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:

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

Здесь $y_i, x_i$ - известные величины, $\theta$ - неизвестный параметр, который мы хотели бы оценить, $\varepsilon_i$ - 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 $\varepsilon_i$.

- $x_i$ - моменты измерения положения объекта,
- $y_i$ - измеренное положение объекта в момент времени $x_i$,
- $\varepsilon_i$ - случайная ошибка измерений,
- $\theta$ - неизвестная скорость объекта.

Можно выделить следующие постановки статистических задач:

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

2) **Интервальное оценивание**.
   Здесь нам не так важно знать точное значение $\theta$. 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 $x_1, \ldots, x_n, y_1, \ldots, y_n$ 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 $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_m$, 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, $a_1, \ldots, a_k, b_1, \ldots, b_m$. После проведения серии независимых экспериментов получены количественные данные $n_{ij}$, $i = 1, \ldots, k$, $j = 1, \ldots, m$, 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, $\mathcal{B}_\mathcal{X}$ - $\sigma$-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 $\mathbf{X} = (X_1, \ldots, X_n)$ есть набор независимых одинаково распределенных случайных величин, то $\mathbf{X}$ называется **выборкой** размера $n$ из некоторого распределения.

Предположим, что у нас есть выборка $\mathbf{X} = (X_1, \ldots, X_n)$ размера $n$ из неизвестного распределения $\mathrm{P}_X$ 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 $\mathrm{P}_n^*$, как функция от $B \in \mathcal{B}(\mathbb{R})$, называется **эмпирическим распределением**, построенным при выборке $X_1, \ldots, X_n$. Это дискретное распределение, оно имеет практически равномерное распределение на элементах $X_1, \ldots, X_n$.

### 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 $\mathrm{P}_X$. 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 $\mathrm{P}_n^*(B) = \dfrac{1}{n}\sum\limits_{i=1}^{n} \mathrm{I}\{X_i \in B\}$ - 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 $X_1,\ldots,X_n \sim U(0,1)$. 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 $\sigma$-алгебре. На самом деле, это выражение почти всегда равно 1, так как эмпирическое распределение - это дискретное распределение, сосредоточенное в числах $X_1,\ldots,X_n$. Если мы возьмем множество $B = [0,1]\setminus\{X_1,\ldots,X_n\}$, 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 $\sigma$-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 $\mathbf{X} = (X_1,\ldots,X_n)$ есть выборка размера $n$ из неизвестного распределения $\mathrm{P}_X$. 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 $X_1,\ldots,X_n$.

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 $X_1 < X_2 < \ldots < X_n$. 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 $\dfrac{1}{n}$ (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.

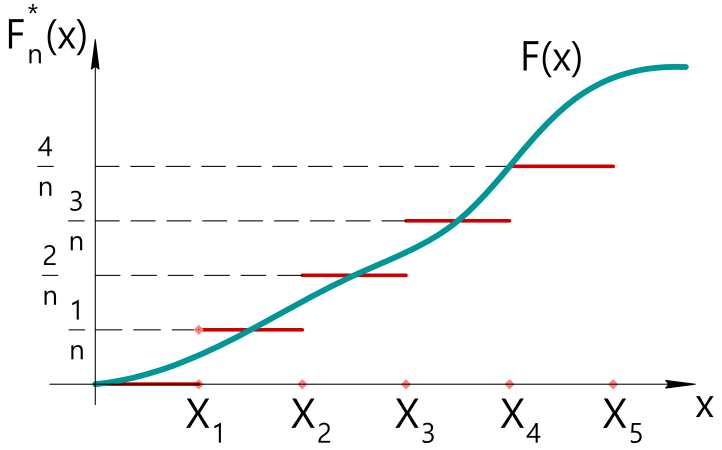


Рис. 1

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 $\omega$ (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

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

$k = 1, \ldots, N-1$ 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 $F_n^*(y) = \dfrac{1}{n}\sum\limits_{i=1}^{n} \mathrm{I}\{X_i \leqslant y\}$, 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 $F_n^*(y-0) = \dfrac{1}{n}\sum\limits_{i=1}^{n} \mathrm{I}\{X_i < y\}$. 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 $\mathrm{P}_\theta$ 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 $\mathbf{X}$= $(X_1, \ldots, X_n)$ - это выборка, то естественно оценивать и параметризовывать распределение $\mathrm{P}_\theta$ 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

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

**Лемма**
Вероятностная мера $\mathrm{P}_\theta^n$ 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 $\mathrm{P}_\theta$. Для каждого $i = 1, \ldots, n$ определим отображение $X_i : \mathcal{X}^n \longrightarrow \mathcal{X}$ по правилу

$$X_i(x_1, \ldots, x_n) = x_i.$$

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

$$\mathrm{P}_\theta^n(X_1 \in B_1, \ldots, X_n \in B_n) = \mathrm{P}_\theta^n(B_1 \times \ldots \times B_n) = \prod_{i=1}^n \mathrm{P}_\theta(B_i) = \prod_{i=1}^n \mathrm{P}_\theta^n(X_i \in B_i)$$

Значит, $X_1, \ldots, X_n$ - 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

- $\mathcal{X}^\infty = \{(x_1, x_2, \ldots) : x_i \in \mathcal{X}\}$ - пространство последовательностей из $\mathcal{X}$;
- $\mathcal{B}_\mathcal{X}^\infty$ - цилиндрическая $\sigma$-алгебра на $\mathcal{X}^\infty$. Пусть $B_n \in \mathcal{B}_\mathcal{X}^n$, тогда определим цилиндр с основанием $B_n$;

  $$\mathcal{F}_n(B_n) = \{x \in \mathcal{X}^\infty, x = (x_1, x_2, \ldots) : (x_1, \ldots, x_n) \in B_n\}$$

  Тогда $\mathcal{B}_\mathcal{X}^\infty$ - $\sigma$-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 $\mathrm{P}_\theta^\infty$ - 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 $\mathrm{P}_\theta^\infty$ 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)

$$X_i(x_1, x_2, \ldots) = x_i$$

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

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

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

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

**Замечание**
В дальнейшем мы будем опускать индексы $n$ и $\infty$ у мер $\mathrm{P}_\theta^n$ и $\mathrm{P}_\theta^\infty$, обозначая их также $\mathrm{P}_\theta$, 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 $\theta$ - 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 $\{\xi_n, n \in \mathbb{N}\}$, $\xi$ - случайные векторы размерности $m$. Тогда последовательность $\xi_n$ сходится к $\xi$

- 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 $\|x\|_2 = \sqrt{x_1^2 + \ldots + x_m^2}$ 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 $\xi = (\xi^{(1)}, \ldots, \xi^{(m)})$, $\xi_n = (\xi_n^{(1)}, \ldots, \xi_n^{(m)})$, то
$$\xi_n \xrightarrow{\text{п.н.}} \xi \Leftrightarrow \forall i = 1, \ldots, m \quad \xi_n^{(i)} \xrightarrow{\text{п.н.}} \xi^{(i)},$$
$$\xi_n \xrightarrow{\mathrm{P}} \xi \Leftrightarrow \forall i = 1, \ldots, m \quad \xi_n^{(i)} \xrightarrow{\mathrm{P}} \xi^{(i)}.$$

**Упражнение (2)**
Для сходимости по распределению векторная сходимость влечет сходимость компонент: если $\xi = (\xi^{(1)}, \ldots, \xi^{(m)})$, $\xi_n = (\xi_n^{(1)}, \ldots, \xi_n^{(m)})$, то
$$\xi_n \xrightarrow{d} \xi \Leftrightarrow \forall i = 1, \ldots, m \quad \xi_n^{(i)} \xrightarrow{d} \xi^{(i)}.$$
Обратное утверждение, вообще говоря, неверно.

**Упражнение (3, взаимоотношение видов сходимостей)**

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

## 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 $(S, \rho)$ - метрическое пространство.

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

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

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

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

### Теорема Александрова

**Утверждение** Если пространство $(S, \rho)$ сепарабельно, то $\mathcal{B}(S)$ является минимальной $\sigma$-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 $(S, \rho)$. Тогда следующие утверждения эквивалентны:

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

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 $\mathrm{P}_\eta$ обозначает распределение случайного вектора $\eta$. 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 $\{\xi_n, n \in \mathbb{N}\}$, $\xi$ - 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

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

**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 $S_n = X_1 + \ldots + X_n$. 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 $\{\xi_n, n \in \mathbb{N}\}$, $\xi$ - случайные векторы размерности $m$. Пусть $h(x) : \mathbb{R}^m \longrightarrow \mathbb{R}^k$ - функция, непрерывная почти всюду относительно распределения $\xi$ (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 $\xi_{n_k}$ 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 $\xi_{n_k}$.

3. Обозначим $\mathrm{Q}_n$ - 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 $\varepsilon$ 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

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

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 $S_n = X_1 + \ldots + X_n$. 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 $\Theta$, 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 $\mathbf{X} = (X_1, \ldots, X_n)$ - 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 $\mathrm{P}_n^*$ - эмпирическое распределение, построенное по выборке $\mathbf{X} = (X_1, \ldots, X_n)$.

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

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

   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 $g_1(x), \ldots, g_k(x)$ - борелевские функции на $\mathbb{R}$, а $h : \mathbb{R}^k \longrightarrow \mathbb{R}$ - борелевская функция нескольких переменных, тогда можно рассмотреть

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

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):

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

Вектор $(X_{(1)}, \ldots, X_{(n)})$ 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.

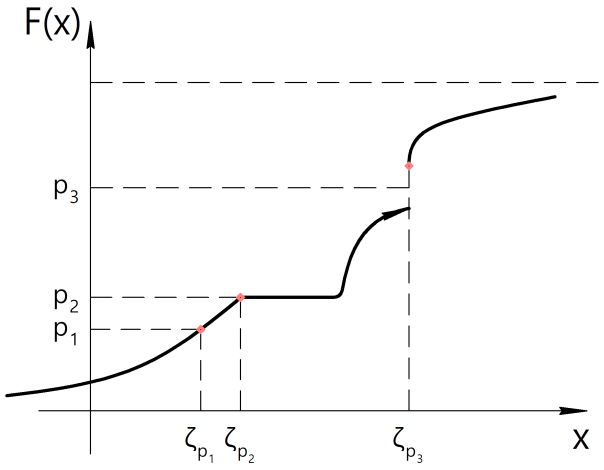


Рис. 2

**Первый случай**: $\zeta_{p_1}$ - 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 - $\zeta_{p_2}$.
**Третий случай**: в точке $p_3$ значение не достигается. Тогда в качестве квантиля мы берем наименьшую точку, в которой значение превышает $p_3$ - $\zeta_{p_3}$.

**Определение 3.4.** Выборочной $p$-квантилью называется $p$-квантиль эмпирического распределения $\mathrm{P}_n^*$, 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 $X_1, \ldots, X_n$ 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 $\theta$, если для любого $\theta\in\Theta$ выполняется равенство

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

Вспомним, что запись $\mathrm{E}_\theta$ означает, что при взятии математического ожидания мы предполагаем, что выборка $\mathbf{X}$ взята из распределения $\mathrm{P}_\theta$:

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

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

### Состоятельность

В асимптотических вопросах мы предполагаем, что $\mathbf{X}$=$(X_1,\ldots,X_n)$ - 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 $\hat{\theta}_n(X_1,\ldots,X_n)$ (точнее - последовательность оценок) называется состоятельной оценкой параметра $\theta$, если для любого $\theta\in\Theta$

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

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

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

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

### Сильная состоятельность

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

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

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

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

**Пример**
Пусть $\mathbf{X}(X_1, \ldots, X_n)$ - 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 $\hat{\theta}_n(X_1, \ldots, X_n)$ (точнее - последовательность оценок) называется асимптотически нормальной оценкой параметра $\theta$, если для любого $\theta \in \Theta$

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

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

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

где $\Phi$ - функция Лапласа (функция распределения стандартного нормального закона).

**Пример**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - 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 $\theta$, но и функции от него, $\tau(\theta)$. Определения свойств при этом сохраняются.

## Наследование свойств

**Вопрос**: пусть $\hat{\theta}_n(X_1, \ldots, X_n)$ хорошо приближает $\theta$, 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 $\hat{\theta}_n(X_1, \ldots, X_n)$ - сильно состоятельная (состоятельная) оценка параметра $\theta$. Если $\tau(\theta)$ непрерывна на $\Theta \subset \mathbb{R}^k$, то $\tau(\hat{\theta}_n(X_1, \ldots, X_n))$ будет сильно состоятельной (состоятельной) оценкой $\tau(\theta)$.

**Доказательство.** Сразу следует из теоремы о наследовании сходимости. ■

**Утверждение (2)**
Пусть $\hat{\theta}_n(x_1, \ldots, X_n)$ - асимптотически нормальная оценка параметра $\theta \in \mathbb{R}$ с асимптотической дисперсией $\sigma^2(\theta)$. Пусть $\tau(\theta)$ дифференцируема на $\Theta \subset \mathbb{R}$.
Тогда $\tau(\hat{\theta}_n(X_1, \ldots, X_n))$ - 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

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

Применяя дельта-метод, получаем, что

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

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 $\tau^{-1}$ и она непрерывна или дифференцируема.

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

**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:

Сильная состоятельность $\Rightarrow$ Состоятельность;

Асимптотическая нормальность $\Rightarrow$ 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 $\theta$ по методу подстановки называется

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

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

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

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

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

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

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

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

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

Другими словами:

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

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

**Замечание (2)**
Обычно используют стандартные пробные функции

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

**Пример**
Пусть $(X_1, \ldots, X_n)$ - выборка из нормального распределения $\mathcal{N}(a, \sigma^2)$, $a \in \mathbb{R}$, $\sigma^2 > 0$, $\theta = (a, \sigma^2)$. Найти оценку $\theta$ 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 $\theta_n^*(X_1, \ldots, X_n) = m^{-1}\left(\overline{g_1(\mathbf{X})}, \ldots, \overline{g_k(\mathbf{X})}\right)$ - корректно определенная оценка по методу моментов. Если функция $m^{-1}$ непрерывна на множестве $m(\Theta)$, то $\theta_n^*$ - сильно состоятельная оценка параметра $\theta$.

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

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

По теореме о наследовании сходимости получаем:

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

■

**Упражнение**
Если в дополнение к условия теоремы для любого $\theta \in \Theta \subset \mathbb{R}^k$ и любого $j = 1, \ldots, k$, конечно $\mathrm{E}_\theta g_j^2(X_1)$, а функция $m^{-1}$ дифференцируема на $m(\Theta)$, то для каждого $j = 1, \ldots, k$ оценка $\theta_{j,n}^*(X_1, \ldots, X_n)$ параметра $\theta_j$ 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 $(X_1, \ldots, X_n)$, 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 $\theta$ не является математическим ожиданием, но оно является $\frac{1}{2}$-квантилью, потому что плотность симметрична относительно $\theta$, соответственно в точке $\theta$ функция распределения будет достигать значения $\frac{1}{2}$. Можем ли мы надеяться на то, что выборочная $\frac{1}{2}$-квантиль хорошо приближает наш параметр $\theta$? 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 $\mathbf{X} = (X_1, \ldots, X_n)$ - 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 $\theta$ является $\frac{1}{2}$-квантилью распределения (медианой распределения).

Напомним, что если $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка, то выборочной $p$-квантилью называется $p$-квантиль эмпирического распределения $\mathrm{P}_n^*$, 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 $\mathbf{X} = (X_1, \ldots, X_n)$ - 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 $X_1, \ldots, X_n$ 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 $\dfrac{1}{2}$-квантиль. Выборочной медианой $\hat{\mu} = \hat{\mu}(X_1, \ldots, X_n)$ выборки $(X_1, \ldots, X_n)$ называется величина

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

**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 $\mathbf{X} = (X_1, \ldots, X_n)$ - 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 $\theta$. Тогда $\theta$ - 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 $\theta$? 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 $\theta$, 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 $\theta_0$ 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 $\theta$, если нет другой оценки $\hat{\theta}_0$, которая оценивает $\theta$ лучше, чем $\theta^*(\mathbf{X})$, в равномерном подходе.

### Байесовский подход

**Правило сравнения**. Пусть Q - некоторое заданное распределение вероятностей на $\Theta$ (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 $\lambda$, 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 $\theta$ 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 $\theta$ в баейсовском подходе для некоторого априорного распределения Q. Тогда $\theta^*(\mathbf{X})$ - это допустимая оценка $\theta$.

### Минимаксный подход

**Правило сравнения**. Пусть $\theta^*(\mathbf{X})$ - оценка $\theta$ 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 $\theta$ 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 $\theta^*$ и $\hat{\theta}$ во всех подходах.

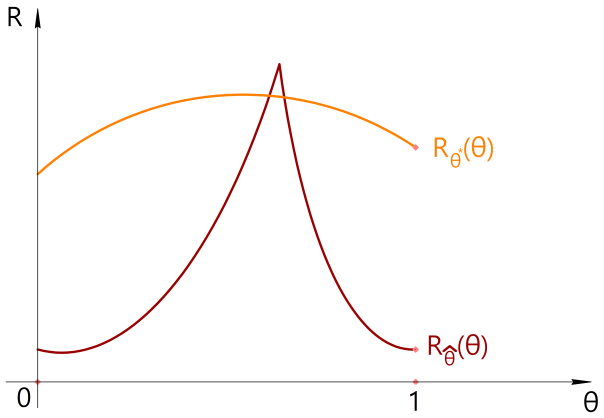


Рис. 3

Мы можем сделать следующие выводы:

- $\theta^*$ 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 $\hat{\theta}_{n,1}(X_1, \ldots, X_n)$ и $\hat{\theta}_{n,2}(X_1, \ldots, X_n)$ - две асимптотически нормальные оценки $\theta \in \mathbb{R}$ с асимптотическими дисперсиями $\sigma_1^2(\theta)$ и $\sigma_2^2(\theta)$. Тогда оценка $\hat{\theta}_{n,1}(X_1, \ldots, X_n)$ лучше оценивает параметр $\theta$, чем оценка $\hat{\theta}_{n,2}(X_1, \ldots, X_n)$, 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 $\hat{\theta}_n(X_1, \ldots, X_n)$ 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 $X_1, \ldots, X_n$ - 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 $\mathrm{P}_\theta$ имеет плотность $p_\theta(x)$ по одной и той же (может быть, $\sigma$-конечной) мере $\mu$ на $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, то семейство $\{\mathrm{P}_\theta, \theta \in \Theta\}$ называется доминируемым относительно $\mu$.

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 $\mu$ - это мера Лебега на $\mathbb{R}$, то $\mathrm{P}_\theta$ - 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 $\mu$ - это считающая мера (мера, которая сопоставляет каждому множеству количество целых точек внутри него) на $\mathbb{Z}$, то $\mathrm{P}_\theta$ - 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 $\mathrm{P}\in\{\mathrm{P}_\theta, \theta\in\Theta\}$, где $\{\mathrm{P}_\theta, \theta\in\Theta\}$ - доминируемое семейство относительно меры $\mu$ 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 $\theta$ (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 $\theta$, 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 $\mathbf{X} = (X_1, \ldots, X_n)$ - 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 $\mathbf{X} = (X_1, \ldots, X_n)$ - 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 $\theta \in \Theta$ $\mathrm{P}_\theta$-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:

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

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 $\theta$ и применяем неравенство, то линейная зависимость между $\hat{\theta}(\mathbf{X})-\tau(\theta)$ и $U_\theta(\mathbf{X})$ для каждого $\theta$ 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:

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

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 $\mathbf{X}=(X_1,\ldots,X_n)$ - выборка из распределения Бернулли $\mathrm{Bin}(1,\theta)$, где $\theta\in(0,1)$. Найдите эффективную оценку $\theta$ 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:

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

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

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 $\theta$.

4) Эффективная оценка $\tau(\theta)$ - это несмещенная оценка $\tau(\theta)$ с равномерно наименьшей дисперсией. Значит она - наилучшая оценка $\theta$ 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 $\mu$ функция $\sqrt{p_\theta(x)}$ непрерывно дифференцируема по $\theta$, а также информация Фишера $I_{\mathbf{X}}(\theta)$ конечна, положительна и непрерывна по $\theta$. 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 $\theta$, то она будет ограничена на любом отрезке внутри $\Theta$. Сузив значение параметра до интервала внутри отрезка, мы получим неравенство Рао-Крамера на интервале. Но раз это верно для любого интервала внутри $\Theta$, то оно будет выполнено и на всем $\Theta$.

## 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 $\lambda$, то есть семейство ее распределений $(G_\theta, \theta \in \Theta)$ является доминируемым относительно $\lambda$. 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 $\theta$, то $I_s(\theta) = 0$.
**Доказательство.**
очевидно, так как в таком случае $\ln g_\theta(S(\mathbf{X}))$ не будет зависеть от $\theta$ 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 $\lambda_1$, а статистика $T(\mathbf{X})$ - плотность $f_\theta(t)$ по мере $\lambda_2$. 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 $\mathbf{X} = (X_1, \ldots, X_n)$ - 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 $\mu$, а $g_\theta(s)$ - плотность статистики $S(\mathbf{X})$ по мере $\lambda$. 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 $\dfrac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X})$ 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 $\mathbf{X} = (X_1, \ldots, X_n)$ - 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) $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)^{\mathrm{T}}$ (вектор-столбец);

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

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

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

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

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

## 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 $\mathrm{E}(\mathbf{H}\mathbf{H}^{\mathrm{T}})$ обратима. Тогда

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

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

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

**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 $\mathbf{A}\mathbf{A}^{\mathrm{T}} \geqslant 0$. Тогда для любой неслучайной матрицы $\mathbf{Z} \in \mathbf{Mat}(n' \times k')$ выполнено следующее:

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

Возьмем математическое ожидание:

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

Теперь раскроем по линейности:

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

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

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

Для критерия равенства остается заметить, что

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

■

## 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 $\boldsymbol{\theta} \in \Theta$

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

$j = 1, \ldots, k$.

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 $\mathrm{E}_{\boldsymbol{\theta}} \hat{\theta}(\mathbf{X}) \hat{\theta}(\mathbf{X})^{\mathrm{T}}$. Тогда для любого $\boldsymbol{\theta} \in \Theta$ выполняется неравенство

$$\mathrm{D}_{\boldsymbol{\theta}} \hat{\theta}(\mathbf{X}) \geqslant \tau'(\boldsymbol{\theta}) I_{\mathbf{X}}^{-1}(\boldsymbol{\theta}) (\tau'(\boldsymbol{\theta}))^{\mathrm{T}},$$

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

$$0 = \mathrm{E}_{\boldsymbol{\theta}} \left( \frac{\partial}{\partial \boldsymbol{\theta}_j} \ln p_{\boldsymbol{\theta}}(\mathbf{X}) \right),$$

$j = 1, \ldots, k$, 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

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

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

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

Согласно матричному неравенству Коши-Буняковского, получаем искомое неравенство:

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

■

### 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 $\mu$ функция $\sqrt{p_{\boldsymbol{\theta}}(x)}$ непрерывно дифференцируема по $\boldsymbol{\theta}_j$, $j = 1, \ldots, k$, а также пусть информационная матрица $I_{\mathbf{X}}(\boldsymbol{\theta})$ положительно определена и непрерывна по $\theta$. 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 $\mu$.
Тогда функцией правдоподобия называется случайная величина $f_\theta(\mathbf{X}) = p_\theta(\mathbf{X})$.

**Замечание**
Если $\mathbf{X} = (X_1, \ldots, X_n)$ - 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 $\theta$ 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 $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из равномерного распределения $\mathrm{U}(0, \theta)$, $\theta > 0$. Найти оценку максимального правдоподобия параметра $\theta$.

**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 $\theta$. Заметим, что $\theta^{-n}$ - это монотонно убывающая функция, поэтому нужно взять такое минимальное $\theta$, что функция правдоподобия не обратится в нуль. Стало быть, $\hat{\theta}(\mathbf{X}) = X_{(n)}$.

**Пример 2**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - 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 $\sigma^2$. 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 $\sigma^2$):

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

Отсюда несложно получить, что решением будут оценки

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

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 $\mu$ и различимыми распределениями, то есть $\mathrm{P}_{\theta_0} = \mathrm{P}_{\theta_1}$ тогда и только тогда, когда $\theta_0 = \theta_1$.

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

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

Напомним, что через $f_\theta(X_1, \ldots, X_n)$ 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

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

**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:

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

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

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

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 $\phi$ 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 $\Theta$ 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** $\Theta$ есть открытый интервал на $\mathbb{R}$.

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

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

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

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

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

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

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.

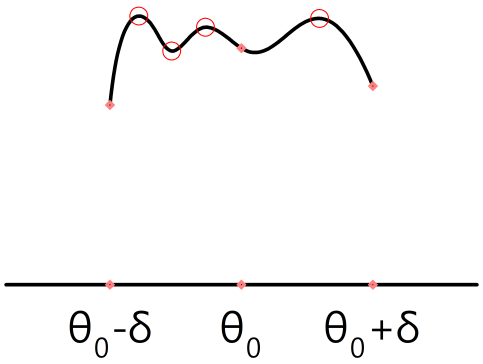


Рис. 4

Рассмотрим наше множество $\Theta$ 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 $\theta_0$.

Допустим, что на этом отрезке есть несколько корней (не обязательно конечное число). Пусть $\tilde{\theta}(\mathbf{X})$ - ближайший к $\theta_0$ 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 $\varepsilon$ 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 $\theta_0$ 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 $\mathbf{X} = (X_1, \ldots, X_n)$ 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 $\theta_0$ - 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 $\theta_0$ 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 $\theta_0$ 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 $\theta$. ■

## 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 $\theta$ 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 $\theta_0 \in \Theta$

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

**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 $\theta$.

Далее, пусть $\theta_0$ есть истинное значение параметра, то есть $\hat{\theta}(\mathbf{X})$ сходится к $\theta_0$ по вероятности $\mathrm{P}_{\theta_0}$. Разложим $\mathcal{L}'(\mathbf{X}, \theta)$ в ряд Тейлора в точке $\theta_0$:

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

где $\tilde{\theta}$ находится между $\theta$ и $\theta_0$. 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 $\theta_0$. 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

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

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

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

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:

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

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

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

Отсюда получаем, что

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

- 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 $\theta_0$ по вероятности. Из этого можно сделать вывод, что $\tilde{\theta}_n$ тоже стремится к $\theta_0$ 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 $\mathbf{X} = (X_1, \ldots, X_n)$ существует единственное решение уравнения правдоподобия, то оно с вероятностью, стремящейся к 1, является оценкой максимального правдоподобия и, значит, ОМП будет асимптотически нормальной оценкой параметра $\theta$ 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 $\theta$ 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 $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из нормального распределения $\mathcal{N}(\theta, 1)$. Введем следующую оценку параметра $\theta$ :

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

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:

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

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 $\theta$ 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 $\theta$ с асимптотической дисперсией $\sigma^2(\theta)$. Если $\sigma^2(\theta)=i^{-1}(\theta)$, то оценка $\hat\theta_n(\mathbf{X})$ называется асимптотически эффективной оценкой $\theta$.

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 $\theta$, 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 $\theta$), 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 $\mu$ и различимыми распределениями, то есть $\mathrm{P}_{\theta_0} = \mathrm{P}_{\theta_1}$ тогда и только тогда, когда $\theta_0 = \theta_1$.

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

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

**R4** Функция $l(x, \theta) = \ln p_\theta(x)$ дважды непрерывно дифференцируема по $\theta$ 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 $\theta$ 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 $\theta$ 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

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

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 $\theta_0$ 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 $\theta_0$:

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

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

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:

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

Докажем, что эта последовательность стремится к нулю почти наверное:

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

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 $\varepsilon$, 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 $\varepsilon^2 > \dfrac{1}{n}$, 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

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

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

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

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

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

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

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

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

- Из доказанного выше следует, что

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

В итоге, по лемме Слуцкого получаем, что

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

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

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

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

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, $\varlimsup\limits_{n\to\infty} \mathrm{P}_{\theta_n}(D_n) \geqslant \dfrac{1}{2}$. 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 $\theta_0$.

**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 $\lambda$ - 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 $\mu$ 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

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

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 $\Theta$.

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 $\theta$ 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 $\xi$ и $\eta$ - 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 $\xi$?

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 $\eta$?

Пусть $(\Omega, \mathcal{F}, \mathrm{P})$ - вероятностное пространство, $\xi$ - случайная величина на нем, а $\mathfrak{C} \subset \mathcal{F}$ под-$\sigma$-алгебра в $\mathcal{F}$.

**Определение 9.1.** Напомним, что $\sigma$-алгеброй, порожденной случайной величиной $\xi$ называется

$$\mathcal{F}_\xi = \{\xi^{-1}(B) - \{\xi \in B\} : B \in \mathcal{B}(\mathcal{R})\}.$$

Случайная величина $\xi$ называется $\mathfrak{C}$-измеримой, если порожденная ею $\sigma$-алгебра входит в $\mathfrak{C}$:

$$\mathcal{F}_\xi \subset \mathfrak{C}.$$

**Определение 9.2.** Условным математическим ожиданием $\xi$ относительно $\sigma$-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:** $\xi$ 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 $\xi$.

## 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 $\nu$ - 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 $\nu$ 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 $\nu$ - 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 $\sigma$-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 $\eta$ 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 $\sigma$-алгебры

Начнем с явной формулы, по которой можно будет вычислить УМО в ситуации, когда $\mathfrak{C}$ - дискретная $\sigma$-алгебра.

**Напоминание**
Дискретная $\sigma$-алгебра $\mathfrak{C}$ - это $\sigma$-алгебра, которая порождена некоторым (не более чем счетным) разбиением $\{D_n, n \in \mathbb{N}\}$ пространства $\Omega$. То есть, в такую $\sigma$-алгебру входят сами $D_n$ и всевозможные их объединения.

**Лемма**
Если $\sigma$-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 $\xi$ 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 $\xi$ на $\Omega$ - функция на отрезке. Что же такое условное математическое ожидание? Из последней формулы мы поняли, что в данной ситуации условным математическим ожиданием $\xi$ относительно $\sigma$-алгебры, порожденной таким разбиением, будет случайная величина, которая является константой на каждом из множеств $D_i$.

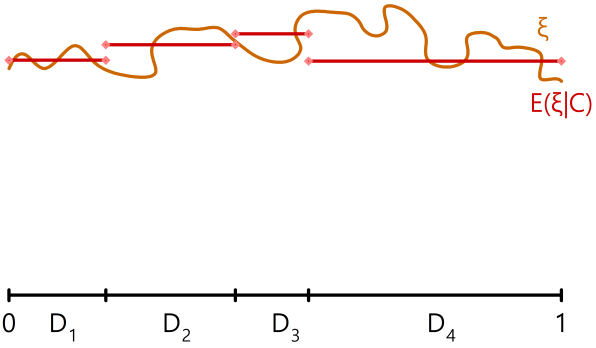


Рис. 5

По сути, условное математическое ожидание - это усреднение нашей случайной величины по множествам из $\sigma$-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 $\xi$ является $\mathfrak{C}$-измеримой, то $\mathrm{E}(\xi|\mathfrak{C}) = \xi$.
**Доказательство.**
Очевидно, что для $\xi$ выполняются свойство измеримости и интегральное свойство. ■

**Свойство (2, линейность)**
Для любых случайных величин $\xi$, $\eta$ и $\sigma$-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 $\xi$ независима с $\mathfrak{C}$, то $\mathrm{E}(\xi|\mathfrak{C}) = \mathrm{E}\xi$.
**Доказательство.**
Напомним, что $\xi$ независима с $\mathfrak{C}$ тогда и только тогда, когда для любого $A \in \mathfrak{C}$ случайная величина $\xi$ независима с $\mathrm{I}_A$. 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 $\delta$

$$\mathrm{E}(\delta \cdot \mathrm{I}\{\delta > 0\}) = \mathrm{E}(\mathrm{E}(\xi|\mathfrak{C}) - \mathrm{E}(\eta|\mathfrak{C}))\mathrm{I}_B).$$

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 $\xi$ - случайная величина, а $\mathfrak{C}_1 \subset \mathfrak{C}_2$ - две $\sigma$-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 $\{\xi_n, n \in \mathbb{N}\}$, $\xi$ - 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 $\xi_n$ 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 $\eta$ 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 $\eta$ - 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 $\eta$. 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 $\eta$ возьмем последовательность простых $\mathfrak{C}$ измеримых случайных величин $\eta_n$ 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 $\varphi$ - 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-$\sigma$-алгебры $\mathfrak{C} \subset \mathcal{F}$ определим

$$\mathrm{P}(A|\mathfrak{C}) := \mathrm{E}(I_A|\mathfrak{C}).$$

Если $\xi$ и $\eta$ - две случайные величины, то полагаем

$$\mathrm{E}(\xi|\eta) := \mathrm{E}(\xi|\mathcal{F}_n),$$

где $\mathcal{F}_n$ - $\sigma$-алгебра, порожденная случайной величиной $\eta$.

**Замечание**
Случайная величина $\zeta$ является $\mathcal{F}_n$-измеримой тогда и только тогда, когда существует такая борелевская функция $\phi(x)$, что $\zeta = \phi(\eta)$.

**Определение 10.2.** Пусть $\xi$ и $\eta$ - 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 $\xi$ 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 $\mathrm{P}_\eta$-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 $\mathrm{P}_\eta$. По теореме Радона-Никодима существует единственная $\mathrm{P}_\eta$-п.н. случайная величина $\varphi$ 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 $\xi$ относительно случайной величины $\eta$ 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 $\xi$ относительно случайной величины $\eta$ называется плотность условного распределения (по некоторой мере $\mu$), 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 $\xi$ относительно случайной величины $\eta$, 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 $\xi$ и $\eta$ 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 $\xi$ по мере $\mu$. Здесь $f_\eta(y)$- это плотность случайной величины $\eta$ по мере $\nu$.

**Доказательство.**
Надо проверить, что функция $P(B, y) = \int\limits_B f_{\xi|\eta}(x|y)\mu(dx)$ является условным распределением $\xi$ относительно $\eta$, 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 $\xi$ и $\eta$.

2) С ее помощью находим $f_{\xi|\eta}(x|y)$ - условную плотность $\xi$ относительно $\eta$.

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 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 $\xi$ и $\eta$ независимы и имеют плотности, то условная плотность $f_{\xi|\eta}(x|y)$ равна просто плотности случайной величины $\xi$.

2) Вычисление $\mathrm{E}(\xi|\eta)$ в случае векторного условия $\eta$ 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 $\xi$ и $\eta$.

**Важный случай**

**Лемма**
Если $\xi$ и $\eta$ - 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)

$$\boxed{=} \int\limits_B \left(\int\limits_{\mathbb{R}} g(x,y)\mathrm{P}_\xi(dx)\right)\mathrm{P}_\eta(dy) = \int\limits_B \mathrm{E}g(\xi,y)\mathrm{P}_\eta(dy).$$

■

## Теорема о наилучшем квадратичном прогнозе

**Теорема 10.4. (о наилучшем квадратичном прогнозе)**
Пусть $\xi$ - случайная величина на $(\Omega,\mathcal{F},\mathrm{P})$ с конечным вторым моментом, $\mathfrak{C}$ - под-$\sigma$-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

$$\mathrm{E}(\xi-\eta)^2 = \mathrm{E}(\xi - \mathrm{E}(\xi|\mathfrak{C}) + \mathrm{E}(\xi|\mathfrak{C}) - \eta)^2 =$$

$$= \mathrm{E}(\xi - \mathrm{E}(\xi|\mathfrak{C}))^2 + \mathrm{E}(\mathrm{E}(\xi|\mathfrak{C}) - \eta)^2 + 2\mathrm{E}\Big((\xi - \mathrm{E}(\xi|\mathfrak{C}))(\mathrm{E}(\xi|\mathfrak{C}) - \eta)\Big).$$

Проверим, что третье слагаемое равно нулю. Тогда $\mathrm{E}(\xi-\eta)^2$ минимизируется, когда второй квадрат равен нулю (т.к. только он зависит от $\eta$). Действительно, по формуле полной вероятности

$$\mathrm{E}\Big((\xi - \mathrm{E}(\xi|\mathfrak{C}))(\mathrm{E}(\xi|\mathfrak{C}) - \eta)\Big) = \mathrm{E}\Big(\mathrm{E}((\xi - \mathrm{E}(\xi|\mathfrak{C}))(\mathrm{E}(\xi|\mathfrak{C}) - \eta)|\mathfrak{C}\Big) \boxed{=}$$

свойство (9) УМО

$$\boxed{=} \mathrm{E}\Big((\mathrm{E}(\xi|\mathfrak{C}) - \eta)\mathrm{E}(\xi - \mathrm{E}(\xi|\mathfrak{C})|\mathfrak{C})\Big) = 0.$$

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 $\Theta$.
- 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 $\lambda$.
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 $\mu$.

В рамках этих предположений:

**Определение 11.1.** Плотность $q(t)$ называется априорной плотностью параметра $\theta$.

## 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 $\theta$ (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 $\theta$ 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 $\theta$ 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

$$\int\limits_{\Theta\times\mathcal{X}} f(t,\mathbf{x})\lambda(dt)\mu(d\mathbf{x}) = \int\limits_{\Theta} q(t)\left(\int\limits_{\mathcal{X}} p_t(\mathbf{x})\mu(d\mathbf{x})\right)\lambda(dt) = \int\limits_{\Theta} q(t)\lambda(dt) = 1.$$

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 $\theta^*$ следующий функционал:

$$\begin{aligned}\int\limits_{\Theta} R_{\theta^*}(t)q(t)\lambda(dt) &= \int\limits_{\Theta} \mathrm{E}_t(\theta^*(\mathbf{X})-t)^2 q(t)\lambda(dt) = \\ &= \int\limits_{\Theta}\left(\int\limits_{\mathcal{X}}(\theta^*(\mathbf{x})-t)^2 p_t(\mathbf{x})\mu(d\mathbf{x})\right)q(t)\lambda(dt) = \\ &= \int\limits_{\Theta\times\mathcal{X}}(\theta^*(\mathbf{x})-t)^2 f(t,\mathbf{x})\lambda(dt)\times\mu(d\mathbf{x}) = \tilde{\mathrm{E}}(\theta^*(\mathbf{X})-\theta)^2.\end{aligned}$$

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 $\theta$;
- Наблюдение $\mathbf{X}$ имеет плотность

$$g(\mathbf{x}) = \int\limits_{\Theta} f(t,\mathbf{x})\lambda(dt);$$

- Условная плотность $\mathbf{X}$ относительно $\theta$ равна $\dfrac{f(t,\mathbf{x})}{q(t)} = p_t(\mathbf{x})$.

Отсюда несложно получить условную плотность $\theta$ 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 $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из нормального распределения $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$. Найдите байесовскую оценку параметра $\theta$, если априорное распределение $\theta$ 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:

$$-\left(\frac{n}{2}+\frac{1}{2\sigma^2}\right)t^2 + t\left(\sum_{i=1}^{n} X_i + \frac{a}{\sigma^2}\right) = -\frac{1}{2}\left(n+\frac{1}{\sigma^2}\right)\left(t - \frac{\overline{\mathbf{X}} + \frac{a}{n\sigma^2}}{1+\frac{1}{n\sigma^2}}\right)^2 + \ldots$$

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 $\theta$, что существует вероятностное распределение Q на $\Theta$ со следующим условием: для всех $\theta \in \Theta$

$$R_{\hat{\theta}}(\theta) \leqslant \int_{\Theta} R_{\hat{\theta}_{\mathrm{Q}}}(t)\mathrm{Q}(dt),$$

где $\hat{\theta}_{\mathrm{Q}}(\mathbf{X})$ - это байесовская оценка $\theta$ 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 $\theta$. Тогда

$$\sup_{\theta\in\Theta} R_{\theta^*}(\theta) \geqslant \int_{\Theta} R_{\theta^*}(t)\mathrm{Q}(dt) \geqslant \int_{\Theta} R_{\hat{\theta}_{\mathrm{Q}}}(t)\mathrm{Q}(dt) \geqslant \sup_{\theta\in\Theta} R_{\hat{\theta}}(\theta).$$

■

**Замечание**
Отметим, что в условиях леммы 1 должно выполняться равенство

$$R_{\hat{\theta}}(\theta) = \int_{\Theta} R_{\hat{\theta}_{\mathrm{Q}}}(t)\mathrm{Q}(dt)$$

почти наверное по мере Q.

**Следствие (1)**
Если оценка $\hat{\theta}(\mathbf{X})$ удовлетворяет двум условиям:

- $\hat{\theta}(\mathbf{X})$ является байесовской оценкой $\theta$ 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 $\theta$;

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 $\theta$.

**Пример**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из распределения Бернулли $\mathrm{Bin}(1, \theta)$, $\theta \in [0, 1]$. Найдите минимаксную оценку $\theta$.

## 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 $\theta$.
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 $\theta$, что существует последовательность априорных распределений $\{\mathrm{Q}_k, k \in \mathbb{N}\}$ на $\Theta$ со следующим условием: для всех $\theta \in \Theta$

$$R_{\hat{\theta}}(\theta) \leqslant \overline{\lim_{k\to+\infty}} \int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}_k}}(t)\mathrm{Q}_k(dt).$$

Тогда оценка $\hat{\theta}(\mathbf{X})$ является минимаксной.
**Доказательство.** Пусть $\theta^*(\mathbf{X})$ - любая оценка $\theta$. 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

$$\sup_{\theta\in\Theta} R_{\theta^*}(\theta) \geqslant \overline{\lim_{k\to+\infty}} \int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}_k}}(t)\mathrm{Q}_k(dt) \geqslant \sup_{\theta\in\Theta} R_{\hat{\theta}}(\theta),$$

что и доказывает минимаксность $\hat{\theta}(\mathbf{X})$. ■

**Пример**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из нормального распределения $\mathcal{N}(\theta, 1), \theta \in \mathbb{R}$. Найдите минимаксную оценку параметра $\theta$.

### 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:

$$\overline{\lim_{k\to+\infty}} \int\limits_{\Theta} R_{\hat{\theta}_{\mathrm{Q}_k}}(t)\mathrm{Q}_k(dt) \geqslant \frac{1}{n}.$$

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 $\theta$.

# 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 $\theta$ устроен так:

$$\mathrm{E}_\theta(\hat{\theta}(\mathbf{X}) - \theta)^2 \to \min_{\hat{\theta}(\mathbf{X})} \text{ равномерно по всем } \theta \in \Theta.$$

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 $\theta$, 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 $\theta$. 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 $\theta$.

## 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 $\theta$, 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 $\mathrm{P}_\theta^S$ - это распределение статистики $S(\mathbf{X})$, когда $\mathbf{X}$ имеет распределение $\mathrm{P}_\theta$.

**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 $\theta$. 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 $\theta$. 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 $\mathrm{E}_\theta$ 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 $\hat\theta(\mathbf{X}) = \theta^*(\mathbf{X})$ $\mathrm{P}_\theta$ - 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

$$a^{\mathrm{T}}\mathrm{D}_\theta\theta^*(\mathbf{X})a \leqslant a^{\mathrm{T}}\mathrm{D}_\theta\hat\theta(\mathbf{X})a.$$

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 $\boldsymbol{\theta}^*(\mathbf{X})$. Тогда $\boldsymbol{\theta}^*(\mathbf{X})$ - оптимальная оценка $\tau(\theta)$.
**Доказательство.**
Пусть оценка $\boldsymbol{\theta}^*(\mathbf{X})$ не оптимальная. Тогда найдется оценка $\hat\theta(\mathbf{X})$, которая будет лучше, т.е. дисперсия $\hat\theta(\mathbf{X})$ будет равномерно не больше дисперсии $\boldsymbol{\theta}^*(\mathbf{X})$ и для некоторого $\theta_0 \in \Theta$

$$\mathrm{D}_{\theta_0}\hat\theta(\mathbf{X}) < \mathrm{D}_{\theta_0}\boldsymbol{\theta}^*(\mathbf{X}).$$

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 $\boldsymbol{\theta}^*(\mathbf{X})$, и $\theta^{**}(\mathbf{X})$ являются функциями от $S(\mathbf{X})$. Следовательно, они равны, и $\boldsymbol{\theta}^*(\mathbf{X})$ не хуже, чем $\hat{\theta}(\mathbf{X})$.
Противоречие. Значит, $\boldsymbol{\theta}^*(\mathbf{X})$ 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 $f(S(\mathbf{X})) = 0$ $\mathrm{P}_\theta$ - 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 $\varphi(S(\mathbf{X})) = \psi(S(\mathbf{X}))$ $\mathrm{P}_\theta$ - 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 $\varphi$) 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 $\mu$ 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 $p_\theta$ в виде

$$p_\theta(\mathbf{x}) = h(\mathbf{x})\psi_\theta(S(x)),$$

где $\psi_\theta$ 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 $\theta$ в силу достаточности $S(\mathbf{X})$, а второй зависит только от $\theta$ 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 $\theta$. 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 $\xi$ 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)

$$= \mathrm{E}\left(\mathrm{I}_\mathrm{A} \cdot \frac{\mathrm{E}(\xi \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C})}{\mathrm{E}(\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C})} \cdot \mathrm{E}\left(\frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C}\right)\right) = \mathrm{E}\left(\mathrm{I}_\mathrm{A} \cdot \mathrm{E}\left(\xi \cdot \frac{d\tilde{\mathrm{P}}}{d\mathrm{P}}|\mathfrak{C}\right)\right) =$$

(интегральное свойство)

$$= \mathrm{E}\left(\mathrm{I}_\mathrm{A} \cdot \xi \cdot \frac{d\tilde{P}}{d\mathrm{P}}\right) = \tilde{E}(\mathrm{I}_\mathrm{A} \cdot \xi).$$

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 $S(\mathbf{X})$ выполнено 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 $\mathrm{A} \in \mathcal{F}_{S(\mathbf{X})}$

$$\mathrm{E}_{\theta,\theta'}(f_\theta(\mathbf{X}) \cdot \mathrm{I}_\mathrm{A}) = \frac{1}{2}(\mathrm{E}_\theta(f_\theta(\mathbf{X})\mathrm{I}_\mathrm{A}) + \mathrm{E}_{\theta'}(f_\theta(\mathbf{X}) \cdot \mathrm{I}_\mathrm{A})) =$$

(интегральное свойство)

$$= \frac{1}{2}(\mathrm{E}_\theta(\mathrm{E}_\theta(f_\theta(\mathbf{X})|S(\mathbf{X})) \cdot \mathrm{I}_\mathrm{A}) + \mathrm{E}_{\theta'}(\mathrm{E}_{\theta'}(f_\theta(\mathbf{X})|S(\mathbf{X})) \cdot \mathrm{I}_\mathrm{A}))$$

$$= \frac{1}{2}(\mathrm{E}_\theta(\mathrm{E}_\theta(f_\theta(\mathbf{X})|S(\mathbf{X}))\cdot \mathrm{I}_\mathrm{A}) + \mathrm{E}_{\theta'}(\mathrm{E}_\theta(f_\theta(\mathbf{X})|S(\mathbf{X}))\cdot \mathrm{I}_\mathrm{A})) = \mathrm{E}_{\theta,\theta'}(\mathrm{E}_\theta(f_\theta(\mathbf{X})|S(\mathbf{X}))\cdot \mathrm{I}_\mathrm{A}).$$

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 $\theta'$:

$$\mathrm{E}_{\theta'}(T(\mathbf{X})|S(\mathbf{X})) = \frac{\mathrm{E}_{\theta,\theta'}(T(\mathbf{X})f'_\theta(\mathbf{X})|S(\mathbf{X}))}{\mathrm{E}_{\theta,\theta'}(f'_\theta(\mathbf{X})|S(\mathbf{X}))} = \mathrm{E}_{\theta,\theta'}(T(\mathbf{X})|S(\mathbf{X})).$$

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 $\theta$. По определению мы получаем, что $S(\mathbf{X})$ - достаточная статистика. ■

## Примеры

**Пример (1)**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из распределения Бернулли $\mathrm{Bin}(1, \theta), \theta \in (0, 1)$. Найдите достаточную статистику.

**Решение**

Плотность имеет вид:

$$p_\theta(X_1,\dots,X_n)=\prod_{i=1}^n p_\theta(X_i)=\prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i}\cdot \mathrm{I}\{X_i\in\{0,1\}\}=$$

$$=\theta^{\sum\limits_{i=1}^n X_i}(1-\theta)^{n-\sum\limits_{i=1}^n X_i}\cdot \mathrm{I}\{\forall X_i\in\{0,1\}\}.$$

Значит, $\sum\limits_{i=1}^n X_i$ - достаточная статистика.

**Пример (2)**
Пусть $\mathbf{X}=(X_1,\dots,X_n)$ - выборка из пуассоновского распределения $\mathrm{Pois}(\theta),\theta>0$. Найдите достаточную статистику.

**Решение**

Плотность имеет вид:

$$p_\theta(X_1,\dots,X_n)=\prod_{i=1}^n p_\theta(X_i)=\prod_{i=1}^n \frac{\theta^{X_i}}{X_i!}\,e^{-\theta}=\left(\frac{1}{\prod\limits_{i=1}^n X_i!}\right)\theta^{\sum\limits_{i=1}^n X_i}\cdot e^{-n\theta}.$$

Значит, $\sum\limits_{i=1}^n X_i$ - достаточная статистика.

**Пример (3)**
Пусть $\mathbf{X}=(X_1,\dots,X_n)$ - 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:

$$p_\theta(X_1,\dots,X_n)=\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\,exp(-\frac{(X_i-a)^2}{2\sigma^2})=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp(-\frac{\sum\limits_{i=1}^n X_i^2-2a\sum\limits_{i=1}^n X_i+na^2}{2\sigma^2}).$$

Значит, $\left(\sum\limits_{i=1}^n X_i^2,\sum\limits_{i=1}^n X_i\right)$ - достаточная статистика.

**Пример (4)**
Пусть $\mathbf{X}=(X_1,\dots,X_n)$ - выборка из равномерного распределения $\mathrm{U}(0,\theta),\theta>0$. Найдите оптимальную оценку $\theta$.

**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 $g(X_{(n)})$:

$$\mathrm{E}_\theta g(X_{(n)}) = \int_0^\theta g(x) \frac{nx^{n-1}}{\theta^n} dx.$$

Теперь предположим, что для всех $\theta > 0$ $\mathrm{E}_\theta\ g(X_{(n)}) = 0$. Это равносильно тому, что

$$\int_0^\theta g(x) x^{n-1} dx = 0.$$

Продифференцируем по $\theta$:

$$g(\theta)\theta^{n-1} = 0 \Longleftrightarrow g(\theta) = 0 \text{ почти наверное для } \theta > 0.$$

Тогда $g(X_{(n)}) = 0$ $\mathrm{P}_\theta$ - 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:

$$\frac{\theta^{n+1}}{n} = \int_0^\theta \varphi(x) x^{n-1} dx.$$

Продифференцируем по $\theta$:

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

Следовательно, оптимальной оценкой $\theta$ 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 $\mu$ 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 $S(\mathbf{X}) = (U_1(\mathbf{X}), \ldots, U_k(\mathbf{X}))$. 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 $\theta$ всего $\Theta$ вектор $\mathbf{a}(\theta) = (a_1(\theta), \ldots, a_k(\theta))$ зачерчивает $k$-мерный параллелепипед или шар (это значит, что внутри множества значений вектора $\mathbf{a}(\theta)$, при $\theta$ пробегающем $\Theta$, содержится некоторый $k$-мерный параллелепипед или шар), то статистика $S(\mathbf{X}) = (U_1(\mathbf{X}), \ldots, U_k(\mathbf{X}))$ будет полной достаточной статистикой для $\{\mathrm{P}_\theta, \theta \in \Theta\}$.

Для доказательства потребуется несколько утверждений.

# Плотность $S(\mathbf{X})$

Доказательство теоремы об экспоненциальном семействе использует комплексный анализ. Нам понадобится теорема единственности для аналитических функций. Но для начала разберемся с вопросом: чему равна плотность $S(\mathbf{X})$? Когда мы говорим о полноте, мы должны проверить, что математическое ожидание функции от $S(\mathbf{X})$ может быть нулем только если эта функция нулевая. Для того, чтобы выписать математическое ожидание в интегральном виде, нам необходимо понимать, а имеет ли $S(\mathbf{X})$ плотность. Оказывается, что плотность у $S(\mathbf{X})$ есть. Обозначим

$$\psi_\theta(S(\mathbf{X})) = \exp\left\{\sum_{i=1}^{k} U_i(\mathbf{X})a_i(\theta) + v(\theta)\right\} = \exp\{\langle \mathbf{a}(\theta), S(\mathbf{X})\rangle + v(\theta)\}.$$

**Лемма (1)**
Статистика $S(\mathbf{X})$ будет иметь плотность $\psi_\theta(\mathbf{s})$ по мере $\nu$ на $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$, заданной соотношением: для $B \in \mathcal{B}(\mathbb{R}^k)$

$$\nu(B) = \int\limits_{S^{-1}(B)} h(\mathbf{x})\mu(d\mathbf{x}).$$

**Доказательство.**
Для существования плотности нужно показать, что распределение статистики $S(\mathbf{X})$ абсолютно непрерывно относительно меры $\nu$. Пусть $G_\theta = \mathrm{P}_\theta(S(\mathbf{X}) \in B)$ есть распределение $S(\mathbf{X})$. Если $\nu(B) = 0$, то $h(\mathbf{x}) = 0$ $\nu$-п.н. на множестве интегрирования $S^{-1}(B)$. Тогда

$$G_\theta(B) = \mathrm{P}_\theta(S(\mathbf{X}) \in B) = \int\limits_{S^{-1}(B)} p_\theta(\mathbf{x})\mu(d\mathbf{x}) = \int\limits_{S^{-1}(B)} h(\mathbf{x})\psi_\theta(S(\mathbf{x}))\mu(d\mathbf{x}) = 0.$$

Следовательно, распределение $G_\theta$ абсолютно непрерывно относительно $\nu$ и по теореме Радона-Никодима существует плотность $\frac{dG_\theta}{d\nu}$. Но в силу определения $\nu$

$$G_\theta(B) = \int\limits_{S^{-1}(B)} h(\mathbf{x})\psi_\theta(S(\mathbf{x}))\mu(d\mathbf{x}) = \int\limits_{B} \psi_\theta(\mathbf{s})\nu(d\mathbf{s}).$$

Значит, $\frac{dG_\theta}{d\nu}(\mathbf{s}) = \psi_\theta(\mathbf{s})$.

■

# Достаточные статистики и информация Фишера

**Замечание**
Лемма (1) выполнена не только для экспоненциального семейства, а всегда, когда имеет место факторизация.

Теперь мы можем доказать и последнее неразобранное свойство информации Фишера.

**Следствие**
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 $\{\mathrm{P}_\theta, \theta \in \Theta\}$.
**Доказательство.**
$\boxed{\Rightarrow}$ Пусть $S(\mathbf{X})$ - достаточная статистика. Тогда по критерию факторизации выполнено

$$p_\theta(\mathbf{x}) = h(\mathbf{x})\psi_\theta(S(\mathbf{x})),$$

где $\psi_\theta(\mathbf{s})$ - плотность $S(\mathbf{X})$. Стало быть, вклады $\mathbf{X}$ и $S(\mathbf{X})$ равны:

$$\frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X}) = \frac{\partial}{\partial\theta}\ln \psi_\theta(S(\mathbf{X})),$$

откуда следует и равенство информаций Фишера, так как информация - второй момент для вклада.

$\boxed{\Leftarrow}$ В обратную сторону почти все следует из доказательства, приведенного на лекции 5. Из равенства информаций Фишера мы получаем, что вклады $\mathbf{X}$ и $S(\mathbf{X})$ должны быть пропорциональны

$$\frac{\partial}{\partial\theta}\ln p_\theta(\mathbf{X}) = a(\theta)\cdot\frac{\partial}{\partial\theta}\ln g_\theta(S(\mathbf{X})),$$

где $g_\theta(\mathbf{s})$ - плотность $S(\mathbf{X})$. Находя из равенства $p_\theta(\mathbf{x})$, мы получаем факторизацию плотности:

$$\ln p_\theta(\mathbf{X}) = \ln \psi_\theta(S(\mathbf{X})) + h(\mathbf{X}).$$

Значит, $S(\mathbf{X})$ - достаточная статистика. ■

# Преобразование Лапласа

**Лемма (2)**
Пусть $G_1$ и $G_2$ - две $\sigma$-конечные меры на $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. Если имеет место равенство конечных интегралов

$$\int\limits_{\mathbb{R}^k} e^{\langle \mathbf{a},\mathbf{u}\rangle} G_1(d\mathbf{u}) = \int\limits_{\mathbb{R}^k} e^{\langle \mathbf{a},\mathbf{u}\rangle} G_2(d\mathbf{u}) < \infty$$

для всех $\mathbf{a} \in \mathbf{I}$, где $\mathbf{I}$ - некоторый параллелепипед в $\mathbb{R}^k$, то $G_1 = G_2$.
**Доказательство.**
Докажем только для $k = 1$ (при $k > 1$ все аналогично, но требует знаний многомерного комплексного анализа).

Сначала считаем, что $\mathbf{I} = \{|x| \leqslant \alpha\}$ для всех $\alpha > 0$. Тогда функции

$$h_i(a) = \int\limits_{\mathbb{R}} e^{au} G_i(du), \quad i = 1, 2,$$

будут аналитическими в интервале $|a| < \alpha$. Рассмотрим теперь $h_i$ как функции комплексного переменного:

$$h_i(z) = \int\limits_{\mathbb{R}} e^{zu} G_i(dz), \quad i = 1, 2, \quad z \in \mathbb{C}.$$

Тогда $h_i(z)$ будут аналитическими в полосе $|\mathrm{Re}(z)| < \alpha$ и совпадающими на интервале $\{|\mathrm{Re}(z)| < \alpha, \mathrm{Im}(z) = 0\}$. По теореме единственности $h_1(z) = h_2(z)$ во всей полосе и, в частности, на мнимой прямой $\mathrm{Re}(z) = 0$. Таким образом, для любого $t \in \mathbb{R}$ выполнено:

$$\int\limits_{\mathbb{R}} e^{itu} G_1(du) = \int\limits_{\mathbb{R}} e^{itu} G_2(du).$$

Видно, что $\int\limits_{\mathbb{R}} G_1(du) = \int\limits_{\mathbb{R}} G_2(du) < +\infty$, поэтому можно считать, что $G_1$ и $G_2$ - это вероятностные меры. Но тогда мы получаем, что совпадают их характеристические функции. Значит, $G_1 = G_2$.

Если же $\mathbf{I} = \{|x - \alpha_0| \leqslant \alpha\}$, то надо перейти к мерам

$$G_i^*(du) = e^{\alpha_0 u} G_i(du), \quad i = 1, 2.$$

■

## Доказательство теоремы об экспоненциальном семействе

**Доказательство.**
Достаточность статистики $S(\mathbf{X})$ следует из критерия факторизации Неймана-Фишера. Докажем ее полноту. Пусть $\varphi(\mathbf{s})$ - такая борелевская функция из $\mathbb{R}^k$ в $\mathbb{R}$, что $\mathrm{E}_\theta\varphi(S(\mathbf{X})) = 0$ для всех $\theta \in \Theta$. Заметим, что по лемме (1) это эквивалентно тому, что для любого $\theta \in \Theta$

$$\mathrm{E}_\theta\varphi(S(\mathbf{X})) = \int\limits_{\mathbb{R}^k} \varphi(\mathbf{s})\psi_\theta(\mathbf{s})\nu(d\mathbf{s}) = 0.$$

Введем две функции: $\varphi^+(\mathbf{s}) = \max\{\varphi(\mathbf{s}), 0\}$ и $\varphi^- = \max\{-\varphi(\mathbf{s}), 0\}$. Тогда для любого $\theta \in \Theta$

$$\int\limits_{\mathbb{R}^k} \varphi^+(\mathbf{s})\psi_\theta(\mathbf{s})\nu(d\mathbf{s}) = \int\limits_{\mathbb{R}^k} \varphi^-(\mathbf{s})\psi_\theta(\mathbf{s})\nu(d\mathbf{s}).$$

Вспомнив, что $\psi_\theta(\mathbf{s}) = e^{\langle \mathbf{a}(\theta), \mathbf{s}\rangle + v(\theta)}$, мы получаем, что для любого $\theta \in \Theta$

$$\int\limits_{\mathbb{R}^k} e^{\langle \mathbf{a}(\theta), \mathbf{s}\rangle} G_+(d\mathbf{s}) = \int\limits_{\mathbb{R}^k} e^{\langle \mathbf{a}(\theta), \mathbf{s}\rangle} G_-(d\mathbf{s}),$$

где $G_\pm(d\mathbf{s}) = \varphi^\pm(\mathbf{s})\nu(d\mathbf{s})$ - новые меры.

Из условия теоремы следует, что для всех $\mathbf{b}$ из некоторого параллелепипеда в $\mathbb{R}^k$ выполняется равенство

$$\int\limits_{\mathbb{R}^k} e^{\langle \mathbf{b}, \mathbf{s}\rangle} G_+(d\mathbf{s}) = \int\limits_{\mathbb{R}^k} e^{\langle \mathbf{b}, \mathbf{s}\rangle} G_-(d\mathbf{s}).$$

По лемме (2) это означает, что $G_+ = G_-$, то есть $\varphi(\mathbf{s}) = 0$ $\nu$-п.н. По определению получаем, что $S(\mathbf{X})$ - это полная статистика для $\{\mathrm{P}_\theta, \theta \in \Theta\}$. ■

## Пример нахождения оптимальной оценки

**Пример**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из нормального распределения $\mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$, $\theta \in \mathbb{R} \times \mathbb{R}_+$. Найдите оптимальную оценку $\theta$.

### Решение

**План решения:**
Необходимо найти полную достаточную статистику. Достаточность мы проверим по критерию факторизации, а полноту по теореме об экспоненциальном семействе. Останется решить уравнение несмещенности.

Начнем с критерия факторизации. Для этого выпишем плотность нашей выборки:

$$p_\theta = \prod_{k=1}^{n} p_\theta(X_k) = \prod_{k=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(X_k-\mu)^2}{2\sigma^2}\right\} = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(X_k-\nu)^2\right\} =$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n} X_k^2 + \frac{\mu}{\sigma^2}\sum_{k=1}^{n} X_k - \frac{n\mu^2}{2\sigma^2} - \frac{n}{2}\ln\sigma^2\right\}.$$

Тогда это будет экспоненциальное семейство распределений, где

$$\mathbf{a}(\theta) = \left(\frac{\mu}{2\sigma^2}, \frac{-1}{2\sigma^2}\right), \quad S(\mathbf{X}) = \left(\sum_{k=1}^{n} X_k, \sum_{k=1}^{n} X_k^2\right),$$

$$v(\theta) = \frac{-n\nu^2}{2\sigma^2} - n\ln\sigma, \quad h(\mathbf{X}) = (2\pi)^{-\frac{n}{2}}.$$

Заметим, что $\mathbf{a}(\theta)$ пробегает $\mathbb{R}\times\mathbb{R}_-$. Тогда по теореме $S(\mathbf{X})$ есть полная достаточная статистика.

Теперь надо решить уравнения несмещенности, причем их можно решать по-отдельности для $\mu$ и $\sigma^2$.

Сразу скажем, что $\mathrm{E}_\theta\overline{\mathbf{X}} = \mu$, то есть выборочное среднее есть оптимальная оценка среднего. Далее,

$$\mathrm{E}_\theta\left(\frac{1}{n}\sum_{k=1}^{n} X_k^2\right) = \mu^2 + \sigma^2.$$

Нужно избавиться от $\mu^2$. Для этого посчитаем второй момент выборочного среднего, пользуясь тем, что $\overline{\mathbf{X}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$:

$$\mathrm{E}_\theta(\overline{\mathbf{X}})^2 = \mu^2 + \frac{\sigma^2}{n}.$$

Следовательно, оптимальной оценкой $\sigma^2$ будет

$$\frac{n}{n-1}\cdot S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{\mathbf{X}})^2.$$

Подведем итог и сформулируем общую схему поиска оптимальных оценок.

Пусть требуется отыскать оптимальную оценку $\tau(\theta)$. Тогда

1) Находим достаточную статистику $S(\mathbf{X})$ с помощью критерия факторизации.
2) Проверяем ее на полноту с помощью теоремы об экспоненциальном семействе.
3) Вычисляем распределение $S(\mathbf{X})$.
4) Решаем уравнение несмещенности:

$$\mathrm{E}_\theta \varphi(S(\mathbf{X})) = \tau(\theta).$$

**Ответ**: $\varphi(S(\mathbf{X}))$ - это оптимальная оценка $\tau(\theta)$.

## Доверительное оценивание

Предыдущие темы нашего курса относились к так называемому "точечному" оцениванию параметров. В его рамках параметру $\theta$ сопоставляется оценка $\hat{\theta}(\mathbf{X})$, то есть точка, в некотором смысле близка к $\theta$. Но насколько близко на самом деле находится $\hat{\theta}(\mathbf{X})$ к $\theta$? Например, нам бы хотелось указать такое $\varepsilon$, что с большой вероятностью выполнено $|\hat{\theta}(\mathbf{X}) - \theta| < \varepsilon$.

**Более общая постановка вопроса:** указать интервал, который накрывает $\theta$ с вероятностью не менее $\gamma$ (обычно берут $\gamma = 0.95$ или $0.99$). Величина $\gamma$ называется уровнем доверия (доверительной вероятностью), она отражает "степень готовности мириться с ошибкой".

**Важный момент:** для выбранного $\gamma$ следует выбрать наименее короткий интервал ( в идеале - его длина должна стремится к нулю с ростом числа наблюдений).

### Доверительные интервалы

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

**Определение 13.2.** Пара статистик $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ называется доверительным интервалом уровня доверия $\gamma$ для параметра $\theta$, если для любого $\theta \in \Theta$ выполняется неравенство:

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

Если данная вероятность равна $\gamma$ для всех $\theta \in \Theta$, то доверительный интервал называется точным.

### Доверительные области

Если же параметр $\theta$ является многомерным, $\Theta \subset \mathbb{R}^k$, $k > 1$, то можно строить доверительные интервалы для его компонент $\theta_i$ или скалярных функций $\tau(\theta) \in \mathbb{R}$. В общем случае вводится понятие доверительной области.

**Определение 13.3.** Подмножество $S(\mathbf{X}) \subset \Theta$ называется доверительной областью уровня доверия $\gamma$ для параметра $\theta$, если для любого $\theta \in \Theta$ выполнено

$$\mathrm{P}_\theta(\theta \in S(\mathbf{X})) \geqslant \gamma.$$

**Вопрос:** как строить доверительные интервалы и области?

## Метод центральной статистики

Предположим, что нашлась такая одномерная функция $G(\mathbf{X}, \theta)$, что ее распределение не зависит от $\theta$, а сама она зависит от $\theta$. Подобная функция называется **центральной статистикой**.

Пусть распределение $G(\mathbf{X}, \theta)$ нам известно, и для заданных $\gamma_1, \gamma_2 \in (0, 1)$ с условием $\gamma_2 - \gamma_1 = \gamma$ мы смогли вычислить величины $g_i$, $i = 1, 2$, - $\gamma_i$-квантили распределения $G(\mathbf{X}, \theta)$. Тогда для любого $\theta \in \Theta$ мы получаем, что

$$\mathrm{P}_\theta(g_1 \leqslant G(\mathbf{X}, \theta) \leqslant g_2) \geqslant \gamma_2 - \gamma_1 = \gamma.$$

**Утверждение**
Множество $S(\mathbf{X}) = \{\theta \in \Theta : \ g_1 \leqslant G(\mathbf{X}, \theta) \leqslant g_2\}$ является доверительной областью уровня доверия $\gamma$ для параметра $\theta$.

**Доказательство.**

$$\mathrm{P}_\theta(\theta \in S(\mathbf{X})) = \mathrm{P}_\theta(g_1 \coloneqq G(\mathbf{X}, \theta) \leqslant g_2) \geqslant \gamma.$$

■

### Идеи по вычислению

Указанный подход имеет сразу несколько проблем.

- Неравенство $g_1 \leqslant G(\mathbf{X}, \theta) \leqslant g_2$ надо решить относительно $\theta$. Если $G(\mathbf{X}, \theta)$ имеет абсолютно непрерывное распределение а сама монотонно и непрерывно зависит от $\theta$, то $\{g_1 < G(\mathbf{X}, \theta) < g_2\}$ станет интервалом вида $\{T_1(\mathbf{X}) < \theta < T_2(\mathbf{X})\}$. Но если зависимость от $\theta$ "плохая" , то может быть достаточно трудно решить данное неравенство относительно $\theta$.

- Поиск центральной статистики - отдельная и сложная задача. Она легко решается, если проста природа параметра $\theta$. Например, это параметр сдвига или масштаба.
- Даже для параметра сдвига или масштаба необходимо вычислить распределение центральной статистики $G(\mathbf{X}, \theta)$ и добиваться малой длины получающегося интервала.

**Пример**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из равномерного распределения $\mathrm{U}(0, \theta)$, $\theta > 0$. Постройте точный доверительный интервал уровня доверия $\gamma$ с помощью статистики $X_{(n)}$.

**Решение**

Ясно, что $\dfrac{X_i}{\theta} \sim \mathrm{U}(0, 1)$. Тогда

$$\mathrm{P}_\theta\left(\frac{X_{(n)}}{\theta} \leqslant x\right) = x^n, \quad x \in [0, 1].$$

Значит,

$$\mathrm{P}_\theta\left(g_1 < \frac{X_{(n)}}{\theta} < g_2\right) = g_2^n - g_1^n = \gamma.$$

Положим $g_2 = 1$, $g_1 = (1 - \gamma)^{\frac{1}{n}}$. Тогда с вероятностью в точности $\gamma$

$$X_{(n)} < \theta < X_{(n)}(1 - \gamma)^{-\frac{1}{n}}.$$

Длина интервала имеет порядок $O(\frac{1}{n})$.

## Асимптотические доверительные интервалы

**Вопрос:** что делать, если центральную статистику не удается быстро найти или ее распределение сложно вычислить?

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка растущего размера с неизвестным распределением $P \in \{\mathrm{P}_\theta, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$.

**Определение 13.4.** Последовательность пар статистик $(T_n^1(X_1, \ldots, X_n), T_n^2(X_1, \ldots, X_n))$ называется асимптотическим доверительным интервалом уровня доверия $\gamma$ для параметра $\theta$, если для любого $\theta \in \Theta$

$$\varliminf_{n \to +\infty} \mathrm{P}_\theta(T_n^1(X_1, \ldots, X_n) < \theta < T_n^2(X_1, \ldots, X_n)) \geqslant \gamma.$$

Если для любого $\theta \in \Theta$

$$\exists \lim_{n \to +\infty} \mathrm{P}_\theta(T_n^1(X_1, \ldots, X_n) < \theta < T_n^2(X_1, \ldots, X_n)) = \gamma,$$

то асимптотический доверительный интервал называется точным.

### Метод построения

Идея построения использует асимптотически нормальные оценки. Свойство асимптотической нормальности оказывается здесь принципиально важным, так как оно позволяет нам строить весьма хорошие асимптотические доверительные интервалы, которые часто используются на практике.

Пусть $\hat{\theta}_n(\mathbf{X}) = \hat{\theta}_n(X_1, \ldots, X_n)$ - это асимптотически нормальная оценка $\theta$ с асимптотической дисперсией $\sigma^2(\theta)$, то есть для любого $\theta \in \Theta$

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

Тогда

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

Надо избавиться от $\theta$ в знаменателе, заменив $\sigma(\theta)$ на ее состоятельную оценку $\hat{\hat{\delta}}_n(\mathbf{X})$. Тогда

$$\sqrt{n} \cdot \frac{\hat{\theta}_n(\mathbf{X}) - \theta}{\hat{\hat{\delta}}_n(\mathbf{X})} = \sqrt{n} \cdot \frac{\hat{\theta}_n(\mathbf{X}) - \theta}{\sigma(\theta)} \cdot \frac{\sigma(\theta)}{\hat{\hat{\delta}}_n(\mathbf{X})}.$$

Первый множитель сходится по распределению к $\mathcal{N}(0, 1)$, а второй - к 1 по вероятности. По лемме Слуцкого получаем, что

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

Теперь мы можем построить асимптотические доверительные интервалы. Обозначим через $u_\alpha$ - $\alpha$-квантиль $\mathcal{N}(0, 1)$. В силу симметрии будет выполнено

$$u_{\frac{1+\gamma}{2}} = -u_{\frac{1-\gamma}{2}}.$$

Тогда для любого $\theta \in \Theta$

$$\mathrm{P}_\theta\left(\sqrt{n} \cdot \left|\frac{\hat{\theta}_n(\mathbf{X}) - \theta}{\hat{\hat{\delta}}_n(\mathbf{X})}\right| < u_{\frac{1+\gamma}{2}}\right) \longrightarrow \gamma \quad \text{при } n \to +\infty.$$

Следовательно, получаем следующий асимптотический доверительный интервал:

$$\left(\hat{\theta}_n(\mathbf{X}) - u_{\frac{1+\gamma}{2}} \cdot \frac{\hat{\hat{\delta}}_n(\mathbf{X})}{\sqrt{n}}, \hat{\theta}_n(\mathbf{X}) + u_{\frac{1+\gamma}{2}} \cdot \frac{\hat{\hat{\delta}}_n(\mathbf{X})}{\sqrt{n}}\right).$$

Его длина имеет порядок $O(\frac{1}{\sqrt{n}})$.

**Замечание (1)**
Длина получаемого интервала равна

$$2u_{\frac{1+\gamma}{2}} \cdot \frac{\hat{\delta}_n(\mathbf{X})}{\sqrt{n}} \sim 2u_{\frac{1+\gamma}{2}} \cdot \frac{\sigma(\theta)}{\sqrt{n}},$$

то есть выбор оценки с меньшей асимптотической дисперсией уменьшает длину асимптотического доверительного интервала.

**Замечание (2)**
Как находить $\hat{\delta}_n(\mathbf{X})$? Заметим, что $\hat{\theta}_n(\mathbf{X}) \xrightarrow{\mathrm{P}_\theta} \theta$. Если $\sigma(\theta)$ непрерывна по $\theta$, то подойдет $\hat{\delta}_n(\mathbf{X}) = \sigma(\hat{\theta}_n(\mathbf{X}))$, которая является состоятельной оценкой $\sigma(\theta)$ в силу теоремы о наследовании сходимости.

Однако, бывает, что $\sigma(\theta)$ не удается явно найти как функцию от параметра. В этом случае может подойти состоятельная оценка "общего вида". Например, если мы знаем, что $\sigma^2(\theta)$ - это дисперсия $X_1$, то ее состоятельно оценивает выборочная дисперсия $S^2$.

**Пример**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из схемы Бернулли $\mathrm{Bin}(1, \theta)$, $\theta \in (0, 1)$. Постройте асимптотический доверительный интервал уровня доверия $\gamma$.

**Решение**

В силу того, что

$$\mathrm{E}_\theta X_1 = \theta,\ \mathrm{D}_\theta X_1 = \theta(1-\theta),$$

оценка $\overline{\mathbf{X}}$ будет асимптотически нормальной оценкой $\theta$ с асимптотической дисперсией $\sigma^2(\theta) = \theta(1-\theta)$. Согласно ЦПТ

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

Статистика $S^2 = \overline{\mathbf{X}^2} - (\overline{\mathbf{X}})^2$ является состоятельной оценкой $\theta(1-\theta)$, поэтому используя описанную схему, получаем следующий доверительный интервал:

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

# 14. Лекция 14

## Постановка задачи линейной регрессии

Разберем классическую задачу линейной регрессии. Она формулируется следующим образом: допустим, что мы хотим узнать какую-то величину $\boldsymbol{l} \in \mathbb{R}^n$ и проводим ее измерение. Но реальные инструменты не могут иметь абсолютную точность, поэтому замеры будут содержать ошибку:

$$\mathbf{X} = \boldsymbol{l} + \boldsymbol{\varepsilon},$$

где $\mathbf{X}$ - результат измерения, а $\varepsilon$ - ошибка измерения. Обычно предполагается, что $\boldsymbol{\varepsilon}$ есть гауссовский вектор, но сначала мы будем рассматривать общий случай. На $\boldsymbol{\varepsilon}$ накладывается следующие ограничения:

$$\mathrm{E}\boldsymbol{\varepsilon} = \mathbf{0},\ \mathrm{D}\boldsymbol{\varepsilon} = \sigma^2 \mathbf{I}_n,$$

где $\mathbf{I}_n$ - единичная матрица размера $n \times n$. В среднем ошибка нулевая, а между собой ошибки измерений разных компонент некоррелированы и имеют одинаковую известную дисперсию $\sigma^2$. В таком виде можно представить очень многие реальные задачи.

**Задача**
По наблюдению $\mathbf{X}$ оценить неизвестные параметры $\boldsymbol{l}$ и $\sigma^2$.

В чем состоит линейность модели? В том, что мы кое-что знаем о векторе $\boldsymbol{l}$: он принадлежит нетривиальному известному линейному подпространству $L$ размерности $k < n$. Но что означает, что мы знаем $L$?

Будем считать, что нам известен его некоторый базис $z_1, \ldots, z_k \in \mathbb{R}^n$. Составим из базисных векторов регрессионную матрицу: $\mathbf{Z} = (z_1, \ldots, z_k) \in \mathrm{Mat}(n \times k)$. Понятно, что в таком случае $\mathbf{Z} = k$. Тогда $\boldsymbol{l}$ можно разложить в линейную комбинацию:

$$\boldsymbol{l} = \sum_{i=1}^{k} \theta_i z_i = \mathbf{Z}\boldsymbol{\theta}, \text{ где } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^{\mathrm{T}}.$$

Смысл вектора $\boldsymbol{\theta}$ очевиден: это неизвестные координаты вектора $\boldsymbol{l}$ в базисе $z_1, \ldots, z_k$. Тогда задача линейной регрессии переформулируется следующим образом:

**Переформулированная задача**
По наблюдению $\mathbf{X}$ нужно оценить параметры $\boldsymbol{\theta}$ и $\sigma^2$.

**Пример**
Пусть объект движется по прямой с постоянной скоростью. Далее, в некоторые моменты времени $t_1, \dots, t_n$ мы измеряем положение объекта:

$$X_i = a + bt_i + \varepsilon_i,\ i = 1, \dots, n,$$

где $a$ - стартовая позиция, $b$ - скорость движения тела и $\varepsilon_i$ - погрешность измерения на $i$-й итерации. Задача состоит в оценке неизвестных $a$ и $b$. Легко видеть, что это задача линейной регрессии, где

$$\boldsymbol{l} = \begin{pmatrix} a + bt_1 \\ \vdots \\ a + bt_n \end{pmatrix} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \mathbf{Z}\boldsymbol{\theta},$$

$\boldsymbol{\theta} = (a, b)^{\mathrm{T}}$. Можно провести аналогичные рассуждения, если рассматривать объект, который движется по параболе. В таком случае будет добавляться член квадратичной зависимости от времени, $ct_i^2$.

## Метод наименьших квадратов

Как было сказано ранее, задача состоит в оценивании пары $(\boldsymbol{\theta}, \sigma^2)$. Для этого максимально удобным оказывается метод наименьших квадратов.

**Определение 14.1.** Оценкой $\boldsymbol{\theta}$ по методу наименьших квадратов называется

$$\hat{\boldsymbol{\theta}}(\mathbf{X}) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^n} \|\mathbf{X} - \mathbf{Z}\boldsymbol{\theta}\|^2.$$

Сама $\hat{\boldsymbol{\theta}}$ называется оценкой наименьших квадратов (о.н.к.).

Геометрический смысл этого определения крайне прост: $\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$ есть проекция $\mathbf{X}$ на $L$. Но тогда $\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$ есть проекция $\mathbf{X}$ на ортогональное дополнение $L^{\perp}$. Отсюда можно получить явную формулу для $\hat{\boldsymbol{\theta}}(\mathbf{X})$

**Лемма (1)**
Оценка наименьших квадратов имеет следующий вид:

$$\hat{\boldsymbol{\theta}}(\mathbf{X}) = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X}.$$

**Замечание**
Заметим, что $\mathbf{Z}^{\mathrm{T}} \in \mathrm{Mat}(k \times k)$ является обратимой матрицей, так как столбцы $\mathbf{Z}$ являются базисными.

**Доказательство.**
Как было сказано выше, $\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$ есть проекция $\mathbf{X}$ на $L^{\mathrm{T}}$. Следовательно, для любого $\boldsymbol{\theta} \in \mathbb{R}^k$ вектор $\mathbf{Z}\boldsymbol{\theta}$ ортогонален $\mathbf{X}$-$\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$:

$$\langle \mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X}), \mathbf{Z}\boldsymbol{\theta} \rangle = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}(\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})) = 0$$

Но второе равенство выполнено для всех $\boldsymbol{\theta} \in \mathbb{R}^k$ тогда и только тогда, когда $\mathbf{Z}^{\mathrm{T}}(\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})) = 0$. Следовательно,

$$\mathbf{Z}^{\mathrm{T}}\mathbf{X} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X}) \Longrightarrow \hat{\boldsymbol{\theta}}(\mathbf{X}) = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X}.$$

■

## Свойства оценки наименьших квадратов

Попытаемся понять, какие свойства могут быть у оценок методом наименьших квадратов. Начнем с самого простого: посчитаем среднее и дисперсию.

**Лемма (2)**
Для оценки наименьших квадратов $\hat{\boldsymbol{\theta}}(\mathbf{X})$ выполнено следующее:

$$\mathrm{E}\hat{\boldsymbol{\theta}}(\mathbf{X}) = \boldsymbol{\theta},\ \mathrm{D}\hat{\boldsymbol{\theta}}(\mathbf{X}) = \sigma^2(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}.$$

**Доказательство.**
Для начала заметим, что матрица $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$ симметрична. Далее, $\mathbf{X} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, поэтому $\mathrm{E}\mathbf{X} = \mathbf{Z}\boldsymbol{\theta}$ и $\mathrm{D}\mathbf{X} = \sigma^2\mathbf{I}_n$. Тогда

$$\mathrm{E}\hat{\boldsymbol{\theta}}(\mathbf{X}) = \mathrm{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathrm{E}\mathbf{X} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\theta} = \boldsymbol{\theta},$$

$$\mathrm{D}\hat{\boldsymbol{\theta}}(\mathbf{X}) = \mathrm{D}[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X}] = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathrm{D}\mathbf{X} \cdot \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1} = \sigma^{-1}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}.$$

■

**Лемма (3)**
Оценка методом наименьших квадратов $\hat{\boldsymbol{\theta}}(\mathbf{X})$ является оптимальной оценкой $\boldsymbol{\theta}$ в классе линейных несмещенных оценок, то есть оценок вида $\mathbf{B}\mathbf{X}$, где $\mathbf{B}$ - неслучайная матрица.
**Доказательство.**
Пусть $\boldsymbol{\theta}^*(\mathbf{X}) = \mathbf{B}\mathbf{X}$ - другая несмещенная оценка $\boldsymbol{\theta}$. Тогда для любого $\boldsymbol{\theta} \in \mathbb{R}^k$

$$\boldsymbol{\theta} = \mathrm{E}_{\boldsymbol{\theta}}\boldsymbol{\theta}^*(\mathbf{X}) = \mathbf{B}\mathrm{E}_{\boldsymbol{\theta}}\mathbf{X} = \mathbf{B}\mathbf{Z}\boldsymbol{\theta}.$$

Но $\mathbf{B} \in \mathrm{Mat}(k \times n), \mathbf{Z} \in \mathrm{Mat}(n \times k)$ и $\mathrm{BZ} \in \mathrm{Mat}(k \times k)$. Стало быть, $\mathrm{BZ} = \mathrm{I}_k$. Далее заметим, что

$$\mathrm{D}_{\boldsymbol{\theta}}\mathrm{B}\mathbf{X} = \mathbf{B}\mathrm{D}_{\boldsymbol{\theta}\mathbf{X}\mathbf{B}^{\mathrm{T}} = \sigma^2\mathbf{B}\mathbf{B}^{\mathrm{T}}}.$$

В итоге нужно доказать, что $\sigma^2\mathbf{B}\mathbf{B}^{\mathrm{T}} \geqslant \sigma^2(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}$. Для этого рассмотрим следующую дисперсию:

$$\mathrm{D}_{\boldsymbol{\theta}}[(\mathbf{B} - (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X})] = (\mathbf{B} - (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}})\sigma^2\mathbf{I}_n(\mathbf{B}^{\mathrm{T}} - \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}) =$$

$$= \sigma^2[\mathbf{B}\mathbf{B}^{\mathrm{T}} - \mathrm{BZ}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1} - (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} + (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}] =$$

$$= \sigma^2\mathbf{B}\mathbf{B}^{\mathrm{T}} - \sigma^2(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1} \geqslant 0.$$

■

### Оценка $\sigma^2$

Отлично, мы получили наилучшую линейную оценку для $\boldsymbol{\theta}$. Но еще есть неизвестный параметр $\sigma^2$, для которого можно предложить следующую хорошую оценку.

**Лемма (4)**
Пусть $\hat{\boldsymbol{\theta}}(\mathbf{X})$ - это оценка наименьших квадратов. Тогда

$$\mathrm{E}_{\boldsymbol{\theta}}\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2 = (n-k)\sigma^2.$$

Тем самым получаем, что несмещенной оценкой $\sigma^2$ является

$$\hat{\sigma}^2(\mathbf{X}) = \frac{1}{n-k}\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2.$$

**Доказательство.**
Для начала заметим, что $\mathrm{E}_{\boldsymbol{\theta}}(\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})) = 0$. Тогда

$$\mathrm{E}_{\boldsymbol{\theta}}\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2 = \sum_{i=1}^{n}\mathrm{D}_{\boldsymbol{\theta}}(X_i - (\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X}))_i) = \mathrm{tr}\ \mathrm{D}_{\boldsymbol{\theta}}(\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})) =$$

$$= \mathrm{tr}\ \mathrm{D}_{\boldsymbol{\theta}}(\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}})\mathbf{X} = \sigma^2\mathrm{tr}\ (\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}})(\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}) =$$

$$= \sigma^2\mathrm{tr}(\mathbf{I}_n - 2\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}} + \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}) =$$

$$= \sigma^2\mathrm{tr}(\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}) = \sigma^2(n - \mathrm{tr}(\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}})) =$$

$$= \sigma^2(n - \mathrm{tr}((\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z})) = \sigma^2(n-k).$$

■

# Гауссовская линейная модель

Гауссовская линейная модель - это линейная регрессионная модель

$$\mathbf{X} = \boldsymbol{l} + \boldsymbol{\varepsilon},$$

в которой

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n) \Longrightarrow \mathbf{X} \sim \mathcal{N}(l, \sigma^2 \mathbf{I}_n) = \mathcal{N}(\mathbf{Z}\boldsymbol{\theta}, \sigma^2 \mathbf{I}_n).$$

Неизвестные параметры: $\boldsymbol{\theta} \in \mathbb{R}^k, \sigma > 0$.
Оказывается, подобное распределение $\boldsymbol{\varepsilon}$ позволяет решить сразу несколько задач:

- найти оптимальные оценки параметров;
- вычислить их распределения;
- построить точные доверительные интервалы и области.

## Достаточные статистики

Пусть $\hat{\boldsymbol{\theta}}(\mathbf{X})$ - оценка наименьших квадратов.

**Теорема 14.1.** В гауссовской линейной регрессионной модели пара $(\hat{\boldsymbol{\theta}}(\mathbf{X}), \|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2)$ является полной достаточной статистикой.

**Доказательство.**
Найдем плотность $\mathbf{X}$. Вспомним, что $\mathbf{X} \sim \mathcal{N}(l, \sigma^2 \mathbf{I}_n)$, то есть компоненты $\mathbf{X}$ независимы. Тогда

$$p(\mathbf{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(X_i - l_i)^2\right\} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{X} - l\|^2\right\} =$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{X} - \mathbf{Z}\boldsymbol{\theta}\|^2\right\} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\|\mathbf{X}\|^2 + \|\mathbf{Z}\boldsymbol{\theta}\|^2 - 2\boldsymbol{\theta}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{X})\right\} =$$

$$= \exp\left\{-\frac{\|\mathbf{X}\|^2}{2\sigma^2} + \sum_{i=1}^{k} \frac{\theta_i}{\sigma^2}(\mathbf{Z}^{\mathrm{T}}\mathbf{X})_i - \frac{\|\mathbf{Z}\boldsymbol{\theta}\|^2}{2\sigma^2} - \frac{n}{2} \cdot \ln(2\pi\sigma^2)\right\}.$$

При перебегании $\boldsymbol{\theta} \in \mathbb{R}^k, \sigma > 0$ вектор $(\frac{\theta_1}{\sigma^2}, \ldots, \frac{\theta_k}{\sigma^2}, -\frac{1}{2\sigma^2})$ зачертит полупространство $\mathbb{R}^{k+1}$. По теореме об экспоненциальном семействе вектор $(\mathbf{Z}^{\mathrm{T}}\mathbf{X}, \|\mathbf{X}\|^2)$ является полной достаточной статистикой.

Построим биекцию между $(\mathbf{Z}^{\mathrm{T}}\mathbf{X}, \|X\|^2)$ и $(\hat{\boldsymbol{\theta}}(\mathbf{X}), \|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2)$. Заметим, что по теореме Пифагора

$$\|\mathbf{X}\|^2 = \|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2 + \|\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2,$$

так как $\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$ лежит в $L^{\perp}$, а $\|\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2$ - в $L$. Далее, согласно формуле для оценки методом наименьших квадратов

$$\mathbf{Z}^{\mathrm{T}}\mathbf{X} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})\hat{\boldsymbol{\theta}}(\mathbf{X}).$$

Тем самым, если мы знаем значения пары $(\hat{\boldsymbol{\theta}}(\mathbf{X}), \|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2)$, то мы знаем значения пары $(\mathbf{Z}^{\mathrm{T}}\mathbf{X}, \|\mathbf{X}\|^2)$ (так как Z тоже известна).
В обратную сторону все аналогично:

$$\hat{\boldsymbol{\theta}}(\mathbf{X}) = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X},$$

$$\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2 = \|\mathbf{X}\|^2 - \|\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2.$$

Тем самым получаем полную биективную измеримую связь между парами. Следовательно, $(\hat{\boldsymbol{\theta}}(\mathbf{X}), \|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2)$ тоже есть полная достаточная статистика. ■

**Следствие**
В линейной гауссовской модели:

1) $\hat{\boldsymbol{\theta}}(\mathbf{X})$ есть оптимальная оценка $\boldsymbol{\theta}$,

2) $\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$ есть оптимальная оценка $\boldsymbol{l}$,

3) $\dfrac{1}{n-k}\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2$ есть оптимальная оценка $\sigma^2$.

**Доказательство.**
Ранее было показано, что эти величины являются несмещенными оценками для соответствующих параметров. Но они все являются функциями от полной и достаточной статистики. Следовательно, они оптимальны. ■

## Доверительные интервалы и области

Теперь мы бы хотели построить доверительные интервалы и области для параметров линейной регрессии. Для этого нам понадобится теорема об ортогональных разложениях, но перед этим вспомним, что такое распределение хи-квадрат.

**Определение 14.2.** Будем говорить, что случайная величина $\xi$ имеет распределение хи-квадрат с $n$ степенями свободы, если ее плотность равна

$$p_\xi(x) = \frac{x^{\frac{n}{2-1}} e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\mathbf{I}\{x \geqslant 0\},$$

то есть $\xi \sim \Gamma(\frac{1}{2}, \frac{n}{2})$. Обозначение: $\xi \sim \chi^2_n$.

**Упражнение**
Если $\xi_1, \ldots, \xi_n$ есть независимые одинаково распределенные $\mathcal{N}(0,1)$ случайные величины, то $\xi_1^2 + \cdots + \xi_n^2 \sim \chi^2_n$.

**Теорема 14.2. (об ортогональных разложениях)** Пусть $\mathbf{X} \sim \mathcal{N}(\boldsymbol{l}, \sigma^2\mathbf{I}_n)$, а $L_1 \oplus L_2 \oplus \cdots \oplus L_r$ есть разложение $\mathbb{R}^n$ в прямую сумму линейных ортогональных подпространств. Далее, пусть $\mathbf{Y}_i$ есть проекция $\mathbf{X}$ на $L_i$. Тогда $\mathbf{Y}_1, \ldots, \mathbf{Y}_r$ независимы в совокупности и

$$\frac{1}{\sigma^2}\|\mathbf{Y}_i - \mathrm{E}\mathbf{Y}_i\|^2 \sim \chi^2_{\dim L_i}.$$

**Доказательство.**
Идея состоит в том, чтобы свести эту теорему к факту, сформулированному выше. Возьмем в $\mathbb{R}^n$ такой ортонормированный базис $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$, что

$$\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{k_1} \text{ - базис } L_1,$$

$$\boldsymbol{f}_{k_1+1}, \ldots, \boldsymbol{f}_{k_1+k_2} \text{ - базис } L_2,$$

$$\vdots$$

$$\boldsymbol{f}_{k_1+\cdots+k_{r-1}+1}, \ldots, \boldsymbol{f}_n \text{ - базис } L_r.$$

В таком случае размерность $L_i$ равна $k_i$ для всех $i = 1, \ldots, r$. Далее, введем коэффициенты $W_i = \langle \mathbf{X}, \boldsymbol{f}_i \rangle$ для всех $i = 1, \ldots, n$. Соберем эти коэффициенты в вектор и представим его в виде линейного преобразования $\mathbf{X}$:

$$\boldsymbol{W} = \begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix} = \mathbf{C}\mathbf{X}, \text{ где } \mathbf{C} = \begin{pmatrix} \boldsymbol{f}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{f}_n^{\mathrm{T}} \end{pmatrix}.$$

Заметим, что матрица $\mathbf{C}$ ортогональна, так как ее строки задают ортонормированный базис $\mathbb{R}^n$. Вектор $\mathbf{X}$ является гауссовским, значит, и $\boldsymbol{W}$ тоже будет гауссовским, так как он получается линейным преобразованием $\mathbf{X}$.
Найдем распределение $W$:

$$\mathrm{E}\boldsymbol{W} = \mathrm{E}\mathbf{C}\mathbf{X} = \mathbf{C}\boldsymbol{l};$$

$$\mathrm{D}\boldsymbol{W} = \mathbf{C} \cdot \mathrm{D}\mathbf{X} \cdot \mathbf{C}^{\mathrm{T}} = \sigma^2\mathbf{C}\mathbf{C}^{\mathrm{T}} = \sigma^2\mathbf{I}_n.$$

Следовательно, $\boldsymbol{W} \sim \mathcal{N}(\mathbf{C}\boldsymbol{l}, \sigma^2\mathbf{I}_n)$ и, значит, компоненты $\boldsymbol{W}$ независимы в совокупности.

Далее, случайные векторы $\mathbf{Y}_j$, которые вычисляются следующим образом:

$$\mathbf{Y}_j = W_{k_1+\cdots+k_{j-1}+1}\boldsymbol{f}_{k_1+\cdots+k_{j-1}+1} + \cdots + W_{k_1+\cdots+K_j}\boldsymbol{f}_{k_1+\cdots+k_j},$$

тоже будут независимы в совокупности.

Теперь найдем искомые распределения их длин. Для этого заметим, что для всех $i = 1, \ldots, n$

$$\frac{W_i - \mathrm{E}W_i}{\sigma} \sim \mathcal{N}(0, 1).$$

В таком случае, в силу ортонормированности базиса $\boldsymbol{f}_j$ получаем, что

$$\frac{1}{\sigma^2}\|\mathbf{Y}_j - \mathrm{E}\mathbf{Y}_j\|^2 = \left\|\sum_{i=1}^{k_j} \frac{W_{k_1+\cdots+k_{j-1}+i} - \mathrm{E}W_{k_1+\cdots+k_{j-1}+i}}{\sigma}\boldsymbol{f}_{k_1+\cdots+k_{j-1}+i}\right\|^2 =$$

$$= \sum_{i=1}^{k_j}\left(\frac{W_{k_1+\cdots+k_{j-1}+i} - \mathrm{E}W_{k_1+\cdots+k_{j-1}+i}}{\sigma}\right)^2 \sim \chi^2_{k_j}.$$

■

### Распределение статистик

Сформулируем очень важное следствие:

**Следствие**

В гауссовской линейной модели оценка $\hat{\boldsymbol{\theta}}(\mathbf{X})$ независима с $\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$ и

$$\frac{1}{\sigma^2}\|\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \mathbf{Z}\theta\|^2 \sim \chi^2_k,$$

$$\frac{1}{\sigma^2}\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2 \sim \chi^2_{n-k}.$$

**Доказательство.**

Мы знаем, что $\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$ есть проекция $\mathbf{X}$ на $L$, а $\mathbf{X}$-$\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$ проецирует $\mathbf{X}$ на $L^\perp$. Согласно теореме об ортогональных разложениях $\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$ и $\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$ независимы. В силу того, что $\mathrm{E}(\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})) = \mathbf{0}, \mathrm{E}\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X}) = \mathbf{Z}\boldsymbol{\theta}$, мы автоматически получаем и утверждения о распределениях.

Осталось заметить, что $\hat{\boldsymbol{\theta}}(\mathbf{X})$ есть функция от $\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$ :
$\hat{\boldsymbol{\theta}}(\mathbf{X}) = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}(\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X}))$. Значит, $\hat{\boldsymbol{\theta}}(\mathbf{X})$ независима с $\mathbf{X}$-$\mathbf{Z}(\mathbf{X})$. ■

### 1. Доверительный интервал для $\sigma^2$

Воспользуемся методом центральной статистики. Заметим, что в данном случае можно использовать $\sigma^{-2}\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2$ в качестве центральной статистики, так как

ранее было доказано, что эта величина имеет распределение хи-квадрат с $n-k$ степенями свободы.
Пусть $c_\alpha$ - это $\alpha$-квантиль распределения $\chi^2_{n-k}$. Тогда

$$\mathrm{P}\left(c_{\frac{1-\gamma}{2}} < \frac{1}{\sigma^2}\|\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2\right) < c_{\frac{1+\gamma}{2}} = \gamma.$$

Следовательно, точный доверительный интервал для $\sigma^2$ уровня доверия $\gamma$ имеет вид:

$$\left(\frac{1}{c_{\frac{1+\gamma}{2}}}\|\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2, \frac{1}{c_{\frac{1-\gamma}{2}}}\|\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2\right).$$

## 2. Доверительный интервал для $\theta_i$

**Определение 14.3.** Пусть с.в. $\xi$ и $\eta$ независимы, $\xi \sim \mathcal{N}(0,1), \eta \sim \xi^2_k$. Тогда случайная величина

$$\zeta = \frac{\xi}{\sqrt{\frac{\eta}{k}}}$$

имеет распределение Стьюдента с $k$ степенями свободы (T-распределение). Обозначение: $\zeta \sim \mathrm{T}_k$.

Построим доверительный интервал для $\theta_i$. Для этого заметим, что $\hat{\boldsymbol{\theta}}(\mathbf{X}) \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathrm{A})$, где $\mathbf{A}= (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})$. Следовательно,

$$\frac{\hat{\theta}_i - \theta_i}{\sqrt{\sigma^2 a_{ii}}} \sim \mathcal{N}(0,1).$$

Далее, $\hat{\boldsymbol{\theta}}_i$ не зависит от $\|\mathbf{X}-\mathbf{Z}\|$, а $\sigma^{-2}\|\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2 \sim \chi^2_{n-k}$. Значит, величина

$$\sqrt{\frac{n-k}{a_{ii}}}\frac{\hat{\theta}_i - \theta_i}{\|\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|} \sim \mathrm{T}_{n-k}$$

имеет распределение Стьюдента $\mathrm{T}_{n-k}$.
Пусть $t_\alpha$ есть $\alpha$-квантиль распределения Стьюдента $\mathrm{T}_{n-k}$. Так как оно симметрично относительно нуля, то можно сразу записать точный доверительный интервал с уровнем доверия $\gamma$ в следующем виде:

$$\theta_i \in \left(\hat{\theta}_i - t_{\frac{1+\gamma}{2}}\sqrt{\frac{a_{ii}}{n-k}}\|\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|, \hat{\theta}_i + t_{\frac{1+\gamma}{2}}\sqrt{\frac{a_{ii}}{n-k}}\|\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|\right).$$

### 3. Доверительный интервал для $\boldsymbol{\theta}$

Осталось построить доверительную область для $\boldsymbol{\theta}$. Здесь возникает третье стандартное распределение, связанное с гауссовским - распределение Фишера.

**Определение 14.4.** Пусть $\xi, \eta$ - независимые случайные величины, причем $\xi \sim \chi^2_k, \eta \sim \chi^2_n$. Тогда случайная величина $\zeta = \frac{\xi/k}{\eta/n}$ имеет распределение Фишера (распределение Снедекора, F-распределение) с $(k, n)$ степенями свободы. Обозначение $\zeta \sim \mathrm{F}_{k,n}$.

Согласно следствию из теоремы об ортогональных разложенях мы знаем, что

$$\frac{1}{\sigma^2}\|\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \mathbf{Z}\boldsymbol{\theta}\|^2 \sim \chi^2_k,$$

$$\frac{1}{\sigma^2}\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2 \sim \chi^2_{n-k},$$

и эти случайные величины независимы.
Значит,

$$\frac{\|\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \mathbf{Z}\boldsymbol{\theta}\|^2}{\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2}\frac{n-k}{k} \sim \mathrm{F}_{k,n-k}.$$

Далее, пусть $f_\alpha$ есть $\alpha$-квантиль распределения $\mathrm{F}_{k,n-k}$. Так как оно неотрицательно, то доверительной областью для всего вектора $\boldsymbol{\theta}$ будет

$$S(\mathbf{X}) = \left\{\boldsymbol{\theta} \in \mathbb{R}^k : \frac{\|\mathbf{Z}\hat{\theta}(\mathbf{X}) - \mathbf{Z}\boldsymbol{\theta}\|^2}{\|\mathbf{X} - \mathbf{Z}\hat{\theta}(\mathbf{X})\|^2}\frac{n-k}{k} < f_\gamma\right\}.$$

Стоит заметить, что геометрически доверительная область будет иметь форму эллипсоида в $\mathbb{R}^k$.

**Упражнение**
Представить выборку из нормального распределения $\mathcal{N}(a, \sigma^2)$ в виде гауссовской линейной модели и построить доверительные интервалы для параметров с помощью общей теории.

# 15. Лекция 15

На этой лекции мы начнем разбирать последнюю большую тему нашего курса - проверку статистических гипотез. Разберемся сначала с постановкой проблемы.

Пусть $\mathbf{X}$ - наблюдение с неизвестным распределением $\mathrm{P} \in \mathcal{P}$, где $\mathcal{P}$ есть некоторое семейство распределений.

Далее, возьмем подсемейство $\mathcal{P}_0 \subset \mathcal{P}$. Поставим следующий вопрос: верно ли то, что $\mathrm{P} \in \mathcal{P}_0$? Ответ на этот вопрос позволяет сузить семейство распределений, что упростит дальнейший анализ.

Если ответ положительный, то $\mathcal{P}$ можно сузить до $\mathcal{P}_0$. Если же ответ отрицательный, то $\mathcal{P}$ сужается до $\mathcal{P} \setminus \mathcal{P}_0$. В общем случае может быть выдвинуто сразу несколько гипотез, из которых нам необходимо выбрать одну, чтобы "локализовать" истинное распределение.

## Гипотезы

**Определение 15.1.** Статистической гипотезой называется предположение вида

$$\mathrm{H}_0 : \mathrm{P} \in \mathcal{P}_0,$$

где $\mathcal{P}_0 \subset \mathcal{P}$ - подсемейство распределений.

**Задача**
По наблюдению $\mathbf{X}$ либо принять гипотезу (тогда $\mathcal{P}$ заменяется на $\mathcal{P}_0$), либо отклонить (тогда $\mathcal{P}$ заменяется на $\mathcal{P} \setminus \mathcal{P}_0$).

Во втором случае мы можем выдвинуть альтернативную гипотезу $\mathrm{H}_1 : \mathrm{P} \in \mathcal{P}_1$, где $\mathcal{P}_1 \subseteq \mathcal{P} \setminus \mathcal{P}_0$. Соответственно, если у нас есть альтернатива, то мы должны проверять гипотезу $\mathrm{H}_0$ против альтернативы $\mathrm{H}_1$. Альтернатив может быть несколько:

$$\begin{aligned}
&\mathrm{H}_0 : \mathrm{P} \in \mathcal{P}_0,\\
&\mathrm{H}_1 : \mathrm{P} \in \mathcal{P}_1,\\
&\vdots\\
&\mathrm{H}_k : \mathrm{P} \in \mathcal{P}_k.
\end{aligned}$$

# Статистические критерии

Методология проверки гипотез следующая.

**Определение 15.2.** Пусть $\mathcal{X}$ - выборочное пространство, то есть множество всех возможных значений наблюдений. Тогда подмножество $S \subset \mathcal{X}$ называется критерием или критическим множеством для проверки гипотезы $\mathrm{H}_0$ (против альтернативы $\mathrm{H}_1$), если правило принятия $\mathrm{H}_0$ выглядит так:

$$\mathrm{H}_0 \text{ отвергается тогда и только тогда, когда } \mathbf{X} \in S.$$

Тем самым, $S$ является "плохим" множеством, и, если $\mathbf{X}$ попало в него, то $\mathrm{H}_0$ признается неверной и выбирается альтернатива.

**Замечание**
Множество $\mathcal{X} \setminus S$ принято называть областью принятия гипотезы.

Естественно, что при принятии решений возможны ошибки. Например, мы можем промахнуться следующим образом: гипотезу $\mathrm{H}_0$ приняли, хотя, на самом деле, она была не верна. Принято различать ошибки первого и второго рода.

### Ошибки первого и второго рода

**Определение 15.3.** Пусть проверяется гипотеза $\mathrm{H}_0$. Ошибкой первого рода называется ситуация, когда $\mathrm{H}_0$ отвергли, но она была верна. Аналогично, ошибкой второго рода называют ситуацию, когда $\mathrm{H}_0$ приняли, но она была неверна.

**Вопрос:** какая из этих ошибок опаснее?

**Методологически считается, что ошибка первого рода опаснее.**
Идея состоит в том, что мы считаем, что при выдвижении гипотезы $\mathrm{H}_0$ мы выбрали наше подсемейство $\mathcal{P}_0$ из тех соображений, что мы не хотим совершать ошибку первого рода. Предположим, что мы неверно отвергли $\mathrm{H}_0$. Тогда мы вынуждены рассматривать некоторую альтернативу $\mathrm{H}_1$, которая заведомо неверна. Последнее приведет к тому, что мы уйдем далеко от правильного ответа, будет проводиться много бесполезной работы.

В случае совершения ошибки второго рода мы принимаем $\mathrm{H}_0$, но она неверна. В таком случае вердикт может измениться, если добавятся новые данные, и мы сможем двинуться в сторону чего-то более правильного. Другими словами, здесь шансы совершить значительное количество лишней работы заметно меньше.

На этой методологии основан принцип сравнения критериев.

## Функция мощности

**Задача**
Нужно предложить статистический критерий с небольшими вероятностями ошибок как первого, так и второго рода.

**Определение 15.4.** Пусть $S$ - критерий для проверки гипотезы $\mathrm{H}_0$. Функцией мощности критерия $S$ называется

$$\beta(\mathrm{Q}, S) = \mathrm{Q}(\mathbf{X} \in S),\ \mathrm{Q} \in \mathcal{P}.$$

Возникает интересный момент: вероятность ошибки не является одной вероятностью, а является набором вероятностей.

**Наблюдение**
Вероятностями ошибок первого рода называется набор $\{\beta(\mathrm{Q}, S) : \mathrm{Q} \in \mathcal{P}_0\}$.
Вероятностями ошибок второго рода называется набор $\{1 - \beta(\mathrm{Q}, S) : \mathrm{Q} \in \mathcal{P}_1\}$ (или $\mathrm{Q} \in \mathcal{P} \setminus \mathcal{P}_0$).

Соответственно, здесь мы получаем следующую постановку: мы хотим критерий, который имел бы маленькие значения функции мощности на основной гипотезе и большие на альтернативе. В идеале, хочется получить нули, но эта задача безнадежна, и нужно на что-то соглашаться. Но на что? И как определить, что один критерий лучше другого? Обсудим это.

## Уровень значимости

Разумно рассматривать только критерии, у которых вероятности ошибки первого рода равномерно ограничены сверху каким-то $\varepsilon$, раз уж мы решили, что ошибка первого рода опаснее.

**Определение 15.5.** Пусть $S \subset \mathcal{X}$ - это критерий для проверки гипотезы $\mathrm{H}_0 : \mathrm{P} \in \mathcal{P}_0$. Будем говорить, что критерий $S$ имеет уровень значимости $\varepsilon$, если для любого $\mathrm{Q} \in \mathcal{P}_0$ выполнено

$$\beta(\mathrm{Q}, S) \leqslant \varepsilon.$$

По сути, если $\varepsilon$ является подходящим уровнем значимости, то и любое число, большее $\varepsilon$, тоже подходит. Поэтому, наряду с уровнем значимости вводится минимальный уровень значимости:

**Определение 15.6.** Минимальным уровнем значимости или размером критерия называют

$$\alpha(S) = \sup_{\mathrm{Q} \in \mathcal{P}_0} \beta(\mathrm{Q}, S).$$

### Несмещенность и состоятельность оценки

Опираясь на функцию мощности можно ввести два весьма полезных свойства.

**Определение 15.7.** Критерий $S$ для проверки гипотезы $\mathrm{H}_0 : \mathrm{P} \in \mathcal{P}_0$ против альтернативы $\mathrm{H}_1 : \mathrm{P} \in \mathcal{P}_1$ называется несмещенным, если

$$\sup_{\mathrm{Q} \in \mathcal{P}_0} \beta(\mathrm{Q}, S) \leqslant \inf_{\mathrm{Q} \in \mathcal{P}_1} \beta(\mathrm{Q}, S).$$

Данное свойство достаточно естественно: мы хотим, чтобы на $\mathcal{P}_0$ значение функции мощности было маленьким, а на $\mathcal{P}_1$ большим.

**Определение 15.8.** Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка растущего размера. Тогда последовательность критериев $\{S_n, n \in \mathbb{N}\}$, $S_n \subset \mathcal{X}^n$, для проверки гипотезы $\mathrm{H}_0 : \mathrm{P} \in \mathcal{P}_0$ против альтернативы $\mathrm{H}_1 : \mathrm{P} \in \mathcal{P}_1$ называется состоятельной, если для любого $\mathrm{Q} \in \mathcal{P}_1$ выполнено

$$\beta(\mathrm{Q}, S_n) \to 1 \text{ при } n \to \infty.$$

**Смысл:** вероятности ошибки второго рода стремятся к нулю.

## Сравнение критериев

Обсудим теперь сравнение критериев. Пусть есть два критерия $S$ и $R$. Как можно сравнить их и понять, какой лучше?

**Важный принцип.** Мы должны выбрать некоторый уровень значимости $\varepsilon$, вероятность ошибки первого рода, на которую мы согласились, и сравнивать только критерии этого уровня значимости.

**Замечание**
Однако, "загонять" $\varepsilon$ в ноль неосмысленно, так как это лишь увеличит вероятность ошибки второго рода. Здесь нужно искать баланс.

**Определение 15.9.** Пусть $S$ и $R$ - критерии уровня значимости $\varepsilon$ для проверки гипотезы $\mathrm{H}_0 : \mathrm{P} \in \mathcal{P}_0$ против альтернативы $\mathrm{H}_1 : \mathrm{P} \in \mathcal{P}_1$. Будем говорить, что $S$ (равномерно) мощнее $R$, если для любого $\mathrm{Q} \in \mathcal{P}_1$ выполнено

$$\beta(\mathrm{Q}, S) \geqslant \beta(\mathrm{Q}, R),$$

то есть вероятность ошибки второго рода у критерия $S$ равномерно меньше, чем у $R$.

### Равномерно наиболее мощные критерии

Если же критерий оказывается мощнее всех остальных, то мы его будем называть равномерно наиболее мощным.

**Определение 15.10.** Критерий $S$ для проверки гипотезы $\mathrm{H}_0 : \mathrm{P} \in \mathcal{P}_0$ против альтернативы $\mathrm{H}_1\mathrm{P} \in \mathcal{P}_1$ называется равномерно наиболее мощным критерием (р.н.м.к.) уровня значимости $\varepsilon$, если $\alpha(S) \leqslant \varepsilon$ и $S$ мощнее любого другого критерия $R$ уровня значимости $\varepsilon$.

**Вопрос:** как строить р.н.м.к.?

Оказывается в случае простых гипотез есть явный ответ.

**Определение 15.11.** Гипотеза вида $\mathrm{H}_0 : \mathrm{P} = \mathrm{P}_0$, где $\mathrm{P}_0$ - некоторое известное распределение, называется простой.

## Лемма Неймана-Пирсона

Рассмотрим проверку простой гипотезы $\mathrm{H}_0 : \mathrm{P} = \mathrm{P}_0$ против простой альтернативы $\mathrm{H}_1 : \mathrm{P} = \mathrm{P}_1$. Далее, предположим, что $\mathrm{P}_0$ и $\mathrm{P}_1$ имеют плотности $p_0(x)$ и $p_1(x)$ по одной и той же мере $\mu$. По сути, оба распределения должны быть взяты из одного доминируемого семейства.

**Замечание**
Например, различить пуассоновское и экспоненциальное распределения крайне просто - считаем, что экспоненциальное. если получили нецелое число, и пуассоновское - иначе. Такой критерий никогда не ошибается.

Для $\lambda > 0$ рассмотрим критерий $S_\lambda$ :

$$S_\lambda = \{x \in \mathcal{X} : p_1(x) - \lambda p_0(x) \geqslant 0\}.$$

Оказывается, такой критерий обладает интересными свойствами.

**Лемма (Нейман-Пирсон)**
Пусть критерий $R$ таков, что $\mathrm{P}_0(\mathbf{X} \in R) \leqslant \mathrm{P}_0(\mathbf{X} \in S_\lambda)$. Тогда

1) $\mathrm{P}_1(\mathbf{X} \in R) \leqslant \mathrm{P}_1(\mathbf{X} \in S_\lambda)$,

2) $\mathrm{P}_0(\mathbf{X} \in S_\lambda) \leqslant \mathrm{P}_1(\mathbf{X} \in S_\lambda)$.

**Доказательство.**

1) Для начала заметим, что для любого $x \in \mathcal{X}$

$$\mathrm{I}_R(x)(p_1(x) - \lambda p_0(x)) \leqslant \mathrm{I}_{S_\lambda}(x)(p_1(x) - \lambda p_0(x)).$$

Это верно для всех $x$: предположим, что $x$ попадает в $S_\lambda$, тогда $\mathrm{I}_{S_\lambda} = 1$, $\mathrm{I}_R \leqslant 1$, а $p_1(x) - \lambda p_0(x) \geqslant 0$. Если же $x$ не попадает в $S_\lambda$, то $\mathrm{I}_{S_\lambda} = 0$, $\mathrm{I}_R \geqslant 0$, а $p_1(x) - \lambda p_0(x) \leqslant 0$.

Проинтегрируем неравенства по мере $\mu$ по всему выборочному пространству:

$$\int\limits_R (p_1(x) - \lambda p_0(x))\mu(dx) \leqslant \int\limits_{S_\lambda} (p_1(x) - \lambda p_0(x))\mu(dx).$$

Раскроем интегралы по линейности, вынесем константы, тогда:

$$\mathrm{P}_1(\mathbf{X} \in R) - \lambda \cdot \mathrm{P}_0(\mathbf{X} \in R) \leqslant \mathrm{P}_1(\mathbf{X} \in S_\lambda) - \lambda \cdot \mathrm{P}_0(\mathbf{X} \in S_\lambda).$$

Следовательно,

$$\mathrm{P}_1(\mathbf{X} \in R) - \mathrm{P}_1(\mathbf{X} \in S_\lambda) \leqslant \lambda \cdot (\mathrm{P}_0(\mathbf{X} \in R) - \mathrm{P}_0(\mathbf{X} \in S_\lambda)) \leqslant 0.$$

Первое неравенство доказано.

2) Проверим, что для любого $\lambda > 0$ выполнено $\mathrm{P}_0(\mathbf{X} \in S_\lambda) \leqslant \mathrm{P}_1(\mathbf{X} \in S_\lambda)$. Рассмотрим два случая: $\lambda \geqslant 1$ и $\lambda < 1$.

- Пусть $\lambda \geqslant 1$. Тогда для $x \in S_\lambda$ выполнено следующее неравенство:

$$p_1(x) \geqslant \lambda p_0(x) \geqslant p_0(x).$$

Следовательно,

$$\mathrm{P}_0(\mathbf{X} \in S_\lambda) = \int\limits_{S_\lambda} p_0(x)\mu(dx) \leqslant \int\limits_{S_\lambda} p_1(x)\mu(dx) = \mathrm{P}_1(\mathbf{X} \in S_\lambda).$$

- Пусть $\lambda \in (0, 1)$. Тогда для $x \notin S_\lambda$ выполнено

$$p_1(x) < \lambda p_0(x) < p_0(x).$$

Следовательно,

$$\mathrm{P}_1(\mathbf{X} \in \overline{S_\lambda}) = \int\limits_{\overline{S_\lambda}} p_1(x)\mu(dx) \leqslant \int\limits_{\overline{S_\lambda}} p_0(x)\mu(dx) = \mathrm{P}_0(\mathbf{X} \in \overline{S_\lambda}),$$

что эквивалентно $\mathrm{P}_0(\mathbf{X} \in S_\lambda) \leqslant \mathrm{P}_1(\mathbf{X} \in S_\lambda)$.

■

### Следствия

У данной леммы есть два важных следствия. Первое следствие говорит нам о том, как мы можем построить равномерно наиболее мощный критерий уровня значимости $\varepsilon$.

**Следствие (1)**
Если $\lambda > 0$ удовлетворяет соотношению $\mathrm{P}_0(\mathbf{X} \in S_\lambda) = \varepsilon$, то $S_\lambda$ является равномерно наиболее мощным критерием уровня значимости $\varepsilon$ для проверки гипотезы $\mathrm{H}_0 : \mathrm{P} = \mathrm{P}_0$ против альтернативы $\mathrm{H}_1 : \mathrm{P} = \mathrm{P}_1$.

**Доказательство.**
Если $R$ - любой другой критерий уровня значимости $\varepsilon > 0$, то
$\mathrm{P}_0(\mathbf{X} \in R) \leqslant \varepsilon = \mathrm{P}_0(\mathbf{X} \in S_\lambda)$. По первому утверждению леммы Неймана-Пирсона получаем, что $\mathrm{P}_1(\mathbf{X} \in R) \leqslant \mathrm{P}_1(\mathbf{X} \in S_\lambda)$, то есть $S_\lambda$ мощнее $R$. ■

**Следствие (2)**
Критерий $S_\lambda$ несмещенный.

**Доказательство.** Это второе утверждение леммы Неймана-Пирсона. ■

Тем самым мы получили способ, с помощью которого можно построить р.н.м.к. для проверки простой гипотезы против простой альтернативы.

Из следствия (1) можно сделать вывод о том, что для нахождения р.н.м.к. необходимо решить (относительно $\lambda$) уравнение

$$\mathrm{P}_0(\mathbf{X} \in S_\lambda) = \int\limits_{S_\lambda} p_0(x)\mu(Dx) = \varepsilon,$$

где $S_\lambda = \{x \in \mathcal{X} : p_1(x) - \lambda p_0(x) \geqslant 0\}$. Оно почти всегда разрешимо в случае абсолютно непрерывных распределений. В дискретном случае, наоборот, оно для почти всех $\varepsilon$ неразрешимо. В этом случае стоит изменить $\varepsilon$ так, чтобы оно стало разрешимо.

**Комментарий**
В общей ситуации можно ввести рандомизированные критерии. Мы не будем подробно на них останавливаться, лишь кратко опишем идею. Для каждого значения наблюдения мы выбираем с какой вероятностью мы примем гипотезу $\mathrm{H}_0$ против альтернативы $\mathrm{H}_1$, то есть получается некоторая функция от $x$, которая принимает

значения на $[0, 1]$. Если значение не 0 или 1, то мы не можем однозначно выбрать гипотезу. Тогда мы "подбрасываем монетку" с вероятностью, равной значению функции.

Польза от такого подхода следующая. Например, у нас дискретное распределение, следовательно вероятность $\mathrm{P}_0$ как-то дискретно меняется, и для некоторого $\lambda$ вероятность меньше $\varepsilon$, а для другого $\lambda$ больше $\varepsilon$, и эти два значения отличаются на событие, что $x$ попал в какие-то дискретные точки. Соответственно, для $\lambda < \varepsilon$ мы будем предпочитать гипотезу $\mathrm{H}_1$, для всех, которые не попали во второй случай, будем предпочитать $\mathrm{H}_0$, а для тех, которые оказались в промежутке, будем пытаться подобрать числа так, чтобы вероятность $\mathrm{P}_0$ была бы в точности равна $\varepsilon$. Это станет более понятно, когда мы дойдем до примеров.

## Монотонное отношение правдоподобия

Мы построили р.н.м.к. для проверки простой гипотезы против простой альтернативы. Возникает естественный вопрос: можно ли обобщить данный результат на случай сложных гипотез?

Оказывается, это можно сделать, когда речь идет о параметрических семействах с монотонным отношением правдоподобия. В параметрической модели $\{\mathrm{P}_\theta, \theta \in \Theta\}$ гипотезы естественно записывать, как предположения о значении параметра. Например,

$$\mathrm{H}_0 : \theta \in \Theta_0.$$

Функцию мощности критерия $S$ также удобно записывать, как функцию от параметра $\beta(\theta, S) = \mathrm{P}_\theta(\mathbf{X} \in S)$.

Итак, пусть $\mathbf{X}$ - это наблюдение с неизвестным распределением P, принадлежащим параметрическому семейству $\{\mathrm{P}_\theta, \theta \in \Theta\}$, причем семейство $\{\mathrm{P}_\theta, \theta \in \Theta\}$ является доминируемым с плотностью $p_\theta(\mathbf{x})$.

**Определение 15.12.** Пусть $\{\mathrm{P}_\theta, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}$ - это доминируемое семейство распределений с плотностью $p_\theta(x)$ по мере $\mu$. Тогда это семейство имеет монотонное отношение правдоподобия по статистике $T(\mathbf{X}) \in \mathbb{R}$, если для любых $\theta_1 < \theta_2$, $\theta_1, \theta_2 \in \Theta$, отношение плотностей имеет вид

$$\frac{p_{\theta_2}(\mathbf{X})}{p_{\theta_1}(\mathbf{X})} = \psi_{\theta_1,\theta_2}(T(\mathbf{X})), \tag{1}$$

где $\psi_{\theta_1,\theta_2}(t)$ - это монотонная функция, причем ее монотонность одинакова для всех $\theta_1 < \theta_2$.

**Замечание**
В этом случае $T(\mathbf{X})$ - это достаточная статистика.

**Лемма**
Пусть $\psi_{\theta_1,\theta_2}(t)$ в (1) всегда не убывает по $T(\mathbf{X})$. Тогда для всех $c \in \mathbb{R}$ и для всех $\theta_1 < \theta_2$, $\theta_1, \theta_2 \in \Theta$ выполняется неравенство

$$\mathrm{P}_{\theta_1}(T(\mathbf{X}) \geqslant c) \leqslant \mathrm{P}_{\theta_2}(T(\mathbf{X}) \geqslant c),$$

то есть функция $\mathrm{P}_\theta(T(\mathbf{X}) \geqslant c)$ не убывает по $\theta$.

**Доказательство.**
Считаем, что $\psi_{\theta_1,\theta_2}(t) \geqslant 0$ и не убывает на $\mathbb{R}$. Введем множество $D = \{\mathbf{x} : T(\mathbf{x}) \geqslant c\}$. Если $\psi_{\theta_1,\theta_2}(c) \geqslant 1$, то для любого $\mathbf{x} \in D$ выполнено

$$p_{\theta_2}(\mathbf{x}) = \psi_{\theta_1,\theta_2}(T(\mathbf{x})) \cdot p_{\theta_1}(\mathbf{x}) \geqslant \psi_{\theta_1,\theta_2}(c) \cdot p_{\theta_1}(\mathbf{x}) \geqslant p_{\theta_1}(\mathbf{x}).$$

Интегрируем это неравенство по мере $\mu$ по множеству $D$:

$$\mathrm{P}_{\theta_2}(T(\mathbf{X}) \geqslant c) = \int\limits_D p_{\theta_2}(\mathbf{x})\mu(d\mathbf{x}) \geqslant \int\limits_D p_{\theta_1}(\mathbf{x})\mu(d\mathbf{x}) = \mathrm{P}_{\theta_1}(T(\mathbf{X}) \geqslant c).$$

Если же $c$ таково, что $\psi_{\theta_1,\theta_2}(c) \in [0,1]$, то рассматриваем множество $\overline{D}$. На нем, наоборот, $p_{\theta_2}(\mathbf{x}) \leqslant p_{\theta_1}(\mathbf{x})$ и, стало быть,

$$\mathrm{P}_{\theta_2}(T(\mathbf{X}) < c) = \int\limits_{\overline{D}} p_{\theta_2}(\mathbf{x})\mu(d\mathbf{x}) \leqslant \int\limits_{\overline{D}} p_{\theta_1}(\mathbf{x})\mu(d\mathbf{x}) = \mathrm{P}_{\theta_1}(T(\mathbf{X}) < c).$$

В итоге, получаем искомое:

$$\mathrm{P}_{\theta_1}(T(\mathbf{X}) \geqslant c) \leqslant \mathrm{P}_{\theta_2}(T(\mathbf{X}) \geqslant c).$$

■

Эта лемма лежит в основе доказательства следующей теоремы, которая позволяет нам строить р.н.м.к. в ситуации, когда мы имеем дело с параметрическими семействами с монотонным отношением правдоподобия, и в ситуации, когда мы рассматриваем "односторонние" гипотезы.

**Теорема 15.1. (о монотонном отношении правдоподобия)**
Пусть $\{\mathrm{P}_\theta, \theta \in \Theta\}$, $\theta \subseteq \mathbb{R}$ - семейство с монотонным отношением правдоподобия по

статистике $T(\mathbf{X})$, причем функция $\psi_{\theta_1,\theta_2}(t)$ в (1) не убывает и непрерывна. Пусть $\theta_0 \in \Theta$. Если $c \in \mathbb{R}$ удовлетворяет соотношению

$$\mathrm{P}_{\theta_0}(T(\mathbf{X}) \geqslant c) = \varepsilon,$$

то критерий $S = \{\mathbf{x} : T(\mathbf{x}) \geqslant c\}$ является равномерно наиболее мощным критерием уровня значимости $\varepsilon$ для проверки гипотезы $\mathrm{H}_0 : \theta \leqslant \theta_0$ против альтернативы $\mathrm{H}_1 : \theta > \theta_0$.

**Доказательство.**
Для начала возьмем какое-нибудь $\theta < \theta_0$. Тогда по лемме

$$\mathrm{P}_{\theta}(T(\mathbf{X}) \geqslant c) \leqslant P_{\theta_0}(T(\mathbf{X}) \geqslant c) = \varepsilon.$$

Из этого следует, что критерий $S$ имеет уровень значимости $\varepsilon$. Проверим теперь, что он является равномерно наиболее мощным.

Пусть $R$ - любой другой критерий с уровнем значимости $\varepsilon$. Возьмем любое $\theta_1 > \theta_0$. Нужно доказать, что

$$\mathrm{P}_{\theta_1}(\mathbf{X} \in R) \leqslant \mathrm{P}_{\theta_1}(\mathbf{X} \in S).$$

Согласно лемме Неймана-Пирсона равномерно наиболее мощный критерий уровня значимости $\varepsilon$ ля проверки гипотезы $\mathrm{H}_0 : \theta = \theta_0$ против гипотезы $\mathrm{H}_1 : \theta = \theta_1$ имеет вид

$$S_\lambda = \{\mathbf{x} : p_{\theta_1}(\mathbf{x}) - \lambda p_{\theta_0}(\mathbf{x}) \geqslant 0\}, \text{ где } \mathrm{P}_{\theta_0}(\mathbf{X} \in S_\lambda) = \varepsilon.$$

Однако в силу непрерывности и неубывания $\psi_{\theta_0,\theta_1}$ для некоторого $\tilde{\lambda}$ выполнено

$$S_\lambda = \left\{\mathbf{x} : \frac{p_{\theta_1}(\mathbf{x})}{p_{\theta_0}(\mathbf{x})} \geqslant \lambda\right\} = \{\mathbf{x} : \psi_{\theta_0,\theta_1}(T(\mathbf{x})) \geqslant \lambda\} = \{\mathbf{x} : T(\mathbf{x}) \geqslant \tilde{\lambda}\}.$$

Заметим, что $\tilde{\lambda} = e$ подходит под условие. Следовательно, для любого $\theta_1 > \theta_0$

$$\mathrm{P}_{\theta_1}(\mathbf{X} \in R) \leqslant \mathrm{P}_{\theta_1}(\mathbf{X} \in S_\lambda) = \mathrm{P}_{\theta_1}(T(\mathbf{X}) \geqslant e) = \mathrm{P}_{\theta_1}(\mathbf{X} \in S).$$

Значит, $S$ является равномерно наиболее мощным критерием для проверки гипотезы $\mathrm{H}_0 : \theta \leqslant \theta_0$ против альтернативы $\mathrm{H}_1 : \theta > \theta_0$. ∎

**Замечание (1)**
Тот же самый критерий $S$ является равномерно наиболее мощным критерием уровня значимости $\varepsilon$ для проверки гипотезы $\mathrm{H}_0 : \theta = \theta_0$ против альтернативы $\mathrm{H}_1 : \theta > \theta_0$.

**Замечание (2)**
Теорема означает невозможность построения р.н.м.к. в случае двухсторонней альтернативы, то есть для $\mathrm{H}_0 : \theta = \theta_0$ против $\mathrm{H}_1 : \theta \neq \theta_0$, когда мы имеем дело с

экспоненциальными семействами.

**Замечание (3)**
В случае убывания функции $\psi_{\theta_1,\theta_2}$ р.н.м.к. будет иметь вид $T(\mathbf{X}) \leqslant e$. Он же будет оптимальным, если основная гипотеза и альтернатива поменяются местами: $\mathrm{H}_0 : \theta \geqslant \theta_0$ против альтернативы $\mathrm{H}_1 : \theta < \theta_0$. В общем случае можно просто решать неравенство

$$p_{\theta_1}(x) - \lambda p_{\theta_0}(x) \gtrless 0,$$

где $\theta_1$ берется из альтернативы, а $\theta_0$ - из основной гипотезы.

На следующей лекции мы разберемся с тем, как можно проверять линейные гипотезы в линейной гауссовской модели, и разберем некоторые примеры.

# 16. Лекция 16

## Примеры применения теоремы о монотонном отношении правдоподобия

Начнем эту лекцию с примеров, которые не были разобраны на прошлой лекции. Также обсудим что можно сделать в ситуации, когда мы не можем честно решить уравнение на $\lambda$, которое возникает при применении монотонного отношения правдоподобия.

**Пример (1)**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из схемы Бернулли $\mathrm{Bin}(1, \theta)$. Постройте р.н.м.к. уровня значимости $\varepsilon$ для проверки гипотезы $\mathrm{H}_0 : \theta \leqslant \theta_0$ против альтернативы $\mathrm{H}_1 : \theta > \theta_0$. Проверьте этот критерий на состоятельность.

**Решение**

Выпишем функцию правдоподобия:

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

Согласно теореме о монотонном отношении правдоподобия необходимо найти отношение правдоподобия: при $\theta_1 > \theta_0$

$$\frac{p_{\theta_1}(\mathbf{X})}{p_{\theta_0}(\mathbf{X})} = \frac{\theta_1^{n\overline{\mathbf{X}}}(1-\theta_1)^{n-n\overline{\mathbf{X}}}}{\theta_0^{n\overline{\mathbf{X}}}(1-\theta_0)^{n-n\overline{\mathbf{X}}}} = \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{n\overline{\mathbf{X}}} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n}.$$

Решаем неравенство:

$$\frac{p_{\theta_1}(\mathbf{X})}{p_{\theta_0}(\mathbf{X})} = \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{n\overline{\mathbf{X}}} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n} \geqslant \lambda.$$

Так как $\theta_1 > \theta_0$, то выражение в левой части возрастает по $\sum\limits_{i=1}^{n} X_i$. Тогда неравенство выше эквивалентно тому, что

$$\sum_{i=1}^{n} X_i \geqslant c_1 = c_1(\lambda, \theta_0, \theta_1).$$

Теперь надо найти $c_1$ из уравнения

$$\mathrm{P}_{\theta_0}\left(\sum_{i=1}^{n} X_i \geqslant c_1\right) = \varepsilon.$$

В силу того, что $\sum\limits_{i=1}^{n} X_i \sim \mathrm{Bin}(n, \theta_0)$, находим, что $c_1$ - это $(1-\varepsilon)$-квантиль распределения $\mathrm{Bin}(n, \theta_0)$.

Мы нашли вид критерия и чему равняется $c_1$. Проверим теперь состоятельность критерия. Пусть $\theta_1 > \theta_0$. Нам нужно проверить, что

$$\mathrm{P}_{\theta_1}\left(\sum_{i=1}^{n} X_i \geqslant c_1\right) \to 1.$$

Мы знаем, что если $\theta_1$ - это истинное значение параметра, то $\sum\limits_{i=1}^{n} X_i \sim \mathrm{Bin}(n, \theta_1)$, но мы не понимаем, чему равно $c_1$. Перепишем условие на вероятность ошибки первого рода:

$$\varepsilon = \mathrm{P}_{\theta_0}\left(\sum_{i=1}^{n} X_i \geqslant c_1\right) = \mathrm{P}_{\theta_0}\left(\sqrt{n}\frac{(\overline{\mathbf{X}} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}} \geqslant \sqrt{n}\frac{\frac{c_1}{n} - \theta_0}{\sqrt{\theta_0(1-\theta_0)}}\right).$$

Левая часть в неравенстве по ЦПТ сходится по распределению к стандартному нормальному закону. Стало быть,

$$\sqrt{n}\frac{\frac{c_1}{n} - \theta_0}{\sqrt{\theta_0(1-\theta_0)}} \to u_{1-\varepsilon}$$

при $n \to +\infty$, где $u_{1-\varepsilon}$ - это $(1-\varepsilon)$-квантиль $\mathcal{N}(0,1)$. Отсюда получаем, что $c_1 = n\theta_0 + O(\sqrt{n})$. Следовательно,

$$\mathrm{P}_{\theta_1}\left(\sum_{i=1}^{n} X_i \geqslant c_1\right) = \mathrm{P}_{\theta_1}\left(\sqrt{n}\frac{(\overline{\mathbf{X}} - \theta_1)}{\sqrt{\theta_1(1-\theta_1)}} \geqslant \sqrt{n}\frac{\frac{c_1}{n} - \theta_1}{\sqrt{\theta_1(1-\theta_1)}}\right) \to 1,$$

так как левая часть снова сходится по распределению к $\mathcal{N}(0,1)$, а правая часть стремится к $-\infty$ в силу условия $\theta_1 > \theta_0$. Следовательно, критерий состоятелен.

На примере биномиального распределения удобно посмотреть на то, как бы работал рандомизированный критерий проверки гипотез.

У нас были гипотезы $\mathrm{H}_0 : \theta \leqslant \theta_0$ и $\mathrm{H}_1 : \theta > \theta_0$, и мы получили критерий $\sum\limits_{i=1}^{n} X_i \geqslant c_1$, где $c_1$ находится из соображений, что $\mathrm{P}_{\theta_0}\left(\sum\limits_{i=1}^{n} X_i \geqslant c_1\right) = \varepsilon$.

Далее, если $\varepsilon$ таково, что ни для какого $c_1$ мы не можем получить точное равенство, то нам следует уменьшить $\varepsilon$, чтобы уравнение стало разрешимым. Но предположим, что мы хотим решить задачу для всех $\varepsilon$. Тогда стоит действовать так: пусть $c \in \mathbb{N}$ такое, что

$$\mathrm{P}_{\theta_0}\left(\sum X_i \geqslant c+1\right) < \varepsilon < \mathrm{P}_{\theta_0}\left(\sum X_i \geqslant c\right).$$

Тогда введем функцию $\psi(X_1,\ldots,X_n) = \begin{cases} 1, & \sum X_i \geqslant c+1 \\ 0, & \sum X_i < c \\ \rho, & \sum X_i = c \end{cases}$. $\rho$ выбирается таким образом, что $\mathrm{E}_{\theta_0}\psi(X_1,\ldots,X_n) = \varepsilon$, то есть $\rho$ находится из уравнения

$$\mathrm{P}_{\theta_0}\left(\sum_{i=1}^{n} X_i \geqslant c+1\right) + \rho \mathrm{P}_{\theta_0}\left(\sum_{i=1}^{n} X_i = c\right) = \varepsilon$$

Тогда при $\rho = 1$ левая часть становится больше $\varepsilon$, а при $\rho = 0$ - меньше $\varepsilon$, поэтому найдется $\rho \in (0,1)$, для которого равенство выполняется. Соответственно, правило принятия гипотезы говорит о том, что если $\psi = 1$, то мы выбираем гипотезу $H_1$, если $\psi = 0$ - гипотезу $H_0$, а если что-то между 0 и 1, то мы "бросаем монетку" , которая с вероятностью $\rho$ предпочитает $H_1$, а с вероятностью $1-\rho$ предпочитает $H_0$. Продолжим разбирать примеры.

**Пример (2)**
Пусть $\mathbf{X} = (X_1,\ldots,X_n)$ - выборка из нормального распределения $\mathcal{N}(0,1)$. Постройте р.н.м.к. уровня значимости $\varepsilon$ для проверки гипотезы $\mathrm{H}_0 : \theta \geqslant \theta_0$ против альтернативы $\mathrm{H}_1 : \theta < \theta_0$.

**Решение**

Выпишем функцию правдоподобия:

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

Необходимо найти отношение правдоподобия, где $\theta_1 < \theta_0$ :

$$\frac{p_{\theta_1}(\mathbf{X})}{p_{\theta_0}(\mathbf{X})} = \frac{\exp\left\{-\sum\limits_{i=1}^{n} \frac{(X_i-\theta_1)^2}{2}\right\}}{\exp\left\{-\sum\limits_{i=1}^{n} \frac{(X_i-\theta_0)^2}{2}\right\}} = \exp\left\{(\theta_1-\theta_0)\sum_{i=1}^{n} X_i + \frac{n}{2}(\theta_0^2-\theta_1^2)\right\}.$$

Решаем неравенство:

$$\frac{p_{\theta_1}(\mathbf{X})}{p_{\theta_0}(\mathbf{X})} = \exp\left\{(\theta_1-\theta_0)\sum_{i=1}^{n} X_i + \frac{n}{2}(\theta_0^2-\theta_1^2)\right\} \geqslant \lambda.$$

Раз $\theta_1 < \theta_0$, то выражение в левой части убывает по $\sum\limits_{i=1}^{n} X_i$. Тогда неравенство выше эквивалентно тому, что $\sum\limits_{i=1}^{n} X_i \leqslant c = c(\lambda,\theta_0,\theta_1)$. Теперь надо найти $c$ из уравнения

$$\mathrm{P}_{\theta_0}\left(\sum_{i=1}^{n} X_i \leqslant c\right) = \varepsilon.$$

В случае нормального распределения всегда удобно переходить к стандартному нормальному распределению. Заметим, что $\sqrt{n}(\overline{\mathbf{X}}-\theta_0) \sim \mathcal{N}(0,1)$, если $\theta_0$ - истинное значение. Отсюда, критерий имеет вид

$$S = \{\sqrt{n}(\overline{\mathbf{X}} - \theta_0) \leqslant u\},$$

где $u$ - это $\varepsilon$-квантиль распределения $\mathcal{N}(0,1)$.

В этом примере видно, что для любого $\varepsilon$ подобное $u$ находится без проблем, так как распределение абсолютно непрерывно.

## Двойственность проверки гипотез и доверительного оценивания

Прежде чем переходить к обсуждению линейных гипотез в линейной гауссовской модели, поговорим о двойственности проверки гипотез и доверительного оценивания. Идея проверить гипотезу и идея оценить параметр, особенно если гипотеза записывается в виде предположения о значении параметра, выглядят очень близкими задачами. Умение строить доверительные интервалы помогает при проверке гипотез, и наоборот. Пусть $\mathbf{X}$ - это наблюдение с неизвестным распределением P, принадлежащим параметрическому семейству $\{\mathrm{P}_\theta, \theta \in \Theta\}$.

**1.** Пусть $S(\mathbf{X}) \subset \Theta$ - доверительная область уровня доверия $\gamma = 1 - \varepsilon$ для параметра $\theta$. Хотим проверить простую гипотезу $\mathrm{H}_0 : \theta = \theta_0$.

**Утверждение (1)**
Критерий $\tilde{S}_{\theta_0} = \{\mathbf{X} : \theta_0 \notin S(\mathbf{X})\}$ является критерием уровня значимости $\varepsilon$ для проверки гипотезы $\mathrm{H}_0$.

**Доказательство.**

$$\mathrm{P}_{\theta_0}(\mathbf{X} \in \tilde{S}_{\theta_0}) = \mathrm{P}_{\theta_0}(\theta_0 \notin S(\mathbf{X})) = 1 - \mathrm{P}_{\theta_0}(\theta_0 \in S(\mathbf{X})) \leqslant \varepsilon.$$

■

**2**. Пусть для любого $\theta_0 \in \Theta$ мы умеем строить критерий $\tilde{S}_{\theta_0}$ уровня значимости $\varepsilon$ для проверки простой гипотезы $\mathrm{H}_0 : \theta = \theta_0$. Хотим построить доверительную область.

**Утверждение (2)**
Область $S(\mathbf{X}) = \{\theta \in \theta : \mathbf{X} \notin \tilde{S}_\theta\}$ является доверительной областью уровня доверия $\gamma = 1 - \varepsilon$ для параметра $\theta$.

**Доказательство.** Для любого $\theta_0 \in \Theta$ выполнено

$$\mathrm{P}_{\theta_0}(\theta_0 \in S(\mathbf{X})) = \mathrm{P}_{\theta_0}(\mathbf{X} \notin \tilde{S}_{\theta_0}) = 1 - \mathrm{P}_{\theta_0}(\mathbf{X} \in \tilde{S}_{\theta_0}) \geqslant 1 - \varepsilon.$$

■

Применим эти соображения при проверке линейных гипотез в линейной гауссовской регрессии.

## Линейные гипотезы

Пусть

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\theta} + \varepsilon$$

- гауссовская линейная модель $\mathbf{Z} \in \mathrm{Mat}(n \times k)$, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Неизвестные параметры: $\boldsymbol{\theta} \in \mathbb{R}^k$, $\sigma > 0$.

**Задача**
Построить критерий для проверки линейной гипотезы:

$$H_0 : \mathbf{A}\boldsymbol{\theta} = \boldsymbol{\tau},$$

где $\mathbf{A} = \mathrm{Mat}(m \times k)$, $\boldsymbol{\tau} \in \mathbb{R}^m$ - известны и $\mathrm{rank}\ \mathbf{A} = m \leqslant k$.

**Решение**

Мы хотим проверить гипотезу следующего вида: изначально $\boldsymbol{\theta} \in \mathbb{R}^k$, но мы хотим проверить, что $\boldsymbol{\theta}$ пробегает лишь некоторое аффинное подпространство, которое задается уравнением $\mathbf{A}\boldsymbol{\theta} = \boldsymbol{\tau}$.

**Как строить?** Рассмотрим оценку наименьших квадратов $\hat{\boldsymbol{\theta}}(\mathbf{X}) = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X}$. Тогда мы знаем, что $\hat{\boldsymbol{\tau}} = \hat{\boldsymbol{\tau}}(\mathbf{X}) = \mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X})$ является оптимальной оценкой $\mathbf{A}\boldsymbol{\theta}$.

**Лемма**
Обозначим $\mathbf{D} = \mathbf{A}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{A}^{\mathrm{T}}$ и положим

$$\mathrm{Q}_A(\mathbf{X}) = (\hat{\boldsymbol{\tau}} - \mathbf{A}\boldsymbol{\theta})^{\mathrm{T}}\mathbf{D}^{-1}(\hat{\boldsymbol{\tau}} - \mathbf{A}\hat{\boldsymbol{\theta}}).$$

Тогда

$$\frac{\mathrm{Q}_A(\mathbf{X})}{\sigma^2} \sim \chi^2_m.$$

**Доказательство.**
Напомним, что $\hat{\boldsymbol{\theta}}(\mathbf{X}) \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1})$. Тогда, конечно,

$$\hat{\boldsymbol{\tau}} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\theta}, \sigma^2\mathbf{A}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{A}^{\mathrm{T}}) = \mathcal{N}(\mathbf{A}\boldsymbol{\theta}, \sigma^2\mathbf{D}).$$

Матрица $\mathbf{D} \in \mathrm{Mat}(m \times m)$ положительно определена. Значит, из нее можно извлечь корень: $\mathbf{D}^{\frac{1}{2}}$, где $(\mathbf{D}^{\frac{1}{2}})^2 = \mathbf{D}$ и $\mathbf{D}^{\frac{1}{2}} = (\mathbf{D}^{\frac{1}{2}})^{\mathrm{T}}$. Тогда

$$\frac{1}{\sigma} \cdot \mathbf{D}^{\frac{1}{2}}(\hat{\boldsymbol{\tau}} - \mathbf{A}\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{D}^{-\frac{1}{2}} \cdot \mathbf{D} \cdot \mathbf{D}^{-\frac{1}{2}}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_m).$$

В итоге,

$$\left|\frac{1}{\sigma}\mathbf{D}^{-\frac{1}{2}}(\hat{\boldsymbol{\tau}} - \mathbf{A}\boldsymbol{\theta})\right| \sim \chi^2_m.$$

Осталось заметить, что

$$\left|\mathbf{D}^{-\frac{1}{2}}(\hat{\boldsymbol{\tau}} - \mathbf{A}\boldsymbol{\theta})\right|^2 = (\hat{\boldsymbol{\tau}} - \mathbf{A}\boldsymbol{\theta})^{\mathrm{T}}\mathbf{D}^{-\frac{1}{2}} \cdot \mathbf{D}^{-\frac{1}{2}}(\hat{\boldsymbol{\tau}} - \mathbf{A}\boldsymbol{\theta}) = (\hat{\boldsymbol{\tau}} - \mathbf{A}\boldsymbol{\theta})^{\mathrm{T}}\mathbf{D}^{-1}(\hat{\boldsymbol{\tau}} - \mathbf{A}\boldsymbol{\theta}) = \mathrm{Q}_A(\mathbf{X}).$$

■

## Построение критерия

Рассмотрим статистику, похожую на $\mathrm{Q}_A(\mathbf{X})$ :

$$\hat{\mathrm{Q}}_A(\mathbf{X}) = (\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})^{\mathrm{T}}\mathbf{D}^{-1}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}).$$

Согласно лемме в условиях гипотезы $\mathrm{H}_0 : \mathbf{A}\boldsymbol{\theta} = \boldsymbol{\tau}$ выполнено

$$\frac{1}{\sigma^2}\hat{\mathrm{Q}}_A(\mathbf{X}) \sim \chi^2_m.$$

Также заметим, что статистика $\hat{\mathrm{Q}}_A(\mathbf{X})$ является функцией от $\hat{\boldsymbol{\theta}}(\mathbf{X})$. Следовательно, она независима со статистикой $\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2$.

Далее, вспомним, что $\frac{1}{\sigma^2}\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2 \sim \chi^2_{n-k}$. Стало быть, если гипотеза $\mathrm{H}_0$ верна, то выполнено

$$\frac{\hat{\mathrm{Q}}_A(\mathbf{X})}{\|\mathbf{X} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2} \cdot \frac{n-k}{m} \sim \mathrm{F}_{m,n-k}.$$

Теперь мы готовы сформулировать F-критерий для проверки гипотезы $\mathrm{H}_0$.

## F-критерий

### F-критерий

Пусть $f_{1-\varepsilon}$ - это $(1-\varepsilon)$-квантиль $F_{m,n-k}$. Если выполнено неравенство

$$\frac{\hat{\mathrm{Q}}_A(\mathbf{X})}{\|\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2}\cdot\frac{n-k}{m}=\frac{(\hat{\boldsymbol{\tau}}-\boldsymbol{\tau})^{\mathrm{T}}\mathbf{D}^{-1}(\hat{\boldsymbol{\tau}}-\boldsymbol{\tau})}{\|\mathbf{X}-\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2}\cdot\frac{n-k}{m}>f_{1-\varepsilon},$$

то гипотеза $\mathrm{H}_0:\mathbf{A}\boldsymbol{\theta}=\boldsymbol{\tau}$ отвергается.

Вероятность ошибки первого рода у F-критерия равна $\varepsilon$.

### Лемма

F-критерий является несмещенным.

### Пример: проверка однородности

Даны две независимые нормальные выборки: $\mathbf{X}=(X_1,\ldots,X_n)\sim\mathcal{N}(a_1,\sigma^2)$ и $\mathbf{Y}=(Y_1,\ldots,Y_m)\sim\mathcal{N}(a_2,\sigma^2)$. Постройте F-критерий для проверки гипотезы однородности $\mathrm{H}_0:a_1=a_2$.

## Решение

Сведем задачу к гауссовской модели. Для этого составим единый вектор наблюдений:

$$\mathbf{W}=(X_1,\ldots,X_n,Y_1,\ldots,Y_m)^{\mathrm{T}}=(a_1,\ldots,a_1,a_2,\ldots,a_2)^{\mathrm{T}}+\varepsilon,$$

где $\varepsilon\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_{n+m})$. Далее, представим:

$$\begin{pmatrix}a_1\\ \vdots\\ a_1\\ a_2\\ \vdots\\ a_2\end{pmatrix}=\underbrace{\begin{pmatrix}1&0\\ \vdots&\vdots\\ 1&0\\ 0&1\\ \vdots&\vdots\\ 0&1\end{pmatrix}}_{\mathbf{Z}}\underbrace{\begin{pmatrix}a_1\\ a_2\end{pmatrix}}_{\boldsymbol{\theta}}.$$

Гауссовская линейная модель построена. Теперь перепишем гипотезу $\mathrm{H}_0:a_1=a_2$ в виде $\mathrm{H}_0:\mathbf{A}\boldsymbol{\theta}=\mathbf{0}$, где $\mathbf{A}=\begin{pmatrix}1&-1\end{pmatrix}\in\mathrm{Mat}(1\times2)$. Остается вычислить участвующие величины.

1) Оценка наименьших квадратов:

$$\hat{\boldsymbol{\theta}}(\mathbf{X})=\begin{pmatrix}\overline{\mathbf{X}}\\ \overline{\mathbf{Y}}\end{pmatrix};$$

2)

$$\mathbf{D} = \mathbf{A}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1} = \frac{1}{n} + \frac{1}{m};$$

3)

$$\hat{\mathrm{Q}}_A(\mathbf{X}) = (\overline{\mathbf{X}} - \overline{\mathbf{Y}})^2 \left(\frac{1}{n} + \frac{1}{m}\right)^{-1};$$

4)

$$\|\mathbf{W} - \mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})\|^2 = \sum_{i=1}^{n}(X_i - \overline{\mathbf{X}})^2 + \sum_{j=1}^{m}(Y_j - \overline{\mathbf{Y}})^2.$$

В итоге, F-критерий имеет вид

$$\frac{(\overline{\mathbf{X}} - \overline{\mathbf{Y}})^2(\frac{1}{n} + \frac{1}{m})^{-1}}{\sum\limits_{i=1}^{n}(X_i - \overline{\mathbf{X}})^2 + \sum\limits_{j=1}^{m}(Y_j - \overline{\mathbf{Y}})^2} \cdot \frac{n+m-2}{1} =$$

$$= \frac{(\overline{\mathbf{X}} - \overline{\mathbf{Y}})^2}{\sum\limits_{i=1}^{n}(X_i - \overline{\mathbf{X}})^2 + \sum\limits_{j=1}^{m}(Y_j - \overline{\mathbf{Y}})^2} \cdot \frac{nm(n+m-2)}{n+m} > f_{1-\varepsilon},$$

где $f_{1-\varepsilon}$ - это $(1-\varepsilon)$-квантиль $F_{1,m+m-2}$.

**Замечание**
Существенным недостатком F-критерия является необходимость обращения матрицы $\mathbf{D} = \mathbf{A}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{A}^{\mathrm{T}}$. Даже в самых простых постановках это бывает невозможно сделать аналитически.

## Обобщенный метод наименьших квадратов

Оказывается, статистику $\hat{\mathrm{Q}}_A(\mathbf{X})$ можно находить по-другому, используя прямую оптимизацию. Обозначим

$$s(\boldsymbol{\theta}) = \|\mathbf{X} - \mathbf{Z}\boldsymbol{\theta}\|^2$$

и положим

$$\boldsymbol{\theta}^*(\mathbf{X}) = \arg\min_{\boldsymbol{\theta}:\mathbf{A}\boldsymbol{\theta}=\boldsymbol{\tau}} \|\mathbf{X} - \mathbf{Z}\boldsymbol{\theta}\|^2.$$

**Лемма**
Выполняется равенство

$$\boldsymbol{\theta}^*(\mathbf{X}) = \hat{\boldsymbol{\theta}}(\mathbf{X}) - (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\tau}).$$

**Доказательство.**
Заметим, что

$$\mathbf{A}\boldsymbol{\theta}^*(\mathbf{X}) = \mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \mathbf{D}\mathbf{D}^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\tau}) = \boldsymbol{\tau},$$

так что $\boldsymbol{\theta}^*$ может быть искомым.

Теперь для любого $\boldsymbol{\theta} \in \mathbb{R}^k$ рассмотрим:

$$s(\boldsymbol{\theta}) = \|\mathbf{X} - \mathbf{Z}\boldsymbol{\theta}\|^2 = \|\mathbf{X} - \mathbf{Z}\boldsymbol{\theta} \pm \mathbf{Z}\boldsymbol{\theta}^*(\mathbf{X})\|^2 =$$

$$= s(\boldsymbol{\theta}^*(\mathbf{X})) + \|\mathbf{Z}\boldsymbol{\theta} - \mathbf{Z}\boldsymbol{\theta}^*(\mathbf{X})\|^2 + 2(\mathbf{Z}\boldsymbol{\theta}^*(\mathbf{X}) - \mathbf{Z}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{X} - \mathbf{Z}\boldsymbol{\theta}^*(\mathbf{X})).$$

Покажем теперь, что если $\mathbf{A}\boldsymbol{\theta} = \boldsymbol{\tau}$, то последнее слагаемое равно нулю. Действительно,

$$(\mathbf{Z}\boldsymbol{\theta}^*(\mathbf{X}) - \mathbf{Z}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{X} - \mathbf{Z}\boldsymbol{\theta}^*(\mathbf{X})) = (\boldsymbol{\theta}^*(\mathbf{X}) - \boldsymbol{\theta})^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{X} - \mathbf{Z}^{\mathrm{T}}\mathbf{Z}\boldsymbol{\theta}^*(\mathbf{X})) \boxed{=}$$

(так как $\mathbf{Z}^{\mathrm{T}}\mathbf{X} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X})$)

$$\boxed{=} (\boldsymbol{\theta}^*(\mathbf{X}) - \boldsymbol{\theta})^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}(\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X})) \boxed{=}$$

(применим к $\boldsymbol{\theta}^*(\mathbf{X})$ во второй скобке формулу из условия Леммы)

$$\boxed{=} (\boldsymbol{\theta}^*(\mathbf{X}) - \boldsymbol{\theta})^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\tau}) \boxed{=}$$

(так как $\mathbf{A}\boldsymbol{\theta} = \boldsymbol{\tau}$)

$$\boxed{=} (\boldsymbol{\theta}^*(\mathbf{X}) - \boldsymbol{\theta})^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{A}(\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}).$$

Остается заметить, что уже первая часть равна нулю:

$$(\boldsymbol{\theta}^*(\mathbf{X}) - \boldsymbol{\theta})^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} = (\mathbf{A}\boldsymbol{\theta}^*(\mathbf{X}) - \mathbf{A}\boldsymbol{\theta})^{\mathrm{T}} = (\boldsymbol{\tau} - \boldsymbol{\tau})^{\mathrm{T}} = 0.$$

Значит,

$$s(\boldsymbol{\theta}) = s(\boldsymbol{\theta}^*(\mathbf{X})) + \|\mathbf{Z}\boldsymbol{\theta} - \mathbf{Z}\boldsymbol{\theta}^*(\mathbf{X})\|^2 \geqslant s(\boldsymbol{\theta}^*(\mathbf{X}))$$

для всех $\boldsymbol{\theta}$ с условием $\mathbf{A}\boldsymbol{\theta} = \boldsymbol{\tau}$. То есть $\boldsymbol{\theta}^*(\mathbf{X})$ решает экстремальную задачу, на ней достигается наименьшее значение функции $s$ на аффинном подпространстве. Причем в силу полноты ранга $\mathbf{Z}$ равенство будет достигаться тогда и только тогда, когда $\boldsymbol{\theta} = \boldsymbol{\theta}^*(\mathbf{X})$. ∎

$\boldsymbol{\theta}^*$ можно находить как решение экстремальной задачи, а также с ее помощью можно вычислять статистику $\hat{\mathrm{Q}}_A$, которая нужна в критерии. Вторая лемма играет ключевую роль в обобщенном методе наименьших квадратов.

**Лемма**
Выполняется равенство

$$\hat{\mathrm{Q}}_A(\mathbf{X}) = s(\boldsymbol{\theta}^*(\mathbf{X})) - s(\hat{\boldsymbol{\theta}}(\mathbf{X})).$$

**Доказательство.**
Из доказательства предыдущей леммы следует, что для любого $\boldsymbol{\theta} \in \mathbb{R}^k$ выполняется равенство:

$$s(\boldsymbol{\theta}) = s(\boldsymbol{\theta}^*(\mathbf{X})) + \|\mathbf{Z}\boldsymbol{\theta} - \mathbf{Z}\boldsymbol{\theta}^*(\mathbf{X})\|^2 + 2(\boldsymbol{\theta}^*(\mathbf{X}) - \boldsymbol{\theta})^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\tau}) =$$

$$= s(\boldsymbol{\theta}^*(\mathbf{X})) + \|\mathbf{Z}\boldsymbol{\theta} - \mathbf{Z}\boldsymbol{\theta}^*(\mathbf{X})\|^2 - 2(\mathbf{A}\boldsymbol{\theta} - \boldsymbol{\tau})^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\tau}).$$

Положим $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\mathbf{X})$, тогда последнее слагаемое будет в точности равно $-2 \cdot \hat{\mathrm{Q}}_A(\mathbf{X})$ (так как $\hat{\mathrm{Q}}_A(\mathbf{X}) = (\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})^{\mathrm{T}}\mathbf{D}^{-1}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})$). Рассмотрим второе слагаемое и попробуем показать, что оно тоже совпадает со статистикой $\hat{\mathrm{Q}}_A$:

$$\|\mathbf{Z}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \mathbf{Z}\boldsymbol{\theta}^*(\mathbf{X})\|^2 = (\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X}))^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}(\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}^*(\mathbf{X})) \boxed{=}$$

(используем формулу для $\boldsymbol{\theta}^*(\mathbf{X})$)

$$\boxed{=} (\mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\tau})^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{A}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1} \cdot \mathbf{Z}^{\mathrm{T}}\mathbf{Z} \cdot (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\tau}) =$$

$$= (\mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\tau})^{\mathrm{T}}\mathbf{D}^{-1} \cdot \mathbf{D} \cdot \mathbf{D}^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\tau}) = (\mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\tau})^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{A}\hat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\tau}) = \hat{\mathrm{Q}}_A(\mathbf{X}).$$

В итоге, $s(\hat{\boldsymbol{\theta}}(\mathbf{X})) = s(\boldsymbol{\theta}^*(\mathbf{X})) - \hat{\mathrm{Q}}_A(\mathbf{X})$. ■

**Вывод**
Для вычисления статистики $\hat{\mathrm{Q}}_A$ достаточно найти $s(\hat{\boldsymbol{\theta}}(\mathbf{X}))$ и $s(\boldsymbol{\theta}^*(\mathbf{X}))$, которые могут быть найдены путем решения экстремальных задач методами оптимизации. Кроме того, итоговая статистика для применения F-критерия будет иметь вид

$$\frac{s(\boldsymbol{\theta}^*(\mathbf{X})) - s(\hat{\boldsymbol{\theta}}(\mathbf{X}))}{s(\hat{\boldsymbol{\theta}}(\mathbf{X}))} \cdot \frac{n-k}{m}.$$

**Пример: проверка однородности**
Даны $k$ независимых нормальных выборок:

$$\mathbf{X}^{(1)} = (X_1^{(1)}, \ldots, X_{n_1}^{(1)}) \sim \mathcal{N}(a_1, \sigma^2), \ldots, \mathbf{X}^{(k)} = (X_1^{(k)}, \ldots, X_{n_k}^{(k)}) \sim \mathcal{N}(a_k, \sigma^2).$$

Постройте F-критерий для проверки гипотезы однородности $\mathrm{H}_0 : a_1 = a_2 = \cdots = a_k$.

**Решение**

Сведем задачу к гауссовской модели. Для этого составим единый вектор наблюдений:

$$\mathbf{W} = (X_1^{(1)}, \ldots, X_{n_1}^{(1)}, X_1^{(2)}, \ldots, X_1^{(k)}, \ldots, X_{n_k}^{(k)})^{\mathrm{T}} = (a_1, \ldots, a_1, a_2, \ldots, a_k)^{\mathrm{T}} + \boldsymbol{\varepsilon},$$

где $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{n_1+\ldots+n_k})$. Представляя вектор среднего в виде $\mathbf{Z}\boldsymbol{\theta}$, где $\boldsymbol{\theta} = (a_1, a_2, \ldots, a_k)^{\mathrm{T}}$, получаем, что

$$s(\boldsymbol{\theta}) = \|\mathbf{W} - \mathbf{Z}\boldsymbol{\theta}\|^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_i^{(j)} - a_j)^2.$$

Далее,

$$s(\boldsymbol{\theta}^*(\mathbf{X})) = \min_{\boldsymbol{\theta}\in\mathbb{R}^k} \|\mathbf{X} - \mathbf{Z}\boldsymbol{\theta}\|^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_i^{(j)}\overline{\mathbf{X}^{(j)}})^2,$$

где $\overline{\mathbf{X}^{(j)}} = \frac{1}{n_j}\sum\limits_{i=1}^{n_j} X_i^{(j)}$.

В свою очередь, минимизируя $s(\boldsymbol{\theta})$ по подпространству $a_1 = a_2 = \cdots = a_k$, получаем:

$$s(\boldsymbol{\theta}^*(\mathbf{X})) = \min_{\boldsymbol{\theta}\in\mathbb{R}^k : a_1=a_2=\cdots=a_k} \|\mathbf{X} - \mathbf{Z}\boldsymbol{\theta}\|^2 =$$

$$= \min_{a\in\mathbb{R}} \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_i^{(j)} - a)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_i^{(j)} - \overline{\mathbf{X}})^2,$$

где $\overline{\mathbf{X}} = \dfrac{1}{n_1 + \cdots + n_k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j} X_i^{(j)}$.

Тем самым, мы нашли $s(\boldsymbol{\theta}^*(\mathbf{X}))$ и, соответственно, $s(\hat{\boldsymbol{\theta}}(\mathbf{X}))$. Осталось применить формулу из вывода, чтобы получить статистику для F-критерия и сравнить ее с квантилью.

На этом мы завершаем разбор проверки линейных гипотез в линейной гауссовской модели.

# 17. Лекция 17

## Критерии согласия

Рассмотрим так называемые критерии согласия.

**Определение 17.1.** Пусть $\mathbf{X}$ - наблюдение с неизвестным распределением P. Далее, введем гипотезы $\mathrm{H}_0 : \mathrm{P} = \mathrm{P}_0$ и $\mathrm{H}_1 : \mathrm{P} \neq \mathrm{P}_0$. Критерий для проверки гипотезы $\mathrm{H}_0$ против полной альтернативы $\mathrm{H}_1$ называется критерием согласия.

Что мы хотели бы от такого критерия? В идеале мы желаем получить что-то наподобие равномерно наиболее мощного критерия. Но эта задача почти всегда безнадежна. Даже в случае хороших распределений, для которых выполнено свойство монотонного отношения правдоподобия, такого критерия может не быть. Например, для распределений $\mathrm{Bin}(1,\theta)$, $\mathrm{Exp}(\theta)$, $\mathcal{N}(\theta,1)$ найти равномерно наиболее мощный критерий для проверки гипотезы $\mathrm{H}_0 : \theta = \theta_0$ против альтернативы $\mathrm{H}_1 : \theta \neq \theta_0$ не получится.

**Упражнение**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из $\mathrm{U}(0,\theta)$. Постройте р.н.м.к. уровня значимости $\varepsilon$ для проверки гипотезы $\mathrm{H}_0 : \theta = \theta_0$ против альтернативы $\mathrm{H}_1 : \theta \neq \theta_0$.

### Асимптотические критерии согласия

Совершенно естественно на этом пути возникает задача поиска асимптотических критериев. Что мы понимаем под асимптотическим критерием? Как и в случае с оценками, если мы говорим об асимптотике, то она берется по размеру выборки. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка растущего размера (для каждого $n$ мы добавляем еще один элемент выборки). Мы хотели бы получить последовательность критериев $S_n = S_n(X_1, \ldots, X_n)$ для проверки гипотезы $\mathrm{H}_0 : \theta = \theta_0$ против альтернативы $\mathrm{H}_1^\theta \neq \theta_0$ такую, что

- Предел вероятности ошибки первого рода стремится к $\varepsilon \in (0,1)$ :

$$\mathrm{P}_{\theta_0}(\mathbf{X} \in S_n) \to \varepsilon.$$

- Последовательность критериев будет состоятельна: для любого $\theta \neq \theta_0$

$$\mathrm{P}_{\theta}(\mathbf{X} \in S_n) \to 1.$$

Несмотря на то, что условия кажутся сложными для выполнения, оказывается, что решения существуют. Рассмотрим первый из таких критериев - критерий хи-квадрат Пирсона для случая дискретных распределений.

# Критерий хи-квадрат

Пусть есть выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из распределения P с конечным числом значений $\{a_1, \ldots, a_m\}$ :

$$p_j = \mathrm{P}(X_1 = a_j) > 0, \quad j = 1, \ldots, m, \quad \sum_{j=1}^{m} p_j = 1.$$

Введем вектор вероятностей $\mathbf{p} = (p_1, \ldots, p_m)$. Гипотеза $\mathrm{H}_0 : \mathrm{P} = \mathrm{P}_0$ тогда сведется к равенству векторов вероятностей: $\mathrm{H}_0 : \mathbf{p} = \mathbf{p}^0$. Альтернатива примет вид $\mathrm{H}_1 : \mathbf{p}\mathbf{p}^0$.

Для каждого $j = 1, \ldots, m$ введем

$$\mu_j = \sum_{i=1}^{n} \mathbf{I}\{X_i = a_j\},$$

- это число элементов выборки, равных $a_j$.

**Определение 17.2.** Статистикой хи-квадрат для проверки гипотезы $\mathrm{H}_0^{\mathbf{p}} = \mathbf{p}^0$ называется

$$\hat{\chi}_n^2(\mathbf{X}) = \sum_{j=1}^{n} \frac{(\mu_j - np_j^0)^2}{np_0^j}.$$

Сам же критерий устроен следующим образом.

**Критерий хи-квадрат**
Пусть $u_{1-\varepsilon}$ - это $(1-\varepsilon)-$квантиль $\chi^2_{m-1}$. Если выполнено равенство

$$\hat{\chi}_n^2(\mathbf{X}) > u_{1-\varepsilon},$$

то гипотеза $\mathrm{H}_0 : \mathbf{p} = \mathbf{p}^0$ отвергается.

Изучим свойство этого критерия, а именно - найдем предельные вероятности ошибок первого и второго рода. Начнем со второго.

**Утверждение**
Критерий хи-квадрат состоятелен.

**Доказательство.**
Для начала немного перепишем статистику хи-квадрат:

$$\hat{\chi}_n^2(\mathbf{X}) = n \sum_{j=1}^{m} \frac{1}{p_j^0} \left( \frac{\mu_j}{n} - p_j^0 \right)^2.$$

Далее, если $\mathbf{p} \neq \mathbf{p}^0$, то найдется индекс $j \in \{1, 2 \ldots, m\}$ такой, что $p_j \neq p_j^0$. По усиленному закону больших чисел $\frac{\mu_j}{n} \to p_j$ почти наверное. Отсюда следует, что $\hat{\chi}_n^2(\mathbf{X})$ линейно стремится к бесконечности, так как

$$\frac{\hat{\chi}_n^2(\mathbf{X})}{n} \xrightarrow{\text{п.н.}} \sum_{j=1}^{m} \frac{(p_j - p_j^0)^2}{p_j^0} > 0.$$

Тем самым, критерий состоятелен: $\mathrm{P}(\hat{\chi}_n^2(\mathbf{X}) > u_{1-\varepsilon}) \to 1$. ■

Разберемся теперь, что происходит с вероятностью ошибки первого рода, то есть что происходит со статистикой, если гипотеза правильная. Название критерия и его построение вытекает из следующей теоремы

**Теорема 17.1. (Пирсон)**
Если выполнена гипотеза $\mathrm{H}_0$, то имеет место следующая сходимость по распределению:

$$\hat{\chi}_n^2(\mathbf{X}) \xrightarrow[n\to\infty]{d} \chi_{m-1}^2.$$

**Доказательство.**
Доказательство состоит в работе с гауссовскими случайными векторами, а распределение хи-квадрат возникает при попытке перейти от многомерного варианта к одномерному. Распределение $\chi_{m-1}^2$ получается как сумма квадратов стандартных нормальных случайных величин. Поймем теперь, где они возникают.

Пусть $\mathbf{p} = \mathbf{p}^0$. Для всех $i = 1, \ldots, n$ введем случайный вектор
$\mathbf{Y}_i = (\mathbf{I}\{X_i = a_1\}, \ldots, \mathbf{I}\{X_i = a_m\})^{\mathrm{T}}$. Понятно, что такие векторы независимы и одинаково распределены, причем $\mathrm{E}\mathbf{Y}_i = \mathbf{p}^0$. Далее, заметим, что

$$\frac{\mathbf{Y}_1 + \ldots \mathbf{Y}_n}{n} = \frac{1}{n} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}.$$

Следовательно, по многомерной центральной предельной теореме

$$\sqrt{n}\left(\frac{\mathbf{Y}_1 + \ldots \mathbf{Y}_n}{n} - \mathbf{p}^0\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

где $\mathbf{\Sigma} = \mathrm{D}\mathbf{Y}_1$. - матрица ковариаций случайного вектора $\mathbf{Y}_1$. Заметим, что

$$\mathrm{cov}(\mathbf{I}\{X_1 = a_i\}, \mathbf{I}\{X_1 = a_j\}) = \mathrm{P}(X_1 = a_i,\ X_1 = a_j) - p_i^0 p_j^0 = \begin{cases} p_i^0 - p_i^0 p_j^0, & i = j; \\ -p_i^0 p_j^0, & i \neq j. \end{cases}$$

Неприятность состоит в том, что это вырожденная матрица. Запишем матрицу ковариаций следующим образом: $\mathbf{\Sigma} = \mathbf{B} - \mathbf{p}^0(\mathbf{p}^0)^{\mathrm{T}}$, где $\mathbf{B} = \mathrm{diag}(\mathbf{p}^0)$. Теперь наша

цель - повернуть вектор так, чтобы матрица ковариаций стала похожа на матрицу стандартного нормального распределения. Положим

$$\boldsymbol{\xi}'_n = \sqrt{n}\left(\frac{\mathbf{Y}_1 + \dots \mathbf{Y}_n}{n} - \mathbf{p}^0\right).$$

Тогда по теореме о наследовании сходимости

$$(\sqrt{\mathbf{B}})^{-1}\boldsymbol{\xi}'_n \xrightarrow[n\to\infty]{d} (\sqrt{\mathbf{B}})^{-1}\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{0}, (\sqrt{\mathbf{B}})^{-1}\boldsymbol{\Sigma}(\sqrt{\mathbf{B}})^{-1}).$$

Новая матрица ковариаций равна $\mathbf{I}_m - \mathbf{z}\mathbf{z}^{\mathrm{T}}$, где

$$\mathbf{z} = (\sqrt{\mathbf{B}})^{-1}\mathbf{p}^0 = \left(\sqrt{p_1^0}, \dots, \sqrt{p_m^0}\right)^{\mathrm{T}}.$$

Заметим, что вектор $\mathbf{z}$ единичный, значит, он может быть базисным. Рассмотрим ортогональную матрицу $\mathbf{C}$, у которой первая строка равна $\mathbf{z}$, а остальные равны чему угодно. Тогда по теореме о наследовании сходимости

$$\boldsymbol{\xi}''_n = \mathbf{C}(\sqrt{\mathbf{B}})^{-1}\boldsymbol{\xi}'_n \xrightarrow[n\to\infty]{d} \mathcal{N}(\mathbf{0}, \mathbf{C}(\mathbf{I}_m - \mathbf{z}\mathbf{z}^{\mathrm{T}})\mathbf{C}^{\mathrm{T}}).$$

Однако $\mathbf{C}\mathbf{z} = (1, 0, \dots, 0)^{\mathrm{T}}$, а $\mathbf{C}\mathbf{C}^{\mathrm{T}} = \mathbf{I}_m$. Поэтому

$$\mathbf{C}(\mathbf{I}_m - \mathbf{z}\mathbf{z}^{\mathrm{T}})\mathbf{C}^{\mathrm{T}} = \mathrm{diag}(0, 1, \dots, 1) = \mathbf{I}'_m.$$

Снова по теореме о наследовании сходимости

$$\|\xi''_n\|^2 \xrightarrow[n\to\infty]{d} \|\mathcal{N}(\mathbf{0}, \mathbf{I}'_m)\|^2 \sim \chi^2_{m-1}.$$

Но, как известно, умножение вектора на ортогональную матрицу не меняет его норму. Тогда

$$\|(\sqrt{\mathbf{B}})^{-1}\boldsymbol{\xi}'_n\|^2 \xrightarrow[n\to\infty]{d} \chi^2_{m-1}.$$

Осталось заметить, что

$$(\sqrt{\mathbf{B}})^{-1}\boldsymbol{\xi}'_n = (\sqrt{\mathbf{B}})^{-1}\sqrt{n}\begin{pmatrix} \dfrac{\mu_1 - np_1^0}{n} \\ \dots \\ \dfrac{\mu_m - np_m^0}{n} \end{pmatrix} = \begin{pmatrix} \dfrac{\mu_1 - np_1^0}{\sqrt{np_1^0}} \\ \dots \\ \dfrac{\mu_m - np_m^0}{\sqrt{np_m^0}} \end{pmatrix}.$$

Тем самым,

$$\hat{\chi}^2_n(\mathbf{X}) = \sum_{j=1}^{m} \frac{(\mu_j - np_j^0)^2}{np_j^0} = \|(\sqrt{\mathbf{B}})^{-1}\boldsymbol{\xi}'_n\|^2 \xrightarrow[n\to\infty]{d} \chi^2_{m-1}.$$

■

**Следствие**
Для критерия хи-квадрат вероятность ошибки первого рода стремится к $\varepsilon$ с ростом $n$.

Пожертвовав точностью вероятности ошибки первого рода, мы получили критерий с очень сильными асимптотическими свойствами.

**Практическая применимость**
Насколько этот метод применим? На практике считается, что должно быть выполнено следующее ограничение: $np_j^0 \geqslant 5$ для всех $j = 1, \ldots, m$.

## Параметрический критерий хи-квадрат

Предположим, что мы хотим проверить более сложную гипотезу. Оказывается, что критерий хи-квадрат можно естественным образом обобщить. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из неизвестного распределения со значениями $\{a_1, \ldots, a_m\}$ :

$$p_i(\theta) = \mathrm{P}_\theta(X_1 = a_i),\ \theta \in \Theta.$$

В данном случае гипотезу формулируется в виде принадлежности к параметрическому семейству: $\mathrm{H}_0 : \mathbf{p} \in \{\mathbf{p}_\theta,\ \theta \in \Theta\}$ против $\mathrm{H}_1 : \mathbf{p} \notin \{\mathbf{p}_\theta,\ \theta \in \Theta\}$. Снова хотелось бы составить статистику хи-квадрат, но вектора $\mathbf{p}^0$ больше нет. Можно ли его чем-то заменить?

Естественно заменить $\mathbf{p}^0$ на некоторую оценку:

$$\hat{\chi}_n^2(\mathbf{X}) = \sum_{i=1}^{m} \frac{(\mu_i - n\hat{p}_i(\mathbf{X}))^2}{n\hat{p}_i(\mathbf{X})}. \tag{1}$$

Какую оценку взять? Самым разумным выбором является оценка максимального правдоподобия: $\hat{p}_i(\mathbf{X}) = p_i(\hat{\theta}(\mathbf{X}))$, где $\hat{\theta}(\mathbf{X})$ - оценка максимального правдоподобия для $\theta$. Оказывается, что в таком случае можно доказать результат, похожий на теорему Пирсона.

### Условия регулярности

Напомним, что ОМП вычисляется как решение задачи

$$\hat{\theta}_n(\mathbf{X}) = \arg\max_{\boldsymbol{\theta} \in \Theta} \prod_{i=1}^{m} (p_i(\boldsymbol{\theta}))^{\mu_i}.$$

Как всегда, если мы имеем дело с ОМП, то необходимы условия регулярности. Пусть выполнены следующие условия регулярности.

1) $\Theta \subseteq \mathbb{R}^s$, $s < m-1$ - открытое множество.

2) Для любого значения параметра все вероятности отделены от нуля: $p_i(\boldsymbol{\theta}) \geqslant c^2 > 0$ для всех $i$ и $\boldsymbol{\theta}$.

3) Будем считать, что $\dfrac{\partial p_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}$ и $\dfrac{\partial^2 p_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_k}$ непрерывны на всем $\Theta$.

4) Матрица $\mathbf{D}(\boldsymbol{\theta}) = \left( \dfrac{\partial p_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j},\ i = 1, \ldots, m,\ j = 1, \ldots, s \right)$ имеет ранг $s$ для любого $\theta \in \Theta$.

Теперь можно сформулировать параметрический аналог теоремы Пирсона.

**Теорема 17.2. (Фишер)**
Пусть выполнены условия регулярности. Введем следующую систему уравнений:

$$\sum_{i=1}^{m} \frac{\mu_i}{p_i(\boldsymbol{\theta})} \frac{\partial p_i(\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1, \ldots, s, \quad \mu_j = \sum_{i=1}^{n} \mathbf{I}\{X_i = a_j\}.$$

Если верна гипотезы $\mathrm{H}_0$, то с вероятностью, стремящейся к 1, данная система имеет единственное решение $\hat{\theta}_n(\mathbf{X})$ такое, что $\hat{\theta}_n(\mathbf{X})$ сходится по вероятности к истинному значению параметра $\boldsymbol{\theta}_0$ и

$$\hat{\chi}^2_n(\mathbf{X}) = \sum_{i=1}^{m} \frac{(\mu_i - np_i(\hat{\theta}_n(\mathbf{X})))^2}{np_i(\hat{\theta}_n(\mathbf{X}))} \xrightarrow[n\to\infty]{d_{\boldsymbol{\theta}_0}} \chi^2_{m-1-s}.$$

Опустим доказательство этой теоремы и посмотрим на ее применение.

**Параметрический критерий хи-квадрат**
Пусть $u_{1-\varepsilon}$ - это $(1-\varepsilon)-$квантиль $\chi^2_{m-1-s}$. Если выполнено неравенство

$$\hat{\chi}^2_n(\mathbf{X}) > u_{1-\varepsilon},$$

то гипотеза $\mathrm{H}_0 : \mathbf{p} \in \{\mathbf{p}_\theta,\ \theta \in \Theta\}$ отвергается.

Рассмотрим применение теоремы и параметрического критерия хи-квадрат к построению критерия независимости хи-квадрат.

## Критерий независимости хи-квадрат

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из векторного распределения, $X_k = (Y_k, Z_k)$, где случайные величины $Y_k$ принимают значения $A_1, \ldots, A_s$, а $Z_k$ - значения $B_1, \ldots, B_l$ ("признаки" ). Это так называемая двухфакторная модель.

Хотим проверить гипотезу независимости признаков: $\mathrm{H}_0$ : признаки независимы.

Обозначим

$$p_i = \mathrm{P}(Y_1 = A_i),\ i = 1, \ldots, s; \quad q_j = \mathrm{P}(Z_1 = B_j),\ j = 1, \ldots, l;$$

$$p_{ij} = \mathrm{P}(X_1 = (A_i, B_j)),\ i = 1, \ldots, s;\ j = 1, \ldots, l.$$

Сведем задачу к параметрической и построим параметрический критерий для проверки гипотезы $\mathrm{H}_0$ против полной альтернативы $\mathrm{H}_1$: признаки зависимы.

Введем параметр $\boldsymbol{\theta} = (p_1, \ldots, p_{s-1}, q_1, \ldots, q_{l-1})$ с условиями

$$0 < p_i, q_i,\ i = 1, \ldots, s,\ j = 1, \ldots, l; \quad \sum_{i=1}^{s-1} p_i < 1,\ \sum_{j=1}^{l-1} q_j < 1.$$

Размерность параметра равна $s + l - 2$. Положим $p_s = 1 - \sum\limits_{i=1}^{s-1} p_i$, $q_l = 1 - \sum\limits_{j=1}^{l-1} q_j$. В параметрическом семействе, отвечающем гипотезе $\mathrm{H}_0$, выполнено

$$p_{ij} = p_i \cdot q_j,\ i = 1, \ldots, s;\ j = 1, \ldots, l.$$

Для построения статистики $\hat{\chi}^2_n(\mathbf{X})$ нам необходимо определить остальные величины.

1) Обозначим

$$\mu_{ij} = \sum_{k=1}^{n} \mathbf{I}\{X_k = (A_i, B_j)\}, \quad \mu_i = \sum_{k=1}^{n} \mathbf{I}\{Y_k = A_i\}, \quad \nu_j = \sum_{k=1}^{n} \mathbf{I}\{Z_k = B_j\}.$$

2) Найдем оценки максимального правдоподобия для параметра $\boldsymbol{\theta} = (p_1, \ldots, p_{s-1}, q_1, \ldots, q_{l-1})$. Заметим, что функция правдоподобия выборки в случае вероятности гипотезы равна

$$f_{\boldsymbol{\theta}}(\mathbf{X}) = \prod_{i=1}^{s} \prod_{j=1}^{l} (p_i \cdot q_j)^{\mu_{ij}} = \prod_{i=1}^{s} p_i^{\mu_i} \cdot \prod_{j=1}^{l} q_j^{\nu_j}.$$

Тем самым, оптимизация дальше идет отдельно по $(p_1, \ldots, p_{s-1})$ и $(q_1, \ldots, q_{l-1})$.

**Утверждение**
ОМП для $(p_1, \ldots, p_{s-1})$ равны

$$\hat{p}_i(\mathbf{X}) = \frac{\mu_i}{n},\ i = 1, \ldots, s-1.$$

**Доказательство.**
Возьмем частную производную по $p_i$ у логарифма $f_{\boldsymbol{\theta}}(\mathbf{X})$ :

$$\frac{\partial}{\partial p_i} \ln f_{\boldsymbol{\theta}}(\mathbf{X}) = \frac{\mu_i}{p_i} - \frac{\mu_s}{1 - \sum\limits_{k=1}^{s-1} p_k}.$$

Равенство нулю всех производных означает, что все величины $\frac{p_i}{\mu_i}$ равны некоторому одному и тому же $\alpha$. Из условия $\sum\limits_{k=1}^{s} p_k = 1$ получаем, что

$$1 = \sum_{k=1}^{s} p_k = \alpha \sum_{k=1}^{s} \mu_k = \alpha \cdot n.$$

Значит, $\hat{p}_i(\mathbf{X}) = \frac{\mu_i}{n}$, $i = 1, \ldots, s$. ■

В итоге обобщенная статистика хи-квадрат равна

$$\hat{\chi}^2_n(\mathbf{X}) = \sum_{i=1}^{s}\sum_{j=1}^{l} \frac{(\mu_{ij} - n\hat{p}_i(\mathbf{X})\hat{q}_j(\mathbf{X}))^2}{n\hat{p}_i(\mathbf{X})\hat{q}_j(\mathbf{X})} = \sum_{i=1}^{s}\sum_{j=1}^{l} \frac{(n\mu_{ij} - \mu_i \cdot \nu_j)^2}{n \cdot \mu_i \cdot \nu_j}.$$

3) Число степеней свободы равно $sl - 1 - (s + l - 2) = (s-1)(l-1)$.

**Критерий независимости хи-квадрат**
Пусть $u_{1-\varepsilon}$ - это $(1-\varepsilon)-$квантиль $\chi^2_{(s-1)(l-1)}$. Если выполнено неравенство

$$\hat{\chi}^2_n(\mathbf{X}) > u_{1-\varepsilon},$$

то гипотеза $\mathrm{H}_0$ : отвергается.

По теореме Фишера данный критерий имеет вероятность ошибки первого рода в пределе равную $\varepsilon$.

**Упражнение**
Докажите, что критерий независимости состоятелен.

## Критерий однородности хи-квадрат

Еще одним обобщением стандартного критерия хи-квадрат является критерий однородности хи-квадрат. В отличие от критерия независимости, критерий однородности не выводится из параметрического случая.

Задача состоит в проверке однородности выборок. Пусть

$$\mathbf{X}^{(1)} = (X_1^1, \dots, X_{n_1}^1), \dots, \mathbf{X}^{(k)} = (X_k^1, \dots, X_k^{n_k})$$

- независимые выборки, принимающие одни и те же значения $\{a_1, \dots, a_m\}$. Обозначим

$$\mathbf{p}^{(j)} = (p_1^{(j)}, \dots, p_m^{(j)}),\ p_i^{(j)} = \mathrm{P}(X_1^{(j)} = a_i),\ j = 1, \dots, k.$$

Мы хотим проверить гипотезу однородности

$$\mathrm{H}_0 : \mathbf{p}^{(1)} = \dots = \mathbf{p}^{(k)}.$$

Для этого также составляется обобщенная статистика хи-квадрат. Для этого обозначим:

$$\mu_{ij} = \sum_{s=1}^{n_j} \mathbf{I}\{X_s^{(j)} = a_i\},\ \mu_i = \sum_{j=1}^{k} \mu_{ij},\ n = n_1 + \dots + n_k.$$

Тогда обобщенная статистика хи-квадрат равна

$$\hat{\chi}_n^2(\mathbf{X}) = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(\mu_{ij} - n_j \cdot \frac{\mu_i}{n})^2}{n_j \cdot \frac{\mu_i}{n}}.$$

**Критерий однородности хи-квадрат**
Пусть $u_{1-\varepsilon}$ - это $(1-\varepsilon)-$квантиль $\chi^2_{(k-1)(m-1)}$. Если выполнено неравенство

$$\hat{\chi}_n^2(\mathbf{X}) > u_{1-\varepsilon},$$

то гипотеза $\mathrm{H}_0$ : отвергается.

**Упражнение**
Докажите, что критерий однородности состоятелен.

## Критерии для непрерывных распределений

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из непрерывного распределения с функцией распределения $F(x)$. Мы хотели бы проверить простую гипотезу $\mathrm{H}_0 : F = F_0$, где $F_0$ - заданная непрерывная функция распределения на $\mathbb{R}$, против полной альтернативы $\mathrm{H}_1 : F \neq F_0$.

**Здравая идея**
Надо дискретизировать выборку $\mathbf{X} = (X_1, \ldots, X_n)$. Разобьем прямую на полуинтервалы $\Delta_1, \ldots, \Delta_m$, $\Delta_j = (a_j, b_j]$ и обозначим

$$p_i^0 = F_0(b_j) - F_0(a_j).$$

Введем выборку $\mathbf{Y} = (Y_1, \ldots, Y_n)$ со значениями $\{1, \ldots, m\}$ по правилу

$$Y_j = s \iff X_j \in \Delta_s.$$

Далее, проверяем критерием хи-квадрат гипотезу о том, что выборка $\mathbf{Y}$ имеет распределение $\mathbf{p}^0 = (p_1^0, \ldots, p_m^0)$.

Однако, этот метод "склеивает" те функции распределения $F$, которые имеют один и тот же вектор $\mathbf{p}^0$ для одного и того же набора полуинтервалов $\Delta_j$. Получается, что критерий не состоятелен, но он различит содержательно те распределения, которые имеют заведомо неодинаковые векторы $\mathbf{p}^0$ для одних и тех же полуинтервалов.

### Теорема Колмогорова

Построим по выборке эмпирическую функцию распределения:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{I}\{X_j \leqslant x\}.$$

Следующая теорема играет важнейшую роль в статистике для обоснования применения критерия Колмогорова.

**Теорема 17.3. (Колмогоров)**
Если $F(x)$ непрерывна, то распределение случайной величины

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$$

не зависит от $F(x)$. Более того,

$$\sqrt{n} \cdot D_n \xrightarrow{d} K,$$

где $K$ - распределение Колмогорова с ф.р. $K(y) = \sum\limits_{k \in \mathbb{Z}} (-1)^k e^{-2k^2y^2}$, $y > 0$.

Доказывать теорему Колмогорова мы будем на следующей лекции. Пока что построим критерий Колмогорова и обсудим его свойства.

### Критерий Колмогорова

Итак, снова пусть $\mathbf{X} = (X_1, \dots, X_n)$ - выборка из непрерывного распределения с функцией распределения $F(x)$. Мы хотели бы проверить простую гипотезу $\mathrm{H}_0 : F = F_0$, где $F_0$ - заданная непрерывная функция распределения на $\mathbb{R}$, против полной альтернативы $\mathrm{H}_1 : F \neq F_0$. Введем статистику

$$\hat{D}_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|.$$

**Критерий Колмогорова**
Пусть $k_{1-\varepsilon}$ - это $(1-\varepsilon)-$квантиль распределения Колмогорова $K$. Если выполнено неравенство

$$\sqrt{n} \cdot \hat{D}_n > k_{1-\varepsilon},$$

то гипотеза $\mathrm{H}_0 : F = F_0$ отвергается.

**Следствие**
Для критерия Колмогорова вероятность ошибки первого рода стремится к $\varepsilon$ с ростом $n$.

**Утверждение**
Критерий Колмогорова состоятелен

**Доказательство.**
Пусть настоящая функций распределения $F(x)$ не равна $F_0(x)$. По теореме Гливенко-Кантелли мы знаем, что с вероятностью 1

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \to 0.$$

Следовательно, обозначив $\sup\limits_{x \in \mathbb{R}} |F(x) - F_0(x)| = \varepsilon_0$, мы получаем, что

$$\mathrm{P}\left(\hat{D}_n > \frac{\varepsilon_0}{2}\right) \to 1 \quad \text{при } n \to +\infty.$$

Но тогда

$$\sqrt{n} \cdot \hat{D}_n \xrightarrow{\mathrm{P}} +\infty.$$

В итоге, получаем, что

$$\mathrm{P}\left(\sqrt{n} \cdot \hat{D}_n > k_{1-\varepsilon}\right) \to 1 \quad \text{при } n \to +\infty.$$

■

## Критерий Колмогорова-Смирнова

Для проверки однородности аналогом теоремы Колмогорова является теорема Смирнова.

**Теорема 17.4. (Смирнов)**
Если гипотеза однородности верна и функция распределения $F(x)$ выборок непрерывна, то распределение статистики $\hat{D}_{n,m} = \hat{D}_{n,m}(\mathbf{X}, \mathbf{Y})$ зависит только от $n$ и $m$ и, кроме того,

$$\sqrt{\frac{nm}{n+m}} \cdot \hat{D}_{n,m} \xrightarrow{d} K, \ n, m \to +\infty.$$

Опираясь на эту теорему можно сформулировать критерий однородности Колмогорова-Смирнова.

**Критерий однородности Колмогорова-Смирнова**
Пусть $k_{1-\varepsilon}$ - это $(1-\varepsilon)-$квантиль распределения Колмогорова $K$. Если выполнено неравенство

$$\sqrt{\frac{nm}{n+m}} \cdot \hat{D}_n > k_{1-\varepsilon},$$

то гипотеза однородности $\mathrm{H}_0 : X_i \stackrel{d}{=} Y_j$ отвергается.

**Следствие**
Для критерия Колмогорова-Смирнова вероятность ошибки первого рода стремится к $\varepsilon$ с ростом $n$.

**Замечание**
Заметим, что квантили распределения Колмогорова весьма невелики, например $K(1.8) = 0.996932\ldots$.

**Упражнение**
Докажите, что критерий однородности Колмогорова-Смирнова состоятелен.

На следующей лекции мы будем разбирать доказательство теоремы Колмогорова, которое опирается на сходимость по распределению случайных процессов. Нам нужно будет вспомнить некоторый материал из теории случайных процессов.

# 18. Лекция 18

## Напоминание из курса теории вероятностей

### Слабая сходимость вероятностных пар

Пусть $(S, \rho)$ - метрическое пространство.

**Определение 18.1.** Борелевской сигма-алгеброй, $\mathcal{B}(S)$, на $(S, \rho)$ называется минимальная $\sigma$-алгебра, содержащая все открытые множества в $S$.

**Определение 18.2.** Пусть задано метрическое пространство $(S, \rho)$ и последовательность $\{\mathrm{Q}_n, n \in \mathbb{N}\}$ вероятностных мер на $(S, \mathcal{B}(S))$. Будем говорить, что $\mathrm{Q}_n$ слабо сходятся к вероятностной мере Q на $(S, \mathcal{B}(S))$, если для любой ограниченной непрерывной функции $f : S \mapsto \mathbb{R}$ выполнено

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

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

### Теорема Александрова

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

**Теорема 18.1. (А.Д. Александров)**
Пусть $\{\mathrm{Q}_n, n \in \mathbb{N}\}$ и Q - вероятностные меры на метрическом пространстве $(S, \rho)$. Тогда следующие утверждения эквиваленты:

1) $\mathrm{Q}_n \xrightarrow{\omega} \mathrm{Q}$,
2) $\overline{\lim\limits_{n\to\infty}} \mathrm{Q}_n(F) \leqslant \mathrm{Q}(F)$ для любого замкнутого множества $F \subset S$,
3) $\underline{\lim\limits_{n\to\infty}}\,(G) \geqslant \mathrm{Q}(G)$ 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 $\mathrm{Q}_n(B) \to \mathrm{Q}(B)$ при $n \to \infty$.

### Сходимость случайных процессов

Будем рассматривать время $T = [0, 1]$.
Пусть $X^n = (X_t^{(n)}, t \in [0,1], n \in \mathbb{N})$ - последовательность действительных случайных процессов на $[0, 1]$. Как можно определить сходимость по распределению такой последовательности?
**Идея 1:** как слабую сходимость их распределений!
**Вопрос:** а что такое распределение случайного процесса?
Для этого мы напомним основные определения случайных процессов.

## Случайные процессы

**Определение 18.3.** Пусть $(\Omega, \mathcal{F}, \mathrm{P})$ - вероятностное пространство. Отображение $\xi : \Omega \to \mathbb{R}$ называется случайной величиной, если оно измеримо, то есть

$$\forall B \in \mathcal{B}(\mathbb{R})\ \xi^{-1}(B) = \{\omega : \xi(\omega) \in B\} \in \mathcal{F}.$$

**Определение 18.4.** Пусть $T$ - некоторое множество. Тогда набор $X = (X_t, t \in T)$ случайных величин $X_t(\omega)$, заданных на одном и том же вероятностном пространстве $(\Omega, \mathcal{F}, \mathrm{P})$ для $\forall t \in T$, называется случайным процессом с множеством времени $T$.

1) Если $T = [a, b], (a, b), [a, +\infty], \mathbb{R}$, то процесс $X$ называется процессом с непрерывным временем.

2) Если $T \subset \mathbb{Z}$, то процесс $X$ называется процессом с дискретным временем.

3) Если $T \subset \mathbb{R}^d, d > 1$, то процесс $X$ называется случайным полем.

### Пространство траекторий

**Определение 18.5.** При фиксированном $\omega = \omega_0$ функция

$$\tilde{X}_{\omega_0}(t) = X_t(\omega)\Big|_{\omega=\omega_0}$$

на $T$ называется траекторией или реализацией случайного процесса $X = (X_t, t \in T)$.

**Определение 18.6.** Множество $S = \underset{t\in T}{\times}\mathbb{R} = \mathbb{R}^T$ называется пространством траекторий случайного процесса $X$ (здесь $\underset{t\in T}{\times}\mathbb{R}$-декартово произведение), то есть формально

$$S = \{y = (y(t), t \in T) : \forall t \in T\ y(t) \in \mathbb{R}\}.$$

Пространство траекторий - это множество значений случайного процесса.

### Цилиндрическая сигма-алгебра

**Определение 18.7.** Для $\forall t \in T$ и $B_t \in \mathcal{B}(\mathbb{R})$ введем элементарный цилиндр с основанием $B_t$ :

$$C(t, B_t) = \{y \in S : y(t) \in B_t\}.$$

Образно говоря, $C(t, B_t)$ состоит из тех функций $y(t)$, которые в точке $t$ проходят через множество $B_t$ (см. рис. 6).

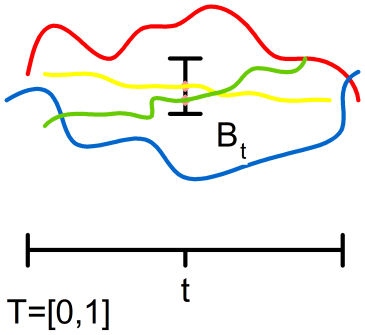


Рис. 6

**Определение 18.8.** Минимальная $\sigma$-алгебра $\mathcal{B}_T$, содержащая все эти элементарные цилиндры, называется цилиндрической $\sigma$-алгеброй на $S$. Формально,

$$\mathcal{B}_T = \sigma\{C(t, B_t) : t \in T, B_t \in \mathcal{B}(\mathbb{R})\}.$$

Для $\mathcal{B}_T$ используется также обозначение $\bigotimes\limits_{t \in T} \mathcal{B}(\mathbb{R})$.

## Измеримые отображения

Стало быть, $(S, \mathcal{B}_T)$ - измеримое пространство. Значит, на $X = (X_t, t \in T)$ можно смотреть как на один случайный элемент со значениями в $(S, \mathcal{B}_T)$ : каждому $\omega \in \Omega$ сопоставляется целая траектория $\tilde{X}_\omega(\cdot)$. Разберемся с вопросом измеримости такого отображения.

**Лемма (об измеримости случайного процесса)**
$X = (X_t, t \in T)$ является случайным процессом на $(\Omega, \mathcal{F}, \mathrm{P})$, то есть семейство $X = (X_t, t \in T)$ есть семейство случайных величин $\Leftrightarrow$ отображение $X : \Omega \to S$

является измеримым относительно цилиндрической $\sigma$-алгебры.
Для доказательства нам понадобится.

**Лемма (достаточное условие измеримости отображения)**
Пусть $(\Omega, \mathcal{F}, \mathrm{P})$ - вероятностное пространство, $(E, \mathcal{E})$ - измеримое пространство, $\mathcal{M} \subset \mathcal{E}$ - подсистема такая, что $\sigma(\mathcal{M}) = \mathcal{E}$. Пусть $X : \Omega \to E$. Если $\forall B \in \mathcal{M}$ $X^{-1}(B) \in \mathcal{F}$, то отображение $X$ измеримо, то есть

$$\forall B \in \mathcal{E}\ X^{-1} \in \mathcal{F}.$$

**Доказательство.(леммы об измеримости случайного процесса)**
($\Longleftarrow$) Надо доказать, что $\forall t$ $X_t$ - случайная величина. Для $\forall B_t \in \mathcal{B}(\mathbb{R})$ :

$$X_t^{-1}(B_t) = \{\omega : X_t(\omega) \in B_t\} = \{\omega : X(\omega) \in C(t, B_t)\} = X^{-1}(C(t, B_t)) \in \mathcal{F},$$

так как $X$ измерим.
($\Longrightarrow$) Воспользуемся достаточным условием измеримости, взяв в качестве $\mathcal{M}$ множество всех элементарных цилиндров.
Тогда $\sigma(\mathcal{M}) = \mathcal{B}_T$ - цилиндрическая $\sigma$-алгебра,

$$X^{-1}(C(t, B_t)) = \{\omega : X(\omega) \in C(t, B_t)\} = \{\omega : X_t(\omega) \in B_t\} = X_t^{-1}(B_t) \in \mathcal{F},$$

так как $X_t$ - случайная величин. ■

## Распределение случайного процесса

**Замечание**
Лемма устанавливает эквивалентное определение случайного процесса как единого случайного элемента со значениями в пространстве траекторий, измеримого относительно цилиндрической $\sigma$-алгебры.

**Определение 18.9.** Распределение $\mathrm{P}_X$ случайного процесса $X = (X_t, t \in T)$ называется вероятностная мера на $(S, \mathcal{B}_T)$ такая, что

$$\forall C \in \mathcal{B}_T\ \mathrm{P}_X(C) = \mathrm{P}(X \in C).$$

Это определение осязаемо только в том случае, когда "время"конечно и мы фактически имеем дело со случайным вектором. Для счетного (или тем более континуального) "времени"это определение весьма непросто для понимания в связи с отсутствием функции распределения.

# Сходимость случайных процессов

Вернемся к сходимостям. Будем рассматривать время $T = [0, 1]$.
Пусть $X^{(n)} = (X_t^{(n)}, t \in [0, 1]), n \in \mathbb{N}$ - последовательность действительных случайных процессов на $[0, 1]$. Как можно было определить сходимость по распределению такой последовательности?
**Идея 1:** как слабую сходимость распределений!

Мы поняли, что каждый случайный процесс измерим относительно цилиндрической $\sigma$-алгебры на пространстве траекторий. Пусть $F[0, 1]$ - множество вещественных функций на $[0, 1]$, а $\mathcal{C}(F[0, 1])$ - цилиндрическая $\sigma$-алгебра на нем. Тогда для процесса $X = (X_t, t \in [0, 1])$ распределение $\mathrm{P}_X$ есть вероятностная мера на $(F[0, 1], \mathcal{C}(F[0, 1]))$, определенная по правилу:

$$\mathrm{P}_X(C) = \mathrm{P}(X \in C) \text{ для любого } C \in \mathcal{C}(F[0, 1]).$$

**Замечание**
Однако пространство $F[0, 1]$ не является метрическим, в $\mathcal{C}(F[0, 1])$ - не является борелевской $\sigma$-алгеброй.

**Идея 2:** а пусть процесс $X = (X_t, t \in [0, 1])$ имеет непрерывные траектории, тогда его пространством траекторий будет $C[0, 1]$, которое отлично метризуется нормой

$$\|x\| = \max_{t \in [0,1]} |x(t)|.$$

**Замечание**

1) $C[0, 1]$ не является элементом цилиндрической $\sigma$-алгебры $\mathcal{C}(F[0, 1])$;

2) $X$, вообще говоря, не измерим относительно $\mathcal{B}(C[0, 1])$.

Решением этих проблем является рассмотрение цилиндрической $\sigma$-алгебра, суженой на $C[0, 1]$:

$$\mathcal{C}(C[0, 1]) = \mathcal{C}(F[0, 1]) \cap C[0, 1] = \{B \cap C[0, 1] : \ B \in \mathcal{C}(F[0, 1])\}.$$

**Лемма**
$\mathcal{B}(C[0, 1]) = \mathcal{C}(C[0, 1])$.

**Доказательство.**
Сначала докажем, что $\mathcal{B}(C[0, 1]) \subseteq \mathcal{C}(C[0, 1])$. Для этого возьмем замкнутый шар

$B_r = \{u \in C[0,1] : \ \|y - x\| \leqslant r\}$. Тогда в силу непрерывности всех функций получаем, что

$$B_r(x) = \bigcap_{t \in \mathbb{Q} \cap [0,1]} \{y \in C[0,1] : \ |y(t) - x(t)| \leqslant r\}.$$

Но это есть ни что иное, как счетное пересечение элементарных цилиндров. Следовательно, $B_r(x) \in \mathcal{C}(C[0,1])$. Шары порождают всю борелевскую $\sigma$-алгебру, поэтому $\mathcal{B}(C[0,1]) \subseteq \mathcal{C}(C[0,1])$.

Теперь докажем обратное вложение. Для этого рассмотрим следующий цилиндр: $t \in [0,1], a < b$,

$$Cyl(t,[a,b]) = \{y \in C[0,1] : \ y(t) \in [a,b]\}.$$

Введем функцию $h_t : C[0,1] \mapsto \mathbb{R}$, действующую по правилу $h_t(x) = x(t)$. Это непрерывная функция, и $Cyl(t,[a,b]) = h_t^{-1}([a,b])$. Отрезок $[a,b]$ замкнут, поэтому $Cyl(t,[a,b])$ является замкнутым множеством в $C[0,1]$.

Следовательно, $Cyl(t,[a,b]) \in \mathcal{B}(C[0,1])$. Подобные цилиндры порождают цилиндрическую $\sigma$-алгебру, поэтому $\mathcal{C}(C[0,1]) \subseteq \mathbb{B}(C[0,1])$. ■

**Определение 18.10.** Пусть $X^{(n)} = (X_t^{(n)}, t \in [0,1]), n \in \mathbb{N}$, и $X = (X_t, t \in [0,1])$ - случайные процессы с непрерывными траекториями. Тогда последовательность $X^{(n)}$ сходится по распределению к $X$, если их распределения сходятся слабо как вероятностные меры на $C[0,1]$ :

$$\mathrm{P}_{X^{(n)}} \xrightarrow{\omega} \mathrm{P}_X \text{ при } n \to \infty.$$

Обозначение: $X^{(n)} \xrightarrow{\mathcal{D}} X$.

**Теорема 18.2. о наследовании сходимости**
Пусть $h : C[0,1] \mapsto \mathbb{R}(\mathbb{R}^k)$ - непрерывная функция и $X^{(n)} \xrightarrow{\mathcal{D}} X$ при $n \to \infty$. Тогда $h(X^{(n)}) \xrightarrow{d} h(X)$ (как случайные величины или векторы).

**Доказательство.**
Пусть $f : \mathbb{R}^k \mapsto \mathbb{R}$ - ограниченная непрерывная функция. Тогда композиция $g = f \odot h : C[0,1] \mapsto \mathbb{R}$ также будет ограниченной непрерывной функцией. Рассмотрим $\mathrm{E}\, f(h(X^{(n)}))$. Оно равно

$$\mathrm{E} f(h(X^{(n)})) = \mathrm{E} g(X^{(n)}) = \int\limits_{C[0,1]} g(x) \mathrm{P}_{X^{(n)}}(dx).$$

Воспользуемся слабой сходимостью распределений:

$$\lim_{n \to \infty} \int\limits_{C[0,1]} g(x) \mathrm{P}_{X^{(n)}}(dx) = \int\limits_{C[0,1]} g(x) \mathrm{P}_X(dx) = \mathrm{E} g(X).$$

Следовательно, для любой ограниченной непрерывной функции $f : \mathbb{R}^k \mapsto \mathbb{R}$ выполнена следующая сходимость: $\mathrm{E} f(h(X^{(n)})) \to \mathrm{E} f(h(X))$ при $n \to \infty$. Но это означает, что $h(X^{(n)}) \xrightarrow{d} h(X)$. ■

# Принцип инвариантности

## Винеровский процесс

Центральным случайным процессом в теории вероятностей является винеровский.

**Определение 18.11.** Случайный процесс $(W_t, t \geqslant 0)$ называется винеровским (процессом броуновского движения), если

1) $W_0 = 0$ почти наверное;

2) $W_t$ имеет независимые приращения, то есть для любого $n \in \mathbb{N}$ и любых $0 \leqslant t_1 < \cdots < t_n$ случайные величины $W_{t_1}, W_{t_2} - W_{t_1}, \ldots, W_{t_n} - W_{t_n - 1}$ независимы в совокупности;

3) $W_t - W_s \sim \mathcal{N}(0, t-s),\ \forall t \geqslant s \geqslant 0.$

4) Траектории $W_t$ непрерывны почти наверное.

Из свойств винеровского процесса отметим, что $W_t$ является гауссовским, то есть для любого $n \in \mathbb{N}$ и любых $t_1, \ldots, t_n \in \mathbb{R}_+$ вектор $(W_{t_1}, W_{t_2}, \ldots, W_{t_n})$ является гауссовским.

Ответ на вопрос о том, где возникает сходимость по распределению случайных процессов дает следующая теорема.

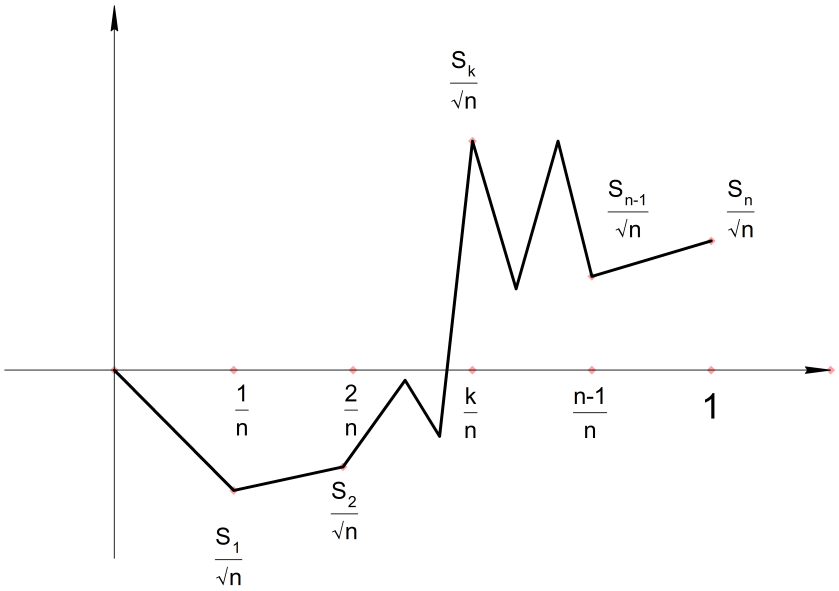


Рис. 7

**Теорема 18.3. (принцип инвариантности Донскера-Прохорова)**
Пусть $\{\xi_n, n \in \mathbb{N}\}$ - последовательность независимых и одинаково распределенных случайных величин с условием $\mathrm{E}\xi_1 = 0$ и $\mathrm{D}\xi_1 = 1$. Введем последовательность случайных величин $\{S_n, n \in \mathbb{N}\}$ по следующему правилу:

$$X^{(n)} = (X_t^{(n)}, t \in [0,1]), \text{ как линейную интерполяцию } S_0, \dots, S_n :$$

$$X^{(n)} = \frac{S_k}{\sqrt{n}}(k+1-nt) + \frac{S_{k+1}}{\sqrt{n}}(nt-k) \text{ при } t \in [\frac{k}{n}, \frac{k+1}{n}], k = 0, \dots, n-1.$$

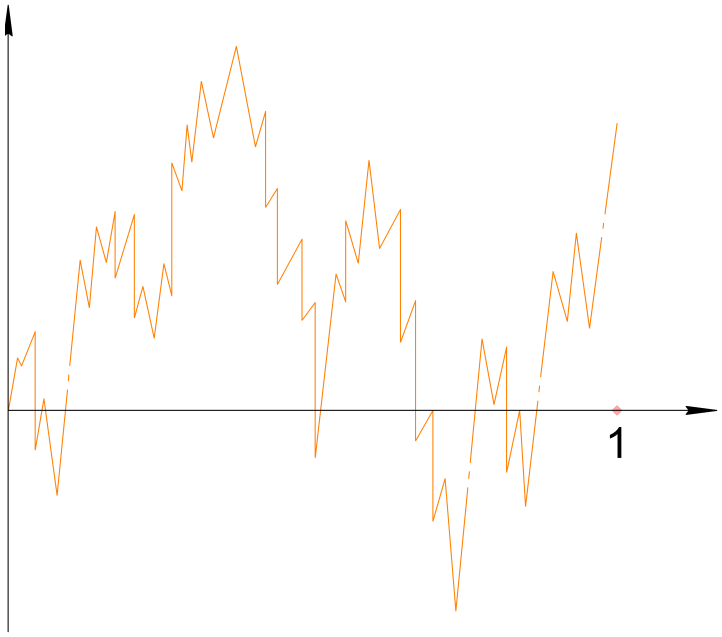


Рис. 8: Винеровский процесс

Тогда $X^{(n)} \xrightarrow{\mathcal{D}} W,$ где $W = (W_t, t \in [0,1])$ - винеровский процесс.

**Замечание**
Центральная предельная теорема является простым следствием принципа инвариантности.

# 19. Лекция 19

## Теорема Колмогорова

Пусть $(X_n, n \in \mathbb{N})$ - выборка неограниченного размера из распределения с функцией распределения $F(x)$. Построим по ней эмпирическую функцию распределения:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}\{X_j \leqslant x\}.$$

Следующая теорема играет важнейшую роль в статистике для обоснования применения критерия Колмогорова.

**Теорема 19.1. (Колмогоров)**
Если $F(x)$ непрерывна, то распределение случайной величины

$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)|$$

не зависит от $F(x)$. Более того,

$$\sqrt{n} \cdot D_n \xrightarrow{d} K,$$

где $K$ - распределение Колмогорова с ф.р. $K(y) = \sum\limits_{k\in\mathbb{Z}} (-1)^k e^{-2k^2y^2}$, $y > 0$.

**Комментарий**
Если сравнивать теорему Колмогорова с теоремой Гливенко-Кантелли, то здесь аналогия примерно та же, что и сравнение УЗБЧ с ЦПТ. То есть, согласно теореме Гливенко-Кантелли мы знаем, что $D_n$ п.н. сходится к нулю. Теперь мы показываем, с какой скоростью статистика $D_n$ стремится к нулю.

### Доказательство теоремы Колмогорова

Разобьем доказательство на несколько шагов.

**Лемма (1)**
Распределение случайной величины $D_n$ не зависит от $F(x)$.

**Доказательство.**
Рассмотрим другую выборку: обозначим $Y_j = F(X_j)$. Тогда случайные величины $Y_1, \ldots, Y_n$ будут независимыми с распределением $\mathrm{U}(0,1)$: для любого $y \in (0,1)$

$$\mathrm{P}(Y_j \leqslant y) = \mathrm{P}(F(X_j) \leqslant y) = \mathrm{P}(X_j \leqslant F^{-1}(y)) = F(F^{-1}(y)) = y,$$

где $F^{-1}(y) = \min\{z : F(z) = y\}$. Далее, заметим, что с вероятностью 1

$$\mathrm{I}\{X_j \leqslant x\} = \mathrm{I}\{F(X_j) \leqslant F(x)\} \text{ для всех } x \in \mathbb{R}.$$

Отсюда с вероятностью 1 получаем равенства

$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| = \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}\{X_j \leqslant x\} - F(x) \right| =$$

$$= \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}\{Y_j \leqslant F(x)\} - F(x) \right| = \sup_{y\in(0,1)} \left| \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}\{Y_j \leqslant y\} - y \right|.$$

Распределение последней случайной величины не зависит от $F(x)$. ■

**Лемма (2)**
Пусть $(W_t, t \in [0,1])$ - винеровский процесс. Тогда

$$\sqrt{n} D_n \xrightarrow{d} \max_{t\in[0,1]} |W_t - tW_1|.$$

**Доказательство.**
В силу леммы (1) можно считать, что $(X_n, n \in \mathbb{N})$ - это выборка из $\mathrm{U}(0,1)$. Поймем, где может достигаться супремум в определении $D_n$. Раз функция $\hat{F}_n(x)$ является кусочно-постоянной, то супремум разности $|\hat{F}_n(x) - x|$ достигается в точках разрыва. Следовательно,

$$D_n = \max\left( \max_{k=1,\dots,n} \left| X_{(k)} - \frac{k}{n} \right|, \max_{k=1,\dots,n} \left| X_{(k)} - \frac{k-1}{n} \right| \right).$$

Попробуем изобразить то, как получается это равенство. Нарисуем графики эмпирической и настоящей функций распределения. Выборка $(X_1, \dots, X_n) \sim \mathrm{U}(0,1)$, поэтому $F(x) = x$. Эмпирическая функция - кусочно-постоянная функция, которая меняет свое значение в точках порядковых статистик нашей выборки. Значения будут равны: $0, \frac{1}{n}, \frac{2}{n}$, и так далее (см. Рис. 9).
Теперь мы хотим найти значение модуля разности. В точке порядковой статистики значение эмпирической функции распределения равно в точности $\frac{k}{n}$. Соответственно, мы получаем разность $|x_{(k)} - \frac{k}{n}|$, где $x_{(k)}$ - значение настоящей функции распределения. С другой стороны, можно рассмотреть предел слева, тогда разность будет иметь вид $|\frac{k}{n} - x_{(k+1)}|$. Видим, что при некотором $k$ одно из этих двух значений должно давать максимум, так как эмпирическая функция распределения постоянна в окрестности, а настоящая функция $F(x) = x$ растет, поэтому, уменьшая или увеличивая $x$, мы можем добиваться больших значений по модулю. Отсюда получается такое представление для статистики $D_n$.

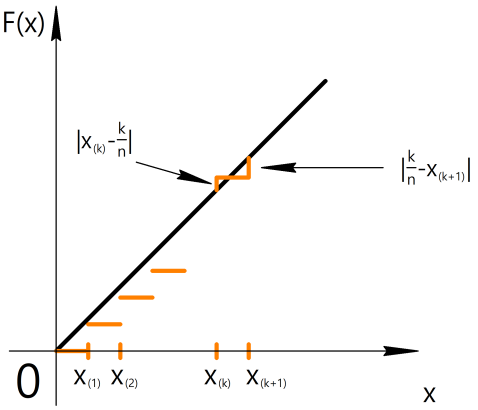


Рис. 9

Обозначим $T_n = \max\limits_{k=1,\dots,n} |X_{(k)} - \frac{k}{n}|$. Тогда $|T_n - D_n| \leqslant \frac{1}{n}$ и нам достаточно показать, что $\sqrt{n} \cdot T_n \xrightarrow{d} \max\limits_{t\in[0,1]} |W_t - tW_1|$.

Воспользуемся далее следующим упражнением.

**Упражнение**
Пусть $\xi_1, \dots, \xi_{n+1}$ - это независимые и одинаково распределенные случайные величины с экспоненциальным распределением Exp(1), а $S_k = \xi_1 + \dots + \xi_k$. Тогда

$$(X_{(1)}, \dots, X_{(n)}) \stackrel{d}{=} \left(\frac{S_1}{S_{n+1}}, \dots, \frac{S_n}{S_{n+1}}\right).$$

Значит, распределение $\sqrt{n} \cdot T_n$ равно распределению

$$\sqrt{n} \cdot \max_{k=1,\dots,n} \left|\frac{S_k}{S_{n+1}} - \frac{k}{n}\right|.$$

Теперь заметим, что в пределе распределение выражения выше будет равно предельному распределению величины

$$\Delta_n = \sqrt{n+1} \cdot \max_{k=1,\dots,n} \left|\frac{S_k}{S_{n+1}} - \frac{k}{n+1}\right|.$$

Преобразуем выражение:

$$\Delta_n = \frac{n+1}{S_{n+1}} \cdot \max_{k=1,\dots,n} \left|\frac{S_k}{\sqrt{n+1}} - \frac{kS_{n+1}}{(n+1)\sqrt{n+1}}\right| =$$

$$= \frac{n+1}{S_{n+1}} \cdot \max_{k=1,\dots,n} \left|\frac{S_k - k}{\sqrt{n+1}} - \frac{k}{n+1}\frac{S_{n+1} - (n+1)}{\sqrt{n+1}}\right|.$$

Теперь введем случайный процесс $Y^{(n)} = (Y_t^{(n)}, t \in [0,1])$ по следующему правилу: если $t \in \left[\frac{k}{n+1}, \frac{k+1}{n+1}\right]$, $k = 0, \ldots, n$, то

$$Y_t^{(n)} = \frac{S_k - k}{\sqrt{n+1}}(k+1-(n+1)t) + \frac{S_{k+1} - (k+1)}{\sqrt{n+1}}((n+1)t - k).$$

Это в точности кусочно-линейный процесс из принципа инвариантности, так как $\mathrm{E}\xi_i = 1$ и $\mathrm{D}\xi_i = 1$.

Принцип инвариантности говорит нам, что данный процесс будет сходиться по распределению к винеровскому процессу, но нам нужно посмотреть на процесс, который участвует в величине $\Delta_n$: $\frac{S_k - k}{\sqrt{n+1}}$ - это значение нашего процесса в точке $\frac{k}{n+1}$, а $\frac{S_{n+1} - (n+1)}{\sqrt{n+1}}$ - значение в 1. То есть мы рассматриваем процесс $Y_t^{(n)} - tY_t^{(n)}$ и берем его максимальные по модулю значения в точках вида $\frac{k}{n+1}$. Поскольку сам $Y_t^{(n)}$ - это кусочно-линейный процесс, где вершины ломанной - точки вида $\frac{k}{n+1}$, то $Y_t^{(n)} - tY_t^{(n)}$ - это тоже процесс, у которого траектория - ломанная линия, причем вершины этой ломанной будут теми же самыми точками $\frac{k}{n_1}$. Максимальное значение по модулю у траектории, которая является ломанной, достигается в одной из вершин, поэтому

$$\Delta_n = \frac{n+1}{S_{n+1}} \max_{t \in [0,1]} |Y_t^{(n)} - tY_1^{(n)}|.$$

Согласно УЗБЧ выполнено

$$\frac{n+1}{S_{n+1}} \xrightarrow{d} 1.$$

Согласно принципу инвариантности выполнено $Y^{(n)} \xrightarrow{\mathcal{D}} W$, а по теореме о наследовании сходимости получаем, что

$$\max_{t \in [0,1]} |Y_t^{(n)} - tY_1^{(n)}| \xrightarrow{d} \max_{t \in [0,1]} |W_t - tW_1|.$$

Наконец, из леммы Слуцкого выводим искомое утверждение:

$$\Delta_n \xrightarrow{d} \max_{t \in [0,1]} |W_t - tW_1|.$$

■

Отметим, что тот процесс, который мы рассматривали в рамках доказательства леммы (2) имеет специальное название.

**Определение 19.1.** Случайный процесс $W^0 = (W_t^0, t \in [0,1])$, где $W_t^0 = W_t - tW_1$, называется броуновским мостом.

Мы ответили на основной вопрос: к чему сходится по распределению статистика, которая участвует в формулировке теоремы Колмогорова. Остается вычислить распределение модуля максимума броуновского моста.

**Вопрос:** как можно вычислить распределение $\max\limits_{t\in[0,1]}|W_t^0|$?

Попробуем понять, как можно по-другому получить распределение броуновского моста, а именно, как условное распределение винеровского процесса при условии, что $W_1 = 0$. Для любого $\varepsilon > 0$ и $B \in \mathcal{B}([0,1])$ положим

$$\mathrm{Q}_\varepsilon(B) = \mathrm{P}(W \in B \mid W_1 \in [-\varepsilon, \varepsilon]).$$

**Лемма (3)**
Пусть $\mathrm{P}_{W^0}$ - это распределение броуновского моста. Тогда

$$\mathrm{Q}_\varepsilon \xrightarrow{w} \mathrm{P}_{W^0} \text{ при } \varepsilon \to 0+.$$

**Доказательство.**
По теореме Александрова достаточно проверить, что для любого замкнутого $F \in \mathcal{B}(C[0,1])$ выполняется неравенство:

$$\overline{\lim_{\varepsilon\to 0+}}\, \mathrm{Q}_\varepsilon(F) \leqslant \mathrm{P}_{W^0}(F) = \mathrm{P}(W^0 \in F).$$

Сначала проверим, что процесс $W^0$ независим с $W_1$. Действительно, для любого $t \in [0,1]$ :

$$\mathrm{cov}(W_t^0, W_1) = \mathrm{cov}(W_t - tW_1, W_1) = 0.$$

Таким образом, для любого $m \in \mathbb{N}$ и для любых $t_1, \ldots, t_m \in [0,1]$ вектор $(W_{t_1}^0, \ldots, W_{t_m}^0)$ независим с $W_1$. Следовательно, процесс $W^0$ независим с $W_1$.

Далее, рассмотрим расстояние от броуновского моста до винеровского процесса:

$$\|W^0 - W\|_{C[0,1]} = \max_{t\in[0,1]} |W_t - (W_t - tW_1)| = \max_{t\in[0,1]} |tW_1| = |W_1|.$$

Следовательно, если $|W_1| < \delta$ и $W \in F$, то $W^0$ обязательно принадлежит множеству $F^\delta = \{y \in C[0,1] : \rho(y, F) \leqslant \delta\}$. Тогда для любого $\varepsilon < \delta$ выполнено

$$\mathrm{Q}_\varepsilon(F) = \mathrm{P}(W \in F \mid |W_1| \leqslant \varepsilon) \leqslant \mathrm{P}(W^0 \in F^\delta \mid |W_1| \leqslant \varepsilon) = \mathrm{P}(W^0 \in F^\delta),$$

где мы в последнем равенстве воспользовались независимостью $W^0$ и $W_1$. Стало быть,

$$\overline{\lim_{\varepsilon\to 0+}}\, \mathrm{Q}_\varepsilon(F) \leqslant \mathrm{P}(W^0 \in F^\delta) = \mathrm{P}_{W^0}(F^\delta).$$

В силу замкнутости $F$ выполнено $F^\delta \downarrow F$ при $\delta \downarrow 0$. По теореме о непрерывности вероятностной меры $\mathrm{P}_{W^0}(F^\delta) \to \mathrm{P}_{W^0}(F)$ при $\delta \to 0$, что и дает искомое неравенство.
■

**Следствие (1)**
Для любого $z > 0$

$$\mathrm{P}\left(\max_{t\in[0,1]}|W_t^0| \leqslant z\right) = \lim_{\varepsilon\to 0+}\mathrm{P}\left(\max_{t\in[0,1]}|W_t| \leqslant z \;\middle|\; W_1 \in [-\varepsilon,\varepsilon]\right) =$$

$$= \lim_{\varepsilon\to 0+}\frac{\mathrm{P}\left(\max\limits_{t\in[0,1]} W_t \leqslant z,\ \min\limits_{t\in[0,1]} W_t \geqslant z,\ W_1 \in [-\varepsilon,\varepsilon]\right)}{\mathrm{P}(W_1 \in [-\varepsilon,\varepsilon])}.$$

Остается найти совместное распределение $(m, M, W_1)$, где

$$m = \min_{t\in[0,1]} W_t, \quad M = \max_{t\in[0,1]} W_t.$$

Для этого мы решим аналогичную задачу для простейшего симметричного случайного блуждания и применим принцип инвариантности. Пусть $(S_n, n \in \mathbb{N})$ - простейшее симметричное случайное блуждание ($S_n$ представим в виде суммы $\xi_1 + \cdots + \xi_n$, где $\xi_i$ - независимые одинаково распределенные случайные величины, которые принимают значение $\pm 1$ с вероятностью $\frac{1}{2}$). Положим

$$m_n = \min_{0\leqslant k\leqslant n} S_k, \quad M_n = \max_{0\leqslant k\leqslant n} S_k.$$

**Лемма (4)**
Для любых $a \leqslant 0 \leqslant b$, $a < b$, $a < v < b$ - целых чисел, выполнено

$$\mathrm{P}(a < m_n \leqslant M_n < b,\ S_n = v) =$$

$$= \sum_{k\in\mathbb{Z}}\mathrm{P}(S_n = v + 2k(b-a)) - \sum_{k\in\mathbb{Z}}\mathrm{P}(S_n = 2b - v + 2k(b-a)), \tag{1}$$

где мы считаем, что $S_0 = 0$.

**Доказательство.**
Будем доказывать индукцией по $n$. Пусть $n = 0$. Тогда

1) Если $v = 0$, $a < 0 < b$, то левая часть равна 1, так как выполняются условия $S_0 = 0$ и $a < 0 \leqslant 0 < b$. В правой части:
   - Первая сумма равна 1, так как $v = 0$, а $2k(b-a)$ равно нулю только в одном случае: $k = 0$.
   - Вторая сумма равна 0, так как $2b > 0$, $v = 0$, и $2k(b-a) \geqslant 0$, поэтому $S_n$ не может быть равно 0.

2) Если $v \neq 0$, $a < 0 < b$, то левая часть равна 0, так как $S_0 \neq 0$. В правой части:
   - Первая сумма равна 0, так как если $k = 0$, то $S_0 = v \neq 0$. Если же $k > 0$, то $S_0 = v + 2k(b-a) > 0$.

- Вторая сумма равна 0, так как $2b - v > 0$, следовательно $S_0 > 0$.

3) Если $a$ или $b$ равны нулю, то левая часть равна нулю, так как в $v \neq 0$. В правой части обе суммы будут одинаковыми, так как, например при $a = 0$, вероятность во второй сумме - это по сути то же самое, что вероятность того, что $S_n = -v + 2kb$, так как добавление $2b$ - сдвиг индекса $k$ на 1, а $S_n$ - симметричное случайное блуждание, поэтому знак при $v$ не имеет значения в этом случае.

Пусть формула (1) выполнена для $n-1$, докажем ее для $n$. Если $a = 0$ или $b = 0$, то снова левая часть равна нулю, а в правой части обе суммы одинаковы.

Так что пусть $a < 0 < b$. Обозначим

$$p_n(a, b, v) = \mathrm{P}(a < m_n \leqslant M_n < b,\ S_n = v), \quad p_n(j) = \mathrm{P}(S_n = j).$$

Тогда заметим, что

$$p_n(a, b, v) = \frac{1}{2}p_{n-1}(a-1, b-1, v-1) + \frac{1}{2}p_{n-1}(a+1, b+1, v+1) \boxed{=}$$

(предположение индукции)

$$\boxed{=} \frac{1}{2}\left[\sum_{k\in\mathbb{Z}} p_{n-1}(v - 1 + 2k(b-a)) - \sum_{k\in\mathbb{Z}} p_{n-1}(2b - v + 1 + 2k(b-a))\right] +$$

$$+\frac{1}{2}\left[\sum_{k\in\mathbb{Z}} p_{n-1}(v + 1 + 2k(b-a)) - \sum_{k\in\mathbb{Z}} p_{n-1}(2b - v - 1 + 2k(b-a))\right] \boxed{=}$$

(так как $p_n(j) = \frac{1}{2}p_{n-1}(j-1) + \frac{1}{2}p_{n-1}(j+1)$)

$$\boxed{=} \sum_{k\in\mathbb{Z}} p_n(v + 2k(b-a)) - \sum_{k\in\mathbb{Z}} p_n(2b - v + 2k(b-a)).$$

■

Выведем следствие из леммы (4). Мы хотим расширить диапазон значений $S_n$, чтобы нам было проще переходить к нормальному закону.

**Следствие (2)**
Для любых целых чисел $a \leqslant 0 \leqslant b$, $a < b$, $a \leqslant u_1 < u_2 \leqslant b$ выполнено

$$\mathrm{P}(a < m_n \leqslant M_n < b,\ u_1 < S_n < u_2) =$$

$$= \sum_{k\in\mathbb{Z}} \mathrm{P}(u_1 + 2k(b-a) < S_n < u_2 + 2k(b-a)) -$$

$$-\sum_{k\in\mathbb{Z}} \mathrm{P}(2b-u_2+2k(b-a)<S_n<2b-u_1+2k(b-a)).$$

Получается суммированием по значениям $v\in(u_1,u_2)$ из леммы (4).

Теперь мы готовы сформулировать лемму о совместном распределении $(m,M,W_1)$.

**Лемма (5)**
Для любых вещественных чисел $a<0<b$, $a<u<v<b$ выполнено

$$\mathrm{P}(a<m\leqslant M<b,\ u<W_1<v)=$$

$$=\sum_{k\in\mathbb{Z}} \mathrm{P}(u+2k(b-a)<\xi<v+2k(b-a))-$$

$$-\sum_{k\in\mathbb{Z}} \mathrm{P}(2b-v+2k(b-a)<\xi<2b-u+2k(b-a)),$$

где $\xi\sim\mathcal{N}(0,1)$.
**Доказательство.**
В силу принципа инвариантности и наследования сходимости получаем, что

$$\left(\frac{m_n}{\sqrt{n}},\frac{M_n}{\sqrt{n}},\frac{S_n}{\sqrt{n}}\right)\xrightarrow{d}(m,M,W_1).$$

Тогда

$$\mathrm{P}(a<m\leqslant M<b,u<W_1<v)=$$

$$=\lim_{n\to+\infty}\mathrm{P}\left(a<\frac{m_n}{\sqrt{n}}\leqslant\frac{M_n}{\sqrt{n}}<b,\ u<\frac{S_n}{\sqrt{n}}<v\right).$$

Сделаем границы целыми: обозначим $a'=[a\sqrt{n}]$, $b'=[b\sqrt{n}]$, $u'=[u\sqrt{n}]$, $v'=[v\sqrt{n}]$. Тогда, согласно следствию, получаем:

$$\mathrm{P}(a<m\leqslant M<b,u<W_1<v)=$$

$$=\lim_{n\to+\infty}\left(\sum_{k\in\mathbb{Z}}\mathrm{P}(u'+2k(b'-a')<S_n<v'+2k(b'-a'))-\right.$$

$$\left.-\sum_{k\in\mathbb{Z}}\mathrm{P}(2b'-u'+2k(b'-a')<S_n<2b'-v'+2k(b'-a'))\right).$$

Заметим, что при фиксированном $k$ в силу центральной предельной теоремы выполнено:

$$\lim_{n\to+\infty}\mathrm{P}(u'+2k(b'-a')<S_n<v'+2k(b'-a'))=$$

$$=\lim_{n\to+\infty}\mathrm{P}\left(\frac{u'+2k(b'-a')}{\sqrt{n}}<\frac{S_n}{\sqrt{n}}<\frac{v'+2k(b'-a')}{\sqrt{n}}\right)=$$

$$= \mathrm{P}(u + 2k(b-a) < \xi < v + 2k(b-a)).$$

Мы хотим переставить предел и суммы местами. Необходимо проверить, что ряды сходятся равномерно по $n$. Мы знаем, что $\frac{S_n}{\sqrt{n}} \xrightarrow{d} \xi$. Тогда для любого $\varepsilon > 0$ найдется такое $k_0 = k_0(\varepsilon)$, что для любого $n$ выполнено:

$$\sum_{k:|k|<k_0} \mathrm{P}\left(\frac{u' + 2k(b'-a')}{\sqrt{n}} < \frac{S_n}{\sqrt{n}} < \frac{v' + 2k(b'-a')}{\sqrt{n}}\right) \leqslant$$

$$\leqslant \mathrm{P}\left(\left|\frac{S_n}{\sqrt{n}}\right| \geqslant 2k_0(b-a) - |a| - 1\right) \leqslant \varepsilon.$$

Переставляя сумму по $k$ и предел по $n$ местами, получаем искомое равенство:

$$\mathrm{P}(a < m \leqslant M < b,\ u < W_1 < v) =$$

$$= \sum_{k\in\mathbb{Z}} \mathrm{P}(u + 2k(b-a) < \xi < v + 2k(b-a)) -$$

$$- \sum_{k\in\mathbb{Z}} \mathrm{P}(2b - v + 2k(b-a) < \xi < 2b - u + 2k(b-a)). -$$

■

Остается подставить правильные параметры в выражение из леммы (5) и посмотреть, куда вероятность при делении на вероятность попадания $W_1$ в отрезок $[-\varepsilon, \varepsilon]$ (хотим получить функцию распределения Колмогорова).

Найдем распределение случайной величины $\max\limits_{t\in[0,1]} |W_t^0|$. Согласно следствию (1) и лемме (5) для любого $z > 0$

$$\mathrm{P}\left(\max_{t\in[0,1]} |W_t^0| \leqslant z\right) = \lim_{\varepsilon\to 0+} \mathrm{P}\left(\max_{t\in[0,1]} |W_t| \leqslant z \;\middle|\; W_1 \in [-\varepsilon, \varepsilon]\right) =$$

$$= \lim_{\varepsilon\to 0+} \left[\frac{1}{\mathrm{P}(|W_1| \leqslant \varepsilon)} \sum_{k\in\mathbb{Z}} \mathrm{P}(4kz - \varepsilon < \xi < 4kz + \varepsilon) -\right.$$

$$\left. - \frac{1}{\mathrm{P}(|W_1| \leqslant \varepsilon)} \sum_{k\in\mathbb{Z}} \mathrm{P}((4k-2)z - \varepsilon < \xi < (4k-2)z + \varepsilon)\right].$$

Заметим, что

$$\mathrm{P}(4kz - \varepsilon < \xi < 4kz + \varepsilon) = \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(y+4kz)^2}{2}} dy =$$

$$= \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}} \cdot e^{-\frac{(4kz)^2}{2} - (4kz)y} dy \leqslant e^{-\frac{(4kz)^2}{2} + |(4kz)|\varepsilon} \cdot \mathrm{P}(|W_1| \leqslant \varepsilon).$$

Ряд $\sum\limits_{k\in\mathbb{Z}} e^{-\frac{(4kz)^2}{2}+|(4kz)|\varepsilon}$ сходится равномерно по $\varepsilon \leqslant 1$. Значит можно переставить местами сумму по $k$ и предел по $\varepsilon$.

Для фиксированного $k \in \mathbb{Z}$ и $\varepsilon \to 0$ имеем:

$$\mathrm{P}(4kz-\varepsilon<\xi<4kz+\varepsilon) \sim \frac{1}{\sqrt{2\pi}}\cdot 2\varepsilon \cdot e^{-\frac{(4kz)^2}{2}}, \text{ а } \mathrm{P}(|W_1| \leqslant \varepsilon) \sim \frac{1}{\sqrt{2\pi}}\cdot 2\varepsilon.$$

В итоге,

$$\mathrm{P}\left(\max_{t\in[0,1]}|W_t^0| \leqslant z\right) = \sum_{k\in\mathbb{Z}} e^{-\frac{(4kz)^2}{2}} - \sum_{k\in\mathbb{Z}} e^{-\frac{(4k-2)z)^2}{2}} =$$

$$= \sum_{k\in\mathbb{Z}} (-1)^k e^{-2k^2z^2} = K(z).$$

Доказательство теоремы Колмогорова завершено.

# 20. Лекция 20

Это последняя лекция курса Математической статистики, на которой мы познакомимся с элементами последовательного анализа - разделом математической статистики, который тесно связан с теорией случайных процессов.

## Мотивировка

Вернемся к задаче, с которой мы начинали рассматривать проверку гипотез. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из неизвестного распределения P, о котором высказаны дле простые гипотезы: $\mathrm{H}_0 : \mathrm{P} = \mathrm{P}_0$ и альтернатива $\mathrm{H}_1 : \mathrm{P} = \mathrm{P}_1$. Относительно распределений $\mathrm{P}_0$ и $\mathrm{P}_1$ известно, что они принадлежат одному доминируемому семейству и имеют плотности $p_0(x)$ и $p_1(x)$.

Казалось бы, все уже изучено. Согласно лемме Неймана-Пирсона р.н.м.к. уровня значимости $\varepsilon_0$ задается видом $S_\lambda = \{\mathbf{X} : p_1(\mathbf{X}) \geqslant \lambda p_0(\mathbf{X})\}$, а параметра $\lambda$ находится из уравнения

$$\mathrm{P}_0(\mathbf{X} \in S_\lambda) = \varepsilon_0.$$

Но есть одна неприятность, которая состоит в том, что мы мало что знаем о вероятности ошибки второго рода. Вероятность ошибки первого рода равна $\varepsilon_0$, однако вероятность ошибки второго рода $\varepsilon_1 = \mathrm{P}_1(\mathbf{X} \notin S_\lambda)$, хоть и имеет минимально возможное значение, может быть достаточно большой. Но в случае состоятельности критерия вероятность ошибки второго рода будет мала при больших $n$.

**Пример**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из схемы Бернулли $\mathrm{Bin}(1, \theta)$. Тогда р.н.м.к. уровня значимости $\varepsilon_0 = 0.05$ для проверки гипотезы $\mathrm{H}_0 : \theta = 0.05$ против альтернативы $\mathrm{H}_1 : \theta = 0.17$ будет иметь вероятность ошибки второго рода $\varepsilon_1 \leqslant 0.1$ начиная с $n = 57$.

Но наблюдения могут быть дороги в получении. С практической точки зрения задача минимизации затрат весьма принципиальна. Для ее решения был предложен так называемый метод последовательного анализа, в рамках которого данные поступают по одному, и мы на каждом шаге должны либо принять одну из гипотез, либо потребовать следующее наблюдение.

**Пример**
Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка, по которой мы хотим различить две гипотезы

$\mathrm{H}_0 : X_i \sim \mathrm{U}(1,3)$ против альтернативы $\mathrm{H}_1 : X_i \sim \mathrm{U}(0,2)$. Но если уже $X_1$ попадает в [2,3] или [0,1], то мы можем однозначно сказать, какая из гипотез верна. Остальные данные вообще не потребуются.

## Последовательные критерии

Итак, пусть P - неизвестное распределение элементов выборки, о котором высказаны две гипотезы: основная $\mathrm{H}_0 : \mathrm{P} = \mathrm{P}_0$ и альтернатива $\mathrm{H}_1 : \mathrm{P} = \mathrm{P}_1$. Пусть $\mathcal{X}$ - выборочное пространство для одного элемента выборки.

**Определение 20.1.** Последовательным критерием $S$ для проверки $\mathrm{H}_0$ против $\mathrm{H}_1$ называется такая последовательность пар $S = ((S_0^n, S_1^n),\ n \in \mathbb{N})$, что для любого $n \in \mathbb{N}$,

$$S_0^n, S_1^n \subset \mathcal{X}^n \text{ и } S_0^n \cap S_1^n = \emptyset.$$

При этом правильно принятия гипотез состоит в следующем. Пусть

$$\tau(S) = \min\{n : \mathbf{X}_n = (X_1, \ldots, X_n) \subset S_0^n \sqcup S_1^n\},$$

тогда гипотеза $\mathrm{H}_1$ принимается, если $\mathbf{X}_{\tau(S)} = (X_1, \ldots, X_{\tau(S)}) \in S_i^{\tau(S)}$, $i = 0, 1$.

**Замечание**
Обычные критерии являются частным случаем последовательных.

Изобразим разницу между обычными и последовательными критериями.

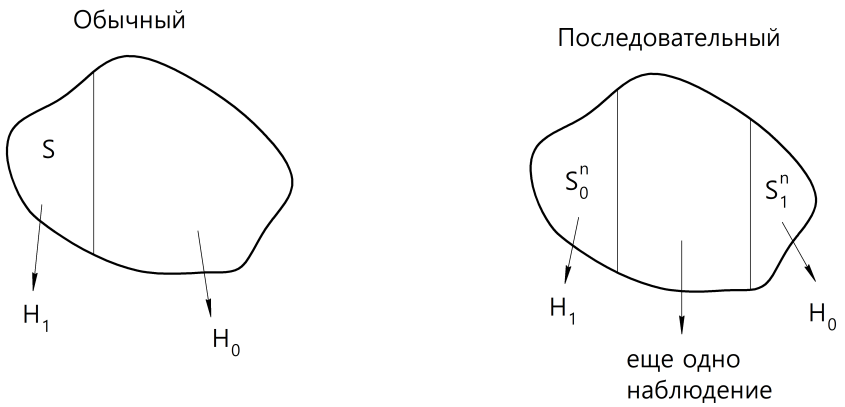


Рис. 10

В обычном критерии мы делили пространство на две части: множество $S$ (критерий), попадая в который, мы принимаем гипотезу $\mathrm{H}_1$, если мы попадаем не в $S$, то мы принимаем гипотезу $\mathrm{H}_0$.

В последовательном критерии выборочное пространство делится на 3 части: $S_0^n$, $S_1^n$ и дополнение к их объединению. Попадаем в $S_0^n$ - принимаем $\mathrm{H}_0$, попадаем в $S_1^n$ - принимаем $\mathrm{H}_1$, в противном случае мы требуем еще одно наблюдение.

Вероятности ошибок первого и второго рода для последовательного критерия равны

$$\alpha_0(S) = \mathrm{P}_0(\mathbf{X}_{\tau(S)} \in S_1^{\tau(S)}), \quad \alpha_1(S) = \mathrm{P}_1(\mathbf{X}_{\tau(S)} \in S_0^{\tau(S)})$$

### Критерий отношения правдоподобия

**Задача**
Для заданных $\varepsilon_0$, $\varepsilon_1 \in (0,1)$ найти последовательный критерий $S$ с условиями

$$\alpha_0(S) = \mathrm{P}_0(\mathbf{X}_{\tau(S)} \in S_1^{\tau(S)}) \leqslant \varepsilon_0, \quad \alpha_1(S) = \mathrm{P}_1(\mathbf{X}_{\tau(S)} \in S_0^{\tau(S)}) \leqslant \varepsilon_1,$$

который одновременно минимизирует величины $\mathrm{E}_i\tau(S)$, $i = 0, 1$.

Для решения задачи оптимальным будет критерий, который следует логике леммы Неймана-Пирсона и использует отношение правдоподобия. Пусть, как и ранее, распределение $\mathrm{P}_0$ и $\mathrm{P}_1$ взяты из одного доминируемого семейства и имеют плотности $p_0(x)$ и $p_1(x)$. Введем плотность выборки:

$$p_i(X_1, \ldots, X_n) = \prod_{j=1}^{n} p_i(X_j), \quad i = 0, 1.$$

**Определение 20.2.** Пусть $0 < \Gamma_0 < 1 < \Gamma_1$ - некоторые числа, $\Gamma = (\Gamma_0, \Gamma_1)$. Последовательным критерием отношения правдоподобия $S(\Gamma)$ называется последовательный критерий, у которого множества $S_0^n$ и $S_1^n$ определены следующим образом:

$$S_0^n = \left\{ \mathbf{X}_n : \frac{p_1(X_1, \ldots, X_n)}{p_0(X_1, \ldots, X_n)} \leqslant \Gamma_0 \right\},$$

$$S_1^n = \left\{ \mathbf{X}_n : \frac{p_1(X_1, \ldots, X_n)}{p_0(X_1, \ldots, X_n)} \geqslant \Gamma_1 \right\}.$$

## Байесовский подход к проверке гипотез

Мы хотим доказать, что для некоторого $\Gamma$ такой критерий будет оптимальным последовательным критерием. Однако, перед этим нам понадобиться установить, что такой же критерий будет оптимальным в байесовском подходе к последовательной проверке гипотез. Пусть у нас есть положительные числа $w_0, w_1, a$, которые имеют смысл "штрафов" :

1) $w_i$ - штраф за совершение ошибки $(i+1)-$го рода, $i = 0, 1$;

2) $a$ - штраф за взятие дополнительного наблюдения.

Пусть далее, у нас есть априорное распределение вероятностей на множестве гипотез. В нашем случае - это просто число $q \in [0, 1]$, так как гипотез всего две.

**Определение 20.3.** Байесовским риском критерия $S$ называется величина

$$R(q, S) = q[\alpha_0(S)w_0 + a\mathrm{E}_0\tau(S)] + (1-q)[\alpha_1(S)w_1 + a\mathrm{E}_1\tau(S)].$$

Последовательный критерий, который минимизирует (для заданных $q$, $w_0$, $w_1$, $a$) байесовский риск называется байесовским.

### Оптимальность в байесовском смысле

**Теорема 20.1.** Для любых $(q, w_0, w_1, a)$ существуют такие $\Gamma_1$ и $\Gamma_2$, что критерий $S(\Gamma)$ является байесовским.

**Доказательство.**
Обозначим через $\delta_i$ критерий, который принимает гипотезу $H_i$, $i = 0, 1$, без проведения наблюдений. Тогда $\tau(\delta_i) = 0$, $\alpha_i(\delta_i) = 0$. Поймем, когда критерий $S$, минимизирующий байесовский риск $R(q, S)$, совпадает с $\delta_i$. Очевидно, что

$$R(q, \delta_0) = (1-q)w_1, \quad R(q, \delta_1) = qw_0.$$

Пусть $\mathcal{K}$ - класс критериев, которые зависят хотя бы от одного наблюдения. Тогда $R(q, S) \geqslant a$ для любого $S \in \mathcal{K}$. Положим

$$R(q) = \inf_{S \in \mathcal{K}} R(q, S).$$

Ясно, что $R(q) < +\infty$, так как величина конечна для любого критерия, который требует ровно одно наблюдение.

Проверим, что $R(q)$ - выпуклая кверху функция. В силу линейности (по $q$) величины $R(q, S)$, получаем следующее: для любого $p \in (0, 1)$

$$R(pq_1 + (1-p)q_2, S) =$$

$$= (pq_1 + (1-p)q_2)[\alpha_0(S)w_0 + a\mathrm{E}_0\tau(S)] + (1 - pq_1 - (1-p)q_2)[\alpha_1(S)w_1 + a\mathrm{E}_1\tau(S)]$$

$$= pR(q_1, S) + (1-p)R(q_2, S).$$

Следовательно,

$$R(pq_1 + (1-p)q_2) = \inf_{S \in \mathcal{K}}[pR(q_1, S) + (1-p)R(q_2, S)] \geqslant pR(q_1) + (1-p)R(q_2).$$

Значит, функция $R(q)$ выпукла кверху. И так как $a \leqslant R(q) < +\infty$ мы получаем, что $R(q)$ непрерывна на $[0,1]$. Сравним теперь риски критериев $\delta_0$, $\delta_1$ и $S \in \mathcal{K}$.

Критерию $\delta_0$ соответствует функция байесовского риска $(1-q)w_1$, критерию $\delta_1$ соответствует $qw_0$, и есть $R(q) \geqslant a$ (значение байесовского риска, которое дает оптимальный критерий для заданного $q$).
Тогда на отрезке $[0,1]$ $\delta_0$ - убывающая функция, которая равна нулю в 1. $\delta_1$ - наоборот, возрастающая функция.
Дальше все зависит от того, как пройдет функция $R(q)$. Она ограничена снизу значением $a$.

1) Если точка пересечения прямых $\delta_0$ и $\delta_1$ попала внутрь $R(q)$, то наблюдений брать не нужно, так как точка пересечения - это точка $\frac{w_1}{w_0+w_1}$, в которой минимален максимальный байесовский риск из пары $\delta_0$,$\delta_1$. Соответственно, если в этой точке значение функции $R(q)$ больше, чем в точке пересечения, то есть

$$R\left(\frac{w_1}{w_0+w_1}\right) \geqslant \frac{w_0 w_1}{w_0+w_1},$$

то в силу выпуклости кверху функция $R(q)$ будет выше, чем минимум из пары рисков $\delta_0, \delta_1$. Это значит, что мы всегда предпочитаем первую или вторую гипотезу в зависимости от $q$: если $q \in [0, \frac{w_1}{w_0+w_1})$, то мы предпочитаем $\mathrm{H}_1$, если $q \in (\frac{w_1}{w_0+w_1}, 1]$, то $\mathrm{H}_0$.

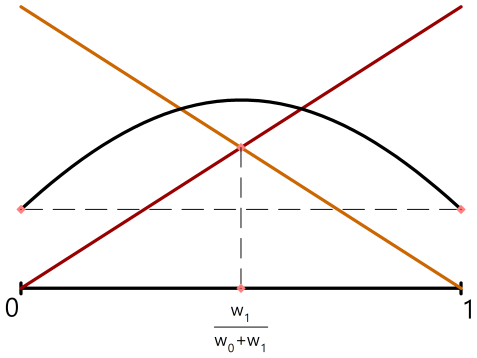


Рис. 11

2) Пусть теперь

$$R\left(\frac{w_1}{w_0+w_1}\right) < \frac{w_0 w_1}{w_0+w_1}.$$

Мы получаем интервал значений $q$, в котором нам предпочтительней взять еще одно наблюдение, чем пытаться выбрать кого-то из пары $\delta_0$, $\delta_1$. Этот интервал ограничен точками, в которых $R(q)$ пересекает прямые. Тогда, попадая левее или правее от этого интервала, мы предпочитаем, соответственно, $\delta_1$ или

$\delta_0$. В интервале оптимальным будет критерий, имеющий хотя-бы одно наблюдение.

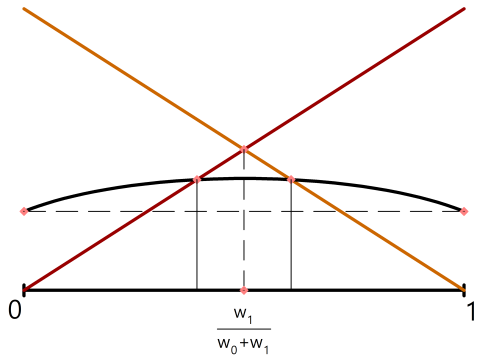


Рис. 12

Более формально эта ситуация выглядит так: существуют решения уравнений $R(q, \delta_i) = R(q)$, $i = 0, 1$. Обозначим их через $1 - \gamma_0 > 1 - \gamma_1$. В первом случае положим

$$1 - \gamma_0 = 1 - \gamma_1 = \frac{w_1}{w_0 + w_1}.$$

Тогда из приведенных рассуждений вытекает следующий оптимальный план действий. По заданным $w_0, w_1, a$ мы вычисляем величины $1 - \gamma_0$, $1 - \gamma_1$.

- Если $q \leqslant 1 - \gamma_1$ (или $\gamma_1 \leqslant 1 - q$), то наименьший риск имеет критерий $\delta_1$ и надо принять $\mathrm{H}_1$.
- Если $q \geqslant 1 - \gamma_0$ (или $\gamma_0 \geqslant 1 - q$), то наименьший риск имеет критерий $\delta_0$ и надо принять $\mathrm{H}_0$.
- Если же $q \in (1 - \gamma_1, 1 - \gamma_0)$ (или $1 - q \in (\gamma_0, \gamma_1)$), то оптимален кто-то из $\mathcal{K}$, то есть надо провести наблюдение.

Проведем теперь индукцию по $n$. Пусть проведено $n$ наблюдений и имеем выборку $(X_1, \ldots, X_n)$ Перед $(n+1)-$м наблюдением мы имеем ту же альтернативу: либо не проводить больше наблюдений и принять одну из гипотез, либо провести еще одно наблюдение. Тот факт, что риск уже увеличен на $an$, роли не играет, так как эти штрафные потери уже не могут быть устранены. Однако, теперь у нас будут другие априорные вероятности для гипотез.

По факту они превращаются в апостериорные вероятности параметра гипотезы при условии заданных $X_1, \ldots, X_n$:

$$q(0|\mathbf{X}_n) = \frac{qp_0(X_1, \ldots, X_n)}{qp_0(X_1, \ldots, X_n) + (1 - q)p_1(X_1, \ldots, X_n)},$$

$$q(1|\mathbf{X}_n) = \frac{(1-q)p_1(X_1,\ldots,X_n)}{qp_0(X_1,\ldots,X_n)+(1-q)p_1(X_1,\ldots,X_n)}.$$

Далее, мы уже знаем оптимальное решение:

- Если $q(0|\mathbf{X}_n) \leqslant 1-\gamma_1$, то надо принять $\mathrm{H}_1$.
- Если $q(0|\mathbf{X}_n) \geqslant 1-\gamma_0$, то надо принять $\mathrm{H}_0$.
- Если же $q(0|\mathbf{X}_n) \in (1-\gamma_1, 1-\gamma_0)$ (или $q(1|\mathbf{X}_n) \in (\gamma_0,\gamma_1)$), то надо провести наблюдение.

Таким образом, мы продолжаем наблюдения, если

$$\gamma_0 < \frac{(1-q)p_1(X_1,\ldots,X_n)}{qp_0(X_1,\ldots,X_n)+(1-q)p_1(X_1,\ldots,X_n)} < \gamma_1$$

Перепишем эти неравенства в терминах ограничений на отношение $p_1$ к $p_0$

$$\frac{q\gamma_0}{(1-q)(1-\gamma_0)} < \frac{p_1(X_1,\ldots,X_n)}{p_0(X_1,\ldots,X_n)} < \frac{q\gamma_1}{(1-q)(1-\gamma_1)}.$$

Но это и есть в точности критерий $S(\Gamma)$ для $\Gamma_i = \frac{q\gamma_i}{(1-q)(1-\gamma_i)}$, $i=0,1$. ■

**Замечание**
Заметим, что числа $\gamma_i = \gamma_i(w_0, w_1, a)$, $i=0,1$ остаются теми же, если все параметры будут умножены на одно и то же положительное число.

Тем самым, у нас есть лишь два параметра. Для удобства будем считать, что $w_1 = 1-w_0$. Важна следующая техническая лемма, которую, впрочем, мы оставим без доказательства.

**Лемма (1)**
Для любых $0<\gamma_0<\gamma_1<1$ найдутся такие $w\in(0,1)$ и $a>0$, что для любого $q\in(1-\gamma_1,1-\gamma_0)$ оптимальное решение байесовской задачи с параметрами $w_0 = 1-w$, $w_1 = w$, $a$ и априорной вероятностью $q$ задается последовательным критерием отношением правдоподобия $S(\Gamma)$, где

$$\Gamma_0 = \frac{q\gamma_0}{(1-q)(1-\gamma_0)}, \quad \Gamma_1 = \frac{q\gamma_1}{(1-q)(1-\gamma_1)}.$$

## Оптимальность последовательного критерия

Теперь мы готовы сформулировать и доказать теорему об оптимальности последовательного критерия отношения правдоподобия.

**Теорема 20.2. (об оптимальности)**
Пусть $\Gamma_0 < 1 < \Gamma_1$. Обозначим через $\varepsilon_i$, $i = 0, 1$, вероятности ошибок первого и второго родов последовательного критерия отношения правдоподобия $S(\Gamma)$. Тогда среди всех последовательных критериев $S$, для которых

$$\alpha_i(S) \leqslant \varepsilon_i, \quad i = 0, 1,$$

критерий $S(\Gamma)$ имеет наименьшие значения величин $\mathrm{E}_i\tau(S)$, $i = 0, 1$.

**Доказательство.**
Сведем задачу к байесовскому случаю. Возьмем произвольное $q \in (0, 1)$ и найдем величины $\gamma_0, \gamma_1$ из равенств:

$$\Gamma_i = \frac{q\gamma_i}{(1-q)(1-\gamma_i)}, \quad i = 0, 1.$$

Заметим, что в силу условия $\Gamma_0 < 1 < \Gamma_1$ мы автоматически получим, что

$$q \in (1 - \gamma_1, 1 - \gamma_0).$$

Тогда согласно лемме (1) найдутся такие параметры $a, w$, что критерий $S(\Gamma)$ будет байесовским в задаче с параметрами $(q, 1 - w, w, a)$.

Пусть $S$ - теперь любой другой критерий с условиями

$$\alpha_i(S) \leqslant \varepsilon_i, \quad i = 0, 1.$$

Сравним теперь $S$ и $S(\Gamma)$ в байесовском подходе с параметрами $(q, 1 - w, w, a)$. Мы знаем, что критерий $S(\Gamma)$ минимизирует байесовский риск, поэтому выполняется неравенство

$$q[\alpha_0(S)w_0 + a\mathrm{E}_0\tau(S)] + (1-q)[\alpha_1(S)w_1 + a\mathrm{E}_1\tau(S)] \geqslant$$

$$\geqslant q[\alpha_0(S(\Gamma))w_0 + a\mathrm{E}_0\tau(S(\Gamma))] + (1-q)[\alpha_1(S(\Gamma))w_1 + a\mathrm{E}_1\tau(S(\Gamma))].$$

Из этого неравенства и условий на $S$ следует, что

$$q\mathrm{E}_0\tau(S) + (1-q)\mathrm{E}_1\tau(S) \geqslant q\mathrm{E}_0\tau(S(\Gamma)) + (1-q)\mathrm{E}_1\tau(S(\Gamma)).$$

В силу произвольности $q \in (0,1)$ мы получаем искомые соотношения:

$$\mathrm{E}_0\tau(S) \geqslant \mathrm{E}_0\tau(S(\Gamma)), \quad \mathrm{E}_1\tau(S) \geqslant \mathrm{E}_1(\tau(S(\Gamma)).$$

■

Приведем важное замечание.

**Замечание**
Можно показать, что если распределения $\mathrm{P}_0$ и $\mathrm{P}_1$ непрерывны, то для любых $\varepsilon_0, \varepsilon_1 \in (0,1)$ найдутся такие $\Gamma_0 < 1 < \Gamma_1$, что последовательный критерий $S(\Gamma)$ будет иметь вероятности ошибок первого и второго рода в точности $\varepsilon_0$ и $\varepsilon_1$.

К сожалению, явное нахождение величин $\Gamma_0 < 1 < \Gamma_1$ при заданных вероятностях ошибок весьма затруднительно. По сути, оно представляет собой явное нахождение границ $(\Gamma_0, \Gamma_1)$, для которых уже заданы вероятности первого выхода случайного блуждания из полосы. Даже для схемы Бернулли необходима правильная связь между распределениями.

**Упражнение**
Рассмотрим задачу построения критерия последовательного отношения правдоподобия для выборки из биномиального распределения $\mathrm{Bin}(1,\theta)$, $\mathrm{H}_0 : \theta = p_0 < \frac{1}{2}$, $\mathrm{H}_1 : \theta = p_1 > \frac{1}{2}$. Положим $q_i = 1 - p_i$. Предположим, что $p_1 = q_0$ и числа $\Gamma_0$, $\Gamma_1$ подобраны так, что

$$\frac{\ln \Gamma_0}{\ln\left(\frac{q_0}{p_0}\right)} = -a\frac{\ln \Gamma_1}{\ln\left(\frac{q_0}{p_0}\right)} = b,$$

где $a$ и $b$ - положительные целые числа. Вычислите вероятности ошибок первого и второго рода для критерия $S(\Gamma)$.

## Приближенные вычисления

Если мы не можем решить задачу точно, то попробуем найти близкие вероятности ошибок, для которых уже можно вычислить границы Г. Например, можно пойти по следующему пути.

Введем события:

$$A_n = \left\{\Gamma_0 < \frac{p_1(X_1,\ldots,X_k)}{p_0(X_1,\ldots,X_k)} < \Gamma_1,\ k = 1,\ldots,n-1,\ \frac{p_1(X_1,\ldots,X_k)}{p_0(X_1,\ldots,X_k)} \leqslant \Gamma_0\right\};$$

$$B_n = \left\{ \Gamma_0 < \frac{p_1(X_1,\dots,X_k)}{p_0(X_1,\dots,X_k)} < \Gamma_1,\ k = 1,\dots,n-1,\ \frac{p_1(X_1,\dots,X_k)}{p_0(X_1,\dots,X_k)} \geqslant \Gamma_1 \right\}.$$

Ясно, что эти события не пересекаются. $\cup_n A_n$ есть событие, что мы принимаем $\mathrm{H}_0$, а $\cup_n B_n$ есть событие, что мы принимаем $\mathrm{H}_1$. Тогда

$$\alpha_0(S(\Gamma)) = \sum_{n=1}^{+\infty} \mathrm{P}_0(B_n) = \sum_{n=1}^{+\infty} \int\limits_{\mathcal{X}^n} \mathrm{I}_{B_n}(x_1,\dots,x_n) p_0(x_1,\dots,x_n)\mu(d\mathbf{x}) \leqslant$$

$$\frac{1}{\Gamma_1} \sum_{n=1}^{+\infty} \int\limits_{\mathcal{X}^n} \mathrm{I}_{B_n}(x_1,\dots,x_n) p_1(x_1,\dots,x_n)\mu(d\mathbf{x}) =$$

$$= \frac{1}{\Gamma_1} \sum_{n=1}^{+\infty} \mathrm{P}_1(B_n) = \frac{1-\alpha_1(S(\Gamma))}{\Gamma_1}.$$

Совершенно аналогично получаем, что

$$\alpha_1(S(\Gamma)) \leqslant \Gamma_0(1-\alpha_0(S(\Gamma))).$$

Обозначим для краткости $\alpha_i = \alpha_i(S(\Gamma))$, $i = 0, 1$. Тогда

$$\Gamma_0 \geqslant \frac{\alpha_1}{1-\alpha_0} = \Gamma_0', \quad \Gamma_1 \leqslant \frac{1-\alpha_1}{\alpha_0} = \Gamma_1'. \tag{1}$$

Мы получили некоторые оценки на границы, которые мы переобозначили штрихами. Рассмотрим теперь критерий $S(\Gamma')$. Его вероятности ошибок $\alpha_i'$, $i = 0, 1$ в силу (1) будут удовлетворять следующим соотношениями:

$$\frac{\alpha_1}{1-\alpha_0} \geqslant \frac{\alpha_1'}{1-\alpha_0'}, \quad \frac{1-\alpha_1}{\alpha_0} \leqslant \frac{1-\alpha_1'}{\alpha_0'}.$$

Следовательно

$$\alpha_0' \leqslant \frac{\alpha_0(1-\alpha_1')}{1-\alpha_1} \leqslant \frac{\alpha_0}{1-\alpha_1}, \qquad \alpha_1' \leqslant \frac{\alpha_1(1-\alpha_0')}{1-\alpha_0} \leqslant \frac{\alpha_1}{1-\alpha_0}$$

$$\alpha_0' + \alpha_1' \leqslant \alpha_0 + \alpha_1.$$

Таким образом, для нового критерия $S(\Gamma')$ вероятности ошибок возможно будут лишь немного больше (при малых $\alpha_0, \alpha_1$), а сумма вероятностей ошибок - даже меньше.

### Численный пример

Вернемся к примеру из начала лекции.

**Пример**

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ - выборка из схемы Бернулли $(1, \theta)$. Тогда р.н.м.к. уровня значимости $\varepsilon_0 = 0.05$ для проверки гипотезы $\mathrm{H}_0 : \theta = 0.05$ против альтернативы $\mathrm{H}_1 : \theta = 0.17$ будет иметь вероятность ошибки второго рода $\varepsilon_1 \leqslant 0.1$ начиная с $n = 57$.

С другой стороны, для данного примера при $\alpha_0 = 0.05$, $\alpha_1 = 0.1$ будет выполнено

$$\Gamma'_0 = 0.105, \quad \Gamma'_1 = 18, \quad \alpha'_0 = 0.031, \quad \alpha'_1 = 0.099,$$

но

$$\mathrm{E}_0\tau(S(\Gamma')) = 31.4, \quad \mathrm{E}_1\tau(S(\Gamma')) = 30.$$

Таким образом, последовательный критерий требует в среднем почти в два раза меньшее число наблюдений при почти тех же вероятностях ошибок первого и второго рода.

Тем самым, мы показали, что последовательный критерий действительно имеет преимущества: мы значительно выигрываем в количестве наблюдений, что крайне удобно с практической точки зрения.

МЕХАНИКО-
МАТЕМАТИЧЕСКИЙ
ФАКУЛЬТЕТ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in

ЛЕКЦИИ УЧЕНЫХ МГУ