубини, так как мы интегрируем совместную плотность по одной переменной, а поскольку совместная плотность интегрируема на квадрате, то по теореме Фубини интеграл по одной координате является измеримой функцией (в нашем случае борелевской) от другой переменной.

Свойство 1 следует из свойств интеграла Лебега.

Проверим свойство 3. Для любых борелевских множеств A, B выполнено

$$\mathrm{P}(\xi \in B, \eta \in A) = \int\limits_{B \times A} f_{\xi,\eta}(x,y)\mu(dx)\nu(dy) =$$

$$= \int\limits_A \left(\int\limits_B \frac{f_{\xi,\eta}(x,y)}{f_\eta(y)}\mu(dx) \right) f_\eta(y)\nu(dy) = \int\limits_A P(B,y) f_\eta(y)\nu(dy) =$$
$$= \int\limits_{\mathbb{R}} P(B,y)\cdot \mathrm{I}\{y \in A\} f_\eta(y)\nu(dy) = \mathrm{E}(P(B,\eta)\cdot \mathrm{I}\{\eta \in A\}).$$

Последнее равенство означает, что функция $\dfrac{f_{\xi,\eta}(x,y)}{f_\eta(y)}$ является условной плотностью. ■

Схема вычисления УМО

Зафиксируем схему вычисления УМО, которая у нас получилась. Необходимо вычислить $\mathrm{E}(g(\xi)|\eta)$.

1) Находим $f_{\xi,\eta}(x,y)$ - совместную плотность ξ и η.

2) С ее помощью находим $f_{\xi|\eta}(x|y)$ - условную плотность ξ относительно η.

3) Вычисляем
$$h(y) = \mathrm{E}(g(\xi)|\eta = y) = \int\limits_{\mathbb{R}} g(x) f_{\xi|\eta}(x|y)dx.$$

4) Подставляем $\mathrm{E}(g(\xi)|\eta) = h(\eta)$.

Пример
Случайные величины X и Y независимые $\mathrm{Exp}(1)$. Найдите

$$\mathrm{E}(X^2|X+Y) = ?$$

Решение

Сначала найдем совместную плотность $(X, X+Y)$. Для $B \in \mathcal{B}(\mathbb{R}^2)$ рассмотрим

$$\mathrm{P}((X, X+Y) \in B) = \int\limits_{(x,y):(x,x+y)\in B} f_{X,Y}(x,y)dxdy =$$

$$= \int\limits_{(x,y):(x,x+y)\in B} e^{-x}e^{-y} \cdot \mathrm{I}\{x, y > 0\}dxd \boxed{=}$$

сделаем замену $\alpha = x$, $\beta = x + y$, якобиан замены $= 1$

$$\boxed{=} \int\limits_{(\alpha,\beta)\in B} e^{-\beta} \cdot \mathrm{I}\{\beta > \alpha > 0\}d\alpha d\beta.$$

Стало быть, функция $p_{X,X+Y}(\alpha, \beta) = e^{-\beta} \cdot \mathrm{I}\{\beta > \alpha > 0\}$ и есть совместная плотность $(X, X+Y)$.

Теперь найдем плотность $X + Y$. Для этого совместную плотность $p_{X,X+Y}(\alpha, \beta)$ проинтегрируем по первой переменной.

$$p_{X+Y}(\beta) = \int\limits_{\mathbb{R}} p(\alpha,\beta)d\alpha = \int\limits_{\mathbb{R}} e^{-\beta}\mathrm{I}\{\beta > \alpha > 0\}d\alpha = e^{-\beta}\mathrm{I}\{\beta > 0\}\int\limits_0^\beta d\alpha = \beta e^{-\beta}\mathrm{I}\{\beta > 0\}.$$

Отсюда условная плотность равна

$$f_{X|X+Y}(\alpha|\beta) = \frac{p_{X,X+Y}(\alpha,\beta)}{p_{X+Y}(\beta)} = \frac{1}{\beta}\mathrm{I}\{\beta > \alpha > 0\}.$$

Остается найти УМО

$$\mathrm{E}(X^2|X+Y=\beta) = \int\limits_{\mathbb{R}} \alpha^2 \cdot \frac{1}{\beta}\mathrm{I}\{\beta > \alpha > 0\}d\alpha = \frac{\beta^2}{3}.$$

В итоге, $\mathrm{E}(X^2|X+Y) = \dfrac{(X+Y)^2}{3}$.

Замечания

1) Если ξ и η независимы и имеют плотности, то условная плотность $f_{\xi|\eta}(x|y)$ равна просто плотности случайной величины ξ.

2) Вычисление $\mathrm{E}(\xi|\eta)$ в случае векторного условия η происходит по тем же формулам.

3) Если условной плотности нет, то формально условное математическое ожидание вычисляется как интеграл по условному распределению
$$E(g(\xi)|\eta = y) = \int\limits_{\mathbb{R}} g(x)Q_y(dx),$$
где $Q_y(B) = \mathrm{P}(\xi \in B|\eta = y)$.

4) Особо важен случай вычисления условных математических ожиданий вида $\mathrm{E}(g(\xi,\eta)|\eta)$ для независимых случайных величин ξ и η.

Важный случай

Лемма
Если ξ и η - независимые случайные величины, то
$$\mathrm{E}(g(\xi,\eta)|\eta = y) = \mathrm{E}g(\xi, y).$$

Доказательство.
По теореме Фубини правая часть есть борелевская функция от y. Проверим интегральное свойство: для любого $B \in \mathcal{B}(\mathbb{R})$
$$\mathrm{E}(g(\xi,\eta)\cdot \mathrm{I}\{\eta \in B\}) = \int\limits_{\mathbb{R}^2} g(x,y)\cdot \mathrm{I}\{y \in B\}\mathrm{P}_{\xi,\eta}(dx,dy) \boxed{=}$$
пользуемся независимостью (совместное распределение - прямое произведение распределений)
$$\boxed{=} \int\limits_B \Bigg(\int\limits_{\mathbb{R}} g(x,y)\mathrm{P}_\xi(dx)\Bigg)\mathrm{P}_\eta(dy) = \int\limits_B \mathrm{E}g(\xi,y)\mathrm{P}_\eta(dy).$$
■

Теорема о наилучшем квадратичном прогнозе

Теорема 10.4. (о наилучшем квадратичном прогнозе)
Пусть ξ - случайная величина на $(\Omega, \mathcal{F}, \mathrm{P})$ с конечным вторым моментом, $\mathfrak{C}$ - под-σ-алгебра в $\mathcal{F}$ю Обозначим через $\mathcal{A}_{\mathfrak{C}}$ - множество всех $\mathfrak{C}$-измеримых случайных величин. Тогда
$$\mathrm{E}(\xi - \mathrm{E}(\xi|\mathfrak{C}))^2 = \min_{\eta\in\mathcal{A}_{\mathfrak{C}}} \mathrm{E}(\xi-\eta)^2,$$
причем $\mathrm{E}(\xi|\mathfrak{C})$ - это единственный минимум.

Доказательство.
Пусть $\eta \in \mathcal{A}_{\mathfrak{C}}$. Тогда

$$\mathrm{E}(\xi-\eta)^2 = \mathrm{E}(\xi - \mathrm{E}(\xi|\mathfrak{C}) + \mathrm{E}(\xi|\mathfrak{C}) - \eta)^2 =$$

$$= \mathrm{E}(\xi - \mathrm{E}(\xi|\mathfrak{C}))^2 + \mathrm{E}(\mathrm{E}(\xi|\mathfrak{C}) - \eta)^2 + 2\mathrm{E}\Bigg((\xi - \mathrm{E}(\xi|\mathfrak{C}))(\mathrm{E}(\xi|\mathfrak{C}) - \eta)\Bigg).$$

Проверим, что третье слагаемое равно нулю. Тогда $\mathrm{E}(\xi-\eta)^2$ минимизируется, когда второй квадрат равен нулю (т.к. только он зависит от η). Действительно, по формуле полной вероятности

$$\mathrm{E}\Bigg((\xi - \mathrm{E}(\xi|\mathfrak{C}))(\mathrm{E}(\xi|\mathfrak{C}) - \eta)\Bigg) = \mathrm{E}\Bigg(\mathrm{E}((\xi - \mathrm{E}(\xi|\mathfrak{C}))(\mathrm{E}(\xi|\mathfrak{C}) - \eta)|\mathfrak{C}\Bigg) \boxed{=}$$

свойство (9) УМО

$$\boxed{=}\ \mathrm{E}\Bigg((\mathrm{E}(\xi|\mathfrak{C}) - \eta)\mathrm{E}(\xi - \mathrm{E}(\xi|\mathfrak{C})|\mathfrak{C})\Bigg) = 0.$$

В итоге,

$$\mathrm{E}(\xi-\eta)^2 = \mathrm{E}(\xi - \mathrm{E}(\xi|\mathfrak{C}))^2 + \mathrm{E}(\mathrm{E}(\xi|\mathfrak{C}) - \eta)^2 \geqslant \mathrm{E}(\xi - \mathrm{E}(\xi|\mathfrak{C}))^2,$$

причем равенство достигается тогда и только тогда, когда $\eta = \mathrm{E}(\xi|\mathfrak{C})$. ■

11. Лекция 11

Постановка задачи

Напомним байесовский подход к сравнению оценок.

- Пусть $\mathbf{X}$ - наблюдение с неизвестным распределением P, принадлежащем параметрическому семейству $\{\mathrm{P}_\theta, \theta \in \Theta\}$.
- Пусть задано некоторое распределение вероятностей Q (которое называется априорным) на множестве значений параметра Θ.
- Пусть задана функция потерь $\rho(x, y), x, y \in \Theta$.
- Для каждой оценки $\theta^*(\mathbf{X})$ определим функцию риска

$$R_{\theta^*}(\theta) = \mathrm{E}_\theta \rho(\theta^*(\mathbf{X}), \theta).$$

Задача Требуется найти такую оценку $\hat{\theta}(\mathbf{X})$, что

$$\hat{\theta}(\mathbf{X}) = \arg \min_{\theta^*(\mathbf{X})} \int_\Theta R_{\theta^*}(t) \mathrm{Q}(dt).$$

Предположения
Оказывается, что при некоторых условиях можно получить явное выражение для наилучшей оценки. Для этого факта сделаем несколько предположений.

1) Будем считать, что параметр числовой: $\Theta \in \mathbb{R}$.
2) Распределение Q имеет плотность $q(t)$ по мере λ.
3) Функция потерь является квадратичной и, значит,

$$R_{\theta^*}(\theta) = \mathrm{E}_\theta (\theta^*(\mathbf{X}) - \theta)^2$$

4) Семейство распределений $\{\mathrm{P}_\theta, \theta \in \Theta\}$ — это доминируемое семейство распределений с плотностью $p_\theta(X)$ по мере μ.

В рамках этих предположений:

Определение 11.1. Плотность $q(t)$ называется априорной плотностью параметра θ.

Байесовские оценки

Определение 11.2. Функция

$$q(t|\mathbf{X}) = \frac{q(t)p_t(\mathbf{X})}{\int\limits_{\Theta} q(s)p_s(\mathbf{X})\lambda(ds)}$$

называется апостериорной плотностью параметра θ (плотностью при заданном $\mathbf{X}$).

Определение 11.3. Оценка

$$\hat{\theta}_{\mathrm{Q}}(\mathbf{X}) = \int\limits_{\Theta} t \cdot q(t|\mathbf{X})\lambda(dt)$$

называется байесовской оценкой параметра θ для заданного априорного распределения Q.

Мы готовы сформулировать теорему о байесовской оценке, которая говорит, что $\hat{\theta}_{\mathrm{Q}}(\mathbf{X})$ это ровно то, что мы ищем.

Теорема 11.1. (о байесовской оценке)
В приведенных условиях байесовская оценка $\hat{\theta}_{\mathrm{Q}}(\mathbf{X})$ является наилучшей оценкой параметра θ в байесовском подходе.

Доказательство.
Идея доказательства:
Доказательство будет сводиться к теореме о наилучшем квадратичном прогнозе. Мы хотим показать, что байесовская оценка - это будет условное математическое ожидание в некотором пространстве и, соответственно, она будет величиной, которая минимизирует квадратичный прогноз. В байесовском подходе мы рассматриваем параметр как случайную величину - у него есть распределение вероятности Q. И идея состоит в том, чтобы рассмотреть пару: параметр и наблюдение как случайный вектор и, соответственно, понять, какую они имеют совместную плотность. Далее, показать, что интегральное значение функции риска для оценки совпадет с математическим ожиданием от квадрата разности "оценка минус параметр но уже в некотором другом вероятностном пространстве, в котором параметр будет случайной величиной.

Итак, рассмотрим на пространстве $\Theta \times \mathcal{X}$ следующую функцию:

$$f(t, \mathbf{x}) = q(t)p_t(\mathbf{x}), t \in \Theta, \mathbf{x} \in \mathcal{X}.$$

Заметим, что

$$\int\limits_{\Theta\times\mathcal{X}} f(t,\mathbf{x})\lambda(dt)\mu(d\mathbf{x}) = \int\limits_{\Theta} q(t)\Bigg(\int\limits_{\mathcal{X}} p_t(\mathbf{x})\mu(d\mathbf{x})\Bigg)\lambda(dt) = \int\limits_{\Theta} q(t)\lambda(dt) = 1.$$

Тогда $f(t,\mathbf{x})$ - это плотность некоторого вероятностного распределения на $\Theta\times\mathcal{X}$ по мере $\lambda\times\mu$. Обозначим через $\tilde{\mathrm{P}}$ соответствующую вероятностную меру.
Далее, посмотрим на вектор $(\theta,\mathbf{X})$, как на случайный вектор на вероятностном пространстве $(\Theta\times\mathcal{X},\tilde{P})$, заданный по правилу:

$$(\theta,\mathbf{X})(t,\mathbf{x}) = (t,\mathbf{x}).$$

Мы хотим минимизировать по θ^* следующий функционал:

$$\int\limits_{\Theta} R_{\theta^*}(t)q(t)\lambda(dt) = \int\limits_{\Theta} \mathrm{E}_t(\theta^*(\mathbf{X})-t)^2 q(t)\lambda(dt) =$$

$$= \int\limits_{\Theta}\Bigg(\int\limits_{\mathcal{X}}(\theta^*(\mathbf{x})-t)^2 p_t(\mathbf{x})\mu(d\mathbf{x})\Bigg)q(t)\lambda(dt) =$$

$$= \int\limits_{\Theta\times\mathcal{X}}(\theta^*(\mathbf{x})-t)^2 f(t,\mathbf{x})\lambda(dt)\times\mu(d\mathbf{x}) = \tilde{\mathrm{E}}(\theta^*(\mathbf{X})-\theta)^2.$$

Тем самым, задача поиска наилучшей оценки в байесовском подходе свелась к минимизации математического ожидания в пространстве $\tilde{\mathrm{E}}$:

$$\tilde{\mathrm{E}}(\theta^*(\mathbf{X})-\theta)^2 \to \min_{\theta^*(\mathbf{X})}.$$

Теорема о наилучшем квадратичном прогнозе говорит, что решением такой задачи является $\tilde{\mathrm{E}}(\theta|\mathbf{X})$. Теперь осталось показать, что $\hat{\theta}_{\mathrm{Q}}(\mathbf{X}) = \tilde{\mathrm{E}}(\theta|\mathbf{X})$.
Для этого заметим, что:

- $q(t)$ есть плотность θ;
- Наблюдение $\mathbf{X}$ имеет плотность

$$g(\mathbf{x}) = \int\limits_{\Theta} f(t,\mathbf{x})\lambda(dt);$$

- Условная плотность $\mathbf{X}$ относительно θ равна $\dfrac{f(t,\mathbf{x})}{q(t)} = p_t(\mathbf{x})$.

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

$$\frac{f(t,\mathbf{x})}{g(x)} = \frac{q(t)p_t(\mathbf{x})}{\int\limits_{\Theta} q(s)p_s(\mathbf{x})\lambda(ds)} = q(t|\mathbf{x}),$$

т.е. это и есть апостериорная плотность.

Применяя формулу для вычисления УМО, получаем, что

$$\tilde{E}(\theta|\mathbf{X}=\mathbf{x}) = \int\limits_{\Theta} tq(t|\mathbf{x})\lambda(dt) = \hat{\theta}_{\mathrm{Q}}(\mathbf{x}).$$

Значит, $\hat{\theta}_{\mathrm{Q}}(\mathbf{x}) = \tilde{\mathrm{E}}(\theta|\mathbf{X})$. ■

Пример

Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из нормального распределения $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. Найдите байесовскую оценку параметра θ, если априорное распределение θ есть $\mathcal{N}(a, \sigma^2)$.

Решение

У нас есть формулы, надо просто посчитать. Найдем плотность выборки:

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

Теперь рассмотрим произведение $p_t(\mathbf{X})q(t)$:

$$p_t(\mathbf{X})q(t) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n} X_i^2 + t\sum_{i=1}^{n} X_i - \frac{n}{2}t^2\right) \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-a)^2}{2\sigma^2}\right).$$

Как функция от t это произведение представляет собой ненормированную плотность нормального распределения.

Найдем параметры этого распределения. Если убрать все слагаемые, не зависящие от t в экспоненте, то останется:

$$-\left(\frac{n}{2} + \frac{1}{2\sigma^2}\right)t^2 + t\left(\sum_{i=1}^{n} X_i + \frac{a}{\sigma^2}\right) = -\frac{1}{2}\left(n + \frac{1}{\sigma^2}\right)\left(t - \frac{\overline{\mathbf{X}} + \frac{a}{n\sigma^2}}{1 + \frac{1}{n\sigma^2}}\right)^2 + \dots$$

Таким образом, апостериорное распределение есть $\mathcal{N}\left(\frac{\overline{\mathbf{X}} + \frac{a}{n\sigma^2}}{1 + \frac{1}{n\sigma^2}}, \left(n + \frac{1}{\sigma^2}\right)^{-1}\right)$.

Байесовская оценка - это математическое ожидание апостериорного распределения, стало быть, она равна:

$$\hat{\theta}_{\mathrm{Q}}(\mathbf{X}) = \frac{\overline{\mathbf{X}} + \frac{a}{n\sigma^2}}{1 + \frac{1}{n\sigma^2}}.$$

Стоит отметить, что с ростом n байесовская оценка становится похожей на $\overline{\mathbf{X}}$

Минимаксный подход

Напомним про минимаксный подход к сравнению оценок. В нём задача поиска наилучшей оценки ставится следующим образом:

$$\hat{\theta}(\mathbf{X}) = \arg\min_{\theta^*(\mathbf{X})} \sup_{\theta \in \Theta} R_{\theta^*}(\theta),$$

т.е. мы хотим минимизировать максимальное значение функции риска.

Оказывается, минимаксный подход тесно связан с байесовским, а минимаксная оценка в некоторых случаях может быть вычислена как байесовская для правильно подобранного априорного распределения Q.

Начнем обсуждение со следующего достаточного условия минимаксности оценки.

Достаточное условие минимаксности оценки

Лемма (1) Пусть $\hat{\theta}(\mathbf{X})$ - такая оценка параметра θ, что существует вероятностное распределение Q на Θ со следующим условием: для всех $\theta \in \Theta$

$$R_{\hat{\theta}}(\theta) \leqslant \int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}}}(t)\mathrm{Q}(dt),$$

где $\hat{\theta}_{\mathrm{Q}}(\mathbf{X})$ - это байесовская оценка θ для априорного распределения Q.
Тогда $\hat{\theta}(\mathbf{X})$ есть минимаксная оценка, то есть это наилучшая оценка в минимаксном подходе.

Доказательство.
Пусть $\theta^*(\mathbf{X})$ - любая оценка θ. Тогда

$$\sup_{\theta \in \Theta} R_{\theta^*}(\theta) \geqslant \int\limits_{\Theta} R_{\theta^*}(t)\mathrm{Q}(dt) \geqslant \int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}}}(t)\mathrm{Q}(dt) \geqslant \sup_{\theta \in \Theta} R_{\hat{\theta}}(\theta).$$

■

Замечание
Отметим, что в условиях леммы 1 должно выполняться равенство

$$R_{\hat{\theta}}(\theta) = \int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}}}(t)\mathrm{Q}(dt)$$

почти наверное по мере Q.

Следствие (1)
Если оценка $\hat{\theta}(\mathbf{X})$ удовлетворяет двум условиям:

- $\hat{\theta}(\mathbf{X})$ является байесовской оценкой θ для некоторого априорного распределения Q;

- существует такая константа c, что функция риска $R_{\hat{\theta}}(\theta) = c$ Q - почти наверное и $R_{\hat{\theta}}(\theta) \leqslant c$ для всех θ;

то оценка $\hat{\theta}(\mathbf{X})$ является минимаксной.

Выводы

Первый вывод.
Минимаксная оценка - это байесовская оценка, которая "выравнивает"функцию риска.

Вопрос: но какое априорное распределение надо взять?

Второй вывод.
Если оценка $\hat{\theta}(\mathbf{X}) = \hat{\theta}_Q(\mathbf{X})$ удовлетворяет условию следствия, то для любого априорного распределения $\tilde{\mathrm{Q}}$ будет выполнено следующее неравенство:

$$c = \int\limits_{\Theta} R_{\hat{\theta}}(t)\mathrm{Q}(dt) \geqslant \int\limits_{\Theta} R_{\hat{\theta}}(t)\tilde{\mathrm{Q}}(dt).$$

Но

$$\int\limits_{\Theta} R_{\hat{\theta}}(t)\tilde{\mathrm{Q}}(dt) \geqslant \int\limits_{\Theta} R_{\hat{\theta}_{\tilde{\mathrm{Q}}}}(t)\tilde{\mathrm{Q}}(dt), \int\limits_{\Theta} R_{\hat{\theta}}(t)\tilde{\mathrm{Q}}(dt) \leqslant \int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}}}(t)\mathrm{Q}(dt).$$

Таким образом, Q - это то распределение, для которого среднее значение функции риска соответствующей байесовской оценки является максимальным.

$$\int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}}}(t)\mathrm{Q}(dt) = \max_{\tilde{\mathrm{Q}}} \int\limits_{\Theta} R_{\hat{\theta}_{\tilde{\mathrm{Q}}}}(t)\tilde{\mathrm{Q}}(dt).$$

Определение 11.4. Подобное априорное распределение называется наихудшим априорным распределением параметра θ.

Пример
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из распределения Бернулли $\mathrm{Bin}(1, \theta)$, $\theta \in [0, 1]$. Найдите минимаксную оценку θ.

Решение

Наш план состоит в следующем:

- найти оценку с постоянной функцией риска,

- доказать, что подобная оценка является байесовской для некоторого априорного распределения.

Разумно искать оценку в виде $\hat{\theta}(\mathbf{X}) = a\overline{\mathbf{X}} + b$ для $a, b \in \mathbb{R}$. Несложными вычислениями можно получить, что

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

Но как искать теперь наихудшее априорное распределение?

Двойственные распределения

Идея: надо посмотреть на функцию правдоподобия как на (ненормированную) плотность от параметра θ.
В нашем примере

$$p_\theta(\mathbf{X}) = \theta^{\sum\limits_{i=1}^{n} X_i} (1-\theta)^{n - \sum\limits_{i=1}^{n} X_i}.$$

По $\theta \in [0, 1]$ - это ненормированная плотность бета-распределения. В связи с этим можно взять в качестве априорного распределения тоже бета-распределение.
Возьмем априорное распределение $\mathrm{Beta}(\alpha, \alpha)$. Тогда априорная плотность будет равна $q(t) = \dfrac{t^{\alpha-1}(1-t)^{\alpha-1}}{B(\alpha, \alpha)}$. Отсюда

$$q(t)p_t(\mathbf{X}) = \frac{1}{B(\alpha, \alpha)} t^{n\overline{\mathbf{X}} + \alpha - 1}(1-t)^{n(1-\overline{\mathbf{X}}) + \alpha - 1}.$$

Следовательно, априорное распределение тоже есть бета-распределение
$\mathrm{Beta}(n\overline{\mathbf{X}} + \alpha, n(1-\overline{\mathbf{X}}) + \alpha)$.
Пример
Следовательно, байесовская оценка будет иметь вид:

$$\hat{\theta}_{\mathrm{Q}}(\mathbf{X}) = \frac{n\overline{\mathbf{X}} + \alpha}{n + 2\alpha} = \frac{\overline{\mathbf{X}} + \frac{\alpha}{n}}{1 + \frac{2\alpha}{n}}.$$

Наша оценка имеет:

$$\hat{\theta}(\mathbf{X}) = \alpha\overline{\mathbf{X}} + b = \left(\frac{1}{1 + \frac{1}{\sqrt{n}}}\right)\overline{\mathbf{X}} + \frac{1}{2(\sqrt{n}+1)} = \frac{\overline{\mathbf{X}} + \frac{1}{2\sqrt{n}}}{1 + \frac{1}{\sqrt{n}}}.$$

Остается заметить, что при $\alpha = \dfrac{\sqrt{n}}{2}$ оценки совпадают, и, значит, $\hat{\theta}(\mathbf{X})$ есть байесовская оценка для некоторого априорного распределения.

Еще одно достаточное условие

Однако, наихудшее априорное распределение есть не всегда! В этом случае может помочь второе достаточное условие того, что оценка минимаксная.

Лемма (2)
Пусть $\hat{\theta}(\mathbf{X})$ - такая оценка параметра θ, что существует последовательность априорных распределений $\{\mathrm{Q}_k, k \in \mathbb{N}\}$ на Θ со следующим условием: для всех $\theta \in \Theta$

$$R_{\hat{\theta}}(\theta) \leqslant \varlimsup_{k\to+\infty} \int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}_k}}(t)\mathrm{Q}_k(dt).$$

Тогда оценка $\hat{\theta}(\mathbf{X})$ является минимаксной.
Доказательство. Пусть $\theta^*(\mathbf{X})$ - любая оценка θ. Тогда для любого $k \in \mathbb{N}$

$$\sup_{\theta\in\Theta} R_{\theta^*}(\theta) \geqslant \int\limits_{\Theta} R_{\theta^*}(t)\mathrm{Q}_k(dt) \geqslant \int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}_k}}(t)\mathrm{Q}_k(dt).$$

Перейдем к пределу по k. Тогда

$$\sup_{\theta\in\Theta} R_{\theta^*}(\theta) \geqslant \varlimsup_{k\to+\infty} \int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}_k}}(t)\mathrm{Q}_k(dt) \geqslant \sup_{\theta\in\Theta} R_{\hat{\theta}}(\theta),$$

что и доказывает минимаксность $\hat{\theta}(\mathbf{X})$. ■

Пример
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из нормального распределения $\mathcal{N}(\theta, 1), \theta \in \mathbb{R}$. Найдите минимаксную оценку параметра θ.

Решение

План решения. Заметим, что $\overline{\mathbf{X}}$ проходит в качестве претендента на минимаксность, так как его функция риска не зависит от параметра.

$$\mathrm{E}_\theta(\overline{\mathbf{X}} - \theta)^2 = \mathrm{D}_\theta\overline{\mathbf{X}} = \frac{1}{n}.$$

Однако, сам $\overline{\mathbf{X}}$ является байесовской оценкой ни для какого априорного распределения, так что мы хотим подобрать подходящую последовательность априорных распределений с условием:

$$\varlimsup_{k\to+\infty} \int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}_k}}(t)\mathrm{Q}_k(dt) \geqslant \frac{1}{n}.$$

Возьмем $\mathrm{Q}_k = \mathcal{N}(0, k)$. Тогда мы знаем, что

$$\hat{\theta}_{\mathrm{Q}_k}(\mathbf{X}) = \frac{\overline{\mathbf{X}}}{1 + \frac{1}{nk}}$$

Найдем искомый интеграл от функции риска:

$$\int\limits_{\mathrm{Q}} R_{\hat{\theta}_{\mathrm{Q}_k}}(t)\mathrm{Q}_k(dt) = \int\limits_{\mathbb{R}} \mathrm{E}_t\left(\frac{\overline{\mathbf{X}}}{1 + \frac{1}{nk}} - t\right)^2 q_k(t)dt =$$

$$= \int\limits_{\mathbb{R}} \left[\int\limits_{\mathbb{R}^n} \left(\frac{\overline{\mathbf{x}}}{1 + \frac{1}{nk}} - t\right)^2 p_t(x_1, \ldots, x_n)dx_1 \ldots dx_n\right] q_k(t)dt =$$

(выразим $q_k(t|\mathbf{x}) = \dfrac{q_k(t)p_t(\mathbf{x})}{\tilde{p}(\mathbf{x})}$, где $\tilde{p}(\mathbf{x}) = \int\limits_{\mathbb{R}} q_k(t)p_t(\mathbf{x})dt$)
(переставим интегралы)

$$= \int\limits_{\mathbb{R}^n} \left[\int\limits_{\mathbb{R}} \left(\frac{\overline{\mathbf{x}}}{1 + \frac{1}{nk}} - t\right)^2 q_k(t|\mathbf{x})dt\right] \tilde{p}(x_1, \ldots, x_n)dx_1 \ldots dx_n =$$

(внутренний интеграл есть дисперсия апостериорного распределения)

$$= \frac{1}{n + \frac{1}{k}} \to \frac{1}{n} \text{ при } k \to +\infty.$$

Стало быть, $\overline{\mathbf{X}}$ - это минимаксная оценка θ.

12. Лекция 12

Понятие оптимальной оценки

Ранее мы подробно рассмотрели асимптотический, байесовский и минимаксный подходы к сравнению оценок. Теперь осталось рассмотреть самый "сильный"подход - равномерный. Как и в случае байесовского и минимаксного подходов будем изучать только квадратичную функцию потерь. Напомним, что в среднеквадратичном подходе поиск наилучшей оценки параметра θ устроен так:

$$\mathrm{E}_\theta(\hat{\theta}(\mathbf{X})-\theta)^2 \to \min_{\hat{\theta}(\mathbf{X})} \quad \theta \in \Theta.$$

Однако, как обсуждалось ранее, если не сузить класс оценок, то эта задача будет бессмысленной. Будем работать в классе несмещённых оценок. В таком случае $\mathrm{E}_\theta(\hat{\theta}(\mathbf{X})-\theta)^2 = \mathrm{D}_\theta\hat{\theta}(\mathbf{X})$. Более того, можно оценивать не только сам θ, но и функции от него.
Все это приводит нас к определению оптимальной оценки

Определение 12.1. Пусть $\mathbf{X}$ - наблюдение с неизвестным распределением $\mathrm{P} \in \{\mathrm{P}_\theta, \theta \in \Theta\}$, а $\tau(\theta) \in \mathbb{R}^k$ - некоторая функция от θ. Тогда оценка $\hat{\theta}(\mathbf{X})$ называется оптимальной оценкой $\tau(\theta)$, если она является несмещённой оценкой $\tau(\theta)$ и имеет равномерно наименьшую дисперсию, т.е. для любой другой несмещённой оценки $\theta^*(\mathbf{X})$ выполнено

$$\mathrm{D}_\theta\hat{\theta}(\mathbf{X}) \leqslant \mathrm{D}_\theta\theta^*(\mathbf{X}) \text{ для всех } \theta \in \Theta.$$

В многомерном случае, т.е. при $\hat{\theta}(\mathbf{X}) \in \mathbb{R}^k, k \geqslant 2$, неравенство дисперсий означает, что разность матриц ковариаций является неотрицательно определенной матрицей:

$$\mathrm{D}_\theta\hat{\theta}(\mathbf{X}) \leqslant \mathrm{D}_\theta\theta^*(\mathbf{X}) \Longleftrightarrow \mathrm{D}_\theta\theta^*(\mathbf{X}) - \mathrm{D}_\theta\hat{\theta}(\mathbf{X}) \text{ неотрицательно определена.}$$

Замечание
Эффективная оценка $\tau(\theta)$ является оптимальной, но она существует далеко не для всех функций от θ.

Достаточные статистики

Определение 12.2. Пусть $\mathbf{X}$ - наблюдение с неизвестным распределением $\mathrm{P} \in \{\mathrm{P}_\theta, \theta \in \Theta\}$. Статистика $S(\mathbf{X})$ называется достаточной для семейства $\{\mathrm{P}_\theta, \theta \in \Theta\}$, если существует вариант условного распределения $\mathrm{P}_\theta(\mathbf{X} \in B | S(\mathbf{X}) = s)$, который не зависит от параметра θ, т.е. существует функция $\mathrm{P}(B, s)$ такая, что для всех $\theta \in \Theta$

$$\mathrm{P}_\theta(\mathbf{X} \in B | S(\mathbf{X}) = s) = \mathrm{P}(B, s) \ \mathrm{P}_\theta^S \text{ - почти наверное},$$

где P_θ^S - это распределение статистики $S(\mathbf{X})$, когда $\mathbf{X}$ имеет распределение P_θ.

Наблюдение: сам $\mathbf{X}$ является достаточной статистикой, но интересен случай, когда размерность $S(\mathbf{X})$ совпадает с размерностью параметра.
Ответ на вопрос, зачем нужны достаточные статистики, дает следующая теорема.

Теорема 12.1. Колмогорова-Блэкуэлла-Рао
Пусть $\hat{\theta}(\mathbf{X})$ - несмещённая оценка $\tau(\theta) \in \mathbb{R}$ с конечным вторым моментом: $\mathrm{E}_\theta(\hat{\theta}(\mathbf{X}))^2 < +\infty$ для всех $\theta \in \Theta$. Далее, пусть $S(\mathbf{X})$ - достаточная статистика для семейства распределений $\{\mathrm{P}_\theta, \theta \in \Theta\}$. Тогда

1) $\theta^*(\mathbf{X}) = \mathrm{E}_\theta(\hat{\theta}(\mathbf{X}) | S(\mathbf{X}))$ есть несмещённая оценка $\tau(\theta)$;
2) для всех $\theta \in \Theta$ выполнено $\mathrm{D}_\theta \theta^*(\mathbf{X}) \leqslant \mathrm{D}_\theta \hat{\theta}(\mathbf{X})$;
3) равенство в неравенстве выше для всех $\theta \in \Theta$ достигается тогда и только тогда, когда $\hat{\theta}(\mathbf{X})$ является $S(\mathbf{X})$-измеримой (т.е. измеримой функцией от $S(\mathbf{X})$).

Тем самым, теорема позволяет "улучшить"оценку, получив другую с равномерно меньшей дисперсией.

Доказательство.

1) Несмещённость $\theta^*(\mathbf{X})$ сразу же следует из формулы полной вероятности:

$$\mathrm{E}_\theta \theta^*(\mathbf{X}) = \mathrm{E}_\theta(\mathrm{E}_\theta(\hat{\theta}(\mathbf{X}) | S(\mathbf{X}))) = \mathrm{E}_\theta \hat{\theta}(\mathbf{X}) = \tau(\theta).$$

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

2) Теперь проверим соотношение между дисперсиями. Для этого заметим, что функция $h(x) = (x - \tau(\theta))^2$ является выпуклой книзу. Тогда можно воспользоваться неравенством Йенсена для УМО:

$$(\theta^*(\mathbf{X}) - \tau(\theta))^2 = (\mathrm{E}_\theta(\hat{\theta}(\mathbf{X}) | S(\mathbf{X})) - \tau(\theta))^2 \leqslant \mathrm{E}_\theta((\hat{\theta}(\mathbf{X}) - \tau(\theta))^2 | S(\mathbf{X})). \quad (1)$$

   Возьмем математическое ожидание E_θ от обеих частей неравенства:

$$\mathrm{D}_\theta \theta^*(\mathbf{X}) = \mathrm{E}_\theta(\theta^*(\mathbf{X}) - \tau(\theta))^2 \leqslant \mathrm{E}_\theta(\hat{\theta}(\mathbf{X}) - \tau(\theta))^2 = \mathrm{D}_\theta \hat{\theta}(\mathbf{X}).$$

3) Осталось понять, когда будет выполняться равенство в данном неравенстве. Заметим, что разность правой и левой части в (1) равна:

$$\mathrm{E}_\theta((\hat\theta(\mathbf{X}))^2|S(\mathbf{X})) - (\theta^*(\mathbf{X}))^2 =$$

$$= \mathrm{E}_\theta((\hat\theta(\mathbf{X}))^2|S(\mathbf{X})) - 2\theta^*(\mathbf{X})\mathrm{E}_\theta(\hat\theta(\mathbf{X})|S(\mathbf{X})) + (\theta^*(\mathbf{X}))^2 =$$

$$= \mathrm{E}_\theta((\hat\theta(\mathbf{X}))^2 - 2\hat\theta(\mathbf{X})\theta^*(\mathbf{X}) + (\theta^*(\mathbf{X}))^2|S(\mathbf{X})) = \mathrm{E}_\theta((\hat\theta(\mathbf{X}) - \theta^*(\mathbf{X}))^2|S(\mathbf{X})).$$

В итоге, разность дисперсий

$$\mathrm{D}_\theta\hat\theta(\mathbf{X}) - \mathrm{D}_\theta\theta^*(\mathbf{X}) = \mathrm{E}_\theta(\mathrm{E}_\theta((\hat\theta(\mathbf{X}) - \theta^*(\mathbf{X}))^2|S(\mathbf{X}))) = \mathrm{E}_\theta(\hat\theta(\mathbf{X}) - \theta^*(\mathbf{X}))^2,$$

равна нулю тогда и только тогда, когда $\hat\theta(\mathbf{X}) = \theta^*(\mathbf{X})$ P_θ - почти наверное для всех $\theta \in \Theta$.

■

Следствие (1)
Теорема Колмогорова-Блэкуэлла-Рао верная и в многомерном случае.
Доказательство.
Пусть $\tau(\theta) \in \mathbb{R}^k$. Возьмем произвольный ненулевой вектор $a \in \mathbb{R}^k$. Тогда $\langle\hat\theta(\mathbf{X}), a\rangle$ есть несмещённая оценка $\langle\tau(\theta), a\rangle$. Следовательно, в силу уже доказанного

$$\mathrm{D}_\theta\langle\theta^*(\mathbf{X}), a\rangle \leqslant \mathrm{D}_\theta\langle\hat\theta(\mathbf{X}), a\rangle.$$

Но это означает, что

$$a^T\mathrm{D}_\theta\theta^*(\mathbf{X})a \leqslant a^T\mathrm{D}_\theta\hat\theta(\mathbf{X})a.$$

Следовательно, $\mathrm{D}_\theta\theta^*(\mathbf{X}) \leqslant \mathrm{D}_\theta\hat\theta(\mathbf{X})$. Критерий равенства доказывается аналогично одномерному случаю. ■

Вопрос: при каких условиях оценка в теореме Колмогорова-Блэкуэлла-Рао является оптимальной?

Следствие (2)
Пусть в условиях теоремы Колмогорова-Блэкуэлла-Рао для $\tau(\theta)$ существует единственная $S(\mathbf{X})$-измеримая несмещённая оценка $\theta^*(\mathbf{X})$. Тогда $\theta^*(\mathbf{X})$ - оптимальная оценка $\tau(\theta)$.
Доказательство.
Пусть оценка $\theta^*(\mathbf{X})$ не оптимальная. Тогда найдется оценка $\hat\theta(\mathbf{X})$, которая будет лучше, т.е. дисперсия $\hat\theta(\mathbf{X})$ будет равномерно не больше дисперсии $\theta^*(\mathbf{X})$ и для некоторого $\theta_0 \in \Theta$

$$\mathrm{D}_{\theta_0}\hat\theta(\mathbf{X}) < \mathrm{D}_{\theta_0}\theta^*(\mathbf{X}).$$

Но тогда по теореме Колмогорова-Блэкуэлла-Рао оценка

$$\theta^{**}(\mathbf{X}) = \mathrm{E}_\theta(\hat\theta(\mathbf{X})|S(\mathbf{X}))$$

будет не хуже, чем $\hat{\theta}(\mathbf{X})$. Однако и $\theta^*(\mathbf{X})$, и $\theta^{**}(\mathbf{X})$ являются функциями от $S(\mathbf{X})$. Следовательно, они равны, и $\theta^*(\mathbf{X})$ не хуже, чем $\hat{\theta}(\mathbf{X})$.
Противоречие. Значит, $\theta^*(\mathbf{X})$ оптимальна. ■

Полные статистики

Определение 12.3. Статистика $S(\mathbf{X})$ называется полной для семейства $\{\mathrm{P}_\theta, \theta \in \Theta\}$, если из того, что

$$\mathrm{E}_\theta f(S(\mathbf{X})) = 0 \text{ для любого } \theta \in \Theta$$

следует, что $f(S(\mathbf{X})) = 0$ P_θ - почти наверное для всех $\theta \in \Theta$.

Смысл определения: несмещённо оценить нуль можно лишь нулевой функцией от $S(\mathbf{X})$.

Оказывается, что если статистика $S(\mathbf{X})$ является и достаточной, и полной, то улучшение оценки в теореме Колмогорова-Блэкуэлла-Рао сразу даст нам оптимальную оценку.

Теорема 12.2. (об оптимальной оценке)
Пусть $S(\mathbf{X})$ - полная достаточная статистика для семейства $\{\mathrm{P}_\theta, \theta \in \Theta\}$. Пусть функция $\varphi(s)$ такова, что $\varphi(S(\mathbf{X}))$ - несмещённая оценка $\tau(\theta)$ и $\mathrm{E}_\theta\varphi^2(S(\mathbf{X})) < +\infty$ для $\forall\theta \in \Theta$. Тогда $\varphi(S(\mathbf{X}))$ является оптимальной оценкой $\tau(\theta)$.

Доказательство. Согласно следствию 2 достаточно проверить, что $\varphi(S(\mathbf{X}))$ является единственной $S(\mathbf{X})$-измеримой несмещённой оценкой $\tau(\theta)$. Пусть $\psi(S(\mathbf{X}))$ - другая несмещённая оценка $\tau(\theta)$. Тогда для любого $\theta \in \Theta$

$$\mathrm{E}_\theta(\varphi(S(\mathbf{X})) - \psi(S(\mathbf{X}))) = 0.$$

В силу полноты статистики $S(\mathbf{X})$ получаем, что $\varphi(S(\mathbf{X})) = \psi(S(\mathbf{X}))$ P_θ - почти наверное. Отсюда получаем, что $\varphi(S(\mathbf{X}))$ - это единственная $S(\mathbf{X})$-измеримая несмещённая оценка $\tau(\theta)$. Следовательно, она оптимальна. ■

Замечания

Замечание (1)
Если $S(\mathbf{X})$ - полная достаточная статистика, то для нахождения оптимальной оценки $\tau(\theta)$ достаточно решить (относительно φ) уравнение несмещённости

$$\mathrm{E}_\theta\varphi(S(\mathbf{X})) = \tau(\theta).$$

Замечание (2)
Если $\tau(\theta) \in \mathbb{R}^k$ при $k > 1$, то оптимальная оценка вектора есть вектор из оптимальных оценок компонент.

Вопрос 1: как находить достаточные статистики?
Вопрос 2: как проверить их на полноту?

Критерий факторизации

На первый вопрос даёт ответ следующая теорема.

Теорема 12.3. (критерий факторизации Неймана-Фишера)
Пусть наблюдение $\mathbf{X}$ имеет неизвестное распределение $\mathrm{P} \in \{\mathrm{P}_\theta, \theta \in \Theta\}$, и это семейство является доминируемым относительно меры μ с плотностью $p_\theta(\mathbf{x})$. Тогда статистика $S(\mathbf{X})$ будет достаточной для семейства $\{\mathrm{P}_\theta, \theta \in \Theta\}$ тогда и только тогда, когда существует представление p_θ в виде

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

где ψ_θ и h - некоторые неотрицательные измеримые функции.

Таким образом, статистика является достаточной тогда и только тогда, когда плотность зависит от параметра через функцию от нее.

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

$$p_\theta(\mathbf{x}) = \mathrm{P}_\theta(\mathbf{X} = \mathbf{x}) = \mathrm{P}_\theta(\mathbf{X} = \mathbf{x}|S(\mathbf{X}))\mathrm{P}_\theta(S(\mathbf{X}) = S(\mathbf{x})).$$

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

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

где $h(x) = \mathrm{P}_\theta(\mathbf{X} = \mathbf{x}|S(\mathbf{X}) = S(\mathbf{x}))$, а $\psi_\theta(S(\mathbf{x})) = \mathrm{P}_\theta(S(\mathbf{X}) = S(\mathbf{x}))$.

$\boxed{\Leftarrow}$ Теперь предположим, что существует представление $\mathrm{P}\theta(\mathbf{X} = \mathbf{x})$ в виде $h(\mathbf{x})\psi_\theta(S(\mathbf{x}))$. Рассмотрим условное распределение $\mathrm{P}_\theta(\mathbf{X} = \mathbf{x}|S(\mathbf{X}) = s)$. Если $S(\mathbf{x}) \neq s$, то такая вероятность равна нулю. Иначе:

$$\mathrm{P}_\theta(\mathbf{X} = \mathbf{x}|S(\mathbf{X}) = s) = \frac{\mathrm{P}_\theta(\mathbf{X} = \mathbf{x}, S(\mathbf{X}) = s)}{\mathrm{P}_\theta(S(\mathbf{X}) = s)} =$$

$$= \frac{\mathrm{P}_\theta(\mathbf{X}=\mathbf{x})}{\mathrm{P}_\theta(S(\mathbf{X})=s)} = \frac{\mathrm{P}_\theta(\mathbf{X}=\mathbf{x})}{\sum\limits_{\mathbf{y}:S(\mathbf{y})=S(\mathbf{x})} \mathrm{P}_\theta(\mathbf{X}=\mathbf{y})} =$$

$$= \frac{h(\mathbf{x})\psi_\theta(S(\mathbf{x}))}{\sum\limits_{\mathbf{y}:S(\mathbf{y})=S(\mathbf{x})} h(\mathbf{y}\psi_\theta(S(\mathbf{y})))} = \frac{h(\mathbf{x})}{\sum\limits_{\mathbf{y}:S(\mathbf{y})=S(\mathbf{x})} h(\mathbf{y})}.$$

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

Формула пересчета УМО

Для доказательства общего случая нам понадобится формула пересчета УМО. Она аналогичная формуле пересчета обычных математических ожиданий (интегралов Лебега): если мера $\tilde{P}$ абсолютно непрерывна относительно P, то

$$\tilde{\mathrm{E}}\xi = \mathrm{E}\left(\xi \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}\right).$$

Теорема 12.4. Пусть $(\Omega, \mathcal{F}, \mathrm{P})$ - вероятностное пространство, $\tilde{\mathrm{P}}$ - другая вероятностная мера, абсолютно непрерывная относительно P ($\tilde{\mathrm{P}} \ll \mathrm{P}$), а $\dfrac{d\tilde{\mathrm{P}}}{d\mathrm{P}}$ - соответствующая производная Радона-Никодима. Тогда для любой случайной величины ξ с условием $\mathrm{E}|\xi| < +\infty$ и любой под$-\sigma$-алгебры $\mathfrak{C} \subset \mathcal{F}$ выполнено

$$\tilde{\mathrm{E}}(\xi|\mathfrak{C}) = \frac{\mathrm{E}\left(\xi \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C}\right)}{\mathrm{E}\left(\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C}\right)}.$$

Доказательство. (формулы пересчета)
Заметим, что $\mathrm{E}(\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C}) \neq 0$ $\tilde{\mathrm{P}}$ - почти наверное. Обозначим $\mathrm{B} = \{\mathrm{E}(\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C}) = 0\} \in \mathfrak{C}$. Тогда, применяя интегральное свойство УМО, получаем:

$$0 = \mathrm{E}\left(\mathrm{I}_B \cdot \mathrm{E}\left(\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C}\right)\right) = \mathrm{E}\left(\mathrm{I}_B \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}\right) = \tilde{\mathrm{E}}\mathrm{I}_B = \tilde{\mathrm{P}}(B).$$

Теперь докажем саму формулу. Правая часть является $\mathfrak{C}$-измеримой случайной величины, проверим интегральное свойство: для любого $\mathrm{A} \in \mathfrak{C}$

$$\tilde{\mathrm{E}}\left(\mathrm{I}_\mathrm{A} \cdot \frac{\mathrm{E}(\xi \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C})}{\mathrm{E}(\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C})}\right) = \mathrm{E}\left(\mathrm{I}_\mathrm{A} \cdot \frac{\mathrm{E}(\xi \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C})}{\mathrm{E}(\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C})}\right) \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}} =$$

(интегральное свойство)

$$= \mathrm{E}\left(\mathrm{I}_\mathrm{A} \cdot \mathrm{E}\left(\frac{\mathrm{E}(\xi \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C})}{\mathrm{E}(\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C})} \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}\middle|\mathfrak{C}\right)\right) =$$

(применяем свойство 9 УМО)

$$= \mathrm{E}\left(\mathrm{I_A} \cdot \frac{\mathrm{E}(\xi \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C})}{\mathrm{E}(\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C})} \cdot \mathrm{E}\left(\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C}\right)\right) = \mathrm{E}\left(\mathrm{I_A} \cdot \mathrm{E}\left(\xi \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C}\right)\right) =$$

(интегральное свойство)

$$= \mathrm{E}\left(\mathrm{I_A} \cdot \xi \cdot \frac{d\tilde{P}}{d\mathrm{P}}\right) = \tilde{E}(\mathrm{I_A} \cdot \xi).$$

Все доказано. ■

Доказательство. (критерия факторизации)
Зафиксируем некоторое $\theta' \in \Theta$ и для любого $\theta \neq \theta'$ введем вероятностную меру

$$\mathrm{P}_{\theta,\theta'} = \frac{1}{2}\mathrm{P}_\theta + \frac{1}{2}\mathrm{P}_{\theta'}.$$

Заметим, что $\mathrm{P}_\theta, \mathrm{P}_{\theta'} \ll \mathrm{P}_{\theta,\theta'}$, поэтому определены плотности:

$$f_\theta(\mathbf{X}) = \frac{d\mathrm{P}_\theta}{d\mathrm{P}_{\theta,\theta'}} = \frac{p_\theta(\mathbf{X})}{\frac{1}{2}(p_\theta(\mathbf{X}) + p_{\theta'}(\mathbf{X}))},$$

$$f'_\theta(\mathbf{X}) = \frac{d\mathrm{P}_{\theta'}}{d\mathrm{P}_{\theta,\theta'}} = \frac{p_{\theta'}(\mathbf{X})}{\frac{1}{2}(p_\theta(\mathbf{X}) + p_{\theta'}(\mathbf{X}))}.$$

$\boxed{\Rightarrow}$ Пусть $S(\mathbf{X})$ - достаточная статистика. Покажем, что $f_\theta(\mathbf{X})$ есть функция от $S(\mathbf{X})$, $f_\theta(\mathbf{X}) = \psi_\theta(S(\mathbf{X}))$. Этого будет достаточно для доказательства существования нужного представления для $p_\theta(\mathbf{x})$:

$$p_\theta(\mathbf{x}) = \frac{f_\theta(\mathbf{x})p_{\theta'}(\mathbf{x})}{2 - f_\theta(\mathbf{x})} = \frac{\psi_\theta(S(\mathbf{x}))}{2 - \psi_\theta(S(\mathbf{x}))} \cdot p_{\theta'}(\mathbf{x}),$$

где $h(x) = p_{\theta'}(\mathbf{x})$.
В силу достаточности $S(\mathbf{X})$ выолнено равенство

$$\mathrm{E}_\theta(f_\theta(\mathbf{X})|S(\mathbf{X})) = \mathrm{E}_{\theta'}(f_\theta(\mathbf{X})|S(\mathbf{X}).)$$

Проверим, что $\mathrm{E}_{\theta,\theta'}(f_\theta(\mathbf{X})|S(\mathbf{X}))$ равно тому же самому. Свойство измеримости уже есть, проверим интегральное. Для любого $\mathrm{A} \in \mathcal{F}_{S(\mathbf{X})}$

$$\mathrm{E}_{\theta,\theta'}(f_\theta(\mathbf{X}) \cdot \mathrm{I_A}) = \frac{1}{2}(\mathrm{E}_\theta(f_\theta(\mathbf{X})\mathrm{I_A}) + \mathrm{E}_{\theta'}(f_\theta(\mathbf{X}) \cdot \mathrm{I_A})) =$$

(интегральное свойство)

$$= \frac{1}{2}(\mathrm{E}_\theta(\mathrm{E}_\theta(f_\theta(\mathbf{X})|S(\mathbf{X})) \cdot \mathrm{I_A}) + \mathrm{E}_{\theta'}(\mathrm{E}_{\theta'}(f_\theta(\mathbf{X})|S(\mathbf{X})) \cdot \mathrm{I_A}))$$

$$= \frac{1}{2}(\mathrm{E}_\theta(\mathrm{E}_\theta(f_\theta(\mathbf{X})|S(\mathbf{X}))\cdot \mathrm{I}_\mathrm{A}) + \mathrm{E}_{\theta'}(\mathrm{E}_\theta(f_\theta(\mathbf{X})|S(\mathbf{X}))\cdot \mathrm{I}_\mathrm{A})) = \mathrm{E}_{\theta,\theta'}(\mathrm{E}_\theta(f_\theta(\mathbf{X})|S(\mathbf{X}))\cdot \mathrm{I}_\mathrm{A}).$$

Далее, по формуле пересчета УМО

$$\mathrm{E}_\theta(f_\theta(\mathbf{X})|S(\mathbf{X})) = \frac{\mathrm{E}_{\theta,\theta'}(f_\theta^2(\mathbf{X})|S(\mathbf{X}))}{\mathrm{E}_{\theta,\theta'}(f_\theta(\mathbf{X})|S(\mathbf{X}))}.$$

Но мы только что доказали, что $\mathrm{E}_\theta(f_\theta(\mathbf{X})|S(\mathbf{X})) = \mathrm{E}_{\theta,\theta'}(f_\theta(\mathbf{X})|S(\mathbf{X}))$, таким образом,

$$\mathrm{E}_{\theta,\theta'}(f_\theta^2(\mathbf{X})|S(\mathbf{X})) = (\mathrm{E}_{\theta,\theta'}(f_\theta(\mathbf{X})|S(\mathbf{X})))^2.$$

Вычитая из левой части правую и беря

$$\mathrm{E}_{\theta,\theta'}f_\theta^2(\mathbf{X}) - \mathrm{E}_{\theta,\theta'}(\mathrm{E}_{\theta,\theta'}(f_\theta(\mathbf{X})|S(\mathbf{X})))^2 = \mathrm{E}_{\theta,\theta'}(f_\theta(\mathbf{X}) - \mathrm{E}_{\theta,\theta'}(f_\theta(\mathbf{X})|S(\mathbf{X})))^2 = 0.$$

Последнее означает, что $f_\theta(\mathbf{X}) = \mathrm{E}_{\theta,\theta'}(f_\theta(\mathbf{X})|S(\mathbf{X}))$, т.е. она $S(\mathbf{X})$-измерима.

$\boxed{\Rightarrow}$ Из факторизации следует, что $f_\theta(\mathbf{X})$ и $f_{\theta'}(\mathbf{X})$ есть функции от $S(\mathbf{X})$. По формуле для пересчета УМО получаем, что для любой статистики $T(\mathbf{X})$ выполнено

$$\mathrm{E}_\theta(T(\mathbf{X})|S(\mathbf{X})) = \frac{\mathrm{E}_{\theta,\theta'}(T(\mathbf{X})f_\theta(\mathbf{X})|S(\mathbf{X}))}{\mathrm{E}_{\theta,\theta'}(f_\theta(\mathbf{X})|S(\mathbf{X}))} = \mathrm{E}_{\theta,\theta'}(T(\mathbf{X})|S(\mathbf{X}))$$

в силу $S(\mathbf{X})$-измеримости $f_\theta(\mathbf{X})$. Совершенно аналогично, получаем такое же равенство для θ':

$$\mathrm{E}_{\theta'}(T(\mathbf{X})|S(\mathbf{X})) = \frac{\mathrm{E}_{\theta,\theta'}(T(\mathbf{X})f_\theta'(\mathbf{X})|S(\mathbf{X}))}{\mathrm{E}_{\theta,\theta'}(f_\theta'(\mathbf{X})|S(\mathbf{X}))} = \mathrm{E}_{\theta,\theta'}(T(\mathbf{X})|S(\mathbf{X})).$$

Следовательно, для любого $\theta \in \Theta$ выполнено

$$\mathrm{E}_\theta(T(\mathbf{X})|S(\mathbf{X})) = \mathrm{E}_{\theta'}(T(\mathbf{X})|S(\mathbf{X})) \text{ почти наверное}$$

Взяв $T(\mathbf{x}) = \mathrm{I}\{\mathbf{x} \in B\}$, получаем, что существует вариант условного распределения $\mathrm{P}_\theta(\mathbf{X} \in B|S(\mathbf{X}))$, который не зависит от θ. По определению мы получаем, что $S(\mathbf{X})$ - достаточная статистика. ■

Примеры

Пример (1)

Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из распределения Бернулли $\mathrm{Bin}(1,\theta), \theta \in (0,1)$. Найдите достаточную статистику.

Решение

Плотность имеет вид:

$$p_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} \cdot \mathrm{I}\{X_i \in \{0,1\}\} =$$

$$= \theta^{\sum\limits_{i=1}^{n} X_i}(1-\theta)^{n-\sum\limits_{i=1}^{n} X_i} \cdot \mathrm{I}\{\forall X_i \in \{0,1\}\}.$$

Значит, $\sum\limits_{i=1}^{n} X_i$ - достаточная статистика.

Пример (2)
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из пуассоновского распределения $\mathrm{Pois}(\theta), \theta > 0$. Найдите достаточную статистику.

Решение

Плотность имеет вид:

$$p_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} \frac{\theta^{X_i}}{X_i!} e^{-\theta} = \left(\frac{1}{\prod\limits_{i=1}^{n} X_i!}\right) \theta^{\sum\limits_{i=1}^{n} X_i} \cdot e^{-n\theta}.$$

Значит, $\sum\limits_{i=1}^{n} X_i$ - достаточная статистика.

Пример (3)
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из нормального распределения $\mathcal{N}(a, \sigma^2), \theta = (a, \sigma^2), a \in \mathbb{R}, \sigma > 0$. Найдите достаточную статистику.

Решение

Плотность имеет вид:

$$p_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(X_i - a)^2}{2\sigma^2}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp(-\frac{\sum\limits_{i=1}^{n} X_i^2 - 2a\sum\limits_{i=1}^{n} X_i + na^2}{2\sigma^2}).$$

Значит, $\left(\sum\limits_{i=1}^{n} X_i^2, \sum\limits_{i=1}^{n} X_i\right)$ - достаточная статистика.

Пример (4)
Пусть $\mathbf{X} = (X_1, ..., X_n)$ - выборка из равномерного распределения $\mathrm{U}(0, \theta), \theta > 0$. Найдите оптимальную оценку θ.

Решение

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

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

Достаточной статистикой в данном случае будет являться $X_{(n)}$, так как в критерии факторизации можно взять $h(\mathbf{X}) = \mathrm{I}\{0 \leqslant X_{(1)} \leqslant X_{(n)}\}, \psi_\theta(S(\mathbf{X})) = \theta^{-n}\mathrm{I}\{X_{(n)} \leqslant \theta\}$. Покажем теперь, что $X_{(n)}$ есть полная статистика. Для этого вспомним, что $P_\theta(X_{(n)} \leqslant x) = (\frac{x}{\theta})^n$ для $x \in [0, \theta]$, поэтому плотность максимума равна

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

Посчитаем математическое ожидание произвольной функции $g(X_{(n)})$:

$$\mathrm{E}_\theta g(X_{(n)}) = \int\limits_0^\theta g(x)\frac{nx^{n-1}}{\theta^n}dx.$$

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

$$\int\limits_0^\theta g(x)x^{n-1}dx = 0.$$

Продифференцируем по θ:

$$g(\theta)\theta^{n-1} = 0 \Longleftrightarrow g(\theta) = 0 \text{ почти наверное для } \theta > 0.$$

Тогда $g(X_{(n)}) = 0$ P_θ - почти наверное для все $\theta > 0$ и $X_{(n)}$ есть полная достаточная статистика.
Осталось решить уравнение несмещённости: $\mathrm{E}_\theta\varphi(X_{(n)}) = 0\theta$. Заметим, что это равносильно тому, что:

$$\frac{\theta^{n+1}}{n} = \int\limits_0^\theta \varphi(x)x^{n-1}dx.$$

Продифференцируем по θ:

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

Следовательно, оптимальной оценкой θ является $\dfrac{n+1}{n}X_{(n)}$.

13. Лекция 13

На этой лекции мы разберемся с тем, как можно проверять полноту достаточной статистики в ситуации, когда мы имеем дело с экспоненциальными семействами распределений, и начнем обсуждать доверительные интервалы (интервальный подход к оценке параметра).

Экспоненциальные семейства

Для достаточности есть общий критерий факторизации, который дает ответ в большинстве разумных случаев. Для полноты такого удобного критерия нет. Однако в некоторых хороших случаях можно вывести простые достаточные условия полноты - например, если семейство распределений является экспоненциальным.

Определение 13.1. Пусть $\{\mathrm{P}_\theta, \theta \in \Theta\}$ - семейство распределений с k-мерным параметром: $\Theta \subseteq \mathbb{R}^k$, $k \in \mathbb{N}$. Далее, пусть оно является доминируемым относительно меры μ с плотностью $p_\theta(x)$. Если эта плотность представима в следующем виде:

$$p_\theta(\mathbf{x}) = h(\mathbf{x}) \exp\left\{\sum_{i=1}^{k} U_i(\mathbf{x}) a_i(\theta) + v(\theta)\right\}, \tag{1}$$

где $h(\mathbf{x}) \geqslant 0$ и $U_i(\mathbf{x})$ - это борелевские функции, то семейство распределений $\{\mathrm{P}_\theta, \theta \in \Theta\}$ называется экспоненциальным.

Теперь предположим, что семейство является экспоненциальным. Какая для него будет достаточная статистика? Понятно, что она будет равна $S(\mathbf{X}) = (U_1(\mathbf{X}), ..., U_k(\mathbf{X}))$. Оказывается, что если функции $a_i\theta$ достаточно хороши, то $S(\mathbf{X})$ будет и полной.

Теорема 13.1. (об экспоненциальном семействе)
Пусть $\mathbf{X}$ - наблюдение с неизвестным распределением, принадлежащим экспоненциальному семейству $\{\mathrm{P}_\theta, \theta \in \Theta\}$ с плотностью вида (1). Если при пробегании θ всего Θ вектор $\mathbf{a}(\theta) = (a_1(\theta), ..., a_k(\theta))$ зачерчивает k-мерный параллелепипед или шар (это значит, что внутри множества значений вектора $\mathbf{a}(\theta)$, при θ пробегающем Θ, содержится некоторый k-мерный параллелепипед или шар), то статистика $S(\mathbf{X}) = (U_1(\mathbf{X}), ..., U_k(\mathbf{X}))$ будет полной достаточной статистикой для $\{\mathrm{P}_\theta, \theta \in \Theta\}$.

Для доказательства потребуется несколько утверждений.

МЕХАНИКО-
МАТЕМАТИЧЕСКИЙ
ФАКУЛЬТЕТ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in
ЛЕКЦИИ УЧЕНЫХ МГУ