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

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

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 упражнением.

4. Сходимость по распределению

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

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

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:

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

   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:

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

$\square$

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:

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

$\square$

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. $\xi_n \xrightarrow{\text{п.н.}} \xi \Rightarrow h(\xi_n) \xrightarrow{\text{п.н.}} h(\xi)$
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 $h(\xi_n) \overset{P}{\not\to} h(\xi)$, то:

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

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 $S_n = \xi_1 + \ldots \xi_n$.

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. $g(x,y)=\langle A(x-y),x-y\rangle=(x-y)^TA^T(x-y)=(x-y)^TA(x-y)$, 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.

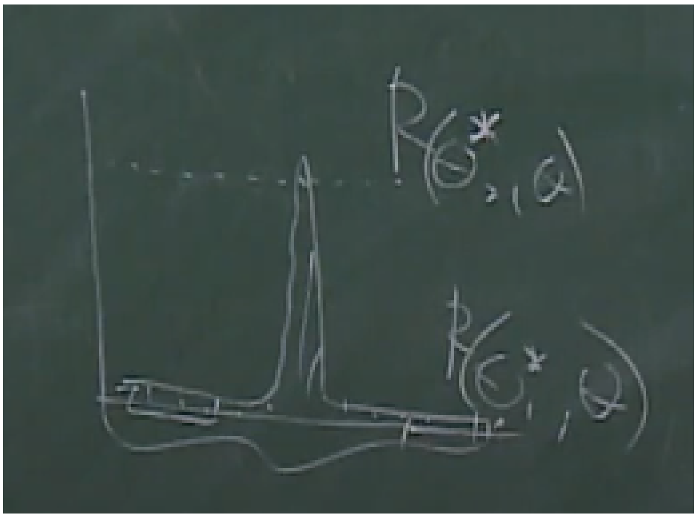
3.2.2 Минимаксный подход

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.



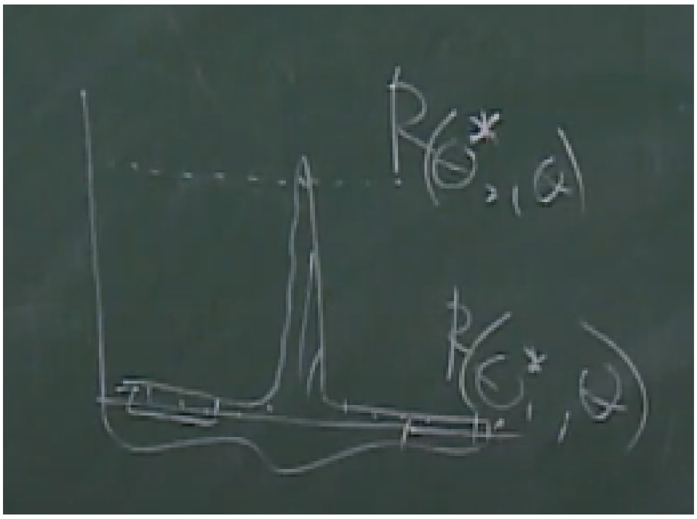
3.2.3 Байесовский подход

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 $X_1, \ldots, X_n$ – 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):

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

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:

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

$$= \sum_{i=1}^n \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln p_\theta(X_i) \right)^2 + \sum_{i,j=1}^n \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln p_\theta(X_i) \right) \left(\frac{\partial}{\partial \theta} \ln p_\theta(X_j) \right) =$$

$$= \sum_{i=1}^n \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln p_\theta(X_i) \right)^2 + \sum_{i,j=1}^n \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln p_\theta(X_i) \right) \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln p_\theta(X_j) \right) =$$

$$= \sum_{i=1}^n i(\theta) + \sum_{i,j=1}^n (\mathbb{E}_\theta U_\theta(X_1))^2 = ni(\theta)$$

□

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:*

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

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 $a(\theta)[\hat{\theta}(X) - \tau(\theta)] = b(\theta)U_\theta(X)\ P_\theta$-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 $\hat{\theta}(X) - \tau(\theta) = c(\theta)U_\theta(X)\ P_\theta$-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 $\alpha(\theta)[\hat{\theta}(X) - \tau(\theta)] + \beta(\theta)U_\theta(X) + \gamma(\theta) = 0\ P_\theta$-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:

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

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 эквивалентно 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:

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

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:

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

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:

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

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.

$\square$