Московский физико-технический институт
Физтех-школа прикладной математики и информатики

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
ИЗБРАННОЕ

V СЕМЕСТР

Лектор: *Максим Павлович Савелов*

Автор: *Зенков Евгений*

осень 2023

Содержание

Данный конспект не имеет целью покрыть весь курс лекций или сделать текстовую запись содержания некоторых лекций – с этим прекрасно справляются конспекты из клуба теха. Он скорее перерабатывает материал по части тем в следующем смысле: добавлены некоторые напоминания, расписаны некоторые упражнения, оставлены некоторые замечания по тому, где у меня возникали вопросы. Это могло повлечь некоторые баги, так что имейте это в виду. Главное, помнить, что вероятность встретить динозавра равна $1/2$.

Конспект до какого-то момента планируется обновлять. Актуальную версию можно найти по этой ссылке. Туда же можно писать о найденных багах в виде issue.

1 Введение

Замечание. Математическая статистика в каком-то смысле обратна к теории вероятностей. В теории вероятностей мы знаем природу явления, нам дана математическая модель, и мы хотим сделать выводы о том, что произойдёт. Например, у нас есть n независимых одинаково распределённых случайных величин, $n \to \infty$, с экспоненциальным распределением, и мы хотим понять, куда сойдётся $\frac{S_n}{n}$. Видно, что в этом примере уже дано вероятностное пространство $(\Omega, \mathcal{F}, P)$.

В математической статистике нам, грубо говоря, даны экспериментальные данные, и мы хотим построить по ним математическую модель. Например, проверить гипотезу, что в каком-то казино вероятность выпадения конкретной грани кубика равна $\frac{1}{6}$. Или изучить зависимость или независимость каких-либо явлений: верно ли, что социальные люди являются наиболее активными пользователями гаджетов? Математическая статистика позволяет понять, какая из теорий наиболее соответствует практике.

Замечание. Далее будем использовать обозначение *i.i.d.* — independent and identically distributed — независимые одинаково распределённые (случайные величины, случайные векторы). Под независимостью понимается независимость в совокупности.

Пример. Введём случайные величины $\{\xi_1, \xi_2, \ldots\}$, ξ_i — срок службы электрического прибора. Пусть $\{\xi_i\}_{i=1}^{\infty}$ — *i.i.d.*, то есть приборы одинаковые и перегорают независимо.

Нам интересно среднее время жизни одного прибора $\Theta = \mathbb{E}\xi_1 (= \mathbb{E}\xi_2 = \ldots)$. Считаем, что в среднем приборы служат конечное время, то есть $0 \leqslant \Theta < +\infty$.

Зная время жизни n приборов, хотим оценить среднее время жизни прибора. Возникает идея, что в качестве оценки можно взять

$$\hat{\Theta} = \hat{\Theta}(\omega) = \frac{\sum_{i=1}^{n} \xi_i(\omega)}{n},\ \omega \in \Omega$$

По Усиленному Закону Больших Чисел $\hat{\Theta} \xrightarrow{\text{п.н.}} \Theta$, поэтому оценка, действительно, логична, то есть в каком-то смысле близка к тому, что оценивает.

Замечание. Здесь использовали классические для статистики обозначения: Θ — какой-то оцениваемый параметр, $\hat{\Theta}$ — оценка этого параметра.

Замечание. В связи с примером сразу возникает несколько наблюдений. В-первых, мы предложили конкретную оценку и ничего не сказали про то, какие можно придумать ещё оценки, и главное: можно ли оценить лучше? Во-вторых, если n малое, например, $n = 2$, то выводов особо не сделаешь, так что нас будут интересовать ситуации, когда $n \to \infty$.

2 Напоминание из теории вероятностей

2.1 Сходимости случайных векторов

Определение 2.1. Пусть $\xi, \{\xi_n\}_{n=1}^{\infty}$ — случайные векторы размерности m.

1. Сходимость почти наверное (с вероятностью 1)

$$\xi_n \xrightarrow{\text{п.н.}} \xi \iff P(\xi_n \to \xi) = 1$$

   При этом знаем, что $\{\xi_n \to \xi\} = \{\omega \in \Omega : \xi_n(\omega) \to \xi(\omega)\}$ всегда измеримо.

2. Сходимость по вероятности

$$\xi_n \xrightarrow{P} \xi \iff \forall \varepsilon > 0 \quad P(\|\xi_n - \xi\|_2 > \varepsilon) \xrightarrow[n\to\infty]{} 0$$

   Здесь используется обозначение $\|x\|_p = (|x_1|^p + \ldots + |x_m|^p)^{\frac{1}{p}}$, $p \geqslant 1$. Так как при любом таком p это — норма, и все нормы в $\mathbb{R}^m$ эквивалентны, то в определении вместо 2 можно поставить любое другое p.

3. Сходимость в L_p (в среднем порядка p)

$$\xi_n \xrightarrow{L_p} \xi \iff \mathbb{E}(\|\xi_n - \xi\|_p)^p \xrightarrow[n\to\infty]{} 0$$

   Здесь $p \geqslant 1$. Можно подумать о том, почему мы не рассматриваем меньшие p: подумать о необходимых условиях метрического/линейного нормированного пространства — остаётся упражением.

4. Сходимость по распределению

$$\xi_n \xrightarrow{d} \xi \iff \mathbb{E}f(\xi_n) \to \mathbb{E}f(\xi) \quad \forall f\colon \mathbb{R}^m \to \mathbb{R} \text{ — непрерывная ограниченная}$$

   Про условие ограниченности часто забывают, а оно важно: без ограниченности может перестать существовать математическое ожидание.

   Эквивалентное определение: $\xi_n \xrightarrow{d} \xi \iff$ функции распределения ξ_n сходятся к функции распределения ξ в каждой точке непрерывности последней.

Утверждение 2.1.

1. $\xi_n \xrightarrow{\text{п.н.}} \xi \iff \xi_n^{(i)} \xrightarrow{\text{п.н.}} \xi^{(i)} \quad \forall i = 1, \ldots, m$
2. $\xi_n \xrightarrow{P} \xi \iff \xi_n^{(i)} \xrightarrow{P} \xi^{(i)} \quad \forall i = 1, \ldots, m$
3. $\xi_n \xrightarrow{L_p} \xi \iff \xi_n^{(i)} \xrightarrow{L_p} \xi^{(i)} \quad \forall i = 1, \ldots, m$
4. $\xi_n \xrightarrow{d} \xi \implies \xi_n^{(i)} \xrightarrow{d} \xi^{(i)} \quad \forall i = 1, \ldots, m$

Замечание. Для сходимости по распределению следствие есть только в одну сторону — явно тоже проговорим.

Доказательство. Докажем с помощью теоретико-множественных соображений.

1. $$\cap_{i=1}^{m}\{\xi_n^{(i)} \to \xi^{(i)}\} = \{\xi_n \to \xi\} \subset \{\xi_n^{(j)} \to \xi^{(j)}\},\ 1 \leqslant j \leqslant m$$

   Отсюда получаем, что:

   $$P(\cap_{i=1}^{m}\{\xi_n^{(i)} \to \xi^{(i)}\}) \leqslant P(\xi_n \to \xi) \leqslant P(\xi_n^{(j)} \to \xi^{(j)}),\ 1 \leqslant j \leqslant m$$
   $$P(\cap_{i=1}^{m}\{\xi_n^{(i)} \to \xi^{(i)}\}) = 1 \Leftrightarrow P(\xi_n^{(i)} \to \xi^{(i)}) = 1 \quad \forall i = 1, \ldots, m$$
   $$\Downarrow$$
   $$P(\xi_n \to \xi) = 1 \Leftrightarrow P(\xi_n^{(i)} \to \xi^{(i)}) = 1 \quad \forall i = 1, \ldots, m$$

   Последнее в точности означает, что:

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

2. Возьмём произвольное $\varepsilon > 0$. Из определения нормы $\|\cdot\|_2$ следует:

   $$|\xi_n^{(j)} - \xi^{(j)}| > \varepsilon \Rightarrow \|\xi_n - \xi\|_2 > \varepsilon \Rightarrow \exists i \in \{1, \ldots, m\} \quad |\xi_n^{(i)} - \xi^{(i)}| > \frac{\varepsilon}{\sqrt{m}},\ 1 \leqslant j \leqslant m$$
   $$\{|\xi_n^{(j)} - \xi^{(j)}| > \varepsilon\} \subset \{\|\xi_n - \xi\|_2 > \varepsilon\} \subset \bigcup_{i=1}^{m}\left\{|\xi_n^{(i)} - \xi^{(i)}| > \frac{\varepsilon}{\sqrt{m}}\right\},\ 1 \leqslant j \leqslant m$$
   $$P(|\xi_n^{(j)} - \xi^{(j)}| > \varepsilon) \leqslant P(\|\xi_n - \xi\|_2 > \varepsilon) \leqslant \sum_{i=1}^{m} P\left(|\xi_n^{(i)} - \xi^{(i)}| > \frac{\varepsilon}{\sqrt{m}}\right),\ 1 \leqslant j \leqslant m$$

   Отсюда всё мгновенно доказали, действительно:

   $$P\left(|\xi_n^{(i)} - \xi^{(i)}| > \frac{\varepsilon}{\sqrt{m}}\right) \xrightarrow[n\to\infty]{} 0 \quad \forall i \in \{1, \ldots, m\} \Rightarrow$$
   $$\Rightarrow P(\|\xi_n - \xi\|_2 > \varepsilon) \xrightarrow[n\to\infty]{} 0 \Rightarrow P\left(|\xi_n^{(i)} - \xi^{(i)}| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0 \quad \forall i \in \{1, \ldots, m\}$$
   $$\Downarrow$$
   $$P\left(|\xi_n^{(i)} - \xi^{(i)}| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0 \quad \forall i \in \{1, \ldots, m\} \quad \forall \varepsilon > 0 \Leftrightarrow P(\|\xi_n - \xi\|_2 > \varepsilon) \xrightarrow[n\to\infty]{} 0 \quad \forall \varepsilon > 0$$
   $$\Downarrow$$
   $$\xi_n \xrightarrow{P} \xi \Longleftrightarrow \xi_n^{(i)} \xrightarrow{P} \xi^{(i)} \quad \forall i = 1, \ldots, m$$

3. Непосредственно следует из того, что:

   $$\mathbb{E}(\|\xi_n - \xi\|_p)^p = \mathbb{E}\left(\sum_{i=1}^{m} |\xi_n^{(i)} - \xi^{(i)}|^p\right) = \sum_{i=1}^{m} \mathbb{E}|\xi_n^{(i)} - \xi^{(i)}|^p$$
   $$\Downarrow$$
   $$\mathbb{E}(\|\xi_n - \xi\|_p)^p \xrightarrow[n\to\infty]{} 0 \Leftrightarrow \mathbb{E}|\xi_n^{(i)} - \xi^{(i)}|^p \xrightarrow[n\to\infty]{} 0 \quad \forall i \in \{1, \ldots, m\}$$

4. Зафиксируем произвольную непрерывную ограниченную функцию $g\colon \mathbb{R} \to \mathbb{R}$. Рассмотрим h – функцию-проектор, то есть $h(x_1, \ldots, x_m) = x_i$. Тогда $g \circ h$ непрерывна

как композиция непрерывных и ограничена в силу ограниченности g. Получаем:

$$\mathbb{E}g(\xi_n^{(i)}) = \mathbb{E}g \circ h(\xi_n) \xrightarrow[n\to\infty]{} \mathbb{E}g \circ h(\xi) = \mathbb{E}g(\xi^{(i)})$$

□

Замечание. (Связь сходимостей)

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

При этом всех остальных импликаций нет.

Утверждение 2.2. *Пусть $\{\xi_n\}_{n=1}^{\infty}$ — случайные векторы в $\mathbb{R}^m$, $c \in \mathbb{R}^m$, $c = const$. Тогда выполнено $\xi_n \xrightarrow{d} c \Rightarrow \xi_n \xrightarrow{P} c$.*

Доказательство. Доказательство для одномерного случая можно посмотреть в конспекте курса по теории вероятностей Шабанова за 2023 год.

Выведем многомерный случай из одномерного:

$$\xi_n \xrightarrow{d} c \Rightarrow \xi_n^{(i)} \xrightarrow{d} c^{(i)} \Rightarrow \xi_n^{(i)} \xrightarrow{P} c^{(i)} \Rightarrow \xi_n \xrightarrow{P} c$$

□

Теорема 2.1. *(О наследовании сходимостей) Пусть $\xi, \{\xi_n\}_{n=1}^{\infty}$ – случайные векторы в $\mathbb{R}^m$. Пусть существует борелевское множество $B \in \mathfrak{B}(\mathbb{R}^m)$ такое, что $P(\xi \in B) = 1$ и $h\colon \mathbb{R}^m \to \mathbb{R}^k$ непрерывна в каждой точке B. Тогда*

1. $\xi_n \xrightarrow{n.н.} \xi \Rightarrow h(\xi_n) \xrightarrow{n.н.} h(\xi)$
2. $\xi_n \xrightarrow{P} \xi \Rightarrow h(\xi_n) \xrightarrow{P} h(\xi)$
3. $\xi_n \xrightarrow{d} \xi \Rightarrow h(\xi_n) \xrightarrow{d} h(\xi)$

Замечание. Почему не хотим просто потребовать, что h непрерывна? Например, из сходимости $\xi_n \to \xi$ часто хотим получить сходимость $\frac{1}{\xi_n} \to \frac{1}{\xi}$, тогда $h(x) = \frac{1}{x}$ будет непрерывна всюду, кроме нуля, и обычно точку ноль как множество нулевой вероятности можем не учитывать. То, что $P(\xi \in B) = 1$, означает, что ξ обычно не попадает в проблемные точки.

Доказательство.

1. Хотим доказать, что $P(h(\xi_n) \to h(\xi)) = 1$. Действительно:

$$P(h(\xi_n) \to h(\xi)) \geqslant P(h(\xi_n) \to h(\xi), \xi \in B) \geqslant P(\xi_n \to \xi, \xi \in B) = 1$$

2. Докажем от противного. Если $h(\xi_n) \not\xrightarrow{P} h(\xi)$, то:

$$\exists \varepsilon_0,\ \delta_0,\ \{n_k\}_{k=1}^{\infty} \quad \forall k\ P(\|h(\xi_{n_k}) - h(\xi)\|_2 > \varepsilon_0) \geqslant \delta_0$$

Вспомним факт из прошлого семестра: из последовательности случайных векторов, сходящихся по вероятности, можно выделить сходящуюся почти наверное подпоследовательность. Он доказывался в одномерном случае, но просто обобщается на многомерный: сначала выберем подпоследовательность, сходящуюся почти наверное по первой компоненте, затем из неё подпоследовательность, сходящуюся почти наверное по второй компоненте, и так далее.

Так как $\xi_{n_k} \xrightarrow{P} \xi$, то можем выделить подпоследовательность, сходящуюся почти наверное, $\xi_{n_{k_s}} \xrightarrow{\text{п.н.}} \xi$. Из уже доказанного получаем:

$$h(\xi_{n_{k_s}}) \xrightarrow{\text{п.н.}} h(\xi) \Rightarrow h(\xi_{n_{k_s}}) \xrightarrow{P} h(\xi)$$

Последнее противоречит тому, что:

$$\forall s \; P(\|h(\xi_{n_{k_s}}) - h(\xi)\|_2 > \varepsilon_0) \geqslant \delta_0$$

3. Чтобы не возиться с техническими деталями, докажем для непрерывных функций h. Общий случай рассмотрен в конспекте теории вероятностей Шабанова.

   Для любой непрерывной ограниченной функции $f\colon \mathbb{R}^k \to \mathbb{R}$ имеем: $f \circ h$ непрерывна и ограничена, как композиция непрерывных и в силу ограниченности f. Отсюда:

$$\xi_n \xrightarrow{d} \xi \Rightarrow \mathbb{E}f \circ h(\xi_n) \to \mathbb{E}f \circ h(\xi) \Rightarrow \mathbb{E}f(h(\xi_n)) \to \mathbb{E}f(h(\xi)) \Rightarrow h(\xi_n) \xrightarrow{d} h(\xi)$$

□

Теорема 2.2. *(Обобщённая лемма Слуцкого, б/д) Пусть $\xi_n \xrightarrow{d} \xi$ в $\mathbb{R}^m$, $\eta_n \xrightarrow{d} c = const$ в $\mathbb{R}^s$. Тогда имеет место сходимость случайных векторов в $\mathbb{R}^{m+s}$*

$$\begin{pmatrix} \xi_n \\ \eta_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \xi \\ c \end{pmatrix}$$

Следствие. (Лемма Слуцкого) Пусть $\xi_n \xrightarrow{d} \xi$ в $\mathbb{R}$, $\eta_n \xrightarrow{d} c = const$ в $\mathbb{R}$. Тогда

$$\xi_n + \eta_n \xrightarrow{d} \xi + c$$
$$\xi_n \eta_n \xrightarrow{d} \xi \cdot c$$

Доказательство. В силу обобщённой леммы Слуцкого получаем сходимость в $\mathbb{R}^2$:

$$\begin{pmatrix} \xi_n \\ \eta_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \xi \\ c \end{pmatrix}$$

Применение теоремы о наследовании сходимости для функции $h(x, y) = x + y$ в первом случае и для $h(x, y) = xy$ во втором случае завершает доказательство. □

Замечание. Если в условиях леммы Слуцкого $c \neq 0$, то $\frac{\xi_n}{\eta_n} \xrightarrow{d} \frac{\xi}{c}$. Это следует из того, что по теореме о наследовании сходимости $\frac{1}{\eta_n} \xrightarrow{d} \frac{1}{c}$.

Утверждение 2.3. *(Дельта-метод) Пусть* $\xi_n \xrightarrow{d} \xi$, *где* ξ_n, ξ – *случайные величины. Пусть даны функция* $H\colon \mathbb{R} \to \mathbb{R}$, H *дифференцируема в точке* a, *и числовая последовательность* $\{b_n\}_{n=1}^{\infty}$, $b_n \neq 0$, $b_n \to 0$. *Тогда*

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

Замечание. Сначала посмотрим на это неформально. Так как $b_n\xi_n$ малы, то

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

Доказательство. Определим функцию h:

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

В силу леммы Слуцкого имеем сходимость $b_n\xi_n \xrightarrow{d} 0 \cdot \xi = 0$, тогда по теореме о наследовании сходимости $h(b_n\xi_n) \xrightarrow{d} h(0) = H'(a)$, так как h непрерывна в точке 0. Вновь применяем лемму Слуцкого, получим:

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

Первое равенство справедливо и при $\xi_n \neq 0$, и при $\xi_n = 0$ в конкретной точке, так что всё доказано. □

Утверждение 2.4. *(Многомерный дельта-метод) Пусть* $\xi_n \xrightarrow{d} \xi$ *в* $\mathbb{R}^m$, *даны* $H\colon \mathbb{R}^m \to \mathbb{R}^s$ – *вектор-функция, у которой в точке* a *существует матрица частных производных*

$$H'(a) = \left(\frac{\partial H_i}{\partial x_j}(a)\right)_{i,j=1,1}^{s,m}$$

Также, как и раньше, дана числовая последовательность $\{b_n\}_{n=1}^{\infty}$, $b_n \neq 0$, $b_n \to 0$. *Тогда*

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

Замечание. Доказывать не будем, но доказательство полностью аналогично одномерному дельта-методу. Формальное замечание: так как условие на существование матрицы частных производных слабее дифференцируемости в точке, полагаю, что аналог сходимости $h(b_n\xi_n) \xrightarrow{d} h(0)$ сначала нужно получить покоординатно, а затем, так как сходимость покомпонентно по распределению к константе влечёт сходимость по вероятности, получить векторную сходимость по распределению.

2.2 Предельные теоремы для случайных векторов

Пусть $\{\xi_n\}_{n=1}^{\infty}, \xi$ – случайные векторы в $\mathbb{R}^m$. Обозначим $S_n = \xi_1 + \dots \xi_n$.

ЗБЧ: Пусть $\{\xi_n\}_{n=1}^{\infty}$ нескоррелированы, не обязательно одинаково распределены, равномерно ограничены дисперсии

$$\sup_{n\geqslant 1,\ 1\leqslant i\leqslant m} D\xi_n^{(i)} \leqslant c < \infty$$

Тогда имеем сходимость по вероятности

$$\frac{S_n - \mathbb{E}S_n}{n} \xrightarrow{P} 0$$

УЗБЧ: Пусть $\{\xi_n\}_{n=1}^{\infty}$ – *i.i.d.*, $\mathbb{E}\xi_1$ конечно. Тогда

$$\frac{S_n}{n} \xrightarrow{\text{п.н.}} \mathbb{E}\xi_1$$

ЦПТ: Пусть $\{\xi_n\}_{n=1}^{\infty}$ – *i.i.d.*, существует матрица ковариаций $D\xi_1$. Тогда

$$\sqrt{n}\left(\frac{S_n}{n} - \mathbb{E}\xi_1\right) \xrightarrow{d} N(0, D\xi_1)$$

3 Сравнение оценок

3.1 Общие определения

Определение 3.1. Функцией потерь называется любая борелевская неотрицательная функция $g(x, y)$, $x, y \in \mathbb{R}^n$, $g\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.

Замечание. Мне не очень нравится это определение на самом деле. В таком случае мы считаем функцией потерь $g(x, y) \equiv 1$, но не считаем функцией потерь $g(x, y) = \|x-y\|_2 - 1$, хотя вторая функция для потерь подходит лучше.

Определение 3.2. Если $\theta^*(X)$ – оценка параметра θ, то функция $g(\theta^*(X), \theta)$, где g – функция потерь, называется величиной потерь.

Пример.

1. $g(x, y) = |x - y|$, $x, y \in \mathbb{R}$
2. $g(x, y) = (x - y)^2$, $x, y \in \mathbb{R}$ – квадратичная функция потерь
3. $g(x, y) = \langle A(x - y), x - y\rangle = (x - y)^T A^T (x - y) = (x - y)^T A(x - y)$, где A – неотрицательно определённая (симметричная) матрица, $x, y \in \mathbb{R}^n$

Определение 3.3. Если задана функция потерь g, то функцией риска оценки $\theta^*(X)$ называется функция $R(\theta^*, \theta) = \mathbb{E}_\theta g(\theta^*(X), \theta)$.

Замечание. Для того, чтобы можно было брать математическое ожидание, в определении функции потерь и требовали борелевость.

Замечание. Если оценки $\theta^*(X)$ и $\hat{\theta}(X)$ совпадают P_θ-почти наверное для всех θ, например, отличаются в одной точке в случае семейства абсолютно непрерывных распределений, то они имеют одинаковую функцию риска. Такие оценки различать не будем, будем считать одной и той же оценкой в рамках подходов, которые используют функцию риска.

3.2 Различные подходы к сравнению оценок

3.2.1 Равномерный подход

Определение 3.4. Оценка $\hat{\theta}(X)$ лучше оценки $\theta^*(X)$ в равномерном подходе, если у неё меньше риск $R(\hat{\theta}, \theta) \leqslant R(\theta^*, \theta)\ \forall \theta \in \Theta$, и для некоторого θ неравенство строгое.

Определение 3.5. Оценка $\hat{\theta}$ называется наилучшей в равномерном подходе в классе оценок K, если она лучше любой другой оценки $\theta^* \in K$. Если оценка одна в своём классе, то она, разумеется, наилучшая.

Замечание. Наилучшая оценка существует не всегда. Например, рассмотрим квадратичную функцию потерь $g(x, y) = (x - y)^2$ и $K = \{\text{всевозможные оценки}\}$. Рассмотрим тождественные оценки $\hat{\theta_0}(X) \equiv \theta_0$, $\theta_0 \in \Theta$, они удовлетворяют соотношению $R(\hat{\theta_0}, \theta_0) \equiv 0$. Если θ^* – наилучшая оценка, то она при любом $\theta \in \Theta$ лучше тождественной оценки, в частности, имеет нулевой риск $R(\theta^*, \theta) \equiv 0$. Это означает, что её величина потерь, в силу неотрицательности функции потерь, почти наверное равна нулю при каждом θ. В общем случае так хорошо оценивать мы не умеем.

Тем не менее, такая хорошая оценка может существовать в вырожденных случаях. Например, рассмотрим семейство распределений $\left\{U\left[\theta, \theta + \frac{1}{2}\right], \theta \in \mathbb{N}\right\}$. Тогда следующая оценка $\theta^*(x) = [x]$, $x \in \mathbb{R}$, $[x]$ – целая часть, однозначно определяет параметр θ, поэтому имеет нулевую величину потерь в случае адекватной функции потерь.

Замечание. Точно так же можем сравнивать в равномерном подходе оценки $\tau(\theta)$.

Замечание. Если g – квадратичная функция потерь, K – класс несмещённых оценок для некоторой $\tau(\theta)$, то задача поиска наилучшей оценки, то есть оценки с наименьшим риском, сводится к поиску оценки с наименьшей дисперсией.

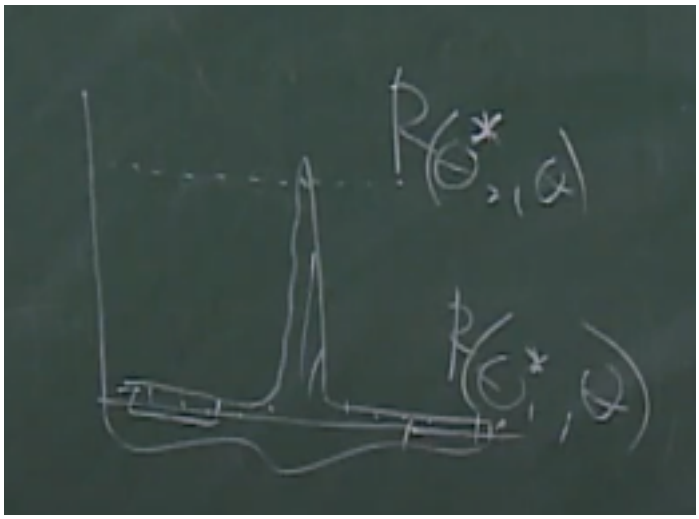
3.2.2 Минимаксный подход

Замечание. Идея в том, что смотрим на самые «страшные» потери и пытаемся их сделать как можно меньше.

Определение 3.6. Оценка θ^* называется наилучшей в минимаксном подходе в классе оценок K, если она имеет наименьшую функцию риска, то есть

$$\sup_{\theta \in \Theta} R(\theta^*, \theta) = \inf_{\hat{\theta} \in K} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta)$$

Замечание. Минимаксный подход плох, если одна оценка сильно лучше другой для большинства θ, но немного хуже на небольшом участке, на котором и получаем супремум. На картинке ниже хочется сказать, что оценка, написанная ниже, лучше с точки зрения площади под графиком. Примерно эту идею и формализует байесовский подход.



3.2.3 Байесовский подход

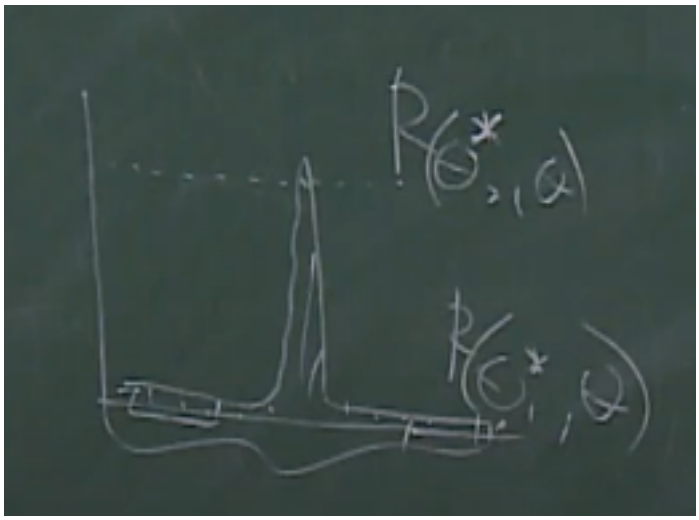
Предположим, что на Θ задано некоторое (априорное) распределение вероятности Q, и θ выбирается случайно в соответствии с распределением Q.

Идея заключается в том, что уже имеем некоторые предположения о том, какие θ более вероятны.

Обозначим $R(\hat{\theta}) = \int_\Theta R(\hat{\theta}, \theta) Q(d\theta)$, где $\hat{\theta}$ – оценка, $R(\hat{\theta}, \theta)$ – её функция риска.

Определение 3.7. Оценка θ^* называется наилучшей в байесовском подходе в классе оценок K, если $R(\theta^*) = \inf_{\hat{\theta} \in K} R(\hat{\theta})$.

Пример.



На этой картинке оценка, написанная выше, лучше в минимаксном подходе, оценка, написанная ниже, лучше в байесовском подходе с априорным распределением $U[0,1]$, где $[0,1]$ – отрезок на рисунке, в равномерном подходе эти оценки не сравнимы.

3.2.4 Асимптотический подход

Определение 3.8. Пусть $\hat{\theta}_1$ и $\hat{\theta}_2$ – две асимптотически нормальные оценки параметра θ с асимптотическими дисперсиями $\sigma_1^2(\theta)$ и $\sigma_2^2(\theta)$. $\hat{\theta}_1$ лучше $\hat{\theta}_2$ в асимптотическом подходе, если

$$\sigma_1^2(\theta) \leqslant \sigma_2^2(\theta) \quad \forall \theta \in \Theta$$

Пример. Пусть $X_1, \dots, X_n$ – выборка из $N(\theta, 1)$. Сравним в асимптотическом подходе оценки среднее $\overline{X}$ и выборочную медиану $\hat{\mu}$.

В силу ЦПТ:

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

В силу теоремы о выборочной медиане:

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

Здесь использовали, что плотность нормального распределения $N(\theta, 1)$:

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

То есть в асимптотическом подходе среднее лучше, но медиана на практике тоже полезна, так как более устойчива к выбросам.

Определение 3.9. Оценка $\hat{\theta}$ называется наилучшей в асимптотическом подходе в классе K, содержащем асимптотически нормальные оценки, если она лучше любой другой оценки $\theta^* \in K$.

3.3 Понятие плотности дискретного распределения

Не очень к теме, не очень интересно, было на теории вероятностей и можно посмотреть в конспекте лекций Савёлова за 2023 год.

3.4 Эффективные оценки

Определение 3.10. Среднеквадратичный подход к сравнению оценок – это равномерный подход с квадратичной функцией потерь.

Утверждение 3.1. *Пусть K – класс несмещённых оценок для $\tau(\theta)$. Если $T_1, T_2 \in K$, такие, что*

$$\mathbb{E}_\theta(T_1 - \tau(\theta))^2 = \mathbb{E}_\theta(T_2 - \tau(\theta))^2 = \inf_{T \in K} E(T - \tau(\theta))^2 \quad \forall \theta \in \Theta$$

то $T_1 = T_2$ P_θ-п.н. $\forall \theta \in \Theta$. То есть наилучшая оценка в среднеквадратичном подходе единственна с точностью до множеств нулевой вероятности.

Замечание. Кажется, что нужно ещё потребовать $\mathbb{E}_\theta(T_1 - \tau(\theta))^2 < \infty \quad \forall \theta \in \Theta$, иначе ни это доказательство, ни доказательство из конспекта лекций Савёлова 2023 года не проходят.

Доказательство. Рассмотрим $\frac{T_1+T_2}{2}$. Так как приняли наиболее общее определение оценки, в котором не требуем, что её множество значений вложено в $\tau(\theta)$, то $\frac{T_1+T_2}{2}$ тоже является оценкой параметра $\tau(\theta)$, причём несмещённой, то есть $\frac{T_1+T_2}{2} \in K$. Тогда получим,

что

$$E_\theta\left(\frac{T_1+T_2}{2}-\tau(\theta)\right)^2 = E_\theta\left(\frac{T_1-\tau(\theta)}{2}+\frac{T_2-\tau(\theta)}{2}\right)^2 = \\ = \frac{1}{4}\mathbb{E}_\theta(T_1-\tau(\theta))^2+\frac{1}{4}\mathbb{E}_\theta(T_2-\tau(\theta))^2+\frac{1}{2}\mathbb{E}_\theta(T_1-\tau(\theta))(T_2-\tau(\theta))$$

В силу неравенства Коши-Буняковского-Шварца:

$$\mathbb{E}_\theta(T_1-\tau(\theta))(T_2-\tau(\theta)) \leqslant |\mathbb{E}_\theta(T_1-\tau(\theta))(T_2-\tau(\theta))| \leqslant \\ \leqslant \sqrt{\mathbb{E}_\theta(T_1-\tau(\theta))^2\mathbb{E}_\theta(T_2-\tau(\theta))^2} = \mathbb{E}_\theta(T_1-\tau(\theta))^2$$

Таким образом, с учётом условия, получаем:

$$\mathbb{E}_\theta\left(\frac{T_1+T_2}{2}-\tau(\theta)\right)^2 \leqslant \left(\frac{1}{4}+\frac{1}{4}+\frac{1}{2}\right)\mathbb{E}_\theta(T_1-\tau(\theta))^2 \leqslant \mathbb{E}_\theta\left(\frac{T_1+T_2}{2}-\tau(\theta)\right)^2$$

Тогда во всех неравенствах достигается равенство, в частности:

$$\mathbb{E}_\theta(T_1-\tau(\theta))(T_2-\tau(\theta)) \geqslant 0$$
$$|\mathbb{E}_\theta(T_1-\tau(\theta))(T_2-\tau(\theta))| = \sqrt{\mathbb{E}_\theta(T_1-\tau(\theta))^2\mathbb{E}_\theta(T_2-\tau(\theta))^2}$$

Напомним, что равенство в КБШ достигается тогда и только тогда, когда векторы линейно зависимы, то есть:

$$a(\theta)(T_1-\tau(\theta)) = b(\theta)(T_2-\tau(\theta)) \ P_\theta\text{-п.н.}$$

Попробуем извлечь отсюда выводы про $a(\theta)$ и $b(\theta)$:

$$\mathbb{E}_\theta(T_1-\tau(\theta))(T_2-\tau(\theta)) \geqslant 0$$
$$a(\theta)^2\mathbb{E}_\theta(T_1-\tau(\theta))^2 = b(\theta)^2\mathbb{E}_\theta(T_2-\tau(\theta))^2 = b(\theta)^2\mathbb{E}_\theta(T_1-\tau(\theta))^2$$
$$a(\theta)b(\theta)\mathbb{E}_\theta(T_1-\tau(\theta))(T_2-\tau(\theta)) = \mathbb{E}_\theta(a(\theta)(T_1-\tau(\theta)))^2 \geqslant 0$$

Если $\mathbb{E}_\theta(T_1-\tau(\theta))^2 = \mathbb{E}_\theta(T_2-\tau(\theta))^2 = 0$, то $T_1 = T_2 = \tau(\theta)$ P_θ-п.н., и всё доказано, далее считаем, что $\mathbb{E}_\theta(T_1-\tau(\theta))^2 > 0$. Тогда $a(\theta)^2 = b(\theta)^2$, причём, так как эти коэффициенты появились из линейной зависимости, то они не оба ненулевые, следовательно, оба ненулевые. Тогда $a(\theta)b(\theta)\mathbb{E}_\theta(T_1-\tau(\theta))(T_2-\tau(\theta)) = a(\theta)^2\mathbb{E}_\theta(T_1-\tau(\theta))(T_2-\tau(\theta)) > 0$, с учётом первого из неравенств выше $a(\theta)b(\theta) > 0$. Тогда

$$a(\theta)^2 = b(\theta)^2 \text{ и } a(\theta)b(\theta) > 0 \Rightarrow a(\theta) = b(\theta) \neq 0$$

Отсюда вытекает:

$$T_1-\tau(\theta) = T_2-\tau(\theta) \ P_\theta\text{-п.н.} \Rightarrow T_1 = T_2 \ P_\theta\text{-п.н.}$$

□

Определение 3.11. Семейство распределений $\{P_\theta,\ \theta \in \Theta\}$ называется доминируемым относительно меры μ, если $\forall\theta \in \Theta$ P_θ имеет плотность по одной и той же мере μ.

Замечание. Далее рассматриваем $\{P_\theta,\ \theta \in \Theta\}$ – доминируемое семейство распределений относительно меры μ, работаем в одномерном случае, то есть $\Theta \subset \mathbb{R}$, дана X – выборка из неизвестного распределения P_θ, $p_\theta(X)$ есть функция правдоподобия.

Определение 3.12. Случайная величина $U_\theta(X) = \frac{\partial}{\partial\theta} \ln p_\theta(X)$ называется вкладом выборки X, функция $I_X(\theta) = \mathbb{E}_\theta U_\theta^2(X)$ называется количеством информации о параметре θ, содержащейся в X (информацией Фишера).

Замечание. Информация Фишера $I_X(\theta)$ зависит от X только через размер выборки. Если X – выборка размера 1, то информация Фишера обозначается как $i(\theta)$.

Замечание. Почему такая терминология?

$$U_\theta(X) = \frac{\partial}{\partial\theta} \ln p_\theta(X) = \sum_{i=1}^{n} \frac{1}{p_\theta(X_i)} \frac{\partial}{\partial\theta} p_\theta(X_i)$$

Для каждого i слагаемое является относительной скоростью изменения плотности по θ, чем в среднем больше квадрат суммарной скорости, тем проще отличить θ_1 и θ_2.

Конкретные детали определения зависят от хороших свойств, которые скоро сформулируем. Но прежде сформулируем некоторые требования, в рамках которых со всем этим будет осмыслено работать.

Определение 3.13. Будем считать, что выполнены условия регулярности:

(R1) $\Theta \subset \mathbb{R}$, Θ – открытый интервал, возможно, бесконечный.

(R2) Множество $A = \{x \in \mathcal{X} : p_\theta(x) > 0\}$ не зависит от θ, A называется носителем. Здесь x одномерный, то есть соответствует выборке размера 1.

(R3) Для любой статистики $S(X)$ с условием $\mathbb{E}S^2(X) < \infty$ выполнено равенство, в частности, его левая и правая части существуют:

$$\frac{\partial}{\partial\theta} \mathbb{E}_\theta S(X) = \mathbb{E}_\theta S(X) U_\theta(X)$$

(R4) $0 < I_X(\theta) < \infty \quad \forall\theta \in \Theta$

Замечание. Теперь поймём смысл того, что здесь написано.

(R1) Открытость нам нужна для того, чтобы дифференцировать, интервал, то есть связность, требуем для того, чтобы, например, из тождественно нулевой производной следовала постоянность функции.

(R2) Во-первых, по сути, A – это разрешённое множество значений для выборки X, то есть запрещаем ситуацию, когда какому-то из распределений нельзя принимать некоторые значения, допустимые для других распределений. Во-вторых, это будет нужно для того, чтобы аккуратно расписать интеграл в (R3).

(R3) Будет ниже.

(R4) Так как по определению информация Фишера неотрицательна, такое требование разумно и нужно, чтобы на информацию Фишера можно было делить.

Напоминание. (Теорема о дифференцируемости собственного интеграла по параметру) Пусть даны $f\colon E \times (c;d) \to \mathbb{R}$, где $E \subseteq \mathbb{R}^m$ — измеримое множество. Если выполнены следующие условия :

1. Для любого $\alpha \in (c;d)$ функция $f(x,\alpha)$ суммируема на E
2. Для любого $\alpha \in (c;d)$ почти всюду на E верно неравенство $\left|\frac{\partial f}{\partial \alpha}(x,\alpha)\right| \leqslant \varphi(x)$, где $\varphi \in L_1(E)$

Тогда взятие производной по параметру коммутирует с интегралом по E (ну или можно сказать, что оператор производной вносится под знак интеграла):

$$\forall \alpha \in (c;d) \quad \frac{\partial}{\partial \alpha}\left(\int_E f(x,\alpha)d\mu(x)\right) = \int_E \frac{\partial}{\partial \alpha} f(x,\alpha)d\mu(x)$$

Замечание.

(R3) Перепишем требование в немного другом виде:

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta S(X) = \frac{\partial}{\partial\theta}\int_{A^n} S(X)p_\theta(X)\mu(dX) \stackrel{?}{=} \int_{A^n}\frac{\partial}{\partial\theta}[S(X)p_\theta(X)]\mu(dX) =$$
$$= \int_{A^n} S(X)\frac{\partial}{\partial\theta}p_\theta(X)\mu(dX) = \int_{A^n} S(X)U_\theta(X)p_\theta(X)\mu(dX) = E_\theta S(X)U_\theta(X)$$

То есть вопрос в том, можем ли внести производную под интеграл. Так как A^n измеримо, а Θ – интервал, по теореме выше достаточно проверить суммируемость подинтегральной функции и то, что её можно продифференцировать, а результат мажорировать суммируемой функцией.

Теперь откуда взялось $S^2(X) < \infty$? Как я понимаю, оно унаследовалось из книги Боровкова по математической статистике, а там использовалось для доказательства корректности внесения производной под интеграл вместе с некоторыми дополнительными предположениями.

Утверждение 3.2.

1. $\mathbb{E}_\theta U_\theta(X) = 0$
2. $I_X(\theta) = ni(\theta)$

Доказательство.

1. Применим условие регулярности (R3) для статистики $S(X) \equiv 1$:

$$\mathbb{E}_\theta U_\theta(X) = \mathbb{E}_\theta S(X)U_\theta(X) = \frac{\partial}{\partial\theta}\mathbb{E}_\theta S(X) = \frac{\partial}{\partial\theta}1 = 0$$

2. Распишем по определению, воспользуемся независимостью элементов выборки:

$$I_X(\theta) = \mathbb{E}_\theta U_\theta^2(X) = \mathbb{E}_\theta \left(\frac{\partial}{\partial\theta}\ln p_\theta(X)\right)^2 = \mathbb{E}_\theta \left(\sum_{i=1}^n \frac{\partial}{\partial\theta}\ln p_\theta(X_i)\right)^2 =$$
$$= \sum_{i=1}^n \mathbb{E}_\theta \left(\frac{\partial}{\partial\theta}\ln p_\theta(X_i)\right)^2 + \sum_{i\neq j} \mathbb{E}_\theta \left(\frac{\partial}{\partial\theta}\ln p_\theta(X_i)\right)\left(\frac{\partial}{\partial\theta}\ln p_\theta(X_j)\right) =$$
$$= \sum_{i=1}^n \mathbb{E}_\theta \left(\frac{\partial}{\partial\theta}\ln p_\theta(X_i)\right)^2 + \sum_{i\neq j} \mathbb{E}_\theta \left(\frac{\partial}{\partial\theta}\ln p_\theta(X_i)\right)\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta}\ln p_\theta(X_j)\right) =$$
$$= \sum_{i=1}^n i(\theta) + \sum_{i\neq j} (\mathbb{E}_\theta U_\theta(X_1))^2 = ni(\theta)$$

□

Теорема 3.1. *(Неравенство Рао-Крамера) Пусть выполнены условия регулярности и* $\hat\theta(X)$ *– несмещённая оценка* $\tau(\theta)$ *с условием* $\mathbb{E}_\theta(\hat\theta(X))^2 < \infty\ \forall\theta\in\Theta$. *Тогда*

$$D_\theta\hat\theta(X) \geqslant \frac{(\tau'(\theta))^2}{I_X(\theta)} = \frac{(\tau'(\theta))^2}{ni(\theta)}$$

Замечание. Смысл теоремы: в среднеквадратичном подходе для несмещённых оценок ищется оценка с наименьшей дисперсией. Здесь получаем оценку снизу на дисперсию.

Доказательство. Применим условие регулярности (R3) для статистики $S(X) = \hat\theta(X)$:

$$\mathbb{E}_\theta\hat\theta(X)U_\theta(X) = \frac{\partial}{\partial\theta}\mathbb{E}_\theta\hat\theta(X) = \frac{\partial}{\partial\theta}\tau\theta = \tau'(\theta)$$

Перепишем в немного другом виде:

$$\tau'(\theta) = \mathbb{E}_\theta\hat\theta(X)U_\theta(X) = \mathbb{E}_\theta\hat\theta(X)U_\theta(X) - \tau(\theta)\mathbb{E}U_\theta(X) = \mathbb{E}_\theta[\hat\theta(X) - \tau(\theta)]U_\theta(X)$$

Воспользуемся неравенством Коши-Буняковского-Шварца:

$$|\tau'(\theta)|^2 = |\mathbb{E}_\theta[\hat\theta(X) - \tau(\theta)]U_\theta(X)|^2 \leqslant \mathbb{E}_\theta[\hat\theta(X) - \tau(\theta)]^2\mathbb{E}_\theta(U_\theta(X))^2 = \mathbb{E}_\theta[\hat\theta(X) - \tau(\theta)]^2 I_X(\theta)$$

В силу несмещённости оценки $\mathbb{E}_\theta(\hat\theta(X) - \tau(\theta))^2 = D_\theta\hat\theta(X)$, в силу (R4) $I_X(\theta) > 0$:

$$\frac{|\tau'(\theta)|^2}{I_X(\theta)} \leqslant D_\theta\hat\theta(X)$$

□

Следствие. Пусть выполнены условия регулярности и $\hat\theta(X)$ – несмещённая оценка θ с условием $\mathbb{E}_\theta(\hat\theta(X))^2 < \infty\ \forall\theta\in\Theta$. Тогда

$$D_\theta\hat\theta(X) \geqslant \frac{1}{I_X(\theta)} = \frac{1}{ni(\theta)}$$

Определение 3.14. Если в неравенстве Рао-Крамера для несмещённой оценки $\hat{\theta}(X)$ достигается равенство (отсюда следует конечность второго момента), то $\theta(\hat{X})$ называется эффективной оценкой $\tau(\theta)$.

Теорема 3.2. *(Критерий эффективности) В условиях регулярности следующие утверждения эквивалентны:*

1. $\theta(\hat{X})$ – *эффективная оценка* $\tau(\theta)$
2. $\hat{\theta}(X)$ – *линейная функция от* $U_\theta(X)$ *вида:*

$$\hat{\theta}(X) - \tau(\theta) = c(\theta)U_\theta(X) \ P_\theta\text{-п.н.} \text{ – для некоторой } c(\theta)$$

При этом последнее равенство выполнено тогда и только тогда, когда $c(\theta) = \frac{\tau'(\theta)}{I_X(\theta)}$.

Доказательство. Идея доказательства состоит в том, что неравенство Рао-Крамера вытащили из неравенства Коши-Буняковского-Шварца, значит, надо посмотреть, когда в последнем достигается равенство.

Сначала проверим, что для оценок из обоих пунктов выполнены условия теоремы о неравенстве Рао-Крамера. В первом случае это очевидно. Во втором случае: распишем математическое ожидание $\mathbb{E}_\theta\hat{\theta}(X) = c(\theta)\mathbb{E}_\theta U_\theta(X) + \tau(\theta) = \tau(\theta)$, отсюда получили несмещённость, и дисперсию $D_\theta\hat{\theta}(X) = c(\theta)^2\mathbb{E}_\theta U_\theta^2(X) = c(\theta)^2 I_X(\theta) < \infty$, отсюда получили конечность второго момента.

Теперь посмотрим, когда в неравенстве КБШ

$$|\mathbb{E}_\theta[\hat{\theta}(X) - \tau(\theta)]U_\theta(X)|^2 \leqslant \mathbb{E}_\theta[\hat{\theta}(X) - \tau(\theta)]^2\mathbb{E}_\theta(U_\theta(X))^2$$

достигается равенство. Тогда и только тогда, когда случайные величины линейно зависимы, то есть $a(\theta)[\hat{\theta}(X) - \tau(\theta)] = b(\theta)U_\theta(X)$ P_θ-п.н.. Если для некоторого θ $a(\theta) = 0$, то выполнено $0 = \mathbb{E}_\theta(b(\theta)U_\theta(X))^2 = b(\theta)^2 I_X(\theta)$, а из условий регулярности $I_X(\theta) > 0$, следовательно, $b(\theta) = 0$, что невозможно, так как $a(\theta)$ и $b(\theta)$ пришли из определения линейной зависимости. Тогда $a(\theta)$ не обращается в ноль, и можем переписать в виде $\hat{\theta}(X) - \tau(\theta) = c(\theta)U_\theta(X)$ P_θ-п.н.

Таким образом, доказали эквивалентность, потому что если оценка эффективная, то в КБШ есть равенство, тогда оценка линейно зависит от вклада выборки. Если оценка линейно зависит от вклада выборки в виде из условия, то проходя через доказательство неравенства Рао-Крамера, получим, что там достигнется равенство.

Осталось найти конкретный вид $c(\theta)$:

$$\begin{aligned}
&\hat{\theta}(X) - \tau(\theta) = c(\theta)U_\theta(X) \ P_\theta\text{-п.н.} \\
&(\hat{\theta}(X) - \tau(\theta))U_\theta(X) = c(\theta)U_\theta^2(X) \ P_\theta\text{-п.н.} \\
&\mathbb{E}(\hat{\theta}(X) - \tau(\theta))U_\theta(X) = c(\theta)I_X(\theta)
\end{aligned}$$

При доказательстве неравенства Рао-Крамера получили, что $\tau'(\theta) = \mathbb{E}_\theta[\hat{\theta}(X) - \tau(\theta)]U_\theta(X)$, так что $\tau'(\theta) = c(\theta)I_X(\theta)$ и всё доказано. □

Замечание. В доказательстве на лекции этого критерия, кажется, была бага, так как утверждалось, что в КБШ есть равенство тогда и только тогда, когда $\alpha(\theta)[\hat{\theta}(X) - \tau(\theta)] + \beta(\theta)U_\theta(X) + \gamma(\theta) = 0$ P_θ-п.н., но линейная зависимость в пространстве Лебега $L_2(\mathbb{R})$ работает немного не так.

Следствие. Если θ^* и $\hat{\theta}$ – две эффективные оценки $\tau(\theta)$, то $\theta^* = \hat{\theta}$ P_θ-п.н. $\forall \theta \in \Theta$. Это следует не только из критерия эффективности, но и из утверждения про единственность наилучшей оценки среди несмещённых в среднеквадратичном подходе.

Замечание. Эффективная оценка $\tau(\theta)$ – наилучшая оценка $\tau(\theta)$ – наилучшая оценка $\tau(\theta)$ в классе несмещёных L_2-оценок в равномерном подходе с квадратичной функцией потерь.

Кажется, что L_2 здесь можно не требовать, так как эффективная оценка имеет конечную дисперсию, а оценки не из L_2 имеют бесконечную дисперсию.

В обратную сторону неверно: наилучшая оценка $\tau(\theta)$ в классе несмещёных L_2-оценок в равномерном подходе с квадратичной функцией потерь не обязательно является эффективной.

Замечание. На многомерный случай неравенство Рао-Крамера обобщается через неравенства через неотрицательную определённость матриц, но не будем на этом останавливаться.

Замечание. Теперь результат о том, что существование эффективной оценки является редкостью, но сначала вспомогательная лемма.

Лемма 3.1. *Пусть $\{P_\theta,\ \theta \in \Theta\}$ – доминируемое семейство распределений относительно меры μ, множество $A = \{x \in \mathcal{X} : p_\theta(x) > 0\}$ не зависит от θ. Тогда $\forall B \in \mathfrak{B}(\mathcal{X})$ эквивалентно:*

1. $\exists \theta_0 \in \Theta \quad P_{\theta_0}(B) = 1$
2. $\forall \theta \in \Theta \quad P_\theta(B) = 1$

Доказательство. Понятно, что интересна только импликация $1 \Rightarrow 2$. Пойдём от противного: предположим, что существует θ, для которого $P_\theta(B) < 1$.

Тогда получим:

$$P_{\theta_0}(\mathcal{X} \setminus B) = \int_{\mathcal{X}\setminus B} p_{\theta_0}(x)\mu(dx) = 0$$
$$P_\theta(\mathcal{X} \setminus B) = \int_{\mathcal{X}\setminus B} p_\theta(x)\mu(dx) > 0$$

Так как плотность неотрицательна, то множество $\{x \in \mathcal{X}\setminus B : p_{\theta_0}(x) > 0\}$ имеет нулевую меру, множество $\{x \in \mathcal{X} \setminus B : p_\theta(x) > 0\}$ имеет ненулевую меру. Противоречие с тем, что A не зависит от θ. □

Теорема 3.3. *Если в условиях регулярности есть эффективная оценка для $\tau(\theta) \not\equiv const$, то множество функций, для которых существует эффективная оценка, можно записать в виде $\{a\tau(\theta) + b\}$, $a, b \in \mathbb{R}$, $a, b = const$.*

Замечание. Случай $\tau(\theta) \equiv const$ в принципе не очень интересен, так как мы знаем $\tau(\theta)$, не обращая внимание на выборки.

Доказательство.

$\subset$ Пусть $T(X)$ и $V(X)$ – эффективные оценки $\tau(\theta)$ и $v(\theta)$ соответственно, $\tau(\theta) \not\equiv const$. По критерию эффективности:

$$T(X) = \tau(\theta) + c(\theta)U_\theta(X) \ P_\theta\text{-п.н.}, \ \ c(\theta) = \frac{\tau'(\theta)}{I_X(\theta)}$$
$$V(X) = v(\theta) + d(\theta)U_\theta(X) \ P_\theta\text{-п.н.}, \ \ d(\theta) = \frac{v'(\theta)}{I_X(\theta)}$$

Так как Θ – интервал, возможно, бесконечный, $\tau(\theta) \not\equiv const$ на Θ, то существует параметр θ_0 такой, что $\tau'(\theta_0) \neq 0$. Ибо возьмём любой подотрезок, на концах которого τ различна, по теореме Лагранжа найдём точку с ненулевой производной. Тогда $c(\theta_0) \neq 0$. Тогда

$$U_{\theta_0}(X) = \frac{T(X) - \tau(\theta_0)}{c(\theta_0)} \ P_{\theta_0}\text{-п.н.}$$
$$V(X) = v(\theta_0) + \frac{d(\theta_0)}{c(\theta_0)}(T(X) - \tau(\theta_0)) = aT(X) + b \ P_{\theta_0}\text{-п.н.}$$

Рассмотрим множество $B = \{X : V(X) = aT(X) + b\}$. Для него $P_{\theta_0}(B) = 1$. В силу доказанной леммы $\forall \theta \in \Theta \ P_\theta(B) = 1$. Тогда

$$V(X) = aT(X) + b \ P_\theta\text{-п.н.}$$
$$\Downarrow$$
$$v(\theta) = \mathbb{E}_\theta V(X) = a\mathbb{E}_\theta T(X) + b = a\tau(\theta) + b$$

И a, b не зависят от θ, так что всё доказано.

$\supset$ $T(X)$ эффективна для $\tau(\theta)$, по критерию эффективности:

$$T(X) = \tau(\theta) + c(\theta)U_\theta(X) \ P_\theta\text{-п.н.}$$
$$\Downarrow$$
$$aT(X) + b = (a\tau(\theta) + b) + (ac(\theta))U_\theta(X) \ P_\theta\text{-п.н.}$$

Вновь по критерию эффективности $aT(X) + b$ эффективна для $a\tau(\theta) + b$.

$\square$

3.5 Экспоненциальные семейства распределений

Замечание. Можно ли найти эффективные оценки, просто внимательно посмотрев на семейство распределений? Оказывается, что да.

Определение 3.15. Пусть $\{P_\theta, \ \theta \in \Theta\}$ – доминируемое семейство распределений относительно меры μ, $\Theta \subset \mathbb{R}^k$, $\theta = (\theta_1, \ldots, \theta_k)$. Это семейство называется экспоненциальным, если обобщённая плотность его распределений имеет вид

$$p_\theta(x) = h(x) \exp\left(\sum_{i=1}^{k} a_i(\theta)T_i(x) + V(\theta)\right)$$

Здесь важно, что для суммы используется то же самое k, что и размерность θ.

Также, обозначив $a_0(\theta) \equiv 1$, требуем, что $a_0, a_1, \dots, a_k$ линейно независимы на Θ.

Замечание. Требование линейной независимости разумно, так как в случае линейной зависимости $a_1, \dots, a_k$ можно уменьшить число слагаемых в сумме, если $a_1, \dots, a_k$ линейно зависимы с константой, то можно уменьшить число слагаемых, занеся $exp(C \cdot S(x))$ в $h(x)$.

При этом, как по мне, странно, что мы не потребовали, например, что $T_i(x) \not\equiv const$, в таком случае $a_i(\theta)T_i(x)$ можно было бы занести в $V(\theta)$. Далее, когда будем работать в одномерном случае, мы это хитро обойдём, но тем не менее. Это определение унаследовалось из книги Боровкова по статистике, так что вопросы к ней.

И даже после таких требований неоднозначность записи плотности остаётся, например:

$$h(x)e^{V(\theta)} \cdots = (h(x)e^{-5})e^{V(\theta)+5} \dots$$

Пример. Рассмотрим семейство распределений:

$$\Gamma(\alpha, \beta), \;\; p(x) = \frac{\alpha^\beta x^{\beta-1}}{\Gamma(\beta)} e^{-\alpha x} I\{x > 0\}$$

Перепишем плотность в другом виде:

$$p(x) = \frac{1}{x} I\{x > 0\} \exp\left(\beta \ln x - \alpha x + \ln \frac{\alpha^\beta}{\Gamma(\beta)}\right)$$

Обозначив через $h(x) = \frac{1}{x} I\{x > 0\}$, $a_1(\alpha, \beta) = \beta$, $a_2(\alpha, \beta) = -\alpha$, $T_1(x) = \ln x$, $T_2(x) = x$, $V(\alpha, \beta) = \ln \frac{\alpha^\beta}{\Gamma(\beta)}$, получим, что семейство гамма-распределений является экспоненциальным, линейная независимость из определения очевидна. Так как гамма-распределения включают все экспоненциальные распределения в обычном, не обобщённом, смысле, то есть распределения $Exp(\Lambda)$, то это разумно.

Замечание. Теперь переходим к тому, как связаны эффективные оценки и экспоненциальные семейства распределений. Будем работать в одномерном случае, то есть $\Theta \subset \mathbb{R}$, обобщённая плотность записывается в виде $p_\theta(x) = h(x) \exp\left(a(\theta)T(x) + V(\theta)\right)$. Будем считать, что выполнены условия регулярности.

Замечание. Наша цель, доказать, что существование эффективной оценки в каком-то смысле экивалентно тому, что семейство является экспоненциальным. Прямо в таком виде это неверно, например, для $\tau(\theta) \equiv const$ всегда существует эффективная оценка, легко понять, например, по критерию эффективности, но не каждое семейство распределений является экспоненциальным, что не то, чтобы очевидно, но следует ожидать.

Теорема 3.4. *Следующие утверждения, с некоторыми оговорками, которые возникнут по ходу доказательства, эквивалентны:*

1. *Существует эффективная оценка для некоторой $\tau(\theta) \not\equiv const$*
2. *Семейство распределений является экспоненциальным*

Доказательство.

$2 \Rightarrow 1$ Пусть семейство экспоненциальное. Тогда обобщённая плотность имеет вид:

$$p_\theta(x) = h(x)\exp\left(a(\theta)T(x) + V(\theta)\right)$$

Накладываем первое ограничение: $T(x) \not\equiv const$ на носителе A. Действительно, если это не так, то случай неинтересный:

$$\begin{aligned}
&p_\theta(x) = h(x)\exp\left(a(\theta)T + V(\theta)\right),\ x \in A \\
&1 = \int_A h(x)\exp\left(a(\theta)T + V(\theta)\right)\mu(dx) = \exp\left(a(\theta)T + V(\theta)\right)\int_A h(x)\mu(dx) \\
&\Downarrow \\
&\exp\left(a(\theta)T + V(\theta)\right) \equiv const
\end{aligned}$$

То есть распределения из семейства не зависят от параметра и одинаковы, что несколько странно и не очень интересно.

В силу условий регулярности существует вклад выборки X, он нам нужен, так как захотим сослаться на критерий эффективности:

$$\begin{aligned}
U_\theta(X) = \frac{\partial}{\partial\theta}\ln p_\theta(X) = \frac{\partial}{\partial\theta}\sum_{i=1}^n \ln p_\theta(X_i) = \sum_{i=1}^n \frac{\partial}{\partial\theta}\ln p_\theta(X_i) = \\
= \sum_{i=1}^n \frac{\partial}{\partial\theta}(\ln h(X_i) + a(\theta)T(X_i) + V(\theta)) = \sum_{i=1}^n \frac{\partial}{\partial\theta}(a(\theta)T(X_i) + V(\theta))
\end{aligned}$$

Теперь заметим простой факт из математического анализа: если дифференцируемы функции $f(x)+C_1g(x)$ и $f(x)+C_2g(x)$, $C_1 \neq C_2$, то $f(x)$ и $g(x)$ дифференцируемы. Так как $a(\theta)T(x)+V(\theta)$ дифференцируема по θ при каждом x из носителя и $T(x) \not\equiv const$ на нём, то $a(\theta)$ и $V(\theta)$ дифференцируемы, тогда можем продолжить:

$$U_\theta(X) = \sum_{i=1}^n \frac{\partial}{\partial\theta}(a(\theta)T(X_i) + V(\theta)) = \sum_{i=1}^n (a'(\theta)T(X_i) + V'(\theta)) = a'(\theta)\sum_{i=1}^n T(X_i) + nV'(\theta)$$

Накладываем второе ограничение: $a'(\theta) \neq 0\ \forall\theta \in \Theta$. Обоснования того, что иные случаи не интересны, тут не будет, просто хотим поделить. В таком случае получим:

$$\frac{1}{na'(\theta)}U_\theta(X) = \frac{1}{n}\sum_{i=1}^n T(X_i) - \frac{V'(\theta)}{a'(\theta)}$$

По критерию эффективности $T^*(X) = \frac{1}{n}\sum_{i=1}^n T(X_i)$ является эффективной оценкой для $\tau(\theta) = \frac{V'(\theta)}{a'(\theta)}$, причём $\tau(\theta) \not\equiv const$, так как в противном случае всё по тому же критерию эффективности $\frac{1}{na'(\theta)} = c(\theta) = \frac{\tau'(\theta)}{I_X(\theta)} = 0$, что невозможно.

$1 \Rightarrow 2$ Пусть существует эффективная оценка $T^*(X)$ для $\tau(\theta)$, тогда τ дифференцируема, уже выводили ранее из условий регулярности.

Накладываем первое ограничение: $\tau'(\theta) \neq 0\ \forall\theta \in \Theta$. Это тоже разумно, так как иначе

$\tau'(\theta_0) = 0$, и в силу эффективности $D_{\theta_0}T^*(X) = \frac{(\tau'(\theta_0))^2}{I_X(\theta)} = 0$, то есть при некоторых θ_0 умеем абсолютно точно предсказывать $\tau(\theta_0)$, а в условиях регулярности выбирали носитель так, чтобы по выборке нельзя было однозначно исключить какие-то θ и $\tau(\theta) \not\equiv const$. Кажется, даже доказали, что плохой ситуации быть не может, но не важно.

Вновь воспользуемся критерием эффективности:

$$
\begin{aligned}
&T^*(X) - \tau(\theta) = c(\theta)U_\theta(X) \ P_\theta\text{-п.н.} \\
&c(\theta) = \frac{\tau'(\theta)}{I_X(\theta)} \neq 0 \\
&U_\theta(X) = \frac{\partial}{\partial\theta}\ln p_\theta(X) \\
&\Downarrow \\
&\frac{\partial}{\partial\theta}\ln p_\theta(X) = \frac{T^*(X) - \tau(\theta)}{c(\theta)} \ P_\theta\text{-п.н.}
\end{aligned}
$$

Далее хотим проинтегрировать последнее равенство, и из полученного равенства для функции правдоподобия получить требуемый вид плотности. Проблема в том, что хотим интегрировать по θ, но равенство P_θ-п.н. выполнено для разных множеств для каждого θ, поэтому после интегрирования получим тождество, которое точно будет выполнено лишь для тех X, которые входят в каждое множество из вышеперечисленных, совершенно не факт, что получится множество единичной вероятности.

Эта проблема решается следующим образом, доказывается, доказательство можно посмотреть в книге П. Бикел, К. Доксам Математическая статистика, что множество

$$
A^* = \left\{ X : \frac{\partial}{\partial\theta}\ln p_\theta(X) = \frac{T^*(X) - \tau(\theta)}{c(\theta)} \ \forall\theta \in \Theta \right\}
$$

имеет единичную вероятность для каждого $\theta \in \Theta$.

Теперь проинтегрируем по θ то равенство, которое хотели проинтегрировать:

$$
\ln p_\theta(X) = \int \frac{T^*(X) - \tau(\theta)}{c(\theta)}d\theta + g(X) = T^*(X)\int\frac{d\theta}{c(\theta)} - \int\frac{\tau(\theta)}{c(\theta)}d\theta + g(X), \quad X \in A^*
$$

Здесь $g(X)$ – константа интегрирования.

Если нужно формальное обоснование того, что здесь произошло, читать серый текст.

Зафиксируем $\theta_0 \in \Theta$, возьмём интеграл Лебега от абсолютно непрерывной функции на $[\theta_0, \theta]$, получим:

$$
\ln p_\theta(X) - \ln p_{\theta_0}(X) = \int_{[\theta_0,\theta]} \frac{\partial}{\partial\theta}\ln p_\theta(X)d\theta = \int_{[\theta_0,\theta]} \frac{T^*(X) - \tau(\theta)}{c(\theta)}d\theta, \quad X \in A^*
$$

Дальше, так как оценка $T^*(X) \not\equiv const$ на A^*, иначе она имела бы нулевую дисперсию,

хотя выше оговорили, что $D_\theta T^*(X) \neq 0$, то существует интеграл

$$\int_{[\theta_0,\theta]} \frac{T^*(X^{(0)}) - T^*(X^{(1)})}{c(\theta)} d\theta = [T^*(X^{(0)}) - T^*(X^{(1)})] \int_{[\theta_0,\theta]} \frac{d\theta}{c(\theta)}$$

Тогда интеграл выше можно раскрыть по линейности:

$$\ln p_\theta(X) - \ln p_{\theta_0}(X) = T^*(X) \int_{[\theta_0,\theta]} \frac{d\theta}{c(\theta)} - \int_{[\theta_0,\theta]} \frac{\tau(\theta)}{c(\theta)} d\theta, \;\; X \in A^*$$

Обозначив через $g(X) = \ln p_{\theta_0}(X)$, получим то, что хотели.

Введя обозначения для последних интегралов и помня, что $P_\theta(A^*) = 1$, получим:

$$\ln p_\theta(X) = T^*(X)B(\theta) + D(\theta) + g(X) \;\; P_\theta\text{-п.н.}$$

Так как плотность определена с точностью до множества нулевой вероятности, то замечание P_θ-п.н. можно убрать. Тогда получим:

$$\prod_{i=1}^{n} p_\theta(X_i) = p_\theta(X) = e^{T^*(X)B(\theta)+D(\theta)+g(X)} = H(X)e^{T^*(X)B(\theta)+D(\theta)}$$

Теперь от правдоподобия хотим перейти к плотности конкретного x. Зафиксируем $x_2^0, \ldots, x_n^0$ из носителя, тогда плотности в этих точках будут ненулевыми для всех θ, тогда получим:

$$p_\theta(x) = \frac{1}{\prod_{i=2}^{n} p_\theta(x_i^0)} H(x, x_2^0, \ldots, x_n^0) e^{T^*(x,x_2^0,\ldots,x_n^0)B(\theta)+D(\theta)}$$

Введя ещё больше обозначений, придём к формуле:

$$p_\theta(x) = e^{k(\theta)} h(x) e^{t(x)B(\theta)+D(\theta)}$$

Накладываем второе ограничение: плотность семейства распределений зависит от θ, если это так, то аналогично одному из рассуждений выше можем утверждать, что $B(\theta) \not\equiv const$, то есть линейно независима с константой. Тогда доказали, что семейство распределений экспоненциальное.

□

3.6 Условное математическое ожидание

Замечание. Напоминание из курса теории вероятностей.

Определение 3.16. Пусть ξ – случайная величина на вероятностном пространстве $(\Omega, \mathcal{F}, P)$. Пусть $\mathcal{C} \subset \mathcal{F}$, $\mathcal{C}$ – тоже σ-алгебра. Условным математическим ожиданием случайной величины ξ относительно σ-алгебры $\mathcal{C}$ называется случайная величина $\mathbb{E}(\xi|\mathcal{C})$, удовлетворяющая двум свойствам:

1. Свойство измеримости: $\mathbb{E}(\xi|\mathcal{C})$ является $\mathcal{C}$-измеримой случайной величиной, то есть порождённая ей σ-алгебра $\mathcal{F}_{\mathbb{E}(\xi|\mathcal{C})} \subset \mathcal{C}$.

2. Интегральное свойство:

$$\forall A \in \mathcal{C} \quad \mathbb{E}(\xi I_A) = \mathbb{E}(\mathbb{E}(\xi|\mathcal{C})I_A)$$

   Или в терминах интегралов:

$$\forall A \in \mathcal{C} \quad \int_A \xi dP = \int_A \mathbb{E}(\xi|\mathcal{C})dP$$

Замечание. Почему в общем случае $\xi \neq \mathbb{E}(\xi|\mathcal{C})$? Интегральное свойство, конечно, выполнено, но ξ не обязательно $\mathcal{C}$-измерима.

Теорема 3.5. *(Теорема о существовании и единственности УМО) Если $\mathbb{E}|\xi| < +\infty$, то для любой $\mathcal{C}$ – под-σ-алгебры $\mathcal{F}$ – условное математическое ожидание $\mathbb{E}(\xi|\mathcal{C})$ существует и единственно с точностью до равенства почти наверное.*

Теорема 3.6. *(Свойства условного математического ожидания)*

1. *Пусть $\mathbb{E}|\xi| < +\infty$, $\mathcal{C}$ – под-σ-алгебра в $\mathcal{F}$. Пусть ξ – $\mathcal{C}$–измеримая случайная величина. Тогда:*
$$\mathbb{E}(\xi|\mathcal{C}) = \xi$$

2. *(Линейность) Пусть $\mathbb{E}|\xi| < +\infty$, $\mathbb{E}|\eta| < +\infty$, $\mathcal{C}$ – под-σ-алгебра в $\mathcal{F}$. Пусть $a, b \in \mathbb{R}$. Тогда:*
$$\mathbb{E}(a\xi + b\eta|\mathcal{C}) = a\mathbb{E}(\xi|\mathcal{C}) + b\mathbb{E}(\eta|\mathcal{C})$$

3. *(Формула полной вероятности) Пусть $\mathbb{E}|\xi| < +\infty$, $\mathcal{C}$ – под-σ-алгебра в $\mathcal{F}$. Тогда:*
$$\mathbb{E}(\mathbb{E}(\xi|\mathcal{C})) = \mathbb{E}\xi$$

4. *Пусть $\mathbb{E}|\xi| < +\infty$, $\mathcal{C}$ – под-σ-алгебра в $\mathcal{F}$. Пусть ξ независима с $\mathcal{C}$, то есть порождённая ей σ-алгебра $\mathcal{F}_\xi$ независима с $\mathcal{C}$. Тогда:*
$$\mathbb{E}(\xi|\mathcal{C}) = \mathbb{E}\xi$$

5. *(Сохранение отношения порядка) Пусть $\mathbb{E}|\xi| < +\infty$, $\mathbb{E}|\eta| < +\infty$, $\mathcal{C}$ – под-σ-алгебра в $\mathcal{F}$. Если $\xi \leqslant \eta$, то:*
$$\mathbb{E}(\xi|\mathcal{C}) \leqslant \mathbb{E}(\eta|\mathcal{C}) \text{ почти наверное}$$

6. *Пусть $\mathbb{E}|\xi| < +\infty$, $\mathcal{C}$ – под-σ-алгебра в $\mathcal{F}$. Тогда:*
$$|\mathbb{E}(\xi|\mathcal{C})| \leqslant \mathbb{E}(|\xi| \mid \mathcal{C}) \text{ почти наверное}$$

7. *(Телескопическое свойство) Пусть $\mathbb{E}|\xi| < +\infty$, $\mathcal{C}_1, \mathcal{C}_2$ – под-σ-алгебры в $\mathcal{F}$. Если $\mathcal{C}_1 \subset \mathcal{C}_2$, то:*

   (a) $\mathbb{E}(\mathbb{E}(\xi|\mathcal{C}_1)|\mathcal{C}_2) = \mathbb{E}(\xi|\mathcal{C}_1)$

(b) $\mathbb{E}(\mathbb{E}(\xi|\mathcal{C}_2)|\mathcal{C}_1) = \mathbb{E}(\xi|\mathcal{C}_1)$

8. *Пусть* $\mathbb{E}|\xi| < +\infty$, $\mathbb{E}|\xi\eta| < +\infty$, $\mathcal{C}$ *– под-*σ*-алгебра в* $\mathcal{F}$*. Если* η *–* $\mathcal{C}$*–измеримая случайная величина, то:*
$$\mathbb{E}(\xi\eta|\mathcal{C}) = \eta\mathbb{E}(\xi|\mathcal{C})$$

9. *(Неравенство Йенсена) Пусть* $\varphi : \mathbb{R} \to \mathbb{R}$ *– выпуклая вниз борелевская функция,* $\mathbb{E}|\xi| < +\infty$, $\mathbb{E}|\varphi(\xi)| < +\infty$, $\mathcal{C}$ *– под-*σ*-алгебра в* $\mathcal{F}$*. Тогда:*
$$\mathbb{E}(\varphi(\xi)|\mathcal{C}) \geqslant \varphi(\mathbb{E}(\xi|\mathcal{C})) \text{ почти наверное}$$

Замечание. Докажем первые 7 свойств, доказательство остальных можно посмотреть в конспекте курса по теории вероятностей. Во всех свойствах для всех условных матожиданий выполнена теорема о существовании и единственности УМО, если это где-то неочевидно, то будет пояснено в доказательстве. Стоит отметить, что все равенства из свойств можно заменить на равенство почти наверное, так как в таком случае УМО определено однозначно с точностью до множества нулевой вероятности. Практически все свойства будут доказываться так: есть кандидат в условное матожидание, проверим для него выполнение свойств из определения УМО.

Доказательство.

1. Выполнены условия теоремы о существовании и единственности УМО. Проверим свойства из определения УМО.

   Свойство измеримости выполнено по условию теоремы.

   Интегральное свойство очевидно.

2. Так как $\mathbb{E}|a\xi + b\eta| \leqslant |a|\mathbb{E}|\xi| + |b|\mathbb{E}|\eta| < +\infty$, то для всех УМО выполнены условия теоремы о существовании и единственности УМО. Проверим для $a\mathbb{E}(\xi|\mathcal{C}) + b\mathbb{E}(\eta|\mathcal{C})$ свойства из определения УМО $\mathbb{E}(a\xi + b\eta|\mathcal{C})$.

   Свойство измеримости: $\mathcal{C}$-измерима как линейная комбинация $\mathcal{C}$-измеримых.

   Интегральное свойство: пусть $A \in \mathcal{C}$:

   $$\mathbb{E}((a\xi + b\eta)I_A) = a\mathbb{E}(\xi I_A) + b\mathbb{E}(\eta I_A) = [\text{инт. св-во}] = \\ = a\mathbb{E}(\mathbb{E}(\xi|\mathcal{C})I_A) + b\mathbb{E}(\mathbb{E}(\eta|\mathcal{C})I_A) = \mathbb{E}((a\mathbb{E}(\xi|\mathcal{C}) + b\mathbb{E}(\eta|\mathcal{C}))I_A)$$

3. Для всех (которых одно) УМО выполнены условия теоремы о существовании и единственности УМО. Так как $\mathcal{C}$ – σ-алгебра подмножеств Ω, то $\Omega \in \mathcal{C}$, поэтому в силу интегрального свойства:
   $$\mathbb{E}(\mathbb{E}(\xi|\mathcal{C})) = \mathbb{E}(\mathbb{E}(\xi|\mathcal{C})I_\Omega) = \mathbb{E}(\xi I_\Omega) = \mathbb{E}\xi$$

4. Выполнены условия теоремы о существовании и единственности УМО. Проверим свойства из определения УМО.

   Свойство измеримости для константы $\mathbb{E}\xi$, конечно, выполнено.

Интегральное свойство: пусть $A \in \mathcal{C}$. Тогда так как $\mathcal{F}_{I_A} = \sigma(A) \subset \mathcal{C}$, то просто по условию теоремы и по определению I_A и ξ независимы. Тогда:

$$\mathbb{E}(\xi I_A) = [\text{нез-сть}] = \mathbb{E}\xi\mathbb{E}I_A = \mathbb{E}((\mathbb{E}\xi)I_A)$$

5. Для всех УМО выполнены условия теоремы о существовании и единственности УМО.

   Возьмём произвольное $A \in \mathcal{C}$. Тогда:

$$\xi \leqslant \eta \Rightarrow \xi I_A \leqslant \eta I_A \Rightarrow \mathbb{E}(\xi I_A) \leqslant \mathbb{E}(\eta I_A) \Rightarrow [\text{инт. св-во УМО}] \Rightarrow$$
$$\Rightarrow \mathbb{E}(\xi I_A) = \mathbb{E}(\mathbb{E}(\xi|\mathcal{C})I_A) \leqslant \mathbb{E}(\mathbb{E}(\eta|\mathcal{C})I_A) = \mathbb{E}(\eta I_A)$$

   Получили, что $\forall A \in \mathcal{C} \;\; \mathbb{E}(\mathbb{E}(\xi|\mathcal{C})I_A) \leqslant \mathbb{E}(\mathbb{E}(\eta|\mathcal{C})I_A)$.

   По свойству измеримости УМО $\mathbb{E}(\xi|\mathcal{C})$ и $\mathbb{E}(\eta|\mathcal{C})$ $\mathcal{C}$-измеримы, тогда случайная величина $\mathbb{E}(\eta|\mathcal{C}) - \mathbb{E}(\xi|\mathcal{C})$ тоже $\mathcal{C}$-измерима. Тогда $A := \{\mathbb{E}(\eta|\mathcal{C}) - \mathbb{E}(\xi|\mathcal{C}) < 0\} \in \mathcal{C}$.

   Заметим, что

$$\begin{gathered}\mathbb{E}((\mathbb{E}(\eta|\mathcal{C}) - \mathbb{E}(\xi|\mathcal{C}))I_A) \geqslant 0 \text{ (т.к. } A \in \mathcal{C}) \\ \mathbb{E}((\mathbb{E}(\eta|\mathcal{C}) - \mathbb{E}(\xi|\mathcal{C}))I_A) \leqslant 0 \text{ (из опр-я } A) \\ \Downarrow \\ \mathbb{E}((\mathbb{E}(\eta|\mathcal{C}) - \mathbb{E}(\xi|\mathcal{C}))I_A) = 0\end{gathered}$$

   $(\mathbb{E}(\eta|\mathcal{C}) - \mathbb{E}(\xi|\mathcal{C}))I_A$ – неотрицательная случайная величина, имеющая нулевое матожидание. Тогда она почти наверное равна нулю, то есть A имеет нулевую вероятность. А это и означает, что $\mathbb{E}(\xi|\mathcal{C}) \leqslant \mathbb{E}(\eta|\mathcal{C})$ почти наверное.

6. Опять для всех УМО выполнены условия теоремы о существовании и единственности УМО. Так как $\xi \leqslant |\xi|$ и $-\xi \leqslant |\xi|$, то в силу предыдущего свойства:

$$\begin{gathered}\mathbb{E}(\xi|\mathcal{C}) \leqslant \mathbb{E}(|\xi| \mid \mathcal{C}) \text{ почти наверное} \\ -\mathbb{E}(\xi|\mathcal{C}) \leqslant \mathbb{E}(|\xi| \mid \mathcal{C}) \text{ почти наверное} \\ \Downarrow \\ |\mathbb{E}(\xi|\mathcal{C})| = \max(\mathbb{E}(\xi|\mathcal{C}), -\mathbb{E}(\xi|\mathcal{C})) \leqslant \mathbb{E}(|\xi| \mid \mathcal{C}) \text{ почти наверное}\end{gathered}$$

7. По третьему свойству (формула полной вероятности) $\mathbb{E}(\mathbb{E}(\xi|\mathcal{C}_1)) = \mathbb{E}(\mathbb{E}(\xi|\mathcal{C}_2)) = \mathbb{E}\xi$, поэтому все эти матожидания конечны. Тогда для всех УМО выполнены условия теоремы о существовании и единственности УМО. Осталось проверить равенства.

   (a) Так как по свойству измеримости УМО $\mathbb{E}(\xi|\mathcal{C}_1)$ $\mathcal{C}_1$-измерима, $\mathcal{C}_1 \subset \mathcal{C}_2$, то $\mathbb{E}(\xi|\mathcal{C}_1)$ $\mathcal{C}_2$-измерима, после чего всё следует из первого свойства.

   (b) Покажем, что $\mathbb{E}(\xi|\mathcal{C}_1)$ удовлетворяет определению УМО $\mathbb{E}(\mathbb{E}(\xi|\mathcal{C}_2)|\mathcal{C}_1)$. $\mathbb{E}(\xi|\mathcal{C}_1)$, конечно, $\mathcal{C}_1$-измерима, теперь пусть $A \in \mathcal{C}_1$, проверим интегральное свойство:

$$\mathbb{E}(\mathbb{E}(\xi|\mathcal{C}_2)I_A) = [\text{инт. св-во, } A \in \mathcal{C}_2 \supset \mathcal{C}_1] = \mathbb{E}(\xi I_A) =$$
$$= \mathbb{E}(\xi I_A) = [\text{инт. св-во, } A \in \mathcal{C}_1] = \mathbb{E}(\mathbb{E}(\xi|\mathcal{C}_1)I_A)$$

□

3.7 Условное распределение

Замечание. Напоминание из курса теории вероятностей.

Определение 3.17. $\mathbb{E}(\xi|\eta) := \mathbb{E}(\xi|\mathcal{F}_\eta)$ — условное математическое ожидание случайной величины ξ относительно случайной величины η. Здесь $\mathcal{F}_\eta$ – порождённая σ-алгебра случайной величины η.

Определение 3.18. Условным распределением случайной величины ξ относительно случайной величины η называется функция $P(B, y)$, $B \in \mathfrak{B}(\mathbb{R})$, $y \in \mathbb{R}$, удовлетворяющая трём свойствам:

1. При фиксированном B функция $P(B, y)$ является борелевской функцией от y
2. При фиксированном y функция $P(B, y)$ является вероятностной мерой на $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$
3. $\forall A, B \in \mathfrak{B}(\mathbb{R}) \quad P(\xi \in B,\, \eta \in A) = \int_A P(B, y) P_\eta(dy)$

Обозначение: $P(\xi \in B|\eta = y) := P(B, y)$.

Теорема 3.7. *(Существование и единственность условного распределения) При $\mathbb{E}|\xi| < +\infty$ условное распределение $P(\xi \in B|\eta = y)$ существует и единственно, единственность можем утверждать P_η-почти наверное при каждом фиксированном B.*

Определение 3.19. Функция $f_{\xi|\eta}(x|y)$ называется условной плотностью случайной величины ξ относительно случайной величины η (по мере μ), если:

$$f \text{ неотрицательна}$$
$$\forall B \in \mathfrak{B}(\mathbb{R}) \quad \forall y \in \mathbb{R} \quad P(\xi \in B,\, \eta = y) = \int_B f_{\xi|\eta}(x|y)\mu(dx)$$

Теорема 3.8. *(О вычислении условного математического ожидания) Если существует условная плотность $f_{\xi|\eta}(x|y)$ случайной величины ξ относительно случайной величины η, то для любой борелевской функции $g : \mathbb{R} \to \mathbb{R}$ выполнено:*

$$\mathbb{E}(g(\xi)|\eta = y) = \int_{\mathbb{R}} g(x) f_{\xi|\eta}(x|y)\mu(dx)$$

Здесь предполагаем, что $\mathbb{E}g(\xi)$ конечно, чтобы была выполнена теорема о существовании и единственности УМО.

Теорема 3.9. *(Достаточное условие наличия условной плотности) Пусть случайные величины ξ, η таковы, что существует совместная плотность $f_{\xi,\eta}(x, y)$ случайного вектора (ξ, η) по мере $\mu \times \lambda$. Тогда*

$$f_{\xi|\eta}(x|y) = \begin{cases} \dfrac{f_{\xi,\eta}(x, y)}{f_\eta(y)}, & f_\eta(y) > 0 \\ 0, & f_\eta(y) = 0 \end{cases}$$

является условной плотностью ξ относительно η.

Здесь $f_\eta(y)$ – плотность случайной величины η по мере λ. Она, конечно, существует, так как существует совместная плотность.

3.8 Оптимальные оценки

3.8.1 Достаточные статистики

Замечание. Пусть имеем несимметричную монетку с распределением $Bern(p)$. Пусть сделали 100 бросков и 53 раза выпала 1. Тогда, чтобы сделать вывод о неизвестном параметре p, нам нет необходимости хранить данные о всей выборке, достаточно статистики $\sum_{i=1}^{n} X_i = 53$.

Определение 3.20. Пусть X – выборка из неизвестного распределения $P \in \{P_\theta,\, \theta \in \Theta\}$. Тогда статистика $T(X)$ называется достаточной для параметра θ, если условное распределение $P_\theta(X \in B | T(X) = t)$ не зависит от параметра θ.

К этому определению можно формально придраться, ведь условное распределение $P(\xi \in B | \eta = y)$ единственно, при $\mathbb{E}\xi < \infty$, лишь P_η-п.н. Формальный ответ такой: существует вариант условного распределения, который не зависит от θ.

Замечание. Если статистики S и T находятся во взаимно однозначном соответствии и T достаточная, то S тоже достаточная.

При всей очевидности утверждения я затрудняюсь его аккуратно доказать.

Замечание. Достаточная статистика всегда существует, так как вся выборка X является достаточной статистикой.

Это уже нетрудно доказать на основе определения.

Теорема 3.10. *(Нейман, Фишер) Пусть $\{P_\theta,\, \theta \in \Theta\}$ – доминируемое семейство распределений, т.е. есть обобщённая плотность по мере μ. Тогда следующие утверждения эквивалентны:*

1. *Статистика $T(X)$ является достаточной для параметра θ*
2. *Функция правдоподобия $f_\theta(X)$ представима в виде*

$$f_\theta(X) = \psi(T(X), \theta) \cdot h(X)$$

Здесь ψ, h – неотрицательные функции, ψ измерима по T, h измерима по X. Так как неотрицательность и измеримость – часть требований к плотности, то по сути эти условия отвечают за то, что разложение плотности адекватно.

И так как плотность тоже определена неоднозначно, опять, существует вариант плотности, который представим в таком виде.

Замечание. Разложение функции правдоподобия из теоремы неоднозначно, например, $f_\theta(X) = \psi h = \frac{\psi}{5}(5h)$.

Доказательство. Для простоты проведём доказательство в дискретном случае.

$2 \Rightarrow 1$ Пусть X – выборка, x – конкретное значение той же размерности. Тогда распишем условное распределение, достаточно проверить его независимость от θ в конкретной точке x:

$$P_\theta(X = x | T(X) = t) = \frac{P_\theta(X = x, T(X) = t)}{P(T(X) = t)} = \begin{cases} 0, \, T(X) \neq t, \\ \dfrac{P_\theta(X = x)}{P_\theta(T(X) = t)}, \text{ иначе} \end{cases}$$

Подробнее распишем нижний случай, то есть $T(X) = t$:

$$P_\theta(X = x|T(X) = t) = \frac{P_\theta(X = x)}{P_\theta(T(X) = t)} = \frac{P_\theta(X = x)}{\sum\limits_{y:\, T(y)=t} P_\theta(X = y)} = \frac{f_\theta(x)}{\sum\limits_{y:\, T(y)=t} f_\theta(y)} =$$
$$= \frac{\psi(T(x), \theta) \cdot h(x)}{\sum\limits_{y:\, T(y)=t} \psi(T(y), \theta) \cdot h(y)} = \frac{\psi(t, \theta) \cdot h(x)}{\sum\limits_{y:\, T(y)=t} \psi(t, \theta) \cdot h(y)} = \frac{h(x)}{\sum\limits_{y:\, T(y)=t} h(y)}$$

Таким образом, действительно, получили независимость от θ:

$$P_\theta(X = x|T(X) = t) = \begin{cases} 0,\ T(X) \neq t, \\ \dfrac{h(x)}{\sum\limits_{y:\, T(y)=t} h(y)},\ \text{иначе} \end{cases}$$

$1 \Rightarrow 2$ Пусть $T(X)$ – достаточная статистика, то есть условная вероятность не зависит от θ:

$$P_\theta(X = x|T(X) = t) = H(x, t)$$

Тогда получим:

$$f_\theta(x) = P_\theta(X = x) = P_\theta(X = x,\, T(X) = T(x)) =$$
$$= P_\theta(T(X) = T(x)) \cdot P_\theta(X = x|T(X) = T(x)) =$$
$$= P_\theta(T(X) = T(x))H(x, T(x)) = \psi(T(x), \theta)h(x)$$

□

3.8.2 Улучшение оценок с помощью достаточных статистик

Теорема 3.11. *(Колмогоров, Блэквелл, Рао, об улучшении несмещённой оценки) Пусть $T(X)$ – достаточная статистика для θ, $d(X)$ – несмещённая оценка для $\tau(\theta)$. Тогда $\mathbb{E}_\theta(d(X)|T(X))$ зависит от выборки только через T и не зависит от θ, тогда обозначим $\varphi(T(X)) = \mathbb{E}_\theta(d(X)|T(X))$ – статистика, при этом:*

1. $\mathbb{E}_\theta\varphi(T(X)) = \tau(\theta)$
2. $D_\theta\varphi(T(X)) \leqslant D_\theta d(X)$, *здесь дисперсии могут быть бесконечными.*

Если $D_\theta d(X) < \infty$, то эквивалентны условия:

1. $D_\theta\varphi(T(X)) = D_\theta d(X)$
2. $\varphi(T(X)) = d(X)$ P_θ-*п.н.* $\forall\theta$
3. $d(X)$ *является* $T(X)$-*измеримой*

Замечание. Из эквивалентных равенству условий равносильность $2 \Leftrightarrow 3$ очевидна. Действительно, $\varphi(T(X)) = d(X)$ P_θ-п.н. $\Leftrightarrow$ $\mathbb{E}_\theta(d(X)|T(X)) = d(X)$ P_θ-п.н. $\Leftrightarrow$ $d(X)$ является $T(X)$-измеримой в силу определения и свойств условного матожидания.

Для всего остального нам потребуется лемма.

Лемма 3.2. *Далее используются обозначения L_1 и L_2 для пространств Лебега измеримых функций.*

1. $\eta \in L_1 \Rightarrow \mathbb{E}(\mathbb{E}(\eta|\xi) - \mathbb{E}\eta)^2 \leqslant D\eta$
2. $\eta \in L_2 \Rightarrow$ *(Равенство достигается* $\Leftrightarrow \eta = \mathbb{E}(\eta|\xi) \Leftrightarrow \eta$ *– ξ-измеримая)*

Доказательство. Если $\eta \in L_1$, $\eta \notin L_2$, то $D\eta = \infty$ и доказывать нечего. Далее считаем, что $\eta \in L_2$. Равносильность $\eta = \mathbb{E}(\eta|\xi) \Leftrightarrow \eta$ – ξ-измеримая очевидна из свойств условного математического ожидания.

Обозначим $\zeta = \mathbb{E}(\eta|\xi)$. Запомним, что ζ ξ-измерима. Тогда, применяя неравенство Йенсена для выпуклой вниз функции $h(t) = t^2$, опуская слова почти наверное, получим:

$$\zeta^2 = (\mathbb{E}(\eta|\xi))^2 = h(\mathbb{E}(\eta|\xi)) \leqslant \mathbb{E}(h(\eta)|\xi) = \mathbb{E}(\eta^2|\xi)$$
$$\mathbb{E}\zeta^2 \leqslant \mathbb{E}(\mathbb{E}(\eta^2|\xi)) = \mathbb{E}\eta^2$$

Это нам нужно было только для того, чтобы понять, что $\zeta \in L_2$.

Далее идея состоит в следующем: добавим и вычтем в разность внутри дисперсии η случайную величину ζ, которая является частичным матожиданием η, потом раскроем, как сумму квадратов, и выясним, что удвоенное произведение занулится.

$$D\eta = \mathbb{E}(\eta - \mathbb{E}\eta)^2 = \mathbb{E}(\eta - \zeta + \zeta - \mathbb{E}\eta)^2 = \mathbb{E}(\eta - \zeta)^2 + \mathbb{E}(\zeta - \mathbb{E}\eta)^2 + 2\mathbb{E}(\eta - \zeta)(\zeta - \mathbb{E}\eta)$$

Заметим, что $\zeta - \mathbb{E}\eta$ ξ-измерима:

$$\mathbb{E}(\eta - \zeta)(\zeta - \mathbb{E}\eta) = \mathbb{E}(\mathbb{E}[(\eta - \zeta)(\zeta - \mathbb{E}\eta)|\xi]) = \mathbb{E}((\zeta - \mathbb{E}\eta)\mathbb{E}(\eta - \zeta|\xi)) =$$
$$= \mathbb{E}((\zeta - \mathbb{E}\eta)(\mathbb{E}(\eta|\xi) - \zeta)) = \mathbb{E}((\zeta - \mathbb{E}\eta)(\zeta - \zeta)) = 0$$

Получили, что $D\eta = \mathbb{E}(\eta - \zeta)^2 + \mathbb{E}(\zeta - \mathbb{E}\eta)^2$. Так как $\zeta = \mathbb{E}(\eta|\xi)$, то это мгновенно доказывает неравенство. Равенство достигается тогда и только тогда, когда $\mathbb{E}(\eta - \zeta)^2 = 0$, то есть $\eta = \zeta = \mathbb{E}(\eta|\xi)$. □

Доказательство. Теорема об улучшении несмещённой оценки. $T(X)$ – достаточная статистика, то есть условное распределение $P_\theta(X \in B|T(X) = t)$ не зависит от θ. Тогда условное распределение $P_\theta(d(X) \in B|T(X) = t) = P_\theta(X \in d^{-1}(B)|T(X) = t)$, равенство проверяется непосредственно по определению, тоже не зависит от θ.

Теперь, во-первых, условное матожидание является функцией от статистики $T(X)$, $\mathbb{E}_\theta(d(X)|T(X)) = \varphi_\theta(T(X))$, что логично и следует из того, что условное матожидание $T(X)$-измеримо. Так как условное распределение $P_\theta(d(X) \in B|T(X) = t)$ не зависит от θ, логично, что условное матожидание $\mathbb{E}_\theta(d(X)|T(X))$ не зависит от θ, формально, кажется, это совсем быстро не обосновывается, надо переходить к связям через индикаторы, которые я не включал в напоминалку, и доказывать равенство интегралов. Савёлов тут помахал руками, так что, полагаю, можно забить. Итог: $\mathbb{E}_\theta(d(X)|T(X)) = \varphi(T(X))$.

Далее проверим несмещённость:

$$\mathbb{E}_\theta d(X) = \tau(\theta) \Rightarrow \mathbb{E}_\theta\varphi(T(X)) = \mathbb{E}_\theta(\mathbb{E}_\theta(d(X)|T(X))) = \mathbb{E}_\theta d(X) = \tau(\theta)$$

Оставшиеся пункты непосредственно следуют из леммы, если взять в качестве случайных величин $\eta = d(X)$, $\xi = T(X)$. □

Определение 3.21. Наилучшую оценку $\tau(\theta)$ в классе несмещённых оценок в равномерном подходе с квадратичной функцией потерь называют оптимальной оценкой.

Замечание. В одномерном случае поиск оптимальной оценки сводится к сравнению дисперсий, на многомерный случай обобщается через сравнение матриц ковариаций с помощью неотрицательной определённости, подробно не будем на этом останавливаться.

Определение 3.22. Статистика $S(X)$ называется полной для параметра θ, если для любой измеримой функции f выполнено следствие:

$$\begin{aligned}&\forall\theta\in\Theta\ \ \mathbb{E}_\theta f(S(X))=0\\&\Downarrow\\&\forall\theta\in\Theta\ \ f(S(X))=0\ P_\theta\text{-п.н.}\end{aligned}$$

Здесь важно, что квантор по θ встречается до и после следствия.

Замечание. Грубо термин полнота появился из следующих соображений: если всегда $\int_{\mathcal{X}} f(S(X))p_\theta(X)dX=0$, рассматриваем случай доминируемого семейства распределений, то есть $f(S(X))$ ортогональна всем плотностям, то она всегда нулевая. Но аналогия здесь далеко не абсолютная.

Теорема 3.12. *(Лемана-Шеффера об оптимальной оценке) Пусть $T(X)$ – полная достаточная статистика для семейства распределений $\{P_\theta,\ \theta\in\Theta\}$, $d(X)$ – несмещённая оценка для $\tau(\theta)$. Тогда $\varphi(T(X))=\mathbb{E}_\theta(d(X)|T(X))$ – несмещённая оценка с равномерно наименьшей дисперсией для $\tau(\theta)$, про оптимальность пока не говорим, так как дисперсия может быть бесконечной. Если $D_\theta\varphi(T(X))<\infty$, то $\varphi(T(X))$ – оптимальная оценка.*

Замечание. То есть теорема Колмогорова-Блэквелла-Рао даёт не просто улучшение оценки, но ещё и её оптимальность.

Доказательство. Так как $\mathbb{E}_\theta\varphi(T(X))=\mathbb{E}_\theta(\mathbb{E}_\theta(d(X)|T(X)))=\mathbb{E}d(X)=\tau(\theta)$, то оценка является несмещённой. Если $\widetilde{d}(X)$ – другая несмещённая оценка, то по теореме Колмогорова-Блэквелла-Рао $\widetilde{\varphi}(T(X))=\mathbb{E}_\theta(\widetilde{d}(X)|T(X))$ не хуже, а в случае конечной дисперсии строго лучше, оценки $\widetilde{d}(X)$. То есть если докажем, что $\varphi(T(X))=\widetilde{\varphi}(T(X))$ P_θ-п.н. $\forall\theta\in\Theta$, то докажем теорему.

Обозначим $h(T(X))=\varphi(T(X))-\widetilde{\varphi}(T(X))$, хотим доказать, что $h(T(X))=0$ P_θ-п.н.. Это является следствием полноты статистики $T(X)$, действительно:

$$\begin{aligned}&\forall\theta\in\Theta\ \ \mathbb{E}_\theta h(X)=\mathbb{E}_\theta\varphi(T(X))-\mathbb{E}_\theta\widetilde{\varphi}(T(X))=\tau(\theta)-\tau(\theta)=0\\&\Downarrow\\&\forall\theta\in\Theta\ \ h(T(X))=0\ P_\theta\text{-п.н.}\end{aligned}$$

□

Следствие. Пусть $T(X)$ – полная достаточная статистика для семейства распределений $\{P_\theta,\ \theta\in\Theta\}$, $\varphi(T(X))$ – несмещённая оценка для $\tau(\theta)$. Тогда $\varphi(T(X))$ – несмещённая оценка с равномерно наименьшей дисперсией для $\tau(\theta)$. Если $D_\theta\varphi(T(X))<\infty$, то $\varphi(T(X))$ – оптимальная оценка.

Теорема 3.13. *(Об экспоненциальном семействе, без доказательства) Пусть X – выборка из экспоненциального семейства распределений с обобщённой плотностью*

$$p_\theta(x) = h(x)\exp\left(\sum_{i=1}^{k} a_i(\theta)T_i(x) + V(\theta)\right)$$

Если область значений вектор-функции $(a_1(\theta), \ldots, a_k(\theta))$ содержит k-мерный параллелипипед в $\mathbb{R}^k$, то

$$T(X) = \left(\sum_{i=1}^{n} T_1(X_j), \ldots, \sum_{i=1}^{n} T_k(X_j)\right)$$

является полной достаточной статистикой для параметра θ.

Замечание. k-мерный параллелепипед можно заменить на открытое множество, так как одно всегда содержит второе и наоборот.

Замечание. Алгоритм поиска оптимальной оценки:

1. Ищем достаточную статистику $T(X)$.
2. Проверяем на полноту, если экспоненциальное семейство распределений, то всё хорошо, иначе, кажется, грустить и проверять по определению.
3. Если полная, то решаем уравнение несмещённости, подбирая g:

$$\mathbb{E}_\theta g(T(X)) = \tau(\theta) \quad \forall\theta \in Theta$$

4. Если $D_\theta g(T(X)) < \infty$, то $g(T(X))$ – оптимальная оценка.

Замечание. Причём, по-хорошему, для конечности дисперсии достаточно одного θ, так как знаем, что при всех остальных θ оценка не хуже.

4 Доверительные интервалы

4.1 Общие определения

Определение 4.1. Пусть X – выборка из неизвестного распределения $P \in \{P_\theta,\ \theta \in \Theta\}$. Пара статистик $(T_1(X), T_2(X))$ называется доверительным интервалом уровня доверия γ для параметра $\theta \in \Theta \subset \mathbb{R}$, если:

$$\forall\theta \in \Theta \quad P_\theta(T_1(X) < \theta < T_2(X)) \geqslant \gamma$$

Если равенство достигается при всех $\theta \in \Theta$, то доверительный интервал называется точным.

Замечание. На практике обычно используют $\gamma = 0.9,\ 0.95,\ 0.99$. Иногда удобно использовать односторонние доверительные интервалы $(-\infty, T_2(X))$ или $(T_1(X), +\infty)$. В случае многомерного $\theta \in \Theta \subset \mathbb{R}^k$ можно аналогично определить понятие доверительного интервала для компонент θ_i вектора $\theta = (\theta_1, \ldots, \theta_k)$. И во всех случаях можно обощить понятие доверительного интервала на скалярные (действующие в $\mathbb{R}$) функции $\tau(\theta)$.

Определение 4.2. Пусть X – выборка из неизвестного распределения $P \in \{P_\theta,\ \theta \in \Theta\}$. Множество $S(X)$ называется доверительным множеством уровня доверия γ для параметра $\theta \in \Theta \subset \mathbb{R}^k$, если:

$$\forall \theta \in \Theta \quad P_\theta(\theta \in S(X)) \geqslant \gamma$$

4.2 Метод центральной статистики

Пример. Пусть $X_1, \ldots, X_n$ – выборка из нормального распределения $N(\theta, 1)$, $\theta \in \mathbb{R}$. Хотим построить доверительный интервал для параметра θ уровня доверия γ. Заметим, что из свойств нормального распределения:

$$\overline{X} \sim N\left(\theta, \frac{1}{n}\right) \Rightarrow \sqrt{n}(\overline{X} - \theta) \sim N(0, 1)$$

Обозначим u_p – p-квантиль $N(0,1)$. Тогда:

$$\forall \theta \in \Theta \quad P\left(u_{\frac{1-\gamma}{2}} < \sqrt{n}(\overline{X} - \theta) < u_{\frac{1+\gamma}{2}}\right) = \gamma$$

$$\forall \theta \in \Theta \quad P\left(-u_{\frac{1+\gamma}{2}} < \sqrt{n}(\overline{X} - \theta) < u_{\frac{1+\gamma}{2}}\right) = \gamma$$

$$\forall \theta \in \Theta \quad P\left(\overline{X} - \frac{u_{\frac{1+\gamma}{2}}}{\sqrt{n}} < \theta < \overline{X} + \frac{u_{\frac{1+\gamma}{2}}}{\sqrt{n}}\right) = \gamma$$

$$\left(\overline{X} - \frac{u_{\frac{1+\gamma}{2}}}{\sqrt{n}}, \overline{X} + \frac{u_{\frac{1+\gamma}{2}}}{\sqrt{n}}\right) \text{ – точный ДИ у.д. } \gamma$$

Эту конструкцию можно обобщить.

Определение 4.3. Пусть X – выборка из неизвестного распределения $P \in \{P_\theta,\ \theta \in \Theta\}$. Пусть существует известная функция $G(X, \theta)$ такая, что её распределение не зависит от параметра θ. Такая функция G называется центральной статистикой.

Отметим, что G не является статистикой в обычном понимании слова, ибо зависит от параметра θ. Но название, видимо, устоялось. То есть центральная статистика не очень центральная и не очень статистика.

Пусть $\gamma_1, \gamma_2 \in (0,1)$ такие, что $\gamma_2 - \gamma_1 = \gamma$. Пусть g_1, g_2 – γ_1 и γ_2-квантили распределения $G(X, \theta)$. Тогда, аккуратно поясним ниже, выполнено:

$$\forall \theta \in \Theta \quad P_\theta(g_1 \leqslant G(X, \theta) \leqslant g_2) \geqslant \gamma_2 - \gamma_1 = \gamma$$

Введём обозначение $S(X) = \{\theta \in \Theta : g_1 \leqslant G(X, \theta) \leqslant g_2\}$. Тогда для любого $\theta \in \Theta$ имеем $P_\theta(\theta \in S(X)) = P_\theta(g_1 \leqslant G(X, \theta) \leqslant g_2) \geqslant \gamma$, то есть $S(X)$ – доверительное множество уровня доверия γ.

Замечание. Поясним здесь неравенство для квантилей, в случае непрерывных распределений оно очевидно и вырождается в равенство, иначе вызывает вопросы. По определению квантилей:

$$g_\gamma = \inf\{x : F(x) \geqslant \gamma\} = \min\{x : F(x) \geqslant \gamma\}$$

Здесь F – функция распределения, инфимум достигается, так как F непрерывна справа.

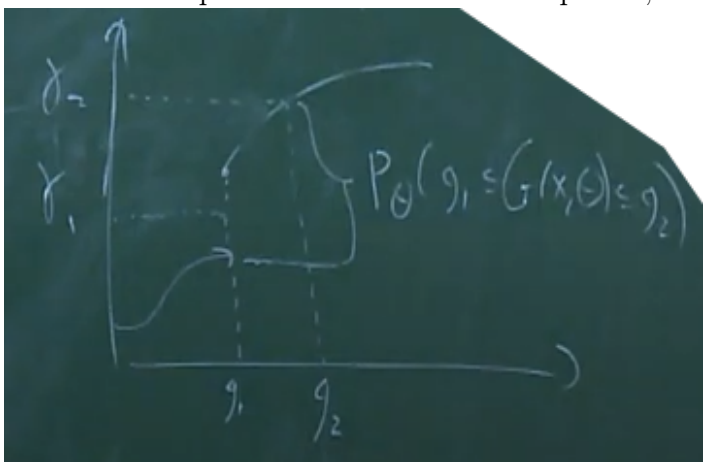
Тогда, используя обозначение $F(x-0)$ для предела слева в точке x, получим:

$$F(g_\gamma) \geqslant \gamma$$
$$F(g_\gamma - \varepsilon) < \gamma \Rightarrow F(g_\gamma - 0) \leqslant \gamma$$

Отсюда уже выводим неравенство:

$$P_\theta(g_1 \leqslant G(X,\theta) \leqslant g_2) = P(G \leqslant g_2) - P(G < g_1) = F(g_2) - F(g_1 - 0) \geqslant \gamma_2 - \gamma_1$$

Замечание. Неравенство может быть строгим, например:



Замечание. Квантили g_1 и g_2 можно найти в статистических таблицах, либо с помощью функций из некоторых библиотек, их вычисляющих.

Замечание. Не всегда центральную статистику можно найти методом пристального взгляда. Для таких случаев есть полезный результат.

Лемма 4.1. *Пусть $X_1, \ldots, X_n$ – независимые одинаково распределённые случайные величины с функцией распределения $F(x)$, $F(x)$ непрерывна. Тогда имеет место распределение:*

$$G(X_1, \ldots, X_n) = -\sum_{i=1}^{n} \ln F(X_i) \sim \Gamma(1, n)$$

Доказательство.

▷ $F(X_i) \sim U[0,1]$

Если функция распределения строго монотонна, в таком случае, с учётом непрерывности, она биективно отображает $\mathbb{R}$ на $(0,1)$, то доказательство простое:

$$\forall y \in (0,1) \quad P(F(X_i) \leqslant y) = P(X_i \leqslant F^{-1}(y)) = F(F^{-1}(y)) = y$$

И получилось, что функция распределения $F(X_i)$ совпадает с функцией распределения $U[0,1]$. В общем случае это рассуждение нужно немного пофиксить:

$$\forall y \in (0,1) \quad P(F(X_i) \leqslant y) = P(X_i \leqslant \max\{x : F(x) \leqslant y\}) = F(\max\{x : F(x) \leqslant y\}) = y$$

Максимум достигается в силу того, что функция распределения непрерывна.

▷ $-\ln U[0,1] \sim Exp(1)$

Пусть $\xi \sim U[0,1]$, тогда:

$$\forall y > 0 \quad P(-\ln \xi \leqslant y) = P(\xi \geqslant e^{-y}) = 1 - P(\xi < e^{-y}) = [e^{-y} \in (0,1)] = 1 - e^{-y}$$
$$P(-\ln \xi \leqslant y) = (1 - e^{-y}) I\{y > 0\}$$

Получили в точности функцию распределения $Exp(1)$.

▷ $G(X_1, \ldots, X_n) \sim \Gamma(1, n)$

Применим накопленные знания, воспользовавшись также формулой для суммы независимых гамма-распределений:

$$G(X_1, \ldots, X_n) = -\sum_{i=1}^{n} \ln F(X_i) \sim -\sum_{i=1}^{n} \ln U[0,1] \sim \sum_{i=1}^{n} Exp(1) \sim \sum_{i=1}^{n} \Gamma(1,1) \sim \Gamma(1,n)$$

□

Следствие. Если $X_1, \ldots, X_n$ – выборка из неизвестного распределения $P \in \{P_\theta,\, \theta \in \Theta\}$, причём для всех $\theta \in \Theta$ функция распределения $F_\theta(x)$ непрерывна по x, то $G(X_1, \ldots, X_n, \theta) = -\sum_{i=1}^{n} \ln F_\theta(X_i)$ является центральной статистикой.

4.3 Асимптотические доверительные интервалы

Определение 4.4. Пусть $(X_n,\, n \geqslant 1)$ – выборка неограниченного размера из неизвестного распределения $P \in \{P_\theta,\, \theta \in \Theta\}$. 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 γ для неизвестного параметра θ, если

$$\forall \theta \in \Theta \quad \varliminf_{n \to \infty} P_\theta(T_n^{(1)}(\ldots) < \theta < T_n^{(2)}(\ldots)) \geqslant \gamma$$

Если при этом

$$\forall \theta \in \Theta \quad \lim_{n \to \infty} P_\theta(T_n^{(1)}(\ldots) < \theta < T_n^{(2)}(\ldots)) = \gamma$$

то асимптотический доверительный интервал называют точным.

Замечание. Нижний предел взяли из тех соображений, что обычный предел может не существовать, в том числе тогда, когда интервал с хорошей вероятностью оценивает параметр. В случае точного АДИ уже разумно брать обычный предел.

Замечание. Асимптотический доверительный интервал можно построить с помощью асимптотически нормальной оценки. Пусть $\hat{\theta}_n(X_1, \ldots, X_n)$ – асимптотически нормальная оценка θ с асимптотической дисперсией $\sigma^2(\theta) > 0$, то есть:

$$\forall \theta \quad \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d_\theta} N(0, \sigma^2(\theta))$$
$$\forall \theta \quad \sqrt{n}\frac{\hat{\theta}_n - \theta}{\sigma(\theta)} \xrightarrow{d_\theta} N(0,1)$$

Стандартно обозначим u_γ – γ-квантиль $N(0,1)$, тогда в силу сходимости по распределению и того, что у $N(0,1)$ функция распределения всюду непрерывна, получим:

$$\lim_{n\to\infty} P_\theta\left(-u_{\frac{1+\gamma}{2}} < \sqrt{n}\frac{\hat\theta_n - \theta}{\sigma(\theta)} < u_{\frac{1+\gamma}{2}}\right) = P_\theta\left(-u_{\frac{1+\gamma}{2}} < N(0,1) < u_{\frac{1+\gamma}{2}}\right) = \gamma$$

Отсюда хочется выразить асимптотический доверительный интервал для θ, но есть одна проблема: знаменатель $\sigma(\theta)$ зависит от θ, поэтому просто выразить не получится. Исправим это, заменив $\sigma(\theta)$ на что-то, зависящее от выборки.

Пусть $\sigma(\theta)$ непрерывна по θ. Из асимптотической нормальности оценки следует её состоятельность, поэтому $\hat\theta_n \xrightarrow{P_\theta} \theta$ и в силу теоремы о наследовании сходимости $\sigma(\hat\theta_n) \xrightarrow{P_\theta} \sigma(\theta)$. Тогда по лемме Слуцкого:

$$\forall\theta \quad \sqrt{n}\frac{\hat\theta_n - \theta}{\sigma(\hat\theta_n)} = \sqrt{n}\frac{\hat\theta_n - \theta}{\sigma(\theta)} \cdot \frac{\sigma(\theta)}{\sigma(\hat\theta_n)} \xrightarrow{d_\theta} N(0,1)\cdot 1 = N(0,1)$$

В таком случае, в соответствии с рассуждениями выше, можем получить асимптотический доверительный интервал для θ уровня доверия γ:

$$\left(\hat\theta_n - u_{\frac{1+\gamma}{2}}\frac{\sigma(\hat\theta_n)}{\sqrt{n}},\ \hat\theta_n + u_{\frac{1+\gamma}{2}}\frac{\sigma(\hat\theta_n)}{\sqrt{n}}\right)$$

4.4 Метод максимального правдоподобия

Определение 4.5. Пусть X – выборка размера 1 из распределения $P \in \{P_\theta,\ \theta\in\Theta\}$, а семейство распределений $\{P_\theta,\ \theta\in\Theta\}$ доминируемо относительно меры μ. Тогда функцией правдоподобия называется

$$f_\theta(X) = p_\theta(X)$$

Здесь $p_\theta(X)$ – плотность P_θ по мере μ. Важно, что функцию правдоподобия мы рассматриваем как функцию от θ при фиксированном X, а не наоборот. Функция правдоподобия, разумеется, не обязана быть вероятностной мерой на Θ, это просто какая-то функция.

Пример. Пусть $X = (X_1,\dots,X_n)$ – выборка из распределения $P \in \{P_\theta,\ \theta\in\Theta\}$, семейство распределений $\{P_\theta,\ \theta\in\Theta\}$ доминируемо относительно меры μ. Тогда функцией правдоподобия является

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

Действительно, если рассматривать X как случайный вектор из многомерного распределения, то именно функция $\prod_{i=1}^n p_\theta(x_i)$ будет плотностью этого распределения, что легко выводится из теоремы Фубини-Тонелли.

Определение 4.6. Пусть X – выборка с функцией правдоподобия $f_\theta(X)$. Оценкой параметра θ по методу максимального правдоподобия (ОМП) называется такая статистика $\hat\theta(X)$, что

$$\hat\theta(X) = \operatorname{argmax}_{\theta\in\Theta} f_\theta(X)$$

Замечание. Заметим, что в определение через термин статистика сразу закладываем

измеримость. Соответственно, проблемы могут заключаться в том, что argmax может не существовать, быть не единственным, и не очевидна измеримость полученной функции.

Пример. Пусть есть монетка с распределением $Bern(p)$, причём известно что p является одним из двух значений $p_1 = \frac{1}{9}$ или $p_2 = \frac{7}{8}$. Пусть есть выборка из трёх бросков монетки, запишем в виде 110. Тогда посмотрим на вероятность такого события в случае разных p:

$$p_1 \colon P_{p_1}(110) = \frac{1}{9} \cdot \frac{1}{9} \cdot \frac{8}{9}$$
$$p_2 \colon P_{p_2}(110) = \frac{7}{8} \cdot \frac{7}{8} \cdot \frac{1}{8}$$

Во втором случае вероятность больше, то есть вторая модель более правдоподобна, гораздо лучше предсказывает то, что произошло в реальности. В этом и заключается идея оценки максимального правдоподобия.

Пример. Пусть есть выборка $X_1, \ldots, X_n \sim U[0, \theta]$. Правдоподобие в таком случае имеет вид:

$$f_\theta(X_1, \ldots, X_n) = \frac{1}{\theta^n} \prod_{i=1}^{n} I\{0 \leqslant X_i \leqslant \theta\} = \frac{1}{\theta^n} \prod_{i=1}^{n} I\{0 \leqslant X_{(1)} \leqslant X_{(n)} \leqslant \theta\}$$

Тогда ОМП есть $\hat{\theta}(X) = X_{(n)}$.

Замечание. Ту теорию, которую будем дальше развивать в этом разделе, можно также посмотреть в книге: Леман Теория точечного оценивания.

Определение 4.7. Пусть $f_\theta(X)$ – функция правдоподобия, тогда $L_\theta(X) = \ln f_\theta(X)$ называется логарифмической функцией правдоподобия. Так как плотность принимает значение ноль на элементах выборки с нулевой вероятностью, ибо интеграл от нуля равен нулю, то можем брать логарифм.

Определение 4.8. Нам вновь понадобятся условия регулярности. На этот раз их будет больше. Опять, сначала сформулируем, потом дадим пояснения.

(R0) $\{P_\theta, \theta \in \Theta\}$ – параметрическое семейство распределений, доминируемое относительно меры μ, $P_{\theta_1} \neq P_{\theta_2}$ при $\theta_1 \neq \theta_2$. Для всех θ обозначим $p_\theta(x)$ – плотность P_θ относительно меры μ.

(R1) Множество $A = \{x \in \mathcal{X} \colon p_\theta(x) > 0\}$ не зависит от θ, A называется носителем.

(R2) X есть выборка из неизвестного распределения $P \in \{P_\theta, \theta \in \Theta\}$.

(R3) Θ – открытый интервал, возможно, бесконечный.

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

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

(R6) Интеграл $\int_A p_\theta(x)\mu(dx)$ трижды дифференцируем по θ под знаком интеграла.

(R7) Информация Фишера $i(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial\theta} \ln p_\theta(X_1)\right)^2 \in (0, +\infty)$

(R8) $\forall \theta_0 \in \Theta \;\; \exists c > 0 \; \exists H(x) \;\; \forall \theta \in (\theta_0 - c, \theta_0 + c) \; \forall x \in A$ выполнено:

$$\left| \frac{\partial^3}{\partial \theta^3} \ln p_\theta(x) \right| \leqslant H(x)$$
$$\mathbb{E}_{\theta_0} H(X_1) < \infty$$

Замечание. Комментарии к некоторым условиям регулярности, те пояснения, которые были к условиям регулярности для эффективных оценок, дублировать не буду:

(R0) Требование $P_{\theta_1} \neq P_{\theta_2}$ при $\theta_1 \neq \theta_2$ важно в следующем смысле, например, имеем семейство распределений $Bern(\theta^2)$ и поняли, что $\theta^2 = \frac{1}{4}$, отсюда никак не сможем сделать вывод о том, чему равно θ, $\frac{1}{2}$ или $-\frac{1}{2}$.

(R2) Просто зафиксировали выборку.

(R5) Условие (R5), конечно, сильнее условия (R4), оба нужны, потому что в разных теоремах будут предполагаться разные условия регулярности. Сначала будут использоваться (R0)-(R2), как самые базовые, в которых осмыслено работать. Затем захотим дифференцировать по θ, будем жить в условиях (R0)-(R4). Затем потребуется раскладывать функцию по формуле Тейлора до некоторого порядка, будут использоваться все условия (R0)-(R8).

(R6) Полагаю, смысл следующий, дифференцируемость подинтегральных функций знаем из предыдущего пункта, здесь утверждение в том, что существуют конечные интегралы $\int_A p_\theta(x)\mu(dx)$, $\int_A \frac{\partial}{\partial\theta} p_\theta(x)\mu(dx)$, $\int_A \frac{\partial^2}{\partial\theta^2} p_\theta(x)\mu(dx)$, $\int_A \frac{\partial^3}{\partial\theta^3} p_\theta(x)\mu(dx)$.

(R8) Здесь Савёлов сказал, что смысл станет понятен при доказательстве соответствующей теоремы, которую потом доказывать не стал. Тем не менее, сейчас некоторую интуицию для этого условия я приведу.

Напоминание. (Теорема о непрерывности собственного интеграла по параметру) Пусть $A \subseteq \mathbb{R}^n$, $E \subseteq \mathbb{R}^m$ — измеримые множества и задана функция $f\colon E \times A \to \mathbb{R}$. Если наложены следующие условия:

1. Для любого $\alpha \in A$ функция $f(x, \alpha)$ измерима на E
2. Почти всюду на E выполнено $|f(x, \alpha)| \leqslant \varphi(x)$, где $\varphi \in L_1(E)$
3. Почти всюду на E имеет место сходимость $f(x, \alpha) \to f(x, \alpha_0)$ при $\alpha \to \alpha_0$, $\alpha \in A$

Тогда интеграл $\int_E f(x, \alpha) d\mu(x)$ непрерывен в точке α_0, то есть имеется предел:

$$\lim_{\alpha \to \alpha_0} \int_E f(x, \alpha) d\mu(x) = \int_E f(x, \alpha_0) d\mu(x)$$

Замечание. Теперь поймём смысл (R8). Выведем из похожего условия (RR8), которое будет сформулировано ниже, в условиях регулярности (R0)-(R4) и (R7) непрерывность информации Фишера $i(\theta)$.

Рассмотрим произвольную точку $\theta_0 \in \Theta$, хотим в ней доказать непрерывность функции:

$$i(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln p_\theta(X_1) \right)^2 = \int_A \left(\frac{\partial}{\partial \theta} \ln p_\theta(x) \right)^2 p_\theta(x) d\mu(x)$$

Рассмотрим последний интеграл в некоторой окрестности $(\theta_0 - c, \theta_0 + c)$ точки θ_0, чтобы доказать непрерывность в θ_0, остальные θ нам не интересны. Хотим применить для множества $A \times (\theta_0 - c, \theta_0 + c)$, рассматривается $x \in A$, $\theta \in (\theta_0 - c, \theta_0 + c)$, и функции на нём $\left(\frac{\partial}{\partial\theta} \ln p_\theta(x)\right)^2 p_\theta(x)$, теорему о непрерывности собственного интеграла по параметру, тогда получим в точности то, что хотим.

Измеримость функции $\left(\frac{\partial}{\partial\theta} \ln p_\theta(x)\right)^2 p_\theta(x)$ как функции от x следует из того, что интеграл по ней, то есть информация Фишера, в силу (R7) существует и конечен. Непрерывность этой функции по θ следует из того, что в силу (R4) $p_\theta(x)$ непрерывно дифференцируема по θ. Осталось мажорируемость интегрируемой функцией, для этого сформулируем условие (RR8):

(RR8) $\forall\theta_0 \in \Theta \;\; \exists c > 0 \; \exists H(x) \;\; \forall\theta \in (\theta_0 - c, \theta_0 + c) \; \forall x \in A$ выполнено:

$$\left(\frac{\partial}{\partial\theta} \ln p_\theta(x)\right)^2 p_\theta(x) \leqslant H(x)$$
$$\int_A H(x)\mu(dx) < \infty$$

Применяем теорему о непрерывности собственного интеграла по параметру, получаем, что информация Фишера $i(\theta)$ непрерывна в произвольной точке θ_0, значит, непрерывна. Запомним этот факт, он нам ещё пригодится.

Теперь сравним условия (R8) и (RR8):

(R8) $\forall\theta_0 \in \Theta \;\; \exists c > 0 \; \exists H(x) \;\; \forall\theta \in (\theta_0 - c, \theta_0 + c) \; \forall x \in A$ выполнено:

$$\left|\frac{\partial^3}{\partial\theta^3} \ln p_\theta(x)\right| \leqslant H(x)$$
$$\mathbb{E}_{\theta_0} H(X_1) < \infty$$

(RR8) $\forall\theta_0 \in \Theta \;\; \exists c > 0 \; \exists H(x) \;\; \forall\theta \in (\theta_0 - c, \theta_0 + c) \; \forall x \in A$ выполнено:

$$\left(\frac{\partial}{\partial\theta} \ln p_\theta(x)\right)^2 p_\theta(x) \leqslant H(x)$$
$$\int_A H(x)\mu(dx) < \infty$$

В (R8) также мажорируется некоторая функция, конкретно, третья производная логарифмического правдоподобия, но рассматривается не интеграл от мажоранты, не зависящий от θ, а интеграл от мажоранты по конкретной плотности, плотности меры P_{θ_0}. Соответственно, беглый просмотр того доказательства, которое в этот раз не приводилось и в котором использовалось условие (R8) показал, что это условие там использовалось совершенно не так, как применялось здесь.

Теорема 4.1. *(Экстремальное свойство правдоподобия) Пусть выполнены условия регулярности (R0)-(R2). Тогда*

$$\forall\theta_0, \theta \in \Theta, \; \theta_0 \neq \theta \;\; P_{\theta_0}(f_{\theta_0}(X_1, \ldots, X_n) > f_\theta(X_1, \ldots, X_n)) \xrightarrow[n\to\infty]{} 1$$

Замечание. В чём смысл теоремы: настоящий параметр θ_0 становится правдоподобнее любого конкурента θ, если увеличивать размер выборки.

Лемма 4.2. *Пусть* $\{\xi_n\}_{n=1}^{\infty}$ *– независимые одинаково распределённые случайные величины. Пусть* $\mathbb{E}\xi_1 \in \mathbb{R} \cup \{-\infty, +\infty\}$. *Обозначим* $S_n = \xi_1 + \cdots + \xi_n$. *Тогда*

$$\frac{S_n}{n} \xrightarrow{\text{п.н.}} \mathbb{E}\xi_1.$$

Доказательство. Если $\mathbb{E}\xi_1$ конечно, то утверждение непосредственно следует из УЗБЧ. Рассмотрим случай $\mathbb{E}\xi_1 = +\infty$, случай $\mathbb{E}\xi_1 = -\infty$ рассматривается аналогично. Возьмём срезки от случайных величин $\xi_{n,[N]} = \min(\xi_n, N)$, тогда $\xi_{n,[N]}$ – независимые одинаково распределённые случайные величины, $-\infty < \mathbb{E}\xi_{1,[N]} \leqslant N < +\infty$. В силу УЗБЧ:

$$\frac{S_{n,[N]}}{n} \xrightarrow{\text{п.н.}} \mathbb{E}\xi_{1,[N]}$$

Теперь, применяя, что $\mathbb{E}\xi_{1,[N]} \xrightarrow[N\to\infty]{} +\infty$, так как $\mathbb{E}\xi_1 = +\infty$, получим:

$$\frac{S_n}{n} \geqslant \frac{S_{n,[N]}}{n} \xrightarrow{\text{п.н.}} \mathbb{E}\xi_{1,[N]} \xrightarrow[N\to\infty]{} +\infty$$

В целом, уже всё доказали, формально доводится следующим образом:

$$\forall K \in \mathbb{N} \quad P\left(\varliminf_{n\to\infty} \frac{S_n}{n} \geqslant K\right) = 1$$

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = +\infty\right) = P\left(\bigcap_{K=1}^{\infty} \varliminf_{n\to\infty} \frac{S_n}{n} \geqslant K\right) = \lim_{K\to\infty} P\left(\varliminf_{n\to\infty} \frac{S_n}{n} \geqslant K\right) = 1$$

Совсем просто сделать нельзя, так как множества для сходимости почти наверное могут отличаться для разных N. □

Доказательство. (теоремы об экстремальном свойстве правдоподобия) Зафиксируем параметры $\theta_0 \neq \theta$. Можем считать, что все элементы выборки $X_1, \ldots, X_n \in A$, т.е. имеют ненулевую плотность. Имеем на это право, ибо такое выполнено почти наверное для всех $P \in \{P_\theta,\, \theta \in \Theta\}$:

$$P(\forall i \in \mathbb{N}\ X_i \in A) = P\left(\bigcap_{i=1}^{\infty}\{X_i \in A\}\right) = \lim_{n\to\infty} P\left(\bigcap_{i=1}^{n}\{X_i \in A\}\right) = \lim_{n\to\infty} 1 = 1$$

Хотим доказать, что $P_{\theta_0}(f_{\theta_0}(X_1, \ldots, X_n) > f_\theta(X_1, \ldots, X_n)) \xrightarrow[n\to\infty]{} 1$. Можем переписать неравенство для функций правдоподобия в виде, к которому можно будет применить доказанную только что лемму об обобщённом УЗБЧ, здесь пользуемся тем, что плотность

не равна нулю:

$$
\begin{aligned}
&f_{\theta_0}(X_1,\ldots,X_n) > f_\theta(X_1,\ldots,X_n)\\
&\Updownarrow\\
&\ln\frac{f_\theta(X_1,\ldots,X_n)}{f_{\theta_0}(X_1,\ldots,X_n)} < 0\\
&\Updownarrow\\
&\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} < 0
\end{aligned}
$$

Далее хотим показать, что $-\infty \leqslant \mathbb{E}_{\theta_0}\ln\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} < 0$. Если это так, то всё доказано, Савёлов это подробнее не обосновывал, мы распишем аккуратно, действительно, по лемме:

$$
\begin{aligned}
&P_{\theta_0}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} = \mathbb{E}_{\theta_0}\ln\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}\right) = 1\\
&P_{\theta_0}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} < 0\right) = 1\\
&P_{\theta_0}\left(\exists N\ \forall n \geqslant N\ \ \frac{1}{n}\sum_{i=1}^{n}\ln\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} < 0\right) = 1\\
&P_{\theta_0}\left(\bigcup_{N=1}^{\infty}\bigcap_{n=N}^{\infty}\left\{\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} < 0\right\}\right) = 1
\end{aligned}
$$

И отсюда получаем требуемую предельную вероятность:

$$
\begin{aligned}
&\lim_{n\to\infty}P_{\theta_0}(f_{\theta_0}(X_1,\ldots,X_n) > f_\theta(X_1,\ldots,X_n)) = \lim_{n\to\infty}P_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} < 0\right) \geqslant\\
&\geqslant \lim_{N\to\infty}P_{\theta_0}\left(\bigcap_{n=N}^{\infty}\left\{\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} < 0\right\}\right) = P_{\theta_0}\left(\bigcup_{N=1}^{\infty}\bigcap_{n=N}^{\infty}\left\{\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} < 0\right\}\right) = 1
\end{aligned}
$$

Осталось доказать, что $\mathbb{E}_{\theta_0}\ln\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} < 0$. Для этого вспомним факт из математического анализа: $\ln(1+x) \leqslant x$ при всех $x > -1$, равенство достигается только при $x = 0$. Дальнейшее доказательство будет состоять из серии хитрых и не очень преобразований:

$$
\begin{aligned}
\mathbb{E}_{\theta_0}\ln\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} &= \int_A \ln\frac{p_\theta(x)}{p_{\theta_0}(x)}p_{\theta_0}(x)\mu(dx) = \int_A \ln\left(1+\frac{p_\theta(x)}{p_{\theta_0}(x)}-1\right)p_{\theta_0}(x)\mu(dx) \leqslant\\
&\leqslant \int_A\left(\frac{p_\theta(x)}{p_{\theta_0}(x)}-1\right)p_{\theta_0}(x)\mu(dx) = \int_A (p_\theta(x)-p_{\theta_0}(x))\,\mu(dx) = 1-1 = 0
\end{aligned}
$$

Пока доказали нестрогую версию неравенства, $\mathbb{E}_{\theta_0}\ln\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)} \leqslant 0$, но равенство может быть лишь в случае $\frac{p_\theta(x)}{p_{\theta_0}(x)} - 1 = 0$ μ-п.н. $\Leftrightarrow p_\theta(x) = p_{\theta_0}(x)$ μ-п.н., то есть $P_{\theta_0} = P_\theta$, тогда $\theta = \theta_0$ в силу одного из условий регулярности, что невозможно в силу выбора $\theta \neq \theta_0$. $\square$