Лекции по математической статистике

2-й курс ЭФ, отделение
«математические методы и исследование операций в экономике»

Н. И. Чернова
cher@nsu.ru

Предлагаемый вашему вниманию курс теоретической статистики содержит материал из классических разделов математической статистики. Речь пойдет об оценке параметров, проверке гипотез, немного о регрессионном анализе. Курс предназначен студентам экономического факультета НГУ, но его могут попробовать освоить студенты математического факультета. Курс не содержит экономических приложений и ни в коей мере не собирается обсуждать применение статистических методов. И то, и другое студенты-экономисты в НГУ изучают в годовом курсе эконометрики (регрессионного анализа).

Оглавление

1. Основные понятия математической статистики

Математическая (или теоретическая) статистика опирается на методы и понятия теории вероятностей, но решает в каком-то смысле обратные задачи.

В теории вероятностей рассматриваются случайные величины с заданным распределением или случайные эксперименты, свойства которых целиком известны. Предмет теории вероятностей — свойства и взаимосвязи этих величин (распределений).

Но часто эксперимент представляет собой черный ящик, выдающий лишь некие результаты, по которым требуется сделать вывод о свойствах самого эксперимента. Наблюдатель имеет набор числовых (или их можно сделать числовыми) результатов, полученных повторением одного и того же случайного эксперимента в одинаковых условиях.

При этом возникают, например, следующие вопросы:

Если мы наблюдаем одну случайную величину — как по набору ее значений в нескольких опытах сделать как можно более точный вывод о ее распределении?

Если мы наблюдаем одновременно проявление двух (или более) признаков, т. е. имеем набор значений нескольких случайных величин — что можно сказать об их зависимости? Есть она или нет? А если есть, то какова эта зависимость?

Часто бывает возможно высказать некие предположения о распределении, спрятанном в «черном ящике», или о его свойствах. В этом случае по опытным данным требуется подтвердить или опровергнуть эти предположения («гипотезы»). При этом надо помнить, что ответ «да» или «нет» может быть дан лишь с определенной степенью достоверности, и чем дольше мы можем продолжать эксперимент, тем точнее могут быть выводы. Наиболее благоприятной для исследования оказывается ситуация, когда можно уверенно утверждать о некоторых свойствах наблюдаемого эксперимента — например, о наличии функциональной зависимости между наблюдаемыми величинами, о нормальности распределения, о его симметричности, о наличии у распределения плотности или о его дискретном характере, и т. д.

Итак, о (математической) статистике имеет смысл вспоминать, если

- имеется случайный эксперимент, свойства которого частично или полностью неизвестны,
- мы умеем воспроизводить этот эксперимент в одних и тех же условиях некоторое (а лучше — какое угодно) число раз.

Примером такой серии экспериментов может служить социологический опрос, набор экономических показателей или, наконец, последовательность гербов и решек при тысячекратном подбрасывании монеты.

1.1. Основные понятия выборочного метода

Пусть $\xi : \Omega \to \mathbb{R}$ — случайная величина, наблюдаемая в случайном эксперименте. Предполагается, что вероятностное пространство задано (и не будет нас интересовать).

Будем считать, что проведя n раз этот эксперимент в одинаковых условиях, мы получили числа X_1, X_2, $\ldots$, X_n — значения этой случайной величины в первом, втором, и т. д. экспериментах. Случайная величина ξ имеет некоторое распределение $\mathcal{F}$, которое нам частично или полностью неизвестно.

Рассмотрим подробнее набор $\mathbf{X} = (X_1, \ldots, X_n)$, называемый выборкой.

В серии уже произведенных экспериментов выборка — это набор чисел. Но если эту серию экспериментов повторить еще раз, то вместо этого набора мы получим новый набор чисел. Вместо числа X_1 появится другое число — одно из значений случайной величины ξ. То есть X_1 (и X_2, и X_3, и т. д.) — переменная величина, которая может принимать те же значения, что и случайная величина ξ, и так же часто (с теми же вероятностями). Поэтому до опыта X_1 — случайная величина, одинаково распределенная с ξ, а после опыта — число, которое мы наблюдаем в данном первом эксперименте, т. е. одно из возможных значений случайной величины X_1.

Выборка $\mathbf{X} = (X_1, \ldots, X_n)$ объема n — это набор из n независимых и одинаково распределенных случайных величин («копий ξ»), имеющих, как и ξ, распределение $\mathcal{F}$.

Что значит «по выборке сделать вывод о распределении»? Распределение характеризуется функцией распределения, плотностью или таблицей, набором числовых характеристик — $\mathsf{E}\,\xi$, $\mathsf{D}\,\xi$, $\mathsf{E}\,\xi^k$ и т. д. По выборке нужно уметь строить приближения для всех этих характеристик.

1.2. Выборочное распределение

Рассмотрим реализацию выборки на одном элементарном исходе ω_0 — набор чисел $X_1 = X_1(\omega_0)$, $\ldots$, $X_n = X_n(\omega_0)$. На подходящем вероятностном пространстве введем случайную величину ξ^*, принимающую значения X_1, $\ldots$, X_n с вероятностями по $1/n$ (если какие-то из значений совпали, сложим вероятности соответствующее число раз). Таблица распределения вероятностей и функция распределения случайной величины ξ^* выглядят так:

ξ^*	X_1	$\ldots$	X_n
P	$1/n$	$\ldots$	$1/n$

$$F_n^*(y) = \sum_{X_i < y} \frac{1}{n} = \frac{\text{количество } X_i \in (-\infty, y)}{n}.$$

Распределение величины ξ^* называют эмпирическим или выборочным распределением. Вычислим математическое ожидание и дисперсию величины ξ^* и введем обозначения для этих величин:

$$\mathsf{E}\,\xi^* = \sum_{i=1}^n \frac{1}{n} X_i = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}, \quad \mathsf{D}\,\xi^* = \sum_{i=1}^n \frac{1}{n}(X_i - \mathsf{E}\,\xi^*)^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = S^2.$$

Точно так же вычислим и момент порядка k

$$\mathsf{E}\,(\xi^*)^k = \sum_{i=1}^n \frac{1}{n} X_i^k = \frac{1}{n}\sum_{i=1}^n X_i^k = \overline{X^k}.$$

В общем случае обозначим через $\overline{g(X)}$ величину

$$\mathsf{E}\,g(\xi^*) = \frac{1}{n}\sum_{i=1}^n g(X_i) = \overline{g(X)}.$$

Если при построении всех введенных нами характеристик считать выборку $X_1, \ldots, X_n$ набором случайных величин, то и сами эти характеристики — $F_n^*(y)$, $\overline{X}$, S^2, $\overline{X^k}$, $\overline{g(X)}$ — станут величинами случайными. Эти характеристики выборочного распределения используют для оценки (приближения) соответствующих неизвестных характеристик истинного распределения.

Причина использования характеристик распределения ξ^* для оценки характеристик истинного распределения ξ (или X_1) — в близости этих распределений при больших n.

Рассмотрим, для примера, n подбрасываний правильного кубика. Пусть $X_i \in \{1, \ldots, 6\}$ — количество очков, выпавших при i-м броске, $i = 1, \ldots, n$. Предположим, что единица в выборке встретится n_1 раз, двойка — n_2 раз и т.д. Тогда случайная величина ξ^* будет принимать значения 1, ..., 6 с вероятностями $\dfrac{n_1}{n}, \ldots, \dfrac{n_6}{n}$ соответственно. Но эти пропорции с ростом n приближаются к $1/6$ согласно закону больших чисел. То есть распределение величины ξ^* в некотором смысле сближается с истинным распределением числа очков, выпадающих при подбрасывании правильного кубика.

Мы не станем уточнять, что имеется в виду под близостью выборочного и истинного распределений. В следующих параграфах мы подробнее познакомимся с каждой из введенных выше характеристик и исследуем ее свойства, в том числе ее поведение с ростом объема выборки.

1.3. Эмпирическая функция распределения, гистограмма

Поскольку неизвестное распределение $\mathcal{F}$ можно описать, например, его функцией распределения $F(y) = \mathsf{P}(X_1 < y)$, построим по выборке «оценку» для этой функции.

Определение 1. Эмпирической функцией распределения, построенной по выборке $\mathbf{X} = (X_1, \ldots, X_n)$ объема n, называется случайная функция $F_n^* : \mathbb{R} \times \Omega \to [0, 1]$, при каждом $y \in \mathbb{R}$ равная

$$F_n^*(y) = \frac{\text{количество } X_i \in (-\infty, y)}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(X_i < y).$$

Напоминание: Случайная функция

$$\mathbf{I}(X_i < y) = \begin{cases} 1, & \text{если } X_i < y, \\ 0 & \text{иначе} \end{cases}$$

называется индикатором события $\{X_i < y\}$. При каждом y это — случайная величина, имеющая распределение Бернулли с параметром $p = \mathsf{P}(X_i < y) = F(y)$. почему?

Иначе говоря, при любом y значение $F(y)$, равное истинной вероятности случайной величине X_1 быть меньше y, оценивается долей элементов выборки, меньших y.

Если элементы выборки $X_1, \ldots, X_n$ упорядочить по возрастанию (на каждом элементарном исходе), получится новый набор случайных величин, называемый вариационным рядом:

$$X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n-1)} \leqslant X_{(n)}.$$

Здесь

$$X_{(1)} = \min\{X_1, \ldots, X_n\}, \quad X_{(n)} = \max\{X_1, \ldots, X_n\}.$$

Элемент $X_{(k)}$, $k = 1, \ldots, n$ называется k-м членом вариационного ряда или k-й порядковой статистикой.

Пример 1. Выборка: $\mathbf{X} = (0;\ 2;\ 1;\ 2{,}6;\ 3{,}1;\ 4{,}6;\ 1;\ 4{,}6;\ 6;\ 2{,}6;\ 6;\ 7;\ 9;\ 9;\ 2{,}6)$. Вариационный ряд: $(0;\ 1;\ 1;\ 2;\ 2{,}6;\ 2{,}6;\ 2{,}6;\ 3{,}1;\ 4{,}6;\ 4{,}6;\ 6;\ 6;\ 7;\ 9;\ 9)$.

Эмпирическая функция распределения имеет скачки в точках выборки, величина скачка в точке X_i равна m/n, где m — количество элементов выборки, совпадающих с X_i.

Можно построить эмпирическую функцию распределения по вариационному ряду:

$$F_n^*(y) = \begin{cases} 0, & \text{если } y \leqslant X_{(1)}, \\ \dfrac{k}{n}, & \text{если } X_{(k)} < y \leqslant X_{(k+1)}, \\ 1 & \text{при } y > X_{(n)}. \end{cases}$$

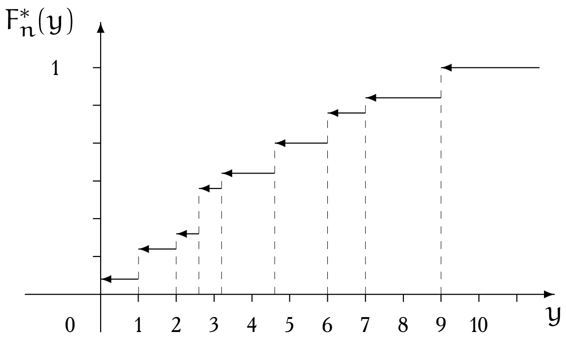


Рис. 1: Пример 1

Другой характеристикой распределения является таблица (для дискретных распределений) или плотность (для абсолютно непрерывных). Эмпирическим, или выборочным аналогом таблицы или плотности является так называемая гистограмма.

Гистограмма строится по группированным данным. Предполагаемую область значений случайной величины ξ (или область выборочных данных) делят независимо от выборки на некоторое количество интервалов (не обязательно одинаковых). Пусть A_1, $\ldots$, A_k — интервалы на прямой, называемые интервалами группировки. Обозначим для $j = 1, \ldots, k$ через ν_j число элементов выборки, попавших в интервал A_j:

$$\nu_j = \{\text{число } X_i \in A_j\} = \sum_{i=1}^{n} \mathbf{I}(X_i \in A_j), \qquad \text{здесь} \quad \sum_{j=1}^{k} \nu_j = n. \tag{1}$$

На каждом из интервалов A_j строят прямоугольник, площадь которого пропорциональна ν_j. Общая площадь всех прямоугольников должна равняться единице. Пусть l_j — длина интервала A_j. Высота f_j прямоугольника над A_j равна

$$f_j = \frac{\nu_j}{n l_j}.$$

Полученная фигура называется гистограммой.

Пример 2. Имеется вариационный ряд (см. пример 1):

$$(0;\ 1;\ 1;\ 2;\ 2{,}6;\ 2{,}6;\ 2{,}6;\ 3{,}1;\ 4{,}6;\ 4{,}6;\ 6;\ 6;\ 7;\ 9;\ 9).$$

Разобьем отрезок $[0, 10]$ на 4 равных отрезка. В отрезок $A_1 = [0; 2{,}5)$ попали 4 элемента выборки, в $A_2 = [2{,}5; 5)$ — 6, в $A_3 = [5; 7{,}5)$ — 3, и в отрезок $A_4 = [7{,}5; 10]$ попали 2 элемента выборки. Строим гистограмму (рис. 2). На рис. 3 — тоже гистограмма для той же выборки, но при разбиении области на 5 равных отрезков.

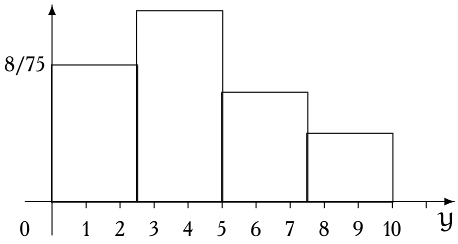


Рис. 2: Пример 2

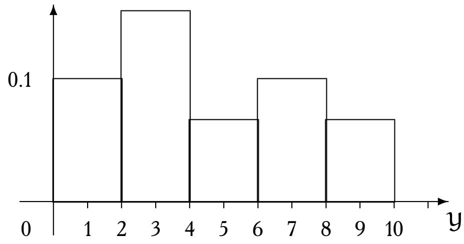


Рис. 3: Пример 2

Замечание 1. В курсе «Эконометрика» утверждается, что наилучшим числом интервалов группировки («формула Стерджесса») является $k = k(n) = 1 + [3.322 \lg n]$.

Здесь $\lg n$ — десятичный логарифм, поэтому $k = 1 + [\log_2 10 \log_{10} n] = 1 + [\log_2 n]$, т. е. при увеличении выборки вдвое число интервалов группировки увеличивается на 1.

Заметим, что чем больше интервалов группировки, тем лучше. Но, если брать число интервалов, скажем, порядка n, то с ростом n гистограмма не будет приближаться к плотности.

Справедливо следующее утверждение:

Если плотность распределения элементов выборки является непрерывной функцией, то при $k(n) \to \infty$ так, что $k(n)/n \to 0$, имеет место поточечная сходимость по вероятности гистограммы к плотности.

Так что выбор логарифма разумен, но не является единственно возможным.

1.4. Выборочные моменты

Знание моментов распределения также многое может сказать о его виде и свойствах. Введем выборочные аналоги неизвестных истинных моментов распределения.

Пусть $\mathsf{E}\,\xi = \mathsf{E}\,X_1 = a$, $\mathsf{D}\,\xi = \mathsf{D}\,X_1 = \sigma^2$, $\mathsf{E}\,\xi^k = \mathsf{E}\,X_1^k = m_k$ — теоретические среднее, дисперсия, k-й момент. Мы уже знакомы с соответствующими характеристиками выборочного распределения $\mathsf{E}\,\xi^* = \overline{X}$, $\mathsf{D}\,\xi^* = S^2$, $\mathsf{E}\,(\xi^*)^k = \overline{X^k}$.

Теоретические характеристики	Эмпирические характеристики
$\mathsf{E}\,\xi = \mathsf{E}\,X_1 = a$	$\overline{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ — выборочное среднее
$\mathsf{D}\,\xi = \mathsf{D}\,X_1 = \sigma^2$	$S^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{X})^2$ — выборочная дисперсия, либо $S_0^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \overline{X})^2$ — несмещенная выборочная дисперсия
$\mathsf{E}\,\xi^k = \mathsf{E}\,X_1^k = m_k$	$\overline{X^k} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^k$ — выборочный k-й момент

Список числовых характеристик и их оценок можно продолжать, рассмотрев, например, центральные, абсолютные и т. п. моменты. В общем случае

момент $\mathsf{E}\,g(\xi)$ будем оценивать величиной $\overline{g(X)} = \frac{1}{n}\sum\limits_{i=1}^{n} g(X_i)$.

1.5. Сходимость эмпирических характеристик к теоретическим

Мы ввели три вида эмпирических характеристик, предназначенных для оценивания неизвестных теоретических характеристик распределения: эмпирическую функцию распределения, гистограмму, выборочные моменты. Если наши оценки удачны, разница между ними и истинными характеристиками должна стремится к нулю с ростом объема выборки. Такое свойство эмпирических характеристик называют состоятельностью. Убедимся, что наши выборочные характеристики таким свойством обладают.

1.5.1. Свойства эмпирической функции распределения

Теорема 1. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка объема n из неизвестного распределения $\mathcal{F}$ с функцией распределения F. Пусть F_n^* — эмпирическая функция распределения, построенная по этой выборке. Тогда для любого $y \in \mathbb{R}$

$$F_n^*(y) \xrightarrow{p} F(y) \quad \text{при} \quad n \to \infty.$$

Замечание 2. $F_n^*(y)$ — случайная величина, так как она является функцией от случайных величин $X_1, \ldots, X_n$. То же самое можно сказать про гистограмму и выборочные моменты.

Доказательство теоремы 1. По определению 1,

$$F_n^*(y) = \frac{\sum_{i=1}^{n} \mathbf{I}(X_i < y)}{n}.$$

Случайные величины $\mathbf{I}(X_1 < y), \mathbf{I}(X_2 < y), \ldots$ независимы и одинаково распределены, их математическое ожидание конечно:

$$\mathsf{E}\,\mathbf{I}(X_1 < y) = 1 \cdot \mathsf{P}\,(X_1 < y) + 0 \cdot \mathsf{P}\,(X_1 \geqslant y) = \mathsf{P}\,(X_1 < y) = F(y) < \infty,$$

поэтому приме́ним ЗБЧ Хинчина, а что это такое? и

$$F_n^*(y) = \frac{\sum_{i=1}^{n} \mathbf{I}(X_i < y)}{n} \xrightarrow{p} \mathsf{E}\,\mathbf{I}(X_1 < y) = F(y).$$

Таким образом, с ростом объема выборки эмпирическая функция распределения сходится (по вероятности) к неизвестной теоретической. □

Верен более общий результат, показывающий, что сходимость эмпирической функции распределения к теоретической имеет «равномерный» характер.

Теорема Гливенко — Кантелли. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка объема n из неизвестного распределения $\mathcal{F}$ с функцией распределения F. Пусть F_n^* — эмпирическая функция распределения, построенная по этой выборке. Тогда

$$\sup_{y \in \mathbb{R}} \left| F_n^*(y) - F(y) \right| \xrightarrow{p} 0 \quad \text{при} \quad n \to \infty.$$

Замечание 3. Более того, в условиях теорем 1 и Гливенко — Кантелли имеет место сходимость не только по вероятности, но и почти наверное.

Если функция распределения F непрерывна, то скорость сходимости к нулю в теореме Гливенко — Кантелли имеет порядок $1/\sqrt{n}$:

Теорема Колмогорова. Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка объема n из неизвестного распределения $\mathcal{F}$ с непрерывной функцией распределения F, а F_n^* — эмпирическая функция распределения. Тогда

$$\sqrt{n} \sup_{y \in \mathbb{R}} \left| F_n^*(y) - F(y) \right| \Rightarrow \eta \quad \text{при} \quad n \to \infty,$$

где случайная величина η имеет распределение Колмогорова с непрерывной функцией распределения

$$K(x) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2} \text{ при } x \geqslant 0, \quad K(x) = 0 \text{ при } x < 0.$$

Следующие свойства эмпирической функции распределения — это хорошо знакомые нам свойства среднего арифметического n независимых слагаемых, имеющих, к тому же, распределение Бернулли.

В первых двух пунктах утверждается, что случайная величина $F_n^*(y)$ имеет математическое ожидание $F(y)$ и дисперсию $\dfrac{F(y)(1-F(y))}{n}$, которая убывает как $1/n$. Третий пункт показывает, что $F_n^*(y)$ сходится к $F(y)$ со скоростью $1/\sqrt{n}$.

Свойство 1. Для любого $y \in \mathbb{R}$

1) $\mathsf{E}\, F_n^*(y) = F(y)$, т. е. $F_n^*(y)$ — «несмещенная» оценка для $F(y)$;

2) $\mathsf{D}\, F_n^*(y) = \dfrac{F(y)(1-F(y))}{n}$;

3) если $0 < F(y) < 1$, то $\sqrt{n}(F_n^*(y) - F(y)) \Rightarrow \mathrm{N}_{0, F(y)(1-F(y))}$, т. е. $F_n^*(y)$ — «асимптотически нормальная» оценка для $F(y)$;

4) случайная величина $n \cdot F_n^*(y)$ имеет биномиальное распределение $\mathrm{B}_{n, F(y)}$.

Доказательство свойства 1. Заметим снова, что $\mathbf{I}(X_1 < y)$ имеет распределение Бернулли $\mathrm{B}_{F(y)}$, поэтому

$$\mathsf{E}\,\mathbf{I}(X_1 < y) = F(y) \quad \text{и} \quad \mathsf{D}\,\mathbf{I}(X_1 < y) = F(y)(1 - F(y)).$$

1) Случайные величины $\mathbf{I}(X_1 < y)$, $\mathbf{I}(X_2 < y)$, ... одинаково распределены, поэтому

где используется одинаковая распределенность?

$$\mathsf{E}\, F_n^*(y) = \mathsf{E}\, \frac{\sum_{i=1}^{n} \mathbf{I}(X_i < y)}{n} = \frac{\sum_{i=1}^{n} \mathsf{E}\,\mathbf{I}(X_i < y)}{n} = \frac{n\mathsf{E}\,\mathbf{I}(X_1 < y)}{n} = F(y).$$

2) Случайные величины $\mathbf{I}(X_1 < y)$, $\mathbf{I}(X_2 < y)$, $\ldots$ независимы и одинаково распределены, поэтому $\boxed{\text{где используется независимость?}}$

$$\mathsf{D}\, F_n^*(y) = \mathsf{D}\, \frac{\sum\limits_{i=1}^{n} \mathbf{I}(X_i < y)}{n} = \frac{\sum\limits_{i=1}^{n} \mathsf{D}\, \mathbf{I}(X_i < y)}{n^2} = \frac{n \mathsf{D}\, \mathbf{I}(X_1 < y)}{n^2} = \frac{F(y)(1 - F(y))}{n}.$$

3) Воспользуемся ЦПТ Ляпунова: $\boxed{\text{а что это такое?}}$

$$\sqrt{n}(F_n^*(y) - F(y)) = \sqrt{n} \left(\frac{\sum\limits_{i=1}^{n} \mathbf{I}(X_i < y)}{n} - F(y) \right) = \frac{\left(\sum\limits_{i=1}^{n} \mathbf{I}(X_i < y) - nF(y) \right)}{\sqrt{n}} =$$

$$= \frac{\left(\sum\limits_{i=1}^{n} \mathbf{I}(X_i < y) - n \mathsf{E}\, \mathbf{I}(X_1 < y) \right)}{\sqrt{n}} \Rightarrow \mathsf{N}_{0, \mathsf{D}\, \mathbf{I}(X_1 < y)} = \mathsf{N}_{0, F(y)(1 - F(y))}.$$

4) Поскольку $\mathbf{I}(X_1 < y)$ (число успехов в одном испытании) имеет распределение Бернулли $\mathsf{B}_{F(y)}$, $\boxed{\text{почему?}}$ то $n \cdot F_n^*(y) = \sum\limits_{i=1}^{n} \mathbf{I}(X_i < y)$ имеет биномиальное распределение $\mathsf{B}_{n, F(y)}$. $\boxed{\text{почему?}}$ $\boxed{\text{а что такое устойчивость по суммированию?}}$

$\square$

Замечание 4. Все определения, как то: «оценка», «несмещенность», «состоятельность», «асимптотическая нормальность» будут даны в главе 2. Но смысл этих терминов должен быть вполне понятен уже сейчас.

1.5.2. Свойства гистограммы

Пусть распределение $\mathcal{F}$ абсолютно непрерывно, f — его истинная плотность. Пусть, кроме того, число k интервалов группировки не зависит от n. Случай, когда $k = k(n)$, отмечен в замечании 1.

Справедлива

Теорема 2. При $n \to \infty$ для любого $j = 1, \ldots, k$

$$l_j \cdot f_j = \frac{\nu_j}{n} \xrightarrow{p} \mathsf{P}\,(X_1 \in A_j) = \int\limits_{A_j} f(x)\, dx.$$

Упражнение. Доказать теорему 2, используя (1) и ЗБЧ.

Теорема утверждает, что площадь столбца гистограммы, построенного над интервалом группировки, с ростом объема выборки сближается с площадью области под графиком плотности над этим же интервалом.

1.5.3. Свойства выборочных моментов

Выборочное среднее $\overline{X}$ является несмещенной, состоятельной и асимптотически нормальной оценкой для теоретического среднего (математического ожидания):

Свойство 2.

1) Если $\mathsf{E}\,|X_1| < \infty$, то $\mathsf{E}\,\overline{X} = \mathsf{E}\,X_1 = a$.

2) Если $\mathsf{E}\,|X_1| < \infty$, то $\overline{X} \xrightarrow{p} \mathsf{E}\,X_1 = a$ при $n \to \infty$.

3) Если $\mathsf{D}\,X_1 < \infty$ и не равна нулю, то $\sqrt{n}\left(\overline{X} - \mathsf{E}\,X_1\right) \Rightarrow \mathsf{N}_{0,\mathsf{D}\,X_1}$.

Доказательство свойства 2.

1) $\mathsf{E}\,\overline{X} = \dfrac{1}{n}\sum\limits_{i=1}^{n}\mathsf{E}\,X_i = \dfrac{1}{n}\cdot n\mathsf{E}\,X_1 = \mathsf{E}\,X_1 = a.$

2) Согласно ЗБЧ в форме Хинчина, $\overline{X} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i \xrightarrow{p} \mathsf{E}\,X_1 = a.$

3) Согласно ЦПТ,

$$\sqrt{n}\left(\overline{X} - \mathsf{E}\,X_1\right) = \frac{\sum\limits_{i=1}^{n} X_i - n\mathsf{E}\,X_1}{\sqrt{n}} \Rightarrow \mathsf{N}_{0,\mathsf{D}\,X_1}.$$

Выборочный k-й момент $\overline{X^k}$ является несмещенной, состоятельной и асимптотически нормальной оценкой для теоретического k-го момента:

Свойство 3.

1) Если $\mathsf{E}\,|X_1|^k < \infty$, то $\mathsf{E}\,\overline{X^k} = \mathsf{E}\,X_1^k = m_k$.

2) Если $\mathsf{E}\,|X_1|^k < \infty$, то $\overline{X^k} \xrightarrow{p} \mathsf{E}\,X_1^k = m_k$ при $n \to \infty$.

3) Если $\mathsf{D}\,X_1^k < \infty$ и не равна нулю, то $\sqrt{n}\left(\overline{X^k} - \mathsf{E}\,X_1^k\right) \Rightarrow \mathsf{N}_{0,\mathsf{D}\,X_1^k}$.

Упражнение. Доказать свойство 3.

В дальнейшем мы не будем оговаривать существование соответствующих моментов. В частности, в первых двух пунктах следующего утверждения предполагается наличие второго момента у случайных величин X_i, а в третьем пункте — четвертого (дисперсии величины X_1^2).

Свойство 4.

1) Выборочные дисперсии S^2 и S_0^2 являются состоятельными оценками для истинной дисперсии: $S^2 \xrightarrow{p} \mathsf{D}\, X_1 = \sigma^2$, $\quad S_0^2 \xrightarrow{p} \mathsf{D}\, X_1 = \sigma^2$.

2) Величина S^2 — смещенная, а S_0^2 — несмещенная оценка дисперсии:

$$\mathsf{E}\, S^2 = \frac{n-1}{n}\mathsf{D}\, X_1 = \frac{n-1}{n}\sigma^2 \neq \sigma^2, \quad \mathsf{E}\, S_0^2 = \mathsf{D}\, X_1 = \sigma^2.$$

3) Выборочные дисперсии S^2 и S_0^2 являются асимптотически нормальными оценками истинной дисперсии: $\sqrt{n}\left(S^2 - \mathsf{D}\, X_1\right) \Rightarrow \mathsf{N}_{0, \mathsf{D}\,(X_1 - \mathsf{E}\, X_1)^2}$.

Доказательство свойства 4.

1) Во первых, раскрыв скобки, полезно убедиться в том, что

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

Из (2) и ЗБЧ следует, что $S^2 = \overline{X^2} - (\overline{X})^2 \xrightarrow{p} \mathsf{E}\, X_1^2 - (\mathsf{E}\, X_1)^2 = \sigma^2$. Кроме того, $\dfrac{n}{n-1} \to 1$, так что $S_0^2 = \dfrac{n}{n-1}\, S^2 \xrightarrow{p} \sigma^2$.

2) Воспользуемся формулой (2):

$$\begin{aligned}
\mathsf{E}\, S^2 &= \mathsf{E}\left(\overline{X^2} - (\overline{X})^2\right) = \mathsf{E}\,\overline{X^2} - \mathsf{E}\,(\overline{X})^2 = (\text{по лемме } 3) = \mathsf{E}\, X_1^2 - \mathsf{E}\,(\overline{X})^2 = \\
&= \mathsf{E}\, X_1^2 - \left((\mathsf{E}\,\overline{X})^2 + \mathsf{D}\,(\overline{X})\right) = \mathsf{E}\, X_1^2 - (\mathsf{E}\, X_1)^2 - \mathsf{D}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \\
&= \sigma^2 - \frac{1}{n^2} n \mathsf{D}\, X_1 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2, \quad \text{откуда сразу следует} \\
\mathsf{E}\, S_0^2 &= \frac{n}{n-1}\mathsf{E}\, S^2 = \sigma^2.
\end{aligned}$$

3) Выборочную дисперсию можно представить в следующем виде:

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

Тогда
$$\begin{aligned}
\sqrt{n}\left(S^2 - \sigma^2\right) &= \sqrt{n}\left(\overline{(X-a)^2} - \left(\overline{X}-a\right)^2 - \sigma^2\right) = \\
&= \sqrt{n}\left(\overline{(X-a)^2} - \mathsf{E}\,(X_1 - a)^2\right) - \sqrt{n}\left(\overline{X} - a\right)^2 = \\
&= \frac{\sum_{i=1}^{n}(X_i - a)^2 - n\mathsf{E}\,(X_1 - a)^2}{\sqrt{n}} - \left(\overline{X} - a\right)\cdot\sqrt{n}\left(\overline{X} - a\right) \Rightarrow \mathsf{N}_{0, \mathsf{D}\,(X_1 - a)^2},
\end{aligned}$$

поскольку первое слагаемое слабо сходится к $\mathsf{N}_{0, \mathsf{D}\,(X_1 - a)^2}$ по ЦПТ, а второе слагаемое $\left(\overline{X} - a\right)\cdot\sqrt{n}\left(\overline{X} - a\right)$ слабо сходится к нулю как произведение сходящейся к нулю по вероятности последовательности и последовательности, слабо сходящейся к $\mathsf{N}_{0, \mathsf{D}\, X_1}$. какое свойство слабой сходимости использовано дважды?

1.6. Группированные данные (некоторые вводные понятия к эконометрии)

Если объем выборки очень велик, часто работают не с элементами выборки, а с группированными данными. Приведем ряд понятий, связанных с группировкой. Для простоты будем делить область выборочных данных на k одинаковых интервалов $A_1, \ldots, A_k$ длины Δ:

$$A_1 = [a_0, a_1), \ldots, A_k = [a_{k-1}, a_k), \quad a_j - a_{j-1} = \Delta.$$

Как прежде, пусть ν_j — число элементов выборки, попавших в интервал A_j, и w_j — частота попадания в интервал A_j (оценка вероятности попадания в интервал):

$$\nu_j = \{\text{число } X_i \in A_j\} = \sum_{i=1}^{n} \mathbf{I}(X_i \in A_j), \quad w_j = \frac{\nu_j}{n}.$$

На каждом из интервалов A_j строят прямоугольник с высотой $f_j = \dfrac{w_j}{\Delta}$ и получают гистограмму.

Рассмотрим середины интервалов: $\overline{a}_j = a_{j-1} + \Delta/2$ — середина A_j. Набор

$$\underbrace{\overline{a}_1, \ldots, \overline{a}_1}_{\nu_1 \text{ раз}}, \quad \ldots, \quad \underbrace{\overline{a}_k, \ldots, \overline{a}_k}_{\nu_k \text{ раз}}$$

можно считать «огрубленной» выборкой, в которой все X_i, попадающие в интервал A_j, заменены на $\overline{a}_j$. По этой выборке можно построить такие же (но более грубые) выборочные характеристики, что и по исходной (обозначим их так же), например выборочное среднее

$$\overline{X} = \frac{1}{n} \sum_{j=1}^{k} \overline{a}_j \nu_j = \sum_{j=1}^{k} \overline{a}_j w_j$$

или выборочную дисперсию

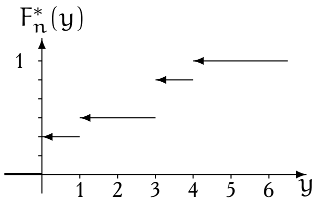
$$S^2 = \frac{1}{n} \sum_{j=1}^{k} (\overline{a}_j - \overline{X})^2 \nu_j = \sum_{j=1}^{k} (\overline{a}_j - \overline{X})^2 w_j.$$

Кривая, соединяющая точки $(a_0, 0)$, $(\overline{a}_1, f_1)$, $\ldots$, $(\overline{a}_k, f_k)$, $(a_k, 0)$, называется полигоном (частот). В отличие от гистограммы полигон — непрерывная функция (ломаная).

1.7. Вопросы и упражнения

1. Можно ли по эмпирической функции распределения, приведенной на рис. 1, восстановить выборку $X_1, \ldots, X_n$, если n известно? А вариационный ряд? Как это сделать? А если n неизвестно?

2. Существует ли выборка $(X_1, \ldots, X_6)$ объёма 6 с нарисованной справа эмпирической функцией распределения? А выборка $(X_1, \ldots, X_{12})$ объёма 12? Если «да», то записать ее и нарисовать эмпирическую функцию распределения выборки $(2X_1, \ldots, 2X_{12})$.



3. Можно ли по гистограмме, приведенной на рис. 2, восстановить выборку $X_1, \ldots, X_n$?

4. Нарисовать эмпирическую функцию распределения, соответствующую выборке объема n из распределения Бернулли B_p. Использовать выборочное среднее $\overline{X}$. Доказать непосредственно, что выполнена теорема Гливенко — Кантелли:

$$\sup_{y \in \mathbb{R}} \left| F_n^*(y) - F(y) \right| \xrightarrow{p} 0 \quad \text{при} \quad n \to \infty.$$

5. Вспомнить, как найти по функции распределения величины X_1 функцию распределения первой и последней порядковой статистики: $X_{(1)} = \min\{X_1, \ldots, X_n\}$, $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Выписать выражения для плотности этих порядковых статистик через функцию распределения и плотность величины X_1.

6. Доказать (или вспомнить), что функция распределения k-й порядковой статистики $X_{(k)}$ имеет вид:

$$\mathsf{P}\left(X_{(k)} < y\right) = \mathsf{P}\left(\text{хотя бы } k \text{ элементов выборки } < y\right) = \sum_{i=k}^{n} C_n^i F(y)^i (1 - F(y))^{n-i},$$

где $F(y)$ — функция распределения величины X_1.

7. Из курса «Эконометрика»: доказать, что среднее степенное

$$\left(\overline{X^k}\right)^{\frac{1}{k}} = \sqrt[k]{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right)}$$

а) стремится к $X_{(1)}$ при $k \to -\infty$ б) стремится к $X_{(n)}$ при $k \to +\infty$
Имеется в виду сходимость для любого набора чисел $X_1, \ldots, X_n$, такого, что среднее степенное определено, т. е. сходимость п. н.
Указание. Вынести $X_{(1)}$ (или $X_{(n)}$) из-под корня, воспользоваться леммой о двух милиционерах и свойствами: $\sqrt[k]{k} \to 1$ при $k \to +\infty$, $\sqrt[k]{1} \to 1$ при $k \to +\infty$, и т.д.

8. В условиях предыдущей задачи доказать, что последовательность

$$\sqrt[k]{\mathsf{E}\,(\xi^*)^k} = \sqrt[k]{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right)}, \quad k = 1, 2, 3, \ldots$$

не убывает по k. **Указание.** Воспользоваться неравенством Йенсена.

2. Точечное оценивание

2.1. Параметрические семейства распределений

Предположим, что имеется выборка объема n, элементы которой $X_1, \ldots, X_n$ независимы, одинаково распределены и имеют распределение $\mathcal{F}_\theta$, известным образом зависящее от неизвестного параметра θ.

Здесь $\mathcal{F}_\theta$ — некий класс распределений, целиком определяющихся значением скалярного или векторного параметра θ. Параметр θ принимает значения из некоторого множества Θ.

Например, для всех $i = 1, \ldots, n$

- X_i имеют распределение Пуассона Π_λ, где $\lambda > 0$ — неизвестный параметр; здесь $\mathcal{F}_\theta = \Pi_\lambda$, $\theta = \lambda$, $\Theta = (0, \infty)$;
- X_i имеют распределение Бернулли B_p, где $p \in (0, 1)$ — неизвестный параметр; здесь $\mathcal{F}_\theta = \mathrm{B}_p$, $\theta = p$, $\Theta = (0, 1)$;
- X_i имеют равномерное распределение $\mathrm{U}_{a,b}$, где $a < b$ — неизвестные параметры; здесь $\mathcal{F}_\theta = \mathrm{U}_{a,b}$, $\theta = (a, b)$, $\Theta = \{(a, b) \;:\; a < b\}$;
- X_i имеют равномерное распределение $\mathrm{U}_{0,\theta}$, где $\theta > 0$ — неизвестный параметр; здесь $\mathcal{F}_\theta = \mathrm{U}_{0,\theta}$, $\Theta = (0, \infty)$;
- X_i имеют нормальное распределение N_{a,σ^2}, где $a \in \mathbb{R}$, $\sigma > 0$ — неизвестные параметры; здесь $\mathcal{F}_\theta = \mathrm{N}_{a,\sigma^2}$, $\theta = (a, \sigma^2)$, $\Theta = \mathbb{R} \times (0, \infty)$;
- X_i имеют нормальное распределение $\mathrm{N}_{a,4}$, где $a \in \mathbb{R}$ — неизвестный параметр; здесь $\mathcal{F}_\theta = \mathrm{N}_{a,4}$, $\theta = a$, $\Theta = \mathbb{R}$.

Такая постановка имеет смысл, поскольку редко о проводимом эксперименте совсем ничего нельзя сказать. Обычно тип распределения ясен заранее, и требуется лишь указать значения параметров этого распределения.

Так, в широких предположениях рост юношей одного возраста имеет нормальное распределение (с неизвестными средним и дисперсией), а число покупателей в магазине в течение часа (не часа пик) — распределение Пуассона, и опять-таки с неизвестной «интенсивностью» λ.

2.2. Точечные оценки. Несмещенность, состоятельность оценок

Итак, пусть $X_1, \ldots, X_n$ — выборка объема n из параметрического семейства распределений $\mathcal{F}_\theta$, $\theta \in \Theta$.

Заметим, что все характеристики случайных величин $X_1, \ldots, X_n$ зависят от параметра θ. Так, например, для X_i с распределением Пуассона Π_λ

$$\mathsf{E}\, X_1 = \lambda, \quad \mathsf{P}\,(X_1 = 2) = \frac{\lambda^2}{2}\, e^{-\lambda}, \quad \mathsf{D}\, X_1 = \lambda \quad \text{и т. д.}$$

Чтобы отразить эту зависимость, будем писать $\mathsf{E}_\theta\, X_1$ вместо $\mathsf{E}\, X_1$ и т.д. Так, $\mathsf{D}_{\theta_1} X_1$ означает дисперсию, вычисленную в предположении $\theta = \theta_1$.

Во многих случаях эта условность необходима. Предположим, что X_i имеют распределение Пуассона Π_λ. В предположении, что $\lambda = 1$, имеем $\mathsf{E}\, X_1 = 1$, тогда как при $\lambda = 7$ имеем $\mathsf{E}\, X_1 = 7$. Таким образом, запись $\mathsf{E}\, X_1$, без указания на распределение X_1, оказывается просто бессмысленной.

Определение 2. Статистикой называется произвольная функция $\theta^* = \theta^*(X_1, \ldots, X_n)$ от элементов выборки.

Замечание 5. Статистика есть функция от эмпирических данных, но никак не от параметра θ. Статистика, как правило, предназаначена именно для оценивания неизвестного параметра θ (поэтому ее иначе называют «оценкой»), и уже поэтому от него зависеть не может.

Конечно, статистика есть не «любая», а «измеримая» функция от выборки (борелевская, для которой прообраз любого борелевского множества из $\mathbb{R}$ есть снова борелевское множество в $\mathbb{R}^n$), но мы никогда встретимся с иными функциями, и более на это обращать внимание не будем.

Определение 3. Статистика $\theta^* = \theta^*(X_1, \ldots, X_n)$ называется несмещенной оценкой параметра θ, если для любого $\theta \in \Theta$ выполнено равенство $\mathsf{E}_\theta\, \theta^* = \theta$.

Определение 4. Статистика $\theta^* = \theta^*(X_1, \ldots, X_n)$ называется состоятельной оценкой параметра θ, если для любого $\theta \in \Theta$ имеет место сходимость $\theta^* \xrightarrow{p} \theta$ при $n \to \infty$.

Несмещенность — свойство оценок при фиксированном n. Означает это свойство отсутствие ошибки «в среднем», т.е. при систематическом использовании данной оценки.

Свойство состоятельности означает, что последовательность оценок приближается к неизвестному параметру при увеличении количества данных. Понятно, что при отсутствии этого свойства оценка совершенно «несостоятельна» как оценка.

Пример 3. Пусть $X_1, \ldots, X_n$ — выборка объема n из нормального распределения N_{a,σ^2}, где $a \in \mathbb{R}$, $\sigma > 0$. Как найти оценки для параметров a и σ^2, если оба эти параметра (можно считать это и одним двумерным параметром) неизвестны?

Мы уже знаем хорошие оценки для математического ожидания и дисперсии любого распределения.

Оценкой для истинного среднего $a = \mathrm{E}_{a,\sigma^2} X_1$ может служить выборочное среднее $a^* = \overline{X}$. Свойство 2 утверждает, что эта оценка несмещенная и состоятельная.

Для дисперсии $\sigma^2 = \mathrm{D}_{a,\sigma^2} X_1$ у нас есть сразу две оценки:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{и} \quad S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

(выборочная дисперсия и несмещенная выборочная дисперсия).

Как показано в свойстве 4, обе эти оценки состоятельны, и одна из них — несмещенная. $\boxed{\text{которая?}}$

Следующий метод получения оценок для неизвестных параметров как раз и предлагает использовать выборочные моменты вместо истинных.

2.3. Методы нахождения оценок: метод моментов

Метод моментов заключается в следующем: любой момент случайной величины X_1 (например, k-й) зависит, часто функционально, от параметра θ. Но тогда и параметр θ может оказаться функцией от теоретического k-го момента. Подставив в эту функцию вместо неизвестного теоретического k-го момента его выборочный аналог, получим вместо параметра θ оценку θ^*.

Пусть $X_1, \ldots, X_n$ — выборка объема n из параметрического семейства распределений $\mathcal{F}_\theta$, где $\theta \in \Theta$. Выберем некоторую функцию $g(y)$ так, чтобы существовал момент

$$\mathrm{E}_\theta\, g(X_1) = h(\theta), \tag{3}$$

и функция h была обратима в области Θ. Тогда в качестве оценки θ^* для θ возьмем решение уравнения

$$\overline{g(X)} = h(\theta^*).$$

Или (что то же самое), сначала решаем уравнение (3) относительно θ, а затем вместо истинного момента берем выборочный:

$$\theta = h^{-1}\left(\mathrm{E}_\theta\, g(X_1)\right), \qquad \theta^* = h^{-1}\left(\overline{g(X)}\right) = h^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right).$$

Чаще всего в качестве функции $g(y)$ берут $g(y) = y^k$. В этом случае

$$\mathsf{E}_\theta X_1^k = h(\theta),$$

и, если функция h обратима в области Θ, то

$$\theta = h^{-1}\left(\mathsf{E}_\theta X_1^k\right), \qquad \theta^* = h^{-1}\left(\overline{X^k}\right) = h^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right).$$

Можно сказать, что мы берем в качестве оценки такое (случайное) значение параметра θ, при котором истинный момент совпадает с выборочным.

Пример 4. Пусть $X_1, \ldots, X_n$ — выборка объема n из равномерного на отрезке $[0, \theta]$ распределения $\mathsf{U}_{0,\theta}$, где $\theta > 0$.

Найдем оценку метода моментов (ОММ) по первому моменту:

$$\mathsf{E}_\theta X_1 = \frac{\theta}{2}, \quad \text{тогда} \quad \theta = 2\mathsf{E}_\theta X_1 \quad \text{и ОММ такова} \quad \theta_1^* = 2\overline{X}.$$

Найдем оценку метода моментов (ОММ) по k-му моменту:

$$\mathsf{E}_\theta X_1^k = \int\limits_0^\theta y^k \frac{1}{\theta}\, dy = \frac{\theta^k}{k+1},$$

тогда

$$\theta = \sqrt[k]{(k+1)\mathsf{E}_\theta X_1^k}, \quad \text{и ОММ такова:} \quad \theta_k^* = \sqrt[k]{(k+1)\overline{X^k}}. \tag{4}$$

Пример 5. Пусть $X_1, \ldots, X_n$ — выборка объема n из распределения Пуассона Π_λ с неизвестным параметром $\lambda > 0$. Введем новый параметр

$$\theta = \theta(\lambda) = \mathsf{P}_\lambda(X_1 = 1) = \lambda\, e^{-\lambda}$$

и найдем оценку метода моментов для θ с помощью функции $g(y) = \mathbf{I}(y = 1)$:

$$\mathsf{E}_\lambda g(X_1) = \mathsf{E}_\lambda \mathbf{I}(X_1 = 1) = \mathsf{P}_\lambda(X_1 = 1) = \lambda\, e^{-\lambda} = \theta,$$

$$\theta^* = \overline{\mathbf{I}(X = 1)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(X_i = 1).$$

Заметим, что оценку для параметра $\lambda > 0$ с помощью функции $g(y) = \mathbf{I}(y = 1)$ найти нельзя: функция $h(\lambda) = \lambda\, e^{-\lambda}$ не является взаимно-однозначной и, следовательно, обратимой по λ в области $\lambda > 0$. Оценку для параметра λ разумно находить по первому моменту: $\mathsf{E}_\lambda X_1 = \lambda$, и $\lambda^* = \overline{X}$ — оценка метода моментов.

Замечание 6. Может случиться так, что $\theta^* = h^{-1}(\overline{g(X)}) \notin \Theta$, тогда как $\theta \in \Theta$. В этом случае оценку корректируют. Например, в качестве ОММ берут ближайшую к $h^{-1}(\overline{g(X)})$ точку из Θ или из замыкания Θ.

Пример 6. Пусть $X_1, \ldots, X_n$ — выборка объема n из нормального распределения $\mathsf{N}_{a,1}$ с неотрицательным средним $a \geqslant 0$. Ищем оценку для a по первому моменту:

$$\mathsf{E}_a X_1 = a, \quad \text{поэтому} \quad a^* = \overline{X}.$$

Однако по условию $a \geqslant 0$, тогда как $\overline{X}$ может быть и отрицательно. Если $\overline{X} < 0$, то в качестве оценки для a более подойдет 0. Если же $\overline{X} > 0$, в качестве оценки нужно брать $\overline{X}$. Итого: $a^* = \max\{0, \overline{X}\}$ — «исправленная» оценка метода моментов.

2.4. Состоятельность оценок метода моментов

Теорема 3. Пусть $\theta^* = h^{-1}\left(\overline{g(X)}\right)$ — оценка параметра θ, полученная по методу моментов, причем функция h^{-1} непрерывна. Тогда θ^* состоятельна.

Доказательство теоремы 3. По ЗБЧ Хинчина имеем:

$$\overline{g(X)} = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{p} \mathsf{E}_\theta\, g(X_1) = h(\theta).$$

Поскольку функция h^{-1} непрерывна, то и

$$\theta_* = h^{-1}\left(\overline{g(X)}\right) \xrightarrow{p} h^{-1}(\mathsf{E}_\theta\, g(X_1)) = h^{-1}(h(\theta)) = \theta.$$

Напоминание: *Для обратимой, т. е. взаимно-однозначной функции* $h : \mathbb{R} \to \mathbb{R}$ непрерывность h и непрерывность h^{-1} эквивалентны.

Если полученные разумным путем оценки обязаны быть состоятельными, то свойство несмещенности — скорее исключение, нежели правило.

Действительно, несмещенность ОММ вида $\theta^* = h^{-1}\left(\overline{g(X)}\right)$ означала бы, что при всех $\theta \in \Theta$ выполнено равенство

$$\mathsf{E}_\theta\, h^{-1}\left(\overline{g(X)}\right) = \theta = h^{-1}\left(h(\theta)\right) = h^{-1}\left(\mathsf{E}_\theta\, \overline{g(X)}\right). \qquad (5)$$

Но функция h^{-1} очень часто оказывается выпуклой или вогнутой. В этом случае неравенство Йенсена утверждает, что между левой и правой частью в (5) равенство возможно лишь если случайная величина $\overline{g(X)}$ вырождена или если функция h^{-1} линейна в области значений этой случайной величины.

Рассмотрим, к примеру, последовательность оценок для неизвестного параметра θ равномерного на отрезке $[0, \theta]$ распределения, полученную в примере 4 и исследуем напрямую их свойства.

Состоятельность:

1. По ЗБЧ, $\theta_1^* = 2\overline{X} \xrightarrow{p} 2\mathsf{E}_\theta X_1 = 2\theta/2 = \theta$, т. е. оценка $\theta_1^* = 2\overline{X}$ состоятельна.

2. Заметим, что по ЗБЧ (или по свойству 3 — только для тех, кто его доказал) при $n \to \infty$

$$\overline{X^k} \xrightarrow{p} \mathsf{E}_\theta X_1^k = \frac{\theta^k}{k+1}.$$

Поскольку функция $\sqrt[k]{(k+1)y}$ непрерывна для всех $y > 0$, то при $n \to \infty$

$$\theta_k^* = \sqrt[k]{(k+1)\overline{X^k}} \xrightarrow{p} \sqrt[k]{(k+1)\frac{\theta^k}{k+1}} = \theta.$$

Упражнение. Зачем нужна ссылка на непрерывность функции $\sqrt[k]{(k+1)y}$?

Несмещенность:

1. По определению,

$$\mathsf{E}_\theta\,\theta_1^* = \mathsf{E}_\theta\,2\overline{X} = 2\mathsf{E}_\theta\,\overline{X} = \text{ (по свойству 2) } = 2\theta/2 = \theta,$$

т. е. оценка $\theta_1^* = 2\overline{X}$ несмещенная.

2. Рассмотрим оценку θ_2^*. Заметим, что

$$\mathsf{E}_\theta\,\theta_2^* = \mathsf{E}_\theta\sqrt{3\overline{X^2}},$$

тогда как по свойству 2

$$\theta = \sqrt{3\mathsf{E}_\theta\,X_1^2} = \sqrt{3\mathsf{E}_\theta\,\overline{X^2}}.$$

Равенство $\mathsf{E}_\theta\,\theta_2^* = \theta$ означало бы, что для случайной величины $\xi = 3\overline{X^2}$ выполнено $\mathsf{E}_\theta\sqrt{\xi} = \sqrt{\mathsf{E}_\theta\,\xi}$, а для величины $\eta = \sqrt{\xi}$ выполнено $\mathsf{E}_\theta\,\eta^2 = (\mathsf{E}_\theta\,\eta)^2$ или $\mathsf{D}_\theta\,\eta = 0$.

Но величина $\eta = \sqrt{3\overline{X^2}}$ имеет невырожденное (более того, абсолютно непрерывное) распределение. Поэтому оценка $\theta_2^* = \sqrt{3\overline{X^2}}$ — смещенная. Такими же смещенными будут и оценки $\theta_k^*, k > 2$. Докажите это, воспользовавшись, как в (5), неравенством Йенсена.

То есть вся последовательность $\{\theta_k^*\}_{k=1}^{\infty} = \left\{\sqrt[k]{(k+1)\overline{X^k}}\right\}$ состоит из состоятельных оценок, при этом только оценка $\theta_1^* = 2\overline{X}$ — несмещенная.

2.5. Методы нахождения оценок: метод максимального правдоподобия

Метод максимального правдоподобия — еще один разумный способ построения оценки неизвестного параметра. Состоит он в том, что в качестве «наиболее правдоподобного» значения параметра берут значение θ, максимизирующее вероятность получить при n опытах данную выборку $\mathbf{X} = (X_1, \ldots, X_n)$. Это значение параметра θ зависит от выборки и является искомой оценкой.

Решим сначала, что такое «вероятность получить данную выборку», т. е. что именно нужно максимизировать. Вспомним, что для абсолютно непрерывных распределений $\mathcal{F}_\theta$ их плотность $f_\theta(y)$ — «почти» (с точностью до dy) вероятность попадания в точку y. А для дискретных распределений $\mathcal{F}_\theta$ вероятность попасть в точку y равна $\mathsf{P}_\theta\,(X_1 = y)$. И то, и другое мы будем называть плотностью распределения $\mathcal{F}_\theta$. Итак,

Определение 5. Функцию

$$f_\theta(y) = \begin{cases} \text{плотность } f_\theta(y), & \text{если распределение } \mathcal{F}_\theta \text{ абсолютно непрерывно,} \\ \mathsf{P}_\theta\,(X_1 = y), & \text{если распределение } \mathcal{F}_\theta \text{ дискретно} \end{cases}$$

мы будем называть плотностью распределения $\mathcal{F}_\theta$.

Для тех, кто знаком с понятием интеграла по мере, нет ничего странного в том, что мы ввели понятие плотности для дискретного распределения. Это — не плотность относительно меры Лебега, но плотность относительно считающей меры.

Если для дискретного распределения величины X_1 со значениями a_1, a_2, ... ввести считающую меру $\#$ на борелевской σ-алгебре как

$$\#(B) = \text{ количество } a_i, \text{ принадлежащих } B, \quad \text{то} \quad \#(B) = \int_B \#(dy) = \sum_{a_i \in B} 1,$$

и тогда $$\mathsf{P}_\theta(X_1 \in B) = \int_B f_\theta(y)\,\#(dy) = \int_B \mathsf{P}_\theta(X_1 = y)\,\#(dy) = \sum_{a_i \in B} \mathsf{P}_\theta(X_1 = a_i).$$

Если же X_1 имеет абсолютно непрерывное распределение, то $f_\theta(y)$ есть привычная плотность относительно меры Лебега $\lambda(dy) = dy$:

$$\mathsf{P}_\theta(X_1 \in B) = \int_B f_\theta(y)\,\lambda(dy) = \int_B f_\theta(y)\,dy.$$

Определение 6. Функция (случайная величина при фиксированном θ)

$$f(\mathbf{X}, \theta) = f_\theta(X_1) \cdot f_\theta(X_2) \cdot \ldots \cdot f_\theta(X_n) = \prod_{i=1}^{n} f_\theta(X_i)$$

называется функцией правдоподобия. Функция (тоже случайная)

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

называется логарифмической функцией правдоподобия.

В дискретном случае функция правдоподобия $f(x_1, \ldots, x_n, \theta)$ есть вероятность выборке X_1, ..., X_n в данной серии экспериментов равняться x_1, ..., x_n. Эта вероятность меняется в зависимости от θ:

$$f(\mathbf{x}, \theta) = \prod_{i=1}^{n} f_\theta(x_i) = \mathsf{P}_\theta(X_1 = x_1) \cdot \ldots \cdot \mathsf{P}_\theta(X_n = x_n) = \mathsf{P}_\theta(X_1 = x_1, \ldots, X_n = x_n).$$

Определение 7. Оценкой максимального правдоподобия $\hat{\theta}$ неизвестного параметра θ называют значение θ, при котором функция $f(\mathbf{X}, \theta)$ достигает максимума (как функция от θ при фиксированных $X_1, \ldots, X_n$):

$$\hat{\theta} = \arg\max_{\theta} f(\mathbf{X}, \theta).$$

Замечание 7. Поскольку функция $\ln y$ монотонна, то точки максимума $f(\mathbf{X}, \theta)$ и $L(\mathbf{X}, \theta)$ совпадают. Поэтому оценкой максимального правдоподобия (ОМП) можно называть точку максимума (по θ) функции $L(\mathbf{x}, \theta)$:

$$\hat{\theta} = \arg\max_{\theta} L(\mathbf{X}, \theta).$$

Напомним, что точки экстремума функции — это либо точки, в которых производная обращается в нуль, либо точки разрыва функции/производной, либо крайние точки области определения функции.

Пример 7. Пусть $X_1, \ldots, X_n$ — выборка объема n из распределения Пуассона Π_λ, где $\lambda > 0$. Найдем ОМП $\hat\lambda$ неизвестного параметра λ.

$$P_\lambda(X_1 = y) = \frac{\lambda^y}{y!}\, e^{-\lambda}, \qquad y = 0, 1, 2, \ldots$$

$$f(\mathbf{X}, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i}}{X_i!}\, e^{-\lambda} = \frac{\lambda^{\Sigma X_i}}{\prod X_i!}\, e^{-n\lambda} = \frac{\lambda^{n\overline{X}}}{\prod X_i!}\, e^{-n\lambda}.$$

Поскольку эта функция при всех $\lambda > 0$ непрерывно дифференцируема по λ, можно искать точки экстремума, приравняв к нулю частную производную по λ. Но удобнее это делать для логарифмической функции правдоподобия:

$$L(\mathbf{X}, \lambda) = \ln f(\mathbf{X}, \lambda) = \ln\left(\frac{\lambda^{n\overline{X}}}{\prod X_i!}\, e^{-n\lambda}\right) = n\overline{X}\ln\lambda - \ln\prod_{i=1}^{n} X_i! - n\lambda.$$

Тогда

$$\frac{\partial}{\partial\lambda} L(\mathbf{X}, \lambda) = \frac{n\overline{X}}{\lambda} - n,$$

и точка экстремума $\hat\lambda$ — решение уравнения: $\dfrac{n\overline{X}}{\lambda} - n = 0$, то есть $\hat\lambda = \overline{X}$.

Упражнение.

1) Убедиться, что $\hat\lambda = \overline{X}$ — точка максимума, а не минимума.

2) Убедиться, что $\hat\lambda = \overline{X}$ совпадает с одной из оценок метода моментов. по какому моменту?

Пример 8. Пусть $X_1, \ldots, X_n$ — выборка объема n из нормального распределения N_{a,σ^2}, где $a \in \mathbb{R}$, $\sigma > 0$; и оба параметра a, σ^2 неизвестны.

Выпишем плотность, функцию правдоподобия и логарифмическую функцию правдоподобия. Плотность:

$$f_{(a,\sigma^2)}(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\, \exp\left(\frac{-(y-a)^2}{2\sigma^2}\right),$$

функция правдоподобия:

$$f(\mathbf{X}, a, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, \exp\left(-\frac{(X_i - a)^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}}\, \exp\left(-\frac{\sum\limits_{i=1}^{n}(X_i - a)^2}{2\sigma^2}\right),$$

логарифмическая функция правдоподобия:

$$L(\mathbf{X}, a, \sigma^2) = \ln f(\mathbf{X}, a, \sigma^2) = -\ln(2\pi)^{n/2} - \frac{n}{2}\ln\sigma^2 - \frac{\sum\limits_{i=1}^{n}(X_i - a)^2}{2\sigma^2}.$$

В точке экстремума (по (a, σ^2)) гладкой функции L обращаются в нуль обе частные производные:

$$\frac{\partial}{\partial a} L(\mathbf{X}, a, \sigma^2) = \frac{2\sum\limits_{i=1}^{n}(X_i - a)}{2\sigma^2} = \frac{n\overline{X} - na}{\sigma^2}; \quad \frac{\partial}{\partial\sigma^2} L(\mathbf{X}, a, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{\sum\limits_{i=1}^{n}(X_i - a)^2}{2\sigma^4}.$$

Оценка максимального правдоподобия $(\hat{a}, \widehat{\sigma^2})$ для (a, σ^2) — решение системы уравнений

$$\frac{n\overline{X} - na}{\sigma^2} = 0; \qquad -\frac{n}{2\sigma^2} + \frac{\sum\limits_{i=1}^{n}(X_i - a)^2}{2(\sigma^2)^2} = 0.$$

Решая, получим хорошо знакомые оценки:

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

Упражнение.

1) Убедиться, что $\hat{a} = \overline{X}$, $\widehat{\sigma^2} = S^2$ — точка максимума, а не минимума.

2) Убедиться, что эти оценки совпадают с некоторыми оценками метода моментов.

Пример 9. Пусть $X_1, \ldots, X_n$ — выборка объема n из равномерного распределения $U_{0,\theta}$, где $\theta > 0$. Тогда $\hat{\theta} = X_{(n)} = \max\{X_1, \ldots, X_n\}$ (см. [3], пример 4.4, с.24 или [1], пример 5, с.91).

Пример 10. Пусть $X_1, \ldots, X_n$ — выборка объема n из равномерного распределения $U_{\theta,\theta+5}$, где $\theta \in \mathbb{R}$ (см. также [1], пример 4, с.91).

Выпишем плотность распределения и функцию правдоподобия. Плотность:

$$f_\theta(y) = \begin{cases} 1/5, & \text{если } y \in [\theta, \theta+5] \\ 0 & \text{иначе,} \end{cases}$$

функция правдоподобия:

$$f(\mathbf{X}, \theta) = \begin{cases} (1/5)^n, & \text{если все } X_i \in [\theta, \theta+5] \\ 0 & \text{иначе} \end{cases} = \begin{cases} (1/5)^n, & \theta \leqslant X_{(1)} \leqslant X_{(n)} \leqslant \theta+5 \\ 0 & \text{иначе} \end{cases} =$$

$$= \begin{cases} (1/5)^n, & \text{если } X_{(n)} - 5 \leqslant \theta \leqslant X_{(1)} \\ 0 & \text{иначе.} \end{cases}$$

Функция правдоподобия достигает своего максимального значения $(1/5)^n$ во всех точках $\theta \in [X_{(n)} - 5, X_{(1)}]$. График этой функции изображен на рис. 4.

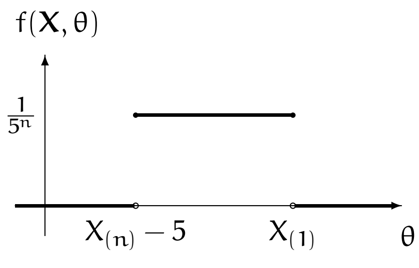


Рис. 4: Пример 10.

Любая точка $\hat{\theta} \in [X_{(n)} - 5, X_{(1)}]$ может служить оценкой максимального правдоподобия. Получаем более чем счетное число оценок вида

$$\hat{\theta}_\alpha = (1-\alpha)(X_{(n)} - 5) + \alpha X_{(1)}$$

при разных $\alpha \in [0,\ 1]$, в том числе $\hat{\theta}_0 = X_{(n)} - 5$ и $\hat{\theta}_1 = X_{(1)}$ — концы отрезка.

Упражнение.

1) Убедиться, что отрезок $[X_{(n)} - 5, X_{(1)}]$ не пуст.

2) Найти оценку метода моментов (по первому моменту) и убедиться, что она иная по сравнению с ОМП.

3) Найти ОМП параметра θ равномерного распределения $U_{\theta,2\theta}$.

2.6. Вопросы и упражнения

1. Дана выборка $X_1, \ldots, X_n$ из распределения Бернулли B_p, где $p \in (0,1)$ — неизвестный параметр. Проверить, что X_1, X_1X_2, $X_1(1-X_2)$ являются несмещенными оценками соответственно для p, p^2, $p(1-p)$. Являются ли эти оценки состоятельными?

2. Дана выборка $X_1, \ldots, X_n$ из распределения Пуассона Π_λ, где $\lambda > 0$ — неизвестный параметр. Проверить, что X_1 и $\mathbf{I}(X_1 = k)$ являются несмещенными оценками соответственно для λ и $\dfrac{\lambda^k}{k!}e^{-\lambda}$. Являются ли эти оценки состоятельными?

3. Дана выборка $X_1, \ldots, X_n$ из равномерного распределения $U_{0,\theta}$, где $\theta > 0$ — неизвестный параметр. Проверить состоятельность и несмещенность оценок $\theta_1^* = X_{(n)}$, $\theta_2^* = 2\overline{X}$, $\theta_3^* = X_{(n)} + X_{(1)}$ для параметра θ.

4. Построить оценки неизвестных параметров по методу моментов для неизвестных параметров следующих семейств распределений:
   а) B_p — по первому моменту, б) Π_λ — по первому и второму моменту, в) $U_{a,b}$ — по первому и второму моменту, г) E_α — по всем моментам, д) $E_{1/\alpha}$ — по первому моменту, е) $U_{-\theta,\theta}$ — как получится, ж) $\Gamma_{\alpha,\lambda}$ — по первому и второму моменту, з) N_{a,σ^2} (для σ^2 при a известном и при a неизвестном).

5. Какие из оценок в задаче 4 несмещенные? состоятельные?

6. Сравнить вид оценок для параметра α, полученных по первому моменту в задачах 4(г) и 4(д). Доказать, что среди них только одна несмещенная. **Указание.** Использовать неравенство Йенсена.

7. Построить оценки неизвестных параметров по методу максимального правдоподобия для следующих семейств распределений: а) B_p, б) $\Pi_{\lambda+1}$, в) $U_{0,2\theta}$, г) $E_{2\alpha+3}$, д) $U_{-\theta,\theta}$, е) N_{a,σ^2} (a известно).

8. Какие из оценок в задаче 7 несмещенные? состоятельные?

3. Сравнение оценок

Используя метод моментов и метод максимального правдоподобия, мы получили для каждого параметра уже достаточно много различных оценок. Каким же образом их сравнивать? Что должно быть показателем «хорошести» оценки?

Понятно, что чем дальше оценка отклоняется от параметра, тем она хуже. Но величина $|\theta^* - \theta|$ для сравнения непригодна: во-первых, параметр θ неизвестен, во-вторых, θ^* — случайная величина, так что эти величины обычно сравнить нельзя. Как, например, сравнивать $|\overline{X}-\theta|$ и $|\overline{X^k}-\theta|$? Или, на одном элементарном исходе, $|2.15-\theta|$ и $|3.1-\theta|$?

Поэтому имеет смысл сравнивать не отклонения как таковые, а средние значения этих отклонений, то есть $E_\theta\,|\theta^* - \theta|$.

Но математическое ожидание модуля с. в. считать обычно затруднительно, поэтому более удобной характеристикой для сравнения оценок считается $E_\theta\,(\theta^*-\theta)^2$. Она удобна еще и тем, что очень чутко реагирует на маловероятные, но большие по абсолютному значению отклонения θ^* от θ (возводит их в квадрат).

Заметим еще, что $E_\theta\,(\theta^* - \theta)^2$ есть функция от θ, так что сравнивать эти «среднеквадратические» отклонения нужно как функции от θ — поточечно. Такой подход к сравнению оценок называется среднеквадратическим.

Разумеется, в зависимости от потребностей исследователя можно пользоваться и другими характеристиками, например, $E_\theta\,(\theta^* - \theta)^4$ или $E_\theta\,|\theta^* - \theta|$.

Существует и так называемый асимптотический подход к сравнению оценок, при котором для сравнения оценок используется некая характеристика «разброса» оценки относительно параметра при больших n.

3.1. Среднеквадратический подход. Эффективность оценок

Пусть $X_1, \ldots, X_n$ — выборка объема n из параметрического семейства распределений $\mathcal{F}_\theta$, где $\theta \in \Theta$.

Определение 8. Говорят, что оценка θ_1^* лучше оценки θ_2^* в смысле среднеквадратического подхода, если для любого $\theta \in \Theta$

$$\mathsf{E}_\theta\,(\theta_1^* - \theta)^2 \leqslant \mathsf{E}_\theta\,(\theta_2^* - \theta)^2,$$

и хотя бы при одном θ это неравенство строгое.

Существует ли среди всех оценок наилучшая в смысле среднеквадратического подхода? Скептик сразу ответит «нет». Покажем, что он прав. Предположим, что мы имеем дело с невырожденной задачей: ни для какой статистики θ^* невозможно тождество: $\theta^* = \theta$ при любых $\theta \in \Theta$.

Теорема 4. В классе всех возможных оценок наилучшей в смысле среднеквадратического подхода оценки не существует.

Доказательство теоремы 4. Пусть, напротив, θ^* — наилучшая, то есть для любой другой оценки θ_1^*, при любом $\theta \in \Theta$ выполнено

$$\mathsf{E}_\theta\,(\theta^* - \theta)^2 \leqslant \mathsf{E}_\theta\,(\theta_1^* - \theta)^2.$$

Пусть θ_1 — произвольная точка Θ. Рассмотрим статистику $\theta_1^* \equiv \theta_1$. Тогда

$$\mathsf{E}_\theta\,(\theta^* - \theta)^2 \leqslant \mathsf{E}_\theta\,(\theta_1 - \theta)^2 \text{ при любом } \theta \in \Theta.$$

В частности, при $\theta = \theta_1$ получим $\mathsf{E}_{\theta_1}(\theta^* - \theta_1)^2 \leqslant \mathsf{E}_{\theta_1}(\theta_1 - \theta_1)^2 = 0$. Поэтому $\mathsf{E}_{\theta_1}(\theta^* - \theta_1)^2 = 0$. Но, поскольку θ_1 произвольно, при любом $\theta \in \Theta$ выполняется $\mathsf{E}_\theta\,(\theta^* - \theta)^2 = 0$. А это возможно только если $\theta^* \equiv \theta$ (оценка в точности отгадывает неизвестный параметр), т. е. для вырожденной с точки зрения математической статистики задачи.

Вырожденными являются, например, следующие задачи:
* для выборки из $\mathbf{I}_\theta$, $\theta \in \mathbb{R}$, выполнено тождество $X_1 \equiv \theta$;
* для выборки из $\mathrm{U}_{\theta,\theta+1}$, $\theta \in \mathbb{Z}$, выполнено тождество $[X_1] \equiv \theta$.

Упражнение. Объяснить словесно доказательство теоремы 4.

Если в классе всех оценок наилучшей не существует, то, возможно, следует разбить класс всех оценок на отдельные подклассы и в каждом искать наилучшую.

Обычно рассматривают оценки, имеющие одинаковое смещение

$$b(\theta) = \mathsf{E}_\theta\,\theta^* - \theta.$$

Обозначим через $K_b = K_{b(\theta)}$ класс оценок, имеющих смещение, равное заданной функции $b(\theta)$:

$$K_b = \{\theta^* \;:\; \mathsf{E}_\theta\,\theta^* = \theta + b(\theta)\}, \quad K_0 = \{\theta^* \;:\; \mathsf{E}_\theta\,\theta^* = \theta\}.$$

Здесь K_0 — класс несмещенных оценок.

Определение 9. Оценка $\theta^* \in K_b$ называется эффективной оценкой в классе K_b, если она лучше (не хуже) всех других оценок класса K_b в смысле среднеквадратического подхода. То есть для любой $\theta_1^* \in K_b$, для любого $\theta \in \Theta$

$$\mathsf{E}_\theta\,(\theta^* - \theta)^2 \leqslant \mathsf{E}_\theta\,(\theta_1^* - \theta)^2.$$

Определение 10. Эффективная оценка в классе K_0 называется просто эффективной.

Замечание 8. Для $\theta^* \in K_0$, по определению дисперсии,

$$\mathsf{E}_\theta\,(\theta^* - \theta)^2 = \mathsf{E}_\theta\,(\theta^* - \mathsf{E}_\theta\,\theta^*)^2 = \mathsf{D}_\theta\,\theta^*,$$

так что сравнение в среднеквадратичном несмещенных оценок — это сравнение их дисперсий. Поэтому эффективную оценку (в классе K_0) часто называют «несмещенной оценкой с равномерно минимальной дисперсией». Равномерность подразумевается по всем $\theta \in \Theta$. Для $\theta^* \in K_b$

$$\mathsf{E}_\theta\,(\theta^* - \theta)^2 = \mathsf{D}_\theta\,(\theta^* - \theta^*) + (\mathsf{E}_\theta\,\theta^* - \theta)^2 = \mathsf{D}_\theta\,\theta^* + b^2(\theta),$$

так что сравнение в среднеквадратичном оценок с одинаковым смещением — это также сравнение их дисперсий.

Упражнение. Мы собираемся искать наилучшую оценку в классе K_b. Объясните, почему доказательство теоремы 4 не пройдет в классе K_b.

3.2. Единственность эффективной оценки в классе с заданным смещением

Теорема 5. Если $\theta_1^* \in K_b$ и $\theta_2^* \in K_b$ — две эффективные оценки в классе K_b, то с вероятностью 1 они совпадают: $\mathsf{P}_\theta\,(\theta_1^* = \theta_2^*) = 1$.

Доказательство теоремы 5. Заметим сначала, что $\mathsf{E}_\theta\,(\theta_1^* - \theta)^2 = \mathsf{E}_\theta\,(\theta_2^* - \theta)^2$. Действительно, так как θ_1^* эффективна в классе K_b, то она не хуже оценки θ_2^*, то есть

$$\mathsf{E}_\theta\,(\theta_1^* - \theta)^2 \leqslant \mathsf{E}_\theta\,(\theta_2^* - \theta)^2,$$

и наоборот. Поэтому $\mathsf{E}_\theta\,(\theta_1^* - \theta)^2 = \mathsf{E}_\theta\,(\theta_2^* - \theta)^2$.

Рассмотрим оценку $\theta^* = \dfrac{\theta_1^* + \theta_2^*}{2}$. Она также принадлежит классу K_b. доказать! Вычислим ее среднеквадратическое отклонение. Заметим, что

$$\left(\frac{a+b}{2}\right)^2 + \left(\frac{a-b}{2}\right)^2 = \frac{a^2+b^2}{2}. \tag{6}$$

Положим $a = \theta_1^* - \theta$, $b = \theta_2^* - \theta$. Тогда $(a+b)/2 = \theta^* - \theta$, $a - b = \theta_1^* - \theta_2^*$. Подставим эти выражения в (6) и возьмем математические ожидания обеих частей:

$$\mathsf{E}_\theta\,(\theta^* - \theta)^2 + \mathsf{E}_\theta\left(\frac{\theta_1^* - \theta_2^*}{2}\right)^2 = \mathsf{E}_\theta\,\frac{(\theta_1^* - \theta)^2 + (\theta_2^* - \theta)^2}{2} = \\ = \mathsf{E}_\theta\,(\theta_1^* - \theta)^2 = \mathsf{E}_\theta\,(\theta_2^* - \theta)^2. \tag{7}$$

Но оценка θ^* принадлежит K_b, то есть она не лучше, например, эффективной оценки θ_1^*. Поэтому

$$\mathsf{E}_\theta\,(\theta^* - \theta)^2 \geqslant \mathsf{E}_\theta\,(\theta_1^* - \theta)^2.$$

Сравнивая это неравенство с равенством (7), видим, что

$$\mathsf{E}_\theta\left(\frac{\theta_1^* - \theta_2^*}{2}\right)^2 = \frac{1}{4}\,\mathsf{E}_\theta\,(\theta_1^* - \theta_2^*)^2 \leqslant 0 \quad \text{и, следовательно,} \quad \mathsf{E}_\theta\,(\theta_1^* - \theta_2^*)^2 = 0.$$

Тогда почему? $\mathsf{P}_\theta\,(\theta_1^* = \theta_2^*) = 1$, что и требовалось доказать.

Для примера рассмотрим сравнение двух оценок. Разумеется, сравнивая оценки попарно между собой, наилучшей оценки в целом классе не найти, но выбрать лучшую из двух тоже полезно. А способами поиска наилучшей в целом классе мы тоже скоро займемся.

Пример 11. Пусть $X_1, \ldots, X_n$ — выборка объема n из равномерного распределения $\mathsf{U}_{0,\theta}$, где $\theta > 0$. В примерах 4 и 9 мы нашли ОМП $\hat\theta = X_{(n)} = \max\{X_1, \ldots, X_n\}$ и ОММ по первому моменту $\theta^* = 2\overline{X}$. Сравним их в среднеквадратичном.

Оценка $\theta^* = 2\overline{X}$ несмещенная, поэтому

$$\mathsf{E}_\theta\,(\theta^* - \theta)^2 = \mathsf{D}_\theta\,\theta^* = \mathsf{D}_\theta\,2\overline{X} = 4\mathsf{D}_\theta\,\overline{X} = 4\frac{\mathsf{D}_\theta\,X_1}{n} = 4\frac{\theta^2}{12n} = \frac{\theta^2}{3n}.$$

Для $\hat\theta = X_{(n)} = \max\{X_1, \ldots, X_n\}$ имеем

$$\mathsf{E}_\theta\,(\hat\theta - \theta)^2 = \mathsf{E}_\theta\,\hat\theta^2 - 2\theta\mathsf{E}_\theta\,\hat\theta + \theta^2.$$

Посчитаем первый и второй момент случайной величины $\hat\theta = X_{(n)}$. Найдем (полезно вспомнить, как это делалось в прошлом семестре!) функцию распределения и плотность $\hat\theta$:

$$\mathsf{P}_\theta\,(X_{(n)} < y) = \mathsf{P}_\theta\,(X_1 < y)^n = \begin{cases} 0, & y < 0, \\ \dfrac{y^n}{\theta^n}, & y \in [0, \theta], \\ 1, & y > \theta, \end{cases}$$

$$f_{X_{(n)}}(y) = \begin{cases} 0, & \text{если } y \notin [0, \theta], \\ n\dfrac{y^{n-1}}{\theta^n}, & \text{если } y \in [0, \theta]. \end{cases}$$

$$\mathsf{E}_\theta\,X_{(n)} = \int\limits_0^\theta y n\frac{y^{n-1}}{\theta^n}\,dy = \frac{n}{n+1}\,\theta, \quad \mathsf{E}_\theta\,X_{(n)}^2 = \int\limits_0^\theta y^2 n\frac{y^{n-1}}{\theta^n}\,dy = \frac{n}{n+2}\,\theta^2.$$

Поэтому

$$\mathsf{E}_\theta\,(X_{(n)} - \theta)^2 = \frac{n}{n+2}\,\theta^2 - 2\frac{n}{n+1}\,\theta^2 + \theta^2 = \frac{2}{(n+1)(n+2)}\,\theta^2.$$

При $n = 1, 2$ квадратические отклонения равны, а при $n > 2$

$$\mathsf{E}_\theta\,(X_{(n)} - \theta)^2 = \frac{2\theta^2}{(n+1)(n+2)} < \frac{\theta^2}{3n} = \mathsf{E}_\theta\,(2\overline{X} - \theta)^2,$$

то есть $X_{(n)}$ лучше, чем $2\overline{X}$. При этом $\mathsf{E}_\theta\,(X_{(n)} - \theta)^2$ стремится к нулю со скоростью n^{-2}, тогда как $\mathsf{E}_\theta\,(2\overline{X} - \theta)^2$ — со скоростью n^{-1}.

Упражнение.

1. Доказать, что $X_{(n)} \in K_b$, где $b(\theta) = -\dfrac{\theta}{n+1}$.
2. Доказать, что $\dfrac{n+1}{n} X_{(n)} \in K_0$ (несмещенная).
3. Сравнить оценки $\dfrac{n+1}{n} X_{(n)}$ и $X_{(n)}$ в среднеквадратичном.

3.3. Асимптотически нормальные оценки (АНО)

Для того, чтобы уметь сравнивать оценки вида $\theta_k^* = \sqrt[k]{(k+1)\overline{X^k}}$ (см. пример 4), среднеквадратического подхода недостаточно: второй момент такой случайной величины посчитать вряд ли удастся. Оценки такого вида (функции от сумм) удается сравнивать с помощью асимптотического подхода. Более точно, этот подход применим к так называемым «асимптотически нормальным» оценкам.

Пусть $X_1, \ldots, X_n$ — выборка объема n из параметрического семейства распределений $\mathcal{F}_\theta$, $\theta \in \Theta$.

Определение 11. Оценка θ^* называется асимптотически нормальной оценкой параметра θ с коэффициентом $\sigma^2(\theta)$, если

$$\sqrt{n}(\theta^* - \theta) \Rightarrow \mathsf{N}_{0,\sigma^2(\theta)}, \qquad \text{или} \qquad \frac{\sqrt{n}(\theta^* - \theta)}{\sigma(\theta)} \Rightarrow \mathsf{N}_{0,1}.$$

Пример 12. Пусть $X_1, \ldots, X_n$ — выборка объема n из равномерного распределения $\mathsf{U}_{0,\theta}$, где $\theta > 0$. Проверим, являются ли оценки $\theta^* = 2\overline{X}$ и $\hat{\theta} = X_{(n)}$ асимптотически нормальными (АНО). По ЦПТ,

$$\sqrt{n}(\theta^* - \theta) = \sqrt{n}(2\overline{X} - \theta) = \sqrt{n}\left(2\frac{\sum\limits_{i=1}^{n} X_i}{n} - \theta\right) = \frac{\sum\limits_{i=1}^{n} 2X_i - n\theta}{\sqrt{n}} =$$

$$= \frac{\sum\limits_{i=1}^{n} 2X_i - n\mathsf{E}_\theta\, 2X_1}{\sqrt{n}} \Rightarrow \mathsf{N}_{0,\mathsf{D}_\theta\, 2X_1} = \mathsf{N}_{0,4\mathsf{D}_\theta\, X_1}.$$

То есть оценка $\theta^* = 2\overline{X}$ асимптотически нормальна с коэффициентом

$$\sigma^2(\theta) = 4\mathsf{D}_\theta\, X_1 = 4\theta^2/12 = \theta^2/3.$$

Для оценки $\hat{\theta} = X_{(n)}$ имеем:

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}(X_{(n)} - \theta) < 0 \text{ с вероятностью } 1. \tag{8}$$

По определению, $\xi_n \Rightarrow F$, если для любой точки x, являющейся точкой непрерывности функции распределения F, имеет место сходимость $F_{\xi_n}(x) = \mathsf{P}(\xi_n < x) \to F(x)$.

Но $\mathsf{P}_\theta\left(\sqrt{n}(X_{(n)} - \theta) < 0\right) = 1$, тогда как для нормального распределения $\mathsf{N}_{0,\sigma^2(\theta)}$ функция распределения всюду непрерывна, и в нуле равна $\Phi_{0,\sigma^2(\theta)}(0) = 0.5$. Но 1 не сходится к 0.5 при $n \to \infty$, поэтому слабая сходимость $\sqrt{n}(X_{(n)} - \theta)$ к $\mathsf{N}_{0,\sigma^2(\theta)}$ места не имеет.

Таким образом, оценка $\hat{\theta} = X_{(n)}$ асимптотически нормальной не является. Осталось ответить на напрашивающиеся вопросы:

1) Куда все же сходится по распределению $\sqrt{n}(X_{(n)}-\theta)$?

Упражнение. Доказать, что $\sqrt{n}(X_{(n)}-\theta) \Rightarrow 0$.

Порядок действий: Выписать определение слабой сходимости. Нарисовать функцию распределения нуля. Найти по определению функцию распределения $\sqrt{n}(X_{(n)}-\theta)$. Убедиться, что она сходится к функции распределения нуля во всех точках непрерывности последней. Не забудьте о существовании замечательных пределов, логарифмов и ряда Тейлора.

2) Если $\sqrt{n}(X_{(n)}-\theta) \Rightarrow 0$, то на какую степень n нужно попробовать умножить $X_{(n)}-\theta$, чтобы получить сходимость к величине, отличной от 0 и ∞?

Упражнение. Доказать, что $-n(X_{(n)}-\theta) \Rightarrow \eta$, где случайная величина η имеет показательное распределение $E_{1/\theta}$.

Порядок действий: прежний.

3) Для оценки $\dfrac{n+1}{n}X_{(n)}$ свойство (8) не выполнено. Может ли эта оценка быть АНО?

Упражнение. Модифицировать рассуждения и доказать, что эта оценка тоже не является асимптотически нормальной.

4) Плохо ли, что оценка $\hat{\theta} = X_{(n)}$ не асимптотически нормальна? Может быть, сходимость $n(X_{(n)}-\theta) \Rightarrow -\eta$ еще лучше?

Попробуем ответить на последний вопрос.

3.4. «Скорость» сходимости оценки к параметру

Теорема 6. Если θ^* — асимптотически нормальная оценка для θ, то θ^* состоятельна.

Доказательство теоремы 6. Вспомним свойство слабой сходимости: произведение двух последовательностей, одна из которых сходится (по вероятности) к постоянной, а другая слабо сходится к некоторой случайной величине, слабо сходится к произведению пределов. Поэтому

$$\theta^*-\theta = \frac{1}{\sqrt{n}} \cdot \sqrt{n}(\theta^*-\theta) \Rightarrow 0 \cdot \xi = 0,$$

где ξ имеет нормальное распределение $N_{0,\sigma^2(\theta)}$. Но слабая сходимость к нулю влечет сходимость к нулю по вероятности. $\square$

Упражнение. Верно ли утверждение теоремы 6, если предельная величина ξ имеет распределение, отличное от нормального?

Таким образом, если θ^* асимптотически нормальна, то $\theta^* \xrightarrow{p} \theta$, или $\theta^*-\theta \xrightarrow{p} 0$. Свойство асимптотической нормальности показывает, в частности, что скорость этой сходимости имеет порядок $\dfrac{1}{\sqrt{n}}$, т. е. расстояние между θ^* и θ ведет себя как $\dfrac{1}{\sqrt{n}}$:

$$\theta^*-\theta \xrightarrow{p} 0, \quad \text{но} \quad \sqrt{n}(\theta^*-\theta) \Rightarrow N_{0,\sigma^2(\theta)}.$$

Взглянем с этой точки зрения на оценку $\hat{\theta} = X_{(n)}$ в примере 12. Для нее (и для тех, кто справился с упражнениями)

$$n(X_{(n)}-\theta) \Rightarrow \xi, \tag{9}$$

где ξ — некоторая случайная величина. Иначе говоря, расстояние между $\hat{\theta}$ и θ ведет себя как $\dfrac{1}{n}$.

Упражнение. Лучше это или хуже?

3.5. Асимптотическая нормальность ОММ

В примере 12 мы видели, что для оценок типа $2\overline{X}$ свойство асимптотической нормальности сразу следует из ЦПТ.

Установим асимптотическую нормальность оценок более сложного вида, какими обычно оказываются оценки метода моментов.

Свойство 5. Пусть функция $g(y)$ такова, что $0 \neq \mathsf{D}_\theta\, g(X_1) < \infty$. Тогда статистика $\overline{g(X)} = \dfrac{1}{n}\sum g(X_i)$ является асимптотически нормальной оценкой для $\mathsf{E}_\theta\, g(X_1)$ с коэффициентом $\sigma^2(\theta) = \mathsf{D}_\theta\, g(X_1)$:

$$\sqrt{n}\,\frac{\overline{g(X)} - \mathsf{E}_\theta\, g(X_1)}{\sqrt{\mathsf{D}_\theta\, g(X_1)}} \Rightarrow \mathsf{N}_{0,1}.$$

Упражнение. Вспомнить ЦПТ и доказать свойство 5.

Следующая теорема утверждает асимптотическую нормальность оценок вида

$$\theta^* = H\left(\overline{g(X)}\right) = H\left(\frac{\sum_1^n g(X_i)}{n}\right).$$

Такие оценки получаются обычно $\boxed{\text{найти примеры!}}$ при использовании метода моментов, при этом всегда $\theta = H\left(\mathsf{E}_\theta\, g(X_1)\right)$.

Теорема 7. Пусть функция $g(y)$ такова, что $0 \neq \mathsf{D}_\theta\, g(X_1) < \infty$, а функция $H(y)$ непрерывно дифференцируема в точке $a = \mathsf{E}_\theta\, g(X_1)$ и $H'(a) = H'(y)\big|_{y=a} \neq 0$.
Тогда оценка $\theta^* = H\left(\overline{g(X)}\right)$ является асимптотически нормальной оценкой для $\theta = H\left(\mathsf{E}_\theta\, g(X_1)\right) = H(a)$ с коэффициентом $\sigma^2(\theta) = \left(H'(a)\right)^2 \cdot \mathsf{D}_\theta\, g(X_1)$.

Доказательство теоремы 7.
Согласно ЗБЧ последовательность $\overline{g(X)}$ стремится к $a = \mathsf{E}_\theta\, g(X_1)$ по вероятности с ростом n. Функция

$$G(y) = \begin{cases} \dfrac{H(y) - H(a)}{y - a}, & y \neq a, \\ H'(a), & y = a \end{cases}$$

по условию непрерывна в точке a. Поскольку сходимость по вероятности сохраняется под действием непрерывной функции, получим, что $G(\overline{g(X)}) \xrightarrow{p} G(a) = H'(a)$.

Заметим также, что по свойству 5 величина $\sqrt{n}\left(\overline{g(X)} - a\right)$ слабо сходится к нормальному распределению $\mathsf{N}_{0,\mathsf{D}_\theta\, g(X_1)}$. Пусть ξ — случайная величина из этого распределения. Тогда

$$\sqrt{n}\left(H\left(\overline{g(X)}\right) - H(a)\right) = \underset{\substack{\Downarrow \\ \xi}}{\sqrt{n}\left(\overline{g(X)} - a\right)} \cdot \underset{\substack{\downarrow p \\ H'(a)}}{G\left(\overline{g(X)}\right)} \Rightarrow \xi \cdot H'(a).$$

Мы использовали свойство слабой сходимости: если $\xi_n \Rightarrow \xi$ и $\eta_n \xrightarrow{p} c = \text{const}$, то $\xi_n\eta_n \Rightarrow c\xi$. Но $\xi \cdot H'(a)$ как раз и имеет распределение $\mathsf{N}_{0,(H'(a))^2 \cdot \mathsf{D}_\theta\, g(X_1)}$.

Пример 13. Пусть $X_1, \ldots, X_n$ — выборка объема n из равномерного распределения $U_{0,\theta}$, где $\theta > 0$. Проверим, являются ли оценки $\theta^*_k = \sqrt[k]{(k+1)\overline{X^k}}$, $k = 1, 2, \ldots$, полученные методом моментов в примере 4, асимптотически нормальными.

Пусть $g(y) = (k+1)y^k$, $H(y) = \sqrt[k]{y}$. Тогда

$$\theta^*_k = \sqrt[k]{(k+1)\overline{X^k}} = \sqrt[k]{\frac{\sum (k+1)X_i^k}{n}} = H\left(\frac{\sum g(X_i)}{n}\right).$$

При этом

$$\theta = H\left(\mathsf{E}_\theta\, g(X_1)\right) = \sqrt[k]{\mathsf{E}_\theta\,(k+1)X_1^k} = \sqrt[k]{(k+1)\ \frac{\theta^k}{k+1}}.$$

Впрочем, иначе быть не могло по определению метода моментов. верно? Проверим другие условия теоремы 7:

$$a = \mathsf{E}_\theta\, g(X_1) = (k+1)\ \frac{\theta^k}{k+1} = \theta^k,$$

дисперсия $\mathsf{D}_\theta\, g(X_1) = \mathsf{E}_\theta\,(k+1)^2 X_1^{2k} - a^2 = (k+1)^2 \dfrac{\theta^{2k}}{2k+1} - \theta^{2k} = \dfrac{k^2}{2k+1}\ \theta^{2k}$

конечна и отлична от нуля. Функция $H(y)$ непрерывно дифференцируема в точке a:

$$H'(y) = \frac{1}{k}\, y^{\frac{1-k}{k}}, \quad \text{и} \quad H'(a) = H'(\theta^k) = \frac{1}{k}\theta^{1-k} \quad \text{непрерывна при } \theta > 0.$$

По теореме 7, оценка θ^*_k — АНО для θ с коэффициентом

$$\sigma^2_k(\theta) = \left(H'(a)\right)^2 \mathsf{D}_\theta\, g(X_1) = \frac{1}{k^2}\,\theta^{2-2k} \cdot \frac{k^2}{2k+1}\ \theta^{2k} = \frac{\theta^2}{2k+1}.$$

В том числе для $\theta^*_1 = 2\overline{X}$ имеем коэффициент $\sigma^2_1(\theta) = \dfrac{\theta^2}{3}$ (см. пример 12).

Осталось понять, при чем тут сравнение оценок и что показывает коэффициент асимптотической нормальности.

3.6. Асимптотический подход к сравнению оценок

Возьмем две случайные величины: ξ из нормального распределения $N_{0,1}$ и $10\,\xi$ из нормального распределения $N_{0,100}$. Если для ξ, например, $0,9973.. = \mathsf{P}\left(|\xi| < 3\right)$, то для $10\,\xi$ уже $0,9973.. = \mathsf{P}\left(|\xi| < 30\right)$. Разброс значений величины $10\,\xi$ гораздо больший, и дисперсия (показатель рассеяния) соответственно больше.

Что показывает коэффициент асимптотической нормальности? Возьмем две АНО с коэффициентами 1 и 100:

$$\sqrt{n}(\theta^*_1 - \theta^*) \Rightarrow N_{0,1} \quad \text{и} \quad \sqrt{n}(\theta^*_2 - \theta^*) \Rightarrow N_{0,100}.$$

При больших n разброс значений величины $\sqrt{n}(\theta^*_2 - \theta^*)$ около нуля гораздо больше, чем у величины $\sqrt{n}(\theta^*_1 - \theta^*)$, поскольку больше предельная дисперсия (она же коэффициент асимптотической нормальности).

Но чем меньше отклонение оценки от параметра, тем лучше. Отсюда — естественный способ сравнения асимптотически нормальных оценок:

Определение 12. Пусть θ^*_1 — АНО с коэффициентом $\sigma^2_1(\theta)$, θ^*_2 — АНО с коэффициентом $\sigma^2_2(\theta)$. Говорят, что θ^*_1 лучше, чем θ^*_2 в смысле асимптотического подхода, если для любого $\theta \in \Theta$

$$\sigma^2_1(\theta) \leqslant \sigma^2_2(\theta),$$

и хотя бы при одном θ это неравенство строгое.

Пример 13 (продолжение). Сравним между собой в асимптотическом смысле оценки в последовательности $\theta_1^*, \theta_2^*, \ldots$. Для θ_k^* коэффициент асимптотической нормальности имеет вид $\sigma_k^2(\theta) = \theta^2/(2k+1)$. Коэффициент тем меньше, чем больше k, то есть каждая следующая оценка в этой последовательности лучше предыдущей.

Оценка θ_∞^*, являющаяся «последней», могла бы быть лучше всех оценок в этой последовательности в смысле асимптотического подхода, если бы являлась асимптотически нормальной. Увы:

Упражнение. См. задачу 7 (6) в разделе 1. Доказать, что $\theta_k^* \to X_{(n)}$ п. н., то есть для любого элементарного исхода ω при $k \to \infty$

$$\sqrt[k]{(k+1)\frac{\sum\limits_{i=1}^{n} X_i^k(\omega)}{n}} \to \max\{X_1(\omega), \ldots, X_n(\omega)\}.$$

Еще раз обращаем внимание читателя, что оценка $\hat{\theta} = X_{(n)}$ оказывается лучше любой асимптотически нормальной оценки: «скорость» ее сходимости к параметру, как показывает (9), равна n^{-1} в отличие от $n^{-1/2}$ для любой АНО.

3.7. Вопросы и упражнения

1. Пусть $X_1, \ldots, X_n$ — выборка объема n из равномерного распределения $U_{\theta,\theta+5}$, где $\theta \in \mathbb{R}$. Сравнить оценки $\hat{\theta}_0 = X_{(n)} - 5$, $\hat{\theta}_1 = X_{(1)}$ из примера 10 в среднеквадратичном смысле. Сравнить с этими оценками оценку метода моментов $\theta^* = \overline{X} - 2{,}5$.

2. Для показательного распределения с параметром α оценка, полученная методом моментов по k-му моменту, имеет вид: $\alpha_k^* = \sqrt[k]{\dfrac{k!}{\overline{X^k}}}$. Сравнить оценки α_k^*, $k = 1, 2, \ldots$ в смысле асимптотического подхода. Доказать, что оценка α_1^* наилучшая.

4. Эффективные оценки

Вернемся к сравнению оценок в смысле среднеквадратического подхода. В классе одинаково смещенных оценок эффективной мы назвали оценку с наименьшим среднеквадратическим отклонением (или наименьшей дисперсией). Но попарное сравнение оценок — далеко не лучший способ отыскания эффективной оценки. Сегодня мы познакомимся с утверждением, позволяющим во многих случаях доказать эффективность оценки (если, конечно, она на самом деле эффективна).

Это утверждение называется неравенством Рао — Крамéра и говорит о том, что в любом классе $K_{b(\theta)}$ существует нижняя граница для среднеквадратического отклонения $\mathsf{E}_\theta\,(\theta^* - \theta)^2$ любой оценки.

Таким образом, если найдется оценка, отклонение которой в точности равно этой нижней границе (самое маленькое), то данная оценка — эффективна, поскольку у прочих оценок отклонение меньше быть не может.

К сожалению, данное неравенство верно лишь для так называемых «регулярных» семейств распределений, к которым не относится, например, большинство равномерных.

4.1. Регулярность семейства распределений

Пусть $X_1, \ldots, X_n$ — выборка объема n из параметрического семейства распределений $\mathcal{F}_\theta$, $\theta \in \Theta$. Пусть $f_\theta(y)$ — плотность $\mathcal{F}_\theta$ (в смысле определения 5). Введем понятие носителя семейства распределений $\mathcal{F}_\theta$.

> Любое множество $C \subseteq \mathbb{R}$ такое, что при всех $\theta \in \Theta$ выполняется равенство $\mathsf{P}_\theta\,(X_1 \in C) = 1$, назовем носителем семейства распределений $\mathcal{F}_\theta$.

Замечание 9. Мы ввели понятие носителя семейства мер в $\mathbb{R}$, отличное от общепринятого. Так, носитель в смысле данного нами определения не единственен, но все эти носители отличаются на множество нулевой вероятности.

Следующее условие назовем условием регулярности.

(R) Существует такой носитель C семейства распределений $\mathcal{F}_\theta$, что при каждом $y \in C$ функция $\sqrt{f_\theta(y)}$ непрерывно дифференцируема по θ **во всех точках** $\theta \in \Theta$.

4.2. «Регулярные» и «нерегулярные» семейства распределений

Пример 14 (регулярное семейство). Рассмотрим показательное распределение E_α с параметром $\alpha > 0$. Плотность этого распределения имеет вид

$$f_\alpha(y) = \begin{cases} \alpha e^{-\alpha y}, & \text{если } y > 0, \\ 0, & \text{если } y \leqslant 0, \end{cases} \qquad \sqrt{f_\alpha(y)} = \begin{cases} \sqrt{\alpha} e^{-\alpha y/2}, & \text{если } y > 0, \\ 0, & \text{если } y \leqslant 0. \end{cases}$$

В качестве множества C можно взять $(0, +\infty)$, поскольку $\mathsf{P}_\alpha(X_1 > 0) = 1$. При любом $y \in C$, т. е. при $y > 0$, существует производная функции $\sqrt{f_\alpha(y)}$ по α, и эта производная непрерывна во всех точках $\alpha > 0$:

$$\frac{\partial}{\partial \alpha}\sqrt{f_\alpha(y)} = \frac{1}{2\sqrt{\alpha}} e^{-\alpha y/2} - \sqrt{\alpha}\frac{y}{2} e^{-\alpha y/2}.$$

Пример 15 (нерегулярное семейство). Рассмотрим равномерное распределение $\mathsf{U}_{0,\theta}$ с параметром $\theta > 0$. Плотность этого распределения имеет вид

$$f_\theta(y) = \begin{cases} \dfrac{1}{\theta}, & \text{если } 0 \leqslant y \leqslant \theta, \\ 0, & \text{если } y \notin [0, \theta] \end{cases} \quad = \quad \begin{cases} \dfrac{1}{\theta}, & \text{если } \theta \geqslant y \text{ и } y > 0, \\ 0 & \text{иначе.} \end{cases}$$

Поскольку параметр θ может принимать любые положительные значения, то никакой ограниченный интервал $(0, x)$ не является носителем этого семейства распределений: $\mathsf{P}_\theta(X_1 \in (0, x)) < 1$ при $\theta > x$. Возьмем $C = (0, +\infty)$ — оно при любом $\theta > 0$ обладает свойством $\mathsf{P}_\theta(X_1 \in C) = 1$. Так что носитель этого семейства распределений — вся положительная полуось (с точностью до множеств нулевой лебеговой меры). Покажем, что условие (R) не выполнено: множество тех $y \in C$, при каждом из которых функция $\sqrt{f_\theta(y)}$ дифференцируема по θ, пусто.

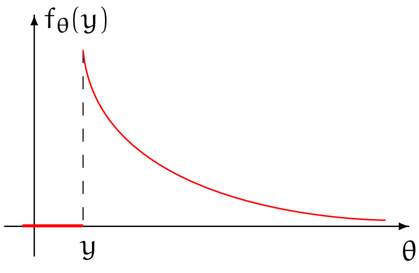


Рис. 5: Пример 15

При фиксированном $y > 0$ изобразим функцию $f_\theta(y)$ (или ее корень — масштаб не соблюден) как функцию переменной θ.
Видим, что какое бы $y \in C$ мы ни взяли, $f_\theta(y)$ даже не является непрерывной по θ, а тем более дифференцируемой. Следовательно, условие (R) не выполнено.

Пример 16 (нерегулярное семейство). Рассмотрим «смещенное» показательное распределение с параметром сдвига $\theta \in \mathbb{R}$ и плотностью

$$f_\theta(y) = \begin{cases} e^{\theta-y}, & \text{если } y > \theta, \\ 0, & \text{если } y \leqslant \theta \end{cases} = \begin{cases} e^{\theta-y}, & \text{если } \theta < y, \\ 0, & \text{если } \theta \geqslant y. \end{cases}$$

Поскольку при любом θ распределение сосредоточено на $(\theta, +\infty)$, а параметр θ может принимать любые вещественные значения, то только $C = \mathbb{R}$ (плюс-минус множество меры нуль) таково, что при любом $\theta > 0$ выполнено $P_\theta(X_1 \in C) = 1$. Покажем, что условие (R) опять не выполнено: множество тех $y \in C$, при каждом из которых функция $\sqrt{f_\theta(y)}$ дифференцируема по θ, столь же пусто, как и в примере 15.

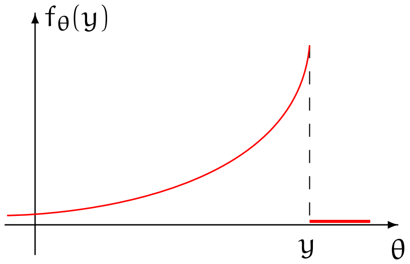


Рис. 6: Пример 16

При фиксированном $y \in \mathbb{R}$ на рисунке 6 изображена функция $f_\theta(y)$ (а может быть, корень из нее) как функция переменной θ.
Какое бы y ни было, $f_\theta(y)$ даже не является непрерывной по θ, а тем более дифференцируемой.

Замечание 10. Вместо непрерывной дифференцируемости $\sqrt{f_\theta(y)}$ можно требовать того же от $\ln f_\theta(y)$.

4.3. Неравенство Рао — Крамера

Пусть $X_1, \ldots, X_n$ — выборка объема n из параметрического семейства распределений $\mathcal{F}_\theta$, $\theta \in \Theta$, и семейство $\mathcal{F}_\theta$ удовлетворяет условию регулярности (R).

Пусть, кроме того, выполнено условие

(RR) «Информация Фишера» $\quad I(\theta) = E_\theta \left(\dfrac{\partial}{\partial\theta} \ln f_\theta(X_1) \right)^2$

существует, положительна и непрерывна по θ во всех точках $\theta \in \Theta$.

Справедливо следующее утверждение.

Неравенство Рао — Крамера.
Пусть семейство распределений $\mathcal{F}_\theta$ удовлетворяет условиям (R) и (RR). Тогда для любой несмещенной оценки $\theta^* \in K_0$, дисперсия которой $D_\theta\,\theta^*$ ограничена на любом компакте в области Θ, справедливо неравенство

$$D_\theta\,\theta^* = E_\theta\,(\theta^* - \theta)^2 \geqslant \frac{1}{nI(\theta)}.$$

Упражнение. Проверить, что для показательного семейства распределений E_α с параметром $\alpha > 0$ дисперсия $D_\alpha X_1$ не ограничена глобально при $\alpha > 0$, но ограничена на компактах.

Неравенство сформулировано для класса несмещенных оценок. В классе оценок с произвольным смещением $b(\theta)$ неравенство Рао — Крамера выглядит так:

Неравенство Рао — Крамера.
Пусть семейство распределений $\mathcal{F}_\theta$ удовлетворяет условиям (R) и (RR). Тогда для любой оценки $\theta^* \in K_{b(\theta)}$, дисперсия которой $\mathsf{D}_\theta\,\theta^*$ ограничена на любом компакте в области Θ, справедливо неравенство

$$\mathsf{E}_\theta\,(\theta^* - \theta)^2 \geqslant \frac{(1+b'(\theta))^2}{nI(\theta)} + b^2(\theta), \quad \text{т. е.} \quad \mathsf{D}_\theta\,\theta^* \geqslant \frac{(1+b'(\theta))^2}{nI(\theta)}.$$

Для доказательства нам понадобится следующее утверждение.

Лемма 1. При выполнении условий (R) и (RR) для любой статистики $T = T(\mathbf{X})$, дисперсия которой ограничена на компактах, имеет место равенство

$$\frac{\partial}{\partial\theta}\mathsf{E}_\theta\,T = \mathsf{E}_\theta\left(T\cdot\frac{\partial}{\partial\theta}L(\mathbf{X},\theta)\right).$$

Упражнение. Вспомнить, что такое функция правдоподобия $f(\mathbf{X},\theta)$, логарифмическая функция правдоподобия $L(\mathbf{X},\theta)$ (определение 6), как они связаны друг с другом, с плотностью X_1 и совместной плотностью выборки.

Доказательство леммы 1.
Напоминание: математическое ожидание функции от нескольких случайных величин есть (многомерный) интеграл от этой функции, помноженной на совместную плотность этих случайных величин. Поэтому

$$\mathsf{E}_\theta\,T(X_1,\ldots,X_n) = \int\limits_{\mathbb{R}^n} T(y_1,\ldots,y_n)\cdot f(y_1,\ldots,y_n,\theta)\,dy_1\ldots dy_n.$$

В следующей цепочке равенство, помеченное $(*)$, мы доказывать не будем, поскольку его доказательство требует знания условий дифференцируемости интеграла по параметру (тема, выходящая за пределы курса МА на ЭФ). Это равенство — смена порядка дифференцирования и интегрирования — то единственное, ради чего введены условия регулярности (см. пример ниже).

$$\frac{\partial}{\partial\theta}\mathsf{E}_\theta\,T(\mathbf{X}) = \frac{\partial}{\partial\theta}\int\limits_{\mathbb{R}^n} T(\mathbf{y})\cdot f(\mathbf{y},\theta)\,d\mathbf{y} \stackrel{*}{=} \int\limits_{\mathbb{R}^n} \frac{\partial}{\partial\theta}T(\mathbf{y})\cdot f(\mathbf{y},\theta)\,d\mathbf{y} =$$

$$= \int\limits_{\mathbb{R}^n} T(\mathbf{y})\cdot\frac{\partial}{\partial\theta}f(\mathbf{y},\theta)\,d\mathbf{y} = \int\limits_{\mathbb{R}^n} T(\mathbf{y})\cdot\left(\frac{\frac{\partial}{\partial\theta}f(\mathbf{y},\theta)}{f(\mathbf{y},\theta)}\right)\cdot f(\mathbf{y},\theta)\,d\mathbf{y} =$$

$$= \int\limits_{\mathbb{R}^n} T(\mathbf{y})\cdot\left(\frac{\partial}{\partial\theta}L(\mathbf{y},\theta)\right)\cdot f(\mathbf{y},\theta)\,d\mathbf{y} = \mathsf{E}_\theta\left(T(\mathbf{X})\cdot\frac{\partial}{\partial\theta}L(\mathbf{X},\theta)\right).$$

Через $\mathbf{y}$ в интегралах обозначен вектор $(y_1,\ldots,y_n)$.

Доказательство неравенства Рао — Крамера.
Мы докажем только неравенство для класса K_0. Необходимые изменения в доказательстве для класса K_b читатель может внести самостоятельно.

Воспользуемся леммой 1. Будем брать в качестве $T(\mathbf{X})$ разные функции и получать забавные формулы, которые потом соберем вместе.

1. Пусть $T(\mathbf{X}) \equiv 1$. Тогда

$$0 = \frac{\partial}{\partial\theta} 1 = \mathsf{E}_\theta \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta).$$

Далее, поскольку $f(\mathbf{X}, \theta) = \prod f_\theta(X_i)$, то $L(\mathbf{X}, \theta) = \sum \ln f_\theta(X_i)$, и

$$0 = \mathsf{E}_\theta \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta) = \mathsf{E}_\theta \sum \frac{\partial}{\partial\theta} \ln f_\theta(X_i) = n \cdot \mathsf{E}_\theta \frac{\partial}{\partial\theta} \ln f_\theta(X_1). \qquad (10)$$

2. Пусть $T(\mathbf{X}) = \theta^* \in K_0$, т. е. $\mathsf{E}_\theta \theta^* = \theta$. Тогда

$$\frac{\partial}{\partial\theta} \mathsf{E}_\theta \theta^* = \frac{\partial}{\partial\theta} \theta = 1 = \mathsf{E}_\theta \theta^* \cdot \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta). \qquad (11)$$

Вспомним свойство коэффициента корреляции:

$$\operatorname{cov}(\xi, \eta) = \mathsf{E}\,\xi\eta - \mathsf{E}\,\xi \mathsf{E}\,\eta \leqslant \sqrt{\mathsf{D}\,\xi \mathsf{D}\,\eta}.$$

Используя свойства (10) и (11), имеем

$$\operatorname{cov}\left(\theta^*, \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta)\right) = \mathsf{E}_\theta \left(\theta^* \cdot \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta)\right) - \mathsf{E}_\theta \theta^* \mathsf{E}_\theta \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta) =$$

$$= \mathsf{E}_\theta \left(\theta^* \cdot \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta)\right) = 1 \leqslant \sqrt{\mathsf{D}_\theta \theta^* \mathsf{D}_\theta \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta)}. \qquad (12)$$

Найдем $\mathsf{D}_\theta \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta)$:

$$\mathsf{D}_\theta \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta) = \mathsf{D}_\theta \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f_\theta(X_i) = n\mathsf{D}_\theta \frac{\partial}{\partial\theta} \ln f_\theta(X_1) = n\mathsf{E}_\theta \left(\frac{\partial}{\partial\theta} \ln f_\theta(X_1)\right)^2 = nI(\theta).$$

Подставляя дисперсию в неравенство (12), получим

$$1 \leqslant \mathsf{D}_\theta \theta^* \cdot nI(\theta) \quad \text{или} \quad \mathsf{D}_\theta \theta^* \geqslant \frac{1}{nI(\theta)},$$

что и требовалось доказать.

Следующий пример показывает, что условие регулярности является существенным для выполнения равенства, помеченного $(*)$ в лемме 1.

Пример 17 (нерегулярное семейство). Рассмотрим равномерное распределение $U_{0,\theta}$ с параметром $\theta > 0$. Выпишем при $n = 1$ какой-нибудь интеграл и сравним производную от него и интеграл от производной: скажем, для $T(X_1) = 1$

$$\frac{\partial}{\partial\theta} \mathsf{E}_\theta T(X_1) = \frac{\partial}{\partial\theta} \int_0^\theta \frac{1}{\theta}\, dy = \frac{\partial}{\partial\theta} 1 = 0; \qquad \int_0^\theta \frac{\partial}{\partial\theta} \frac{1}{\theta}\, dy = -\frac{1}{\theta} \neq 0.$$

Заметим, что и само утверждение неравенства Рао — Крамера для данного семейства распределений не выполнено: найдется оценка, дисперсия которой ведет себя как $1/n^2$, а не как $1/n$ в неравенстве Рао — Крамера.

Упражнение. Проверить, что в качестве этой «выдающейся» из неравенства Рао — Крамера оценки можно брать, скажем, смещенную оценку $X_{(n)}$ или несмещенную оценку $\dfrac{n+1}{n}X_{(n)} \in K_0$.

4.4. Неравенство Рао — Крамера и эффективность оценок

Сформулируем очевидное следствие из неравенства Рао — Крамера.

Следствие 1. Если семейство распределений $\mathcal{F}_\theta$ удовлетворяет условиям регулярности (R) и (RR), и оценка $\theta^* \in K_{b(\theta)}$ такова, что в неравенстве Рао — Крамера достигается равенство:

$$\mathsf{E}_\theta\,(\theta^*-\theta)^2 = \frac{(1+b'(\theta))^2}{nI(\theta)} + b^2(\theta) \quad \text{или} \quad \mathsf{D}_\theta\,\theta^* = \frac{(1+b'(\theta))^2}{nI(\theta)},$$

то оценка θ^* эффективна в классе $K_{b(\theta)}$.

Оценку, для которой в неравенстве Рао — Крамера достигается равенство, иногда называют R-эффективной оценкой. Следствие 1 можно сформулировать так: если оценка R-эффективна, то она эффективна в соответствующем классе.

Пример 18. Для выборки $X_1, \ldots, X_n$ из распределения Бернулли B_p несмещенная оценка $p^* = \overline{X}$ эффективна, так как для нее достигается равенство в неравенстве Рао — Крамера (см. [3], пример 13.20, с. 67).

Пример 19. Пусть $X_1, \ldots, X_n$ — выборка объема n из нормального распределения N_{a,σ^2}, где $a \in \mathbb{R}$, $\sigma > 0$. Проверим, является ли оценка $a^* = \overline{X} \in K_0$ эффективной (см. также [3], пример 13.6, с. 64).

Найдем информацию Фишера относительно параметра a (считая, что имеется один неизвестный параметр — a).

$$f_{(a,\sigma^2)}(X_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_1-a)^2}{2\sigma^2}\right),$$

$$\ln f_{(a,\sigma^2)}(X_1) = -\ln(2\pi\sigma^2)^{1/2} - \frac{(X_1-a)^2}{2\sigma^2},$$

$$\frac{\partial}{\partial a}\ln f_{(a,\sigma^2)}(X_1) = \frac{(X_1-a)}{\sigma^2},$$

$$I(a) = \mathsf{E}_{(a,\sigma^2)}\left(\frac{\partial}{\partial a}\ln f_{(a,\sigma^2)}(X_1)\right)^2 = \frac{\mathsf{E}_{(a,\sigma^2)}(X_1-a)^2}{\sigma^4} = \frac{\mathsf{D}_{(a,\sigma^2)}X_1}{\sigma^4} = \frac{1}{\sigma^2}.$$

Итак, $I(a) = 1/\sigma^2$. Найдем дисперсию оценки $\overline{X}$.

$$\mathsf{D}_{(a,\sigma^2)}\overline{X} = \frac{1}{n}\mathsf{D}_{(a,\sigma^2)}X_1 = \frac{\sigma^2}{n}.$$

Далее, сравнивая левую и правую части в неравенстве Рао — Крамера, получаем равенство:

$$\mathsf{D}_{(a,\sigma^2)}\overline{X} = \frac{\sigma^2}{n} = \frac{1}{nI(a)}.$$

То есть оценка $a^* = \overline{X}$ эффективна (обладает наименьшей дисперсией среди несмещенных оценок).

Пример 20. Пусть $X_1, \ldots, X_n$ — выборка объема n из нормального распределения N_{0,σ^2}, где $\sigma > 0$. Проверим, является ли оценка $\sigma^{2*} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^2 = \overline{X^2} \in K_0$ эффективной.

Упражнение. Получить эту оценку методом моментов и методом максимального правдоподобия.

Найдем информацию Фишера относительно параметра σ^2.

$$f_{\sigma^2}(X_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{X_1^2}{2\sigma^2}\right),$$

$$\ln f_{\sigma^2}(X_1) = -\ln(2\pi)^{1/2} - \frac{1}{2}\ln\sigma^2 - \frac{X_1^2}{2\sigma^2},$$

$$\frac{\partial}{\partial\sigma^2}\ln f_{\sigma^2}(X_1) = -\frac{1}{2\sigma^2} + \frac{X_1^2}{2\sigma^4},$$

$$I(\sigma^2) = \mathsf{E}_{\sigma^2}\left(\frac{\partial}{\partial\sigma^2}\ln f_{\sigma^2}(X_1)\right)^2 = \mathsf{E}_{\sigma^2}\left(\frac{X_1^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right)^2 =$$

$$= \frac{1}{4\sigma^8}\mathsf{E}_{\sigma^2}(X_1^2 - \sigma^2)^2 = \frac{1}{4\sigma^8}\mathsf{D}_{\sigma^2}X_1^2.$$

Осталось найти $\mathsf{D}_{\sigma^2}X_1^2 = \mathsf{E}_{\sigma^2}X_1^4 - (\mathsf{E}_{\sigma^2}X_1^2)^2 = \mathsf{E}_{\sigma^2}X_1^4 - \sigma^4$. Для тех, кто помнит некоторые формулы вероятности: величина $\xi = X_1/\sigma$ имеет стандартное нормальное распределение, и для нее

$$\mathsf{E}\,\xi^{2k} = (2k-1)!! = (2k-1)(2k-3)\cdot\ldots\cdot 3\cdot 1,$$

Тогда $X_1 = \xi\cdot\sigma$ и

$$\mathsf{E}X_1^4 = \mathsf{E}\,\xi^4\cdot\sigma^4 = 3\sigma^4.$$

Те, кто не помнит, считаем заново:

$$\mathsf{E}_{\sigma^2}X_1^4 = \int\limits_{-\infty}^{\infty} y^4 \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{y^2}{2\sigma^2}}\,dy = 2\sigma^4\int\limits_0^{\infty}\left(\frac{y}{\sigma}\right)^4 \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2\sigma^2}}\,d\left(\frac{y}{\sigma}\right) =$$

$$= 2\sigma^4\int\limits_0^{\infty} t^4\frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}\,dt = -2\sigma^4\frac{1}{\sqrt{2\pi}}\int\limits_0^{\infty} t^3\, de^{-\frac{t^2}{2}} = -2\sigma^4\frac{1}{\sqrt{2\pi}}\left(t^3\, e^{-\frac{t^2}{2}}\Big|_0^{\infty} - \int\limits_0^{\infty} e^{-\frac{t^2}{2}}\,dt^3\right) =$$

$$= 2\sigma^4\frac{1}{\sqrt{2\pi}}\cdot 3\int\limits_0^{\infty} t^2\, e^{-\frac{t^2}{2}}\,dt = 3\sigma^4\int\limits_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}t^2\, e^{-\frac{t^2}{2}}\,dt = 3\sigma^4\cdot\mathsf{D}\,\xi = 3\sigma^4\cdot 1,$$

где ξ имеет стандартное нормальное распределение.

Итак, $\mathsf{D}_{\sigma^2}X_1^2 = \mathsf{E}_{\sigma^2}X_1^4 - \sigma^4 = 2\sigma^4$,

$$I(\sigma^2) = \frac{1}{4\sigma^8}\mathsf{D}_{\sigma^2}X_1^2 = \frac{1}{4\sigma^8}2\sigma^4 = \frac{1}{2\sigma^4}.$$

Найдем дисперсию оценки $\sigma^{2*} = \overline{X^2}$.

$$\mathsf{D}_{\sigma^2}\overline{X^2} = \frac{1}{n^2}\mathsf{D}_{\sigma^2}\sum_1^n X_i^2 = \frac{1}{n}\mathsf{D}_{\sigma^2}X_1^2 = \frac{2\sigma^4}{n}.$$

Сравнивая левую и правую части в неравенстве Рао — Крамера, получаем равенство:

$$\mathsf{D}_{\sigma^2}\overline{X^2} \quad = \quad \frac{2\sigma^4}{n} \quad = \quad \frac{1}{nI(\sigma^2)}.$$

Таким образом, оценка $\sigma^{2*} = \overline{X^2}$ эффективна.

Упражнение. Пусть $X_1, \ldots, X_n$ — выборка объема n из нормального распределения N_{a,σ^2}, где a известно, $\sigma > 0$. Проверить, является ли эффективной оценка $\sigma^{2*} = \dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i - a)^2 = \overline{(X-a)^2}$. Принадлежит ли эта оценка классу K_0? Какими методами получена? Является ли состоятельной и асимптотически нормальной?

Пример 21. Пусть $X_1, \ldots, X_n$ — выборка объема n из показательного распределения $\mathsf{E}_{1/\alpha}$ с параметром $1/\alpha$, где $\alpha > 0$. Проверим, является ли оценка $\alpha^* = \overline{X} \in K_0$ (оценка для параметра α!) эффективной.

Найдем информацию Фишера относительно параметра α

$$I(\alpha) = \mathsf{E}_\alpha\left(\frac{\partial}{\partial\alpha}\ln f_\alpha(X_1)\right)^2.$$

Плотность данного показательного распределения имеет вид:

$$f_\alpha(y) = \begin{cases} \dfrac{1}{\alpha}e^{-\frac{y}{\alpha}}, & \text{если } y > 0, \\ 0, & \text{если } y \leqslant 0. \end{cases}$$

Тогда
$$f_\alpha(X_1) = \frac{1}{\alpha}e^{-\frac{X_1}{\alpha}}, \qquad \ln f_\alpha(X_1) = -\ln\alpha - \frac{X_1}{\alpha},$$
$$\frac{\partial}{\partial\alpha}\ln f_\alpha(X_1) = -\frac{1}{\alpha} + \frac{X_1}{\alpha^2} = \frac{1}{\alpha^2}(X_1 - \alpha),$$
$$I(\alpha) = \mathsf{E}_\alpha\left(\frac{\partial}{\partial p}\ln f_\alpha(X_1)\right)^2 = \frac{\mathsf{E}_\alpha(X_1-\alpha)^2}{\alpha^4} = \frac{\mathsf{D}_\alpha X_1}{\alpha^4} = \frac{\alpha^2}{\alpha^4} = \frac{1}{\alpha^2}.$$

Итак, $I(\alpha) = 1/\alpha^2$. Найдем дисперсию оценки $\overline{X}$.

$$\mathsf{D}_\alpha\overline{X} = \frac{1}{n}\mathsf{D}_\alpha X_1 = \frac{\alpha^2}{n}.$$

Подставив дисперсию и информацию Фишера в неравенство Рао — Крамера, получаем равенство:

$$\mathsf{D}_\alpha\overline{X} \quad = \quad \frac{\alpha^2}{n} \quad = \quad \frac{1}{nI(\alpha)}.$$

То есть оценка $\alpha^* = \overline{X}$ — эффективная оценка параметра α.

Упражнение. Получить эту оценку методом моментов и методом максимального правдоподобия. Она действительно несмещенная? А еще какими свойствами обладает?

Упражнение. Проверьте, что для несмещенной оценки $\alpha^{**} = X_1$ равенство в неравенстве Рао — Крамера не достигается. Объясните, почему, исходя только из этого, нельзя сделать вывод о ее неэффективности в классе K_0. Сделайте этот вывод на основании того, что оценки $\alpha^* = \overline{X}$ и $\alpha^{**} = X_1$ принадлежат классу оценок с одинаковым смещением, и одна из них эффективна. Используйте теорему 5.

Отсутствие равенства в неравенстве Рао — Крамера вовсе не означает неэффективность оценки. Приведем пример оценки, которая является эффективной, но для которой не достигается равенство в неравенстве Рао — Крамера. В эффективности оценки из этого примера мы хотели бы, но не сможем убедиться.

Пример 22. Пусть $X_1, \ldots, X_n$ — выборка объема n из показательного распределения E_α с параметром α, где $\alpha > 0$. Возьмем чуть поправленную оценку метода моментов

$$\alpha^* = \frac{n-1}{n} \cdot \frac{1}{\overline{X}} = \frac{n-1}{\sum_{i=1}^n X_i}.$$

Убедимся, что это — несмещенная оценка. Согласно свойству устойчивости по суммированию для Г-распределения, сумма $\sum\limits_{i=1}^{n} X_i$ случайных величин с распределением $E_\alpha = \Gamma_{\alpha,1}$ имеет распределение $\Gamma_{\alpha,n}$ с плотностью

$$\gamma_{\alpha,n}(y) = \begin{cases} \dfrac{\alpha^n}{\Gamma(n)}\, y^{n-1}\, e^{-\alpha y}, & y > 0, \\ 0, & y \leqslant 0. \end{cases}$$

Напомним, что $\Gamma(n) = (n-1)!$ Вычислим математическое ожидание

$$E_\alpha \alpha^* = E_\alpha\left(\frac{n-1}{\sum X_i}\right) = (n-1)\int\limits_0^\infty \frac{1}{y}\,\frac{\alpha^n}{(n-1)!}\, y^{n-1}\, e^{-\alpha y}\, dy =$$

$$= \frac{(n-1)}{(n-1)!} \cdot \alpha \cdot \int\limits_0^\infty (\alpha y)^{n-2}\, e^{-\alpha y}\, d(\alpha y) = \frac{\alpha}{(n-2)!} \cdot \Gamma(n-1) = \alpha.$$

Итак, оценка α^* принадлежит классу K_0. Найдем информацию Фишера относительно параметра α:

$$f_\alpha(X_1) = \alpha\, e^{-\alpha X_1}, \qquad \ln f_\alpha(X_1) = \ln\alpha - \alpha X_1, \qquad \frac{\partial}{\partial\alpha}\ln f_\alpha(X_1) = \frac{1}{\alpha} - X_1,$$

$$I(\alpha) = E_\alpha\left(\frac{\partial}{\partial p}\ln f_\alpha(X_1)\right)^2 = E_\alpha(X_1 - \alpha)^2 = D_\alpha X_1 = \frac{1}{\alpha^2}.$$

Итак, $I(\alpha) = 1/\alpha^2$. Найдем второй момент и дисперсию оценки α^*.

$$E_\alpha(\alpha^*)^2 = E_\alpha\frac{(n-1)^2}{(\sum X_i)^2} = (n-1)^2\int\limits_0^\infty \frac{1}{y^2}\,\frac{\alpha^n}{(n-1)!}\, y^{n-1}\, e^{-\alpha y}\, dy =$$

$$= \frac{(n-1)^2}{(n-1)!} \cdot \alpha^2 \cdot \int\limits_0^\infty (\alpha y)^{n-3}\, e^{-\alpha y}\, d(\alpha y) = \frac{(n-1)}{(n-2)!} \cdot \alpha^2 \cdot \Gamma(n-2) = \frac{n-1}{n-2}\,\alpha^2.$$

Тогда

$$D_\alpha\alpha^* = E_\alpha(\alpha^*)^2 - (E_\alpha\alpha^*)^2 = \frac{n-1}{n-2}\,\alpha^2 - \alpha^2 = \frac{\alpha^2}{n-2}.$$

Подставив дисперсию и информацию Фишера в неравенство Рао — Крамера, получаем, что при любом n есть строгое неравенство:

$$D_\alpha\alpha^* \quad = \quad \frac{\alpha^2}{n-2} \quad > \quad \frac{\alpha^2}{n} = \frac{1}{nI(\alpha)}.$$

Тем не менее, оценка α^* является эффективной, но доказывать мы это не будем.

4.5. Наилучшие линейные несмещенные оценки

В плане подготовки к курсу «Эконометрика» полезно заметить следующее: в практической статистике часто рассматривают оценки, являющиеся линейными (и по возможности несмещенными) функциями от выборки, то есть оценки вида $\theta^* = \sum\limits_{i=1}^{n} a_i X_i$. В классе таких оценок наилучшая в смысле среднеквадратического подхода оценка обычно находится и без неравенства Рао — Крамера (что особенно полезно для нерегулярных семейств) — достаточно минимизировать $\sum a_i^2$ при заданной $\sum a_i$. Такую оценку принято называть «наилучшей линейной несмещенной оценкой», или, по английски, BLUE (“best linear unbiased estimate”).

Так, скажем, для распределения $U_{0,\theta}$ оценка $\theta_0^* = 2\overline{X}$ является BLUE, так как ее дисперсия найти! или вспомнить пример 11 не больше доказать! дисперсии любой оценки вида $\theta^* = \sum\limits_{i=1}^{n} a_i X_i$, где $\sum\limits_{i=1}^{n} a_i = 2$. почему это гарантирует несмещенность?

Справедливости ради следует добавить (см. пример 11), что оценка $\theta_0^* = 2\overline{X}$, хоть и является BLUE, не может конкурировать в среднеквадратичном смысле с нелинейной оценкой $\hat{\theta} = \frac{n+1}{n} X_{(n)}$ (которая является эффективной в классе несмещенных оценок, но этого мы доказывать не станем).

4.6. Вопросы и упражнения

1. Проверить эффективность ОМП для следующих распределений:
а) B_p, б) Π_λ, в) $N_{a,1}$, г) $B_{m,p}$ (биномиальное), $0 < p < 1$, при известном m.

2. Выполнить все упражнения, содержащиеся в тексте главы 4.

5. Интервальное оценивание

Пусть, как обычно, имеется выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из распределения $\mathcal{F}_\theta$ с неизвестным параметром $\theta \in \Theta \subseteq \mathbb{R}$. До сих пор мы занимались «точечным оцениванием» неизвестного параметра — находили число («оценку»), способную, в некотором смысле, заменить параметр.

Существует другой подход к оцениванию, при котором мы указываем интервал, накрывающий параметр с заданной наперед вероятностью. Такой подход называется «интервальным оцениванием». Сразу заметим: чем больше уверенность в том, что параметр лежит в интервале, тем шире интервал. Так что мечтать найти диапазон, в котором θ лежит с вероятностью 1, бессмысленно — это вся область Θ.

Определение 13. Пусть $0<\varepsilon<1$. Интервал $(\theta^-, \theta^+) = (\theta^-(\mathbf{X}, \varepsilon),\ \theta^+(\mathbf{X}, \varepsilon))$ называется доверительным интервалом для параметра θ уровня доверия $1-\varepsilon$, если для любого $\theta \in \Theta$

$$\mathsf{P}_\theta\left(\theta^- < \theta < \theta^+\right) \geqslant 1-\varepsilon.$$

Определение 14. Пусть $0<\varepsilon<1$. Интервал $(\theta^-, \theta^+) = (\theta^-(\mathbf{X}, \varepsilon),\ \theta^+(\mathbf{X}, \varepsilon))$ называется асимптотическим доверительным интервалом для параметра θ (асимптотического) уровня доверия $1-\varepsilon$, если для любого $\theta \in \Theta$

$$\liminf_{n\to\infty} \mathsf{P}_\theta\left(\theta^- < \theta < \theta^+\right) \geqslant 1-\varepsilon.$$

На самом деле в определении 14 речь идет, конечно, не об одном интервале, но о последовательности интервалов, зависящих от объема выборки n.

Замечание 11. Случайны здесь границы интервала (θ^-, θ^+), поэтому читают формулу $\mathsf{P}_\theta\left(\theta^- < \theta < \theta^+\right)$ как «интервал (θ^-, θ^+) накрывает параметр θ», а не как «θ лежит в интервале...».

Замечание 12. Знак «$\geqslant$» $1-\varepsilon$ обычно соответствует дискретным распределениям, когда нельзя обязаться добиться равенства: например, для $\xi \in B_{1/2}$ при любом x равенство $\mathsf{P}(\xi < x) = 0{,}25$ невозможно, а неравенство имеет смысл:

$$\mathsf{P}(\xi < x) \geqslant 0{,}25 \quad \text{для} \quad x > 0.$$

Если вероятность доверительному интервалу накрывать параметр в точности равна $1-\varepsilon$ (или стремится к $1-\varepsilon$), интервал называют точным (или асимптотически точным) доверительным интервалом уровня доверия $1-\varepsilon$.

Прежде чем рассматривать какие-то регулярные способы построения точных и асимптотических ДИ (доверительных интервалов), разберем два примера, предлагающих очень похожие способы. Далее мы попробуем извлечь из этих примеров некоторую общую философию построения точных и асимптотически точных доверительных интервалов. Начнем с нормального распределения как с наиболее важного и часто встречающегося.

Пример 23. Пусть $X_1, \ldots, X_n$ — выборка объема n из нормального распределения N_{a,σ^2}, где $a \in \mathbb{R}$ — неизвестный параметр, а $\sigma > 0$ известно. Требуется построить точный ДИ для параметра a уровня доверия $1-\varepsilon$.

Вспомним, что нормальное распределение устойчиво по суммированию: доказать бы!

Свойство 6. Пусть ξ_1 имеет нормальное распределение $\mathsf{N}_{a_1,\sigma_1^2}$, ξ_2 имеет нормальное распределение $\mathsf{N}_{a_2,\sigma_2^2}$, и эти случайные величины независимы. Тогда $\eta = b\xi_1 + c\xi_2 + d$ имеет нормальное распределение с параметрами

$$\mathsf{E}\eta = b\,a_1 + c\,a_2 + d, \qquad \mathsf{D}\eta = b^2\sigma_1^2 + c^2\sigma_2^2.$$

Поэтому

$$\sum_1^n X_i \text{ имеет распределение } \mathsf{N}_{na,n\sigma^2},$$

$$\sum_1^n X_i - na \text{ имеет распределение } \mathsf{N}_{0,n\sigma^2},$$

$$\frac{\sum_1^n X_i - na}{\sqrt{n}\sigma} = \sqrt{n}\frac{\overline{X}-a}{\sigma} \text{ имеет распределение } \mathsf{N}_{0,1}.$$

Итак, величина $\eta = \sqrt{n}\dfrac{\overline{X}-a}{\sigma}$ имеет стандартное нормальное распределение. По заданному $\varepsilon \in (0,1)$ найдем число $c > 0$ такое, что $\mathsf{P}(-c < \eta < c) = 1-\varepsilon$.

Число c — квантиль уровня $1-\varepsilon/2$ стандартного нормального распределения:

$$\mathsf{P}\left(-c<\eta<c\right)=\Phi_{0,1}(c)-\Phi_{0,1}(-c)=$$
$$=\Phi_{0,1}(c)-(1-\Phi_{0,1}(c))=2\Phi_{0,1}(c)-1=1-\varepsilon,$$

или $\Phi_{0,1}(c)=1-\frac{\varepsilon}{2}$.

Напоминание:

Определение 15. Пусть распределение $\mathcal{F}$ с функцией распределения F абсолютно непрерывно. Число τ_δ называется квантилью уровня δ распределения $\mathcal{F}$, если $F(\tau_\delta)=\delta$. Если функция F монотонна, квантиль определяется единственным образом.

Итак, $c=\tau_{1-\varepsilon/2}$, или $-c=\tau_{\varepsilon/2}$ (квантили стандартного нормального распределения).

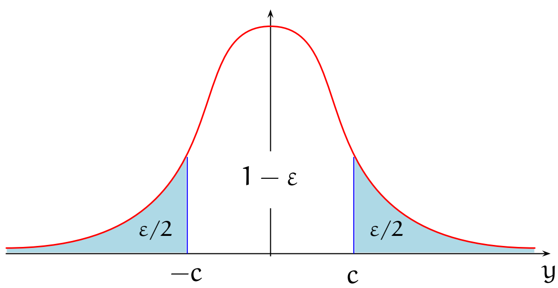


Рис. 7: Плотность стандартного нормального распределения и квантили.

Разрешив неравенство $-c<\eta<c$ относительно a, получим точный доверительный интервал

$$1-\varepsilon=\mathsf{P}_a(-c<\eta<c)=\mathsf{P}_a\left(-c<\sqrt{n}\frac{\overline{X}-a}{\sigma}<c\right)=$$
$$=\mathsf{P}_a\left(\overline{X}-\frac{c\sigma}{\sqrt{n}}<a<\overline{X}+\frac{c\sigma}{\sqrt{n}}\right). \quad (13)$$

Можно подставить $c=\tau_{1-\varepsilon/2}$:

$$\mathsf{P}_a\left(\overline{X}-\frac{\tau_{1-\varepsilon/2}\sigma}{\sqrt{n}}<a<\overline{X}+\frac{\tau_{1-\varepsilon/2}\sigma}{\sqrt{n}}\right)=1-\varepsilon.$$

Итак, искомый точный доверительный интервал уровня доверия $1-\varepsilon$ имеет вид

$$\left(\overline{X}-\frac{\tau_{1-\varepsilon/2}\sigma}{\sqrt{n}},\ \overline{X}+\frac{\tau_{1-\varepsilon/2}\sigma}{\sqrt{n}}\right).$$

Вопросы, на которые стоит себе ответить.

1. Зачем мы брали симметричные квантили? Почему не брать границы для η вида $\mathsf{P}\left(\tau_{\varepsilon/3}<\eta<\tau_{1-2\varepsilon/3}\right)=1-\varepsilon$? Изобразить эти квантили на графике плотности. Как изменилось расстояние между квантилями? Как изменится длина ДИ?

2. Какой из двух ДИ одного уровня доверия и разной длины следует предпочесть?

3. Какова середина полученного в примере 23 ДИ? Какова его длина? Что происходит с границами ДИ при $n\to\infty$?

Пример 24. Пусть $X_1, \ldots, X_n$ — выборка объема n из показательного распределения E_α, где $\alpha > 0$. Требуется построить асимптотический (асимптотически точный) ДИ для параметра α уровня доверия $1-\varepsilon$.

Вспомним ЦПТ:

$$\frac{\sum\limits_1^n X_i - n\,E_\alpha X_1}{\sqrt{n\,D_\alpha X_1}} = \sqrt{n}\,\frac{\overline{X} - 1/\alpha}{1/\alpha} = \sqrt{n}\,\left(\alpha\overline{X} - 1\right) \Rightarrow \eta,$$

где случайная величина η имеет стандартное нормальное распределение. По определению слабой сходимости, при $n \to \infty$

$$P_\alpha\left(-c < \sqrt{n}\,\left(\alpha\overline{X} - 1\right) < c\right) \to P_\alpha(-c < \eta < c) = 1 - \varepsilon \quad \text{при } c = \tau_{1-\varepsilon/2}.$$

То есть

$$P_\alpha\left(-\tau_{1-\varepsilon/2} < \sqrt{n}\,\left(\alpha\overline{X} - 1\right) < \tau_{1-\varepsilon/2}\right) =$$
$$= P_\alpha\left(\frac{1}{\overline{X}} - \frac{\tau_{1-\varepsilon/2}}{\sqrt{n}\,\overline{X}} < \alpha < \frac{1}{\overline{X}} + \frac{\tau_{1-\varepsilon/2}}{\sqrt{n}\,\overline{X}}\right) \to 1 - \varepsilon \quad \text{при } n \to \infty.$$

Итак, искомый асимптотический ДИ уровня доверия $1-\varepsilon$ имеет вид

$$\left(\frac{1}{\overline{X}} - \frac{\tau_{1-\varepsilon/2}}{\sqrt{n}\,\overline{X}},\ \ \frac{1}{\overline{X}} + \frac{\tau_{1-\varepsilon/2}}{\sqrt{n}\,\overline{X}}\right).$$

Сформулируем общий принцип построения точных ДИ:

1. Найти функцию $G(\mathbf{X}, \theta)$, распределение которой $\mathcal{G}$ не зависит от параметра θ. Необходимо, чтобы $G(\mathbf{X}, \theta)$ была обратима по θ при любом фиксированном $\mathbf{X}$.
2. Пусть числа g_1 и g_2 — квантили распределения $\mathcal{G}$ такие, что
$$1 - \varepsilon = P_\theta\left(g_1 < G(\mathbf{X}, \theta) < g_2\right).$$
3. Разрешив неравенство $g_1 < G(\mathbf{X}, \theta) < g_2$ относительно θ (если это возможно), получим точный ДИ.

Совершенно аналогично выглядит общий принцип построения асимптотических ДИ:

1. Найти функцию $G(\mathbf{X}, \theta)$, слабо сходящуюся к распределению $\mathcal{G}$, не зависящему от параметра θ. Необходимо, чтобы $G(\mathbf{X}, \theta)$ была обратима по θ при любом фиксированном $\mathbf{X}$.
2. Пусть g_1 и g_2 — квантили распределения $\mathcal{G}$ такие, что
$$P_\theta\left(g_1 < G(\mathbf{X}, \theta) < g_2\right) \to P_\theta\left(g_1 < \eta < g_2\right) = 1 - \varepsilon.$$
3. Разрешив неравенство $g_1 < G(\mathbf{X}, \theta) < g_2$ относительно θ, получим асимптотический ДИ.

Замечание 13. Часто в качестве g_1 и g_2 берут квантили уровня $\varepsilon/2$ и $1-\varepsilon/2$ распределения $\mathcal{G}$. Но, вообще говоря, квантили следует выбирать так, чтобы получить наиболее короткий ДИ.

Пример 25. Попробуем, пользуясь приведенной выше схемой, построить точный доверительный интервал для параметра $\theta > 0$ равномерного на $[\theta, 2\theta]$ распределения.

Мы знаем, что если X_i имеют распределение $U_{\theta,2\theta}$, то $Y_i = \dfrac{X_i}{\theta} - 1$ имеют распределение $U_{0,1}$. Тогда величина

$$Y_{(n)} = \max\{Y_1, \ldots, Y_n\} = \frac{\max\{X_1, \ldots, X_n\}}{\theta} - 1 = \frac{X_{(n)}}{\theta} - 1 = G(\mathbf{X}, \theta)$$

распределена так же, как максимум из n независимых равномерно распределенных на $[0,\,1]$ случайных величин, то есть имеет не зависящую от параметра θ функцию распределения

$$F_{Y_{(n)}}(y) = \mathsf{P}_\theta(\eta < y) = \begin{cases} 0, & y < 0 \\ y^n, & y \in [0,\,1] \\ 1, & y > 1. \end{cases}$$

Для любых положительных g_1 и g_2

$$\mathsf{P}_\theta\left(g_1 < G(\mathbf{X}, \theta) < g_2\right) = \mathsf{P}_\theta\left(g_1 < \frac{X_{(n)}}{\theta} - 1 < g_2\right) = \\ = \mathsf{P}_\theta\left(\frac{X_{(n)}}{g_2+1} < \theta < \frac{X_{(n)}}{g_1+1}\right). \quad (14)$$

Длина доверительного интервала равна $X_{(n)} \cdot (g_2 - g_1)/((g_1+1)(g_2+1))$ и уменьшается с ростом g_1 и g_2 и с их сближением.

Плотность распределения $Y_{(n)}$ на отрезке $[0,\,1]$ равна ny^{n-1} и монотонно возрастает. Поэтому самые большие значения квантилей g_1 и g_2 при самом маленьком расстоянии между ними и при фиксированной площади под графиком плотности достигается выбором $g_2 = 1$, а g_1 такого, чтобы $1 - \varepsilon = \mathsf{P}_\theta(g_1 < Y_{(n)} < 1)$.

$$\mathsf{P}_\theta(g_1 < Y_{(n)} < 1) = F_{Y_{(n)}}(1) - F_{Y_{(n)}}(g_1) = 1 - g_1^n = 1 - \varepsilon, \quad \text{т. е.} \quad g_1 = \sqrt[n]{\varepsilon}.$$

Подставим найденные квантили в (14):

$$1 - \varepsilon = \mathsf{P}_\theta\left(\sqrt[n]{\varepsilon} < Y_{(n)} < 1\right) = \mathsf{P}_\theta\left(\frac{X_{(n)}}{2} < \theta < \frac{X_{(n)}}{1 + \sqrt[n]{\varepsilon}}\right).$$

Упражнение. Можно ли, пользуясь схемой примера 23, построить точный ДИ для σ при известном a, если разрешить неравенство $-c < \eta < c$ в (13) относительно σ? Можно предположить, например, что $\overline{X} - a > 0$. Чем плох интервал бесконечной длины? А получился ли у Вас интервал бесконечной длины?

Из упражнения видно, что функция G вида $\sqrt{n}\,\dfrac{\overline{X} - a}{\sigma}$ не годится для построения точного ДИ для σ при известном a, а тем более при неизвестном a. В следующей главе мы займемся поиском подходящих функций. Следующий пример (как и пример 24) показывает, что ЦПТ дает универсальный вид функции G для построения асимптотических ДИ.

Пример 26. Пусть $X_1, \ldots, X_n$ — выборка объема n из распределения Пуассона Π_λ, где $\lambda > 0$. Требуется построить асимптотический ДИ для параметра λ уровня доверия $1 - \varepsilon$.

Вспомним ЦПТ:

$$\frac{\sum\limits_{1}^{n} X_i - n\mathsf{E}_\lambda X_1}{\sqrt{n\mathsf{D}_\lambda X_1}} = \sqrt{n}\,\frac{\overline{X} - \lambda}{\sqrt{\lambda}} \Rightarrow \eta,$$

где η имеет стандартное нормальное распределение. По определению слабой сходимости, при $n \to \infty$

$$\mathsf{P}_\lambda\left(-c < \sqrt{n}\,\frac{\overline{X} - \lambda}{\sqrt{\lambda}} < c\right) \to \mathsf{P}_\lambda(-c < \eta < c) = 1 - \varepsilon \text{ при } \ c = \tau_{1-\varepsilon/2}.$$

Но разрешить неравенство под знаком вероятности относительно λ не просто — получается квадратное неравенство из-за корня в знаменателе. Не испортится ли сходимость, если мы заменим $\sqrt{\lambda}$ на $\sqrt{\overline{X}}$?

По свойствам слабой сходимости, если $\xi_n \xrightarrow{p} 1$ и $\eta_n \Rightarrow \eta$, то $\xi_n\eta_n \Rightarrow \eta$. Оценка $\lambda^* = \overline{X}$ состоятельна, поэтому

$$\frac{\lambda}{\overline{X}} \xrightarrow{p} 1.$$

Тогда

$$\sqrt{\frac{\lambda}{\overline{X}}} \cdot \sqrt{n}\,\frac{\overline{X} - \lambda}{\sqrt{\lambda}} = \sqrt{n}\,\frac{\overline{X} - \lambda}{\sqrt{\overline{X}}} \Rightarrow \eta.$$

Поэтому и

$$\mathsf{P}_\lambda\left(-\tau_{1-\varepsilon/2} < \sqrt{n}\,\frac{\overline{X} - \lambda}{\sqrt{\overline{X}}} < \tau_{1-\varepsilon/2}\right) \to \mathsf{P}_\lambda(-\tau_{1-\varepsilon/2} < \eta < \tau_{1-\varepsilon/2}) = 1 - \varepsilon.$$

Разрешая неравенство под знаком вероятности относительно λ, получим

$$\mathsf{P}_\lambda\left(\overline{X} - \frac{\tau_{1-\varepsilon/2}\sqrt{\overline{X}}}{\sqrt{n}} < \lambda < \overline{X} + \frac{\tau_{1-\varepsilon/2}\sqrt{\overline{X}}}{\sqrt{n}}\right) \to 1 - \varepsilon \text{ при } \ n \to \infty.$$

Итак, искомый асимптотический ДИ уровня доверия $1 - \varepsilon$ имеет вид

$$\left(\overline{X} - \frac{\tau_{1-\varepsilon/2}\sqrt{\overline{X}}}{\sqrt{n}}, \ \ \overline{X} + \frac{\tau_{1-\varepsilon/2}\sqrt{\overline{X}}}{\sqrt{n}}\right).$$

Вместо ЦПТ для построения асимптотических ДИ можно использовать асимптотически нормальные оценки (что по сути — та же ЦПТ): если θ^* — АНО для параметра θ с коэффициентом $\sigma^2(\theta)$, то

$$G(\mathbf{X}, \theta) = \sqrt{n}\,\frac{\theta^* - \theta}{\sigma(\theta)} \Rightarrow \eta,$$

где η имеет стандартное нормальное распределение.

Замечание 14. Если $\sigma(\theta)$ в знаменателе мешает, то, как в примере 26, ее можно заменить состоятельной оценкой $\sigma(\theta^*)$. Достаточно, чтобы функция $\sigma(\theta)$ была непрерывной во всей области Θ. Требуется лишь ответить себе: почему θ^* — состоятельная оценка для θ?

6. Распределения, связанные с нормальным

В предыдущей главе мы построили (в числе других) точный доверительный интервал для параметра a нормального распределения при известном σ^2. Для этого мы рассмотрели функцию от выборки и неизвестного параметра a

$$G(\mathbf{X}, a) = \sqrt{n}\,\frac{\overline{X} - a}{\sigma},$$

имеющую при любом a стандартное нормальное распределение.

Остались нерешенными следующие проблемы:

2) построить точный ДИ для σ при известном a,

3) построить точный ДИ для a при неизвестном σ,

4) построить точный ДИ для σ при неизвестном a.

Как мы уже видели, для решения этих задач требуется отыскать функции от выборки и параметров, распределение которых было бы известно. При этом в задаче (3) искомая функция не должна зависеть от неизвестного σ, а в задаче (4) — от a.

Такой особый интерес к нормальному распределению связан, разумеется, с центральной предельной теоремой — по большому счету все в этом мире нормально (или близко к тому).

Займемся поэтому распределениями, связанными с нормальным распределением, их свойствами и свойствами выборок из нормального распределения.

6.1. Гамма-распределение и его свойства

С определением гамма-распределения мы познакомились в курсе теории вероятностей. вспомнить! Нам понадобится свойство устойчивости по суммированию этого распределения, которое до сих пор было доказано только в частном случае — когда независимые слагаемые имеют одно и то же показательное распределение $\mathrm{E}_\alpha = \Gamma_{\alpha,1}$.

Свойство 7. Пусть $\xi_1, \ldots, \xi_n$ независимы, и ξ_i имеет гамма-распределение $\Gamma_{\alpha,\lambda_i}$, $i = 1, \ldots, n$. Тогда

$$S_n = \sum_{i=1}^{n} \xi_i \text{ имеет распределение } \Gamma_{\alpha, \sum_1^n \lambda_i}.$$

Доказательство свойства устойчивости по суммированию (свойства 7).
Воспользуемся свойствами характеристических функций. Характеристическая функция гамма-распределения $\Gamma_{\alpha,\lambda}$ вычислена нами в курсе теории вероятностей и равна

$$\varphi_\xi(t) = \mathrm{E}\, e^{it\xi} = \left(1 - \frac{it}{\alpha}\right)^{-\lambda}.$$

Характеристическая функция суммы независимых случайных величин есть произведение характеристических функций слагаемых:

$$\varphi_{S_n}(t) = \prod_{i=1}^{n} \varphi_{\xi_i}(t) = \prod_{i=1}^{n} \left(1 - \frac{it}{\alpha}\right)^{-\lambda_i} = \left(1 - \frac{it}{\alpha}\right)^{-\sum_1^n \lambda_i}$$

— х.ф. распределения $\Gamma_{\alpha, \sum_1^n \lambda_i}$.

Свойство 8. Если ξ имеет стандартное нормальное распределение, то ξ^2 имеет гамма-распределение $\Gamma_{1/2,1/2}$.

Доказательство. Найдем производную функции распределения величины ξ^2 и убедимся, что она является плотностью распределения. При $y \leqslant 0$

$$F_{\xi^2}(y) = \mathsf{P}\,(\xi^2 < y) = 0, \text{ поэтому и плотность } f_{\xi^2}(y) = 0.$$

При $y > 0$

$$F_{\xi^2}(y) = \mathsf{P}\,(\xi^2 < y) = \mathsf{P}\,(-\sqrt{y} < \xi < \sqrt{y}) = F_\xi(\sqrt{y}) - F_\xi(-\sqrt{y}).$$

Тогда

$$f_{\xi^2}(y) = (F_{\xi^2}(y))' = F'_\xi(\sqrt{y}) \cdot (\sqrt{y})' - F'_\xi(-\sqrt{y}) \cdot (-\sqrt{y})' =$$
$$= (f_\xi(\sqrt{y}) + f_\xi(-\sqrt{y})) \cdot \frac{1}{2\sqrt{y}} = \frac{f_\xi(\sqrt{y})}{\sqrt{y}} = \frac{1}{\sqrt{2\pi y}}\, e^{-y/2}.$$

Но функция $f_{\xi^2}(y)$, равная 0 при $y \leqslant 0$, и равная

$$f_{\xi^2}(y) = \frac{1}{\sqrt{2\pi y}}\, e^{-y/2} = \frac{(1/2)^{1/2}}{\Gamma(1/2)}\, y^{1/2-1}\, e^{-y/2}$$

при $y > 0$, является плотностью гамма-распределения $\Gamma_{1/2,1/2}$.

6.2. Распределение «хи-квадрат» и его свойства

Из свойства 7 и свойства 8 непосредственно следует утверждение:

Следствие 2. Если $\xi_1, \ldots, \xi_k$ независимы и имеют стандартное нормальное распределение, то случайная величина

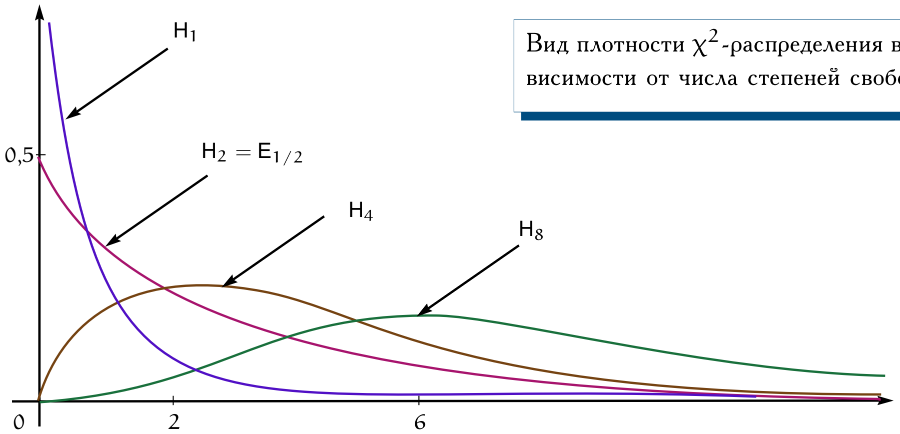
$$\chi_k^2 = \xi_1^2 + \ldots + \xi_k^2$$

имеет гамма-распределение $\Gamma_{1/2,k/2}$.

Определение 16. Распределение суммы k квадратов независимых стандартных нормальных случайных величин называют распределением «хи-квадрат» с k степенями свободы и обозначают H_k. Согласно следствию 2, распределение H_k совпадает с $\Gamma_{1/2,k/2}$.

На графике ниже изображены плотности распределения $H_k = \Gamma_{1/2,k/2}$ при k равном 1, 2, 4 и 8.

Упражнение. Доказать, что при $k \geqslant 2$ максимум плотности распределения H_k достигается в точке $k - 2$.



Вид плотности χ^2-распределения в зависимости от числа степеней свободы

Мы часто будем обозначать через χ_k^2 случайную величину с распределением H_k.

Рассмотрим свойства χ^2-распределения:

1. Устойчивость по суммированию. Пусть случайная величина χ_k^2 имеет распределение H_k, случайная величина χ_m^2 имеет распределение H_m, причем эти случайные величины независимы. Тогда их сумма $\chi_k^2 + \chi_m^2$ имеет распределение H_{k+m}.

Доказательство. Пусть $\xi_1, \xi_2, \ldots$ независимы и имеют стандартное нормальное распределение. Тогда

$$\chi_k^2 \text{ распределено как } \xi_1^2 + \ldots + \xi_k^2, \quad \chi_m^2 \text{ распределено как } \xi_{k+1}^2 + \ldots + \xi_{k+m}^2,$$

а их сумма — как $\xi_1^2 + \ldots + \xi_{k+m}^2$, т. е. имеет распределение H_{k+m}.

2. Моменты распределения χ^2. Если χ_k^2 имеет распределение H_k, то

$$\mathsf{E}\,\chi_k^2 = k \quad \text{и} \quad \mathsf{D}\,\chi_k^2 = 2k.$$

Доказательство. Пусть $\xi_1, \xi_2, \ldots$ независимы и имеют стандартное нормальное распределение. Тогда

$$\mathsf{E}\,\xi_1^2 = 1, \qquad \mathsf{D}\,\xi_1^2 = \mathsf{E}\,\xi_1^4 - (\mathsf{D}\,\xi_1^2)^2 = 3 - 1 = 2$$

(см. пример 20). Поэтому

$$\mathsf{E}\,\chi_k^2 = \mathsf{E}\,(\xi_1^2 + \ldots + \xi_k^2) = k, \quad \mathsf{D}\,\chi_k^2 = \mathsf{D}\,(\xi_1^2 + \ldots + \xi_k^2) = 2k. \quad \square$$

Следствие 3. Если $\xi_1, \ldots, \xi_k$ независимы и имеют нормальное распределение N_{a,σ^2}, то

$$\chi_k^2 = \sum_{i=1}^{k} \left(\frac{\xi_i - a}{\sigma} \right)^2$$

имеет χ^2-распределение H_k с k степенями свободы.

Упражнение. Доказать следствие 3.

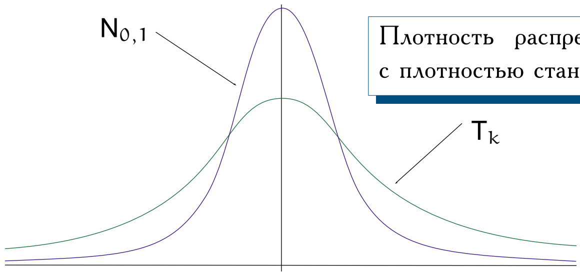
Упражнение. Как, пользуясь таблицей стандартного нормального распределения, найти квантиль заданного уровня для χ^2-распределения с одной степенью свободы?

6.3. Распределение Стьюдента и его свойства

Определение 17. Пусть ξ_0, ξ_1, $\ldots$, ξ_k независимы и имеют стандартное нормальное распределение. Распределение случайной величины

$$t_k = \frac{\xi_0}{\sqrt{\frac{1}{k}(\xi_1^2 + \ldots + \xi_k^2)}} = \frac{\xi_0}{\sqrt{\frac{\chi_k^2}{k}}}$$

называют распределением Стьюдента с k степенями свободы и обозначают T_k.



Плотность распределения Стьюдента по сравнению с плотностью стандартного нормального распределения

Плотность распределения Стьюдента с k степенями свободы равна разглядеть как следует!

$$f_k(y) = \frac{\Gamma((k+1)/2)}{\sqrt{\pi k}\,\Gamma(k/2)} \left(1 + \frac{y^2}{k}\right)^{-(k+1)/2}. \tag{15}$$

Мы не станем выводить эту формулу, предложив читателю-математику либо вывести ее самостоятельно, либо посмотреть вывод в [1, п.6-7 §2 главы 2].

Свойства распределения Стьюдента:

1. Симметричность. Если случайная величина t_k имеет распределение Стьюдента T_k с k степенями свободы, то и $-t_k$ имеет такое же распределение.

Упражнение. Доказать.

2. Асимптотическая нормальность. Распределение Стьюдента T_k слабо сходится к стандартному нормальному распределению при $k \to \infty$.

Доказательство. Пусть ξ_0, ξ_1, $\ldots$, ξ_k независимы и имеют стандартное нормальное распределение. Тогда $\mathsf{E}\,\xi_1^2 = 1$, и по ЗБЧ

$$\frac{\chi_k^2}{k} = \frac{\xi_1^2 + \ldots + \xi_k^2}{k} \xrightarrow{p} 1 \text{ при } k \to \infty.$$

Тогда и

$$t_k = \frac{\xi_0}{\sqrt{\frac{1}{k}(\xi_1^2 + \ldots + \xi_k^2)}} \xrightarrow{p} \xi_0,$$

откуда следует и слабая сходимость последовательности случайных величин t_k с распределением Стьюдента к ξ_0, имеющей стандартное нормальное распределение. То есть $T_k \Rightarrow N_{0,1}$. □

3. Распределение Стьюдента с одной степенью свободы есть стандартное распределение Коши.

Упражнение. Как получить случайную величину с распределением Коши, имея две независимые стандартные нормальные случайные величины?

Доказательство. Подставим $k = 1$ в плотность (15), используя $\Gamma(1/2) = \sqrt{\pi}$ и $\Gamma(1) = 1$, и получим плотность распределения Коши:

$$f_1(y) = \frac{1}{\pi}\left(1 + y^2\right)^{-1}. \quad \square$$

4. У распределения Стьюдента существуют только моменты порядка $m < k$, и не существуют моменты порядка $m \geqslant k$. При этом все существующие моменты нечетного порядка равны нулю.

Упражнение. Посмотреть на плотность (15) и убедиться в сходимости или расходимости на бесконечности при соответствующих m интегралов

$$C(k) \cdot \int_{-\infty}^{\infty} |y|^m \cdot \frac{1}{(k + y^2)^{(k+1)/2}}\, dy.$$

Отметим, что и распределение χ^2, и распределение Стьюдента табулированы, так что если в каких-то доверительных интервалах появятся квантили этих распределений, то мы найдем их по таблице.

Следущее распределение тоже тесно связано с нормальным распределением, но понадобится нам не при построения доверительных интервалов, а чуть позже — в задачах проверки гипотез. Там же мы поймем, почему его называют часто распределением дисперсионного отношения. Призываем математиков сравнить определение [1, п.6 §2 гл.2] с нашим определением и учесть, что в статистических таблицах всегда табулируется распределение Фишера в том виде, как мы его сейчас определим.

$\boxed{\text{что было раньше - курица или яйцо?}}$

6.4. Распределение Фишера

Определение 18. Пусть χ_k^2 имеет распределение H_k, а χ_m^2 — распределение H_m, причем эти случайные величины независимы. Распределение случайной величины

$$f_{k,m} = \frac{\chi_k^2/k}{\chi_m^2/m} = \frac{m \cdot \chi_k^2}{k \cdot \chi_m^2}$$

называют распределением Фишера с k, m степенями свободы и обозначают $F_{k,m}$.

Свойства распределения Фишера (или Фишера — Снедекора):

1. Если $h_{k,m}$ имеет распределение Фишера $F_{k,m}$, то $1/h_{k,m}$ имеет распределение Фишера $F_{m,k}$.

2. Распределение Фишера $F_{k,m}$ слабо сходится к вырожденному в точке 1 распределению $\mathbf{I}_1$ при любом стремлении k и m к бесконечности.

Доказательство. Убедитесь по ЗБЧ, что любая последовательность случайных величин $h_{k,m}$, распределение которой совпадает с распределением отношения двух средних арифметических

$$\frac{\xi_1^2 + \ldots + \xi_k^2}{k} \quad \text{и} \quad \frac{\eta_1^2 + \ldots + \eta_m^2}{m},$$

сходится к 1 по вероятности при $k \to \infty$, $m \to \infty$. Здесь $\xi_1, \xi_2, \ldots$ и $\eta_1, \eta_2, \ldots$ — независимые последовательности, составленные из независимых случайных величин со стандартным нормальным распределением.

6.5. Преобразования нормальных выборок

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка из $N_{0,1}$ (набор независимых и одинаково распределенных величин). Пусть C — ортогональная матрица $(n \times n)$, т.е.

$$CC^T = E = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix},$$

и $\mathbf{Y} = C \cdot \mathbf{X}$ — вектор с координатами $Y_i = \sum_{j=1}^{n} C_{ij} X_j$.

Какое распределение имеют координаты вектора $\mathbf{Y}$? Зависимы ли они? Чтобы ответить на этот вопрос, выясним, как изменится плотность распределения вектора после умножения его на произвольную невырожденную матрицу.

Вспомним, как найти плотность распределения случайной величины $c \cdot \xi$ по плотности распределения ξ:

$$f_{c\xi}(y) = |c|^{-1} \cdot f_\xi\left(c^{-1} \cdot y\right).$$

Поверим, без доказательства, аналогичному утверждению в многомерном случае. Те, кто знаком с заменой переменных в многомерном интеграле и не боится термина «якобиан», могут доказать его самостоятельно.

Изменение плотности совместного распределения при линейном преобразовании вектора. Пусть случайный вектор $\mathbf{X}$ имеет плотность распределения $f_{\mathbf{X}}(y_1, \ldots, y_n) = f_{\mathbf{X}}(\mathbf{y})$, и C — невырожденная матрица. Тогда вектор $\mathbf{Y} = C \cdot \mathbf{X}$ имеет плотность распределения

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{C \cdot \mathbf{X}}(\mathbf{y}) = |\det C|^{-1} \cdot f_{\mathbf{X}}\left(C^{-1} \cdot \mathbf{y}\right). \qquad (16)$$

Докажем самое удивительное свойство нормального распределения.

Свойство 9. Пусть вектор $\mathbf{X}$ состоит из независимых случайных величин со стандартным нормальным распределением, C — ортогональная матрица, и $\mathbf{Y} = C \cdot \mathbf{X}$. Тогда и координаты вектора $\mathbf{Y}$ независимы и имеют стандартное нормальное распределение.

Доказательство. Запишем плотность совместного распределения координат вектора $\mathbf{X}$. В силу независимости это есть произведение плотностей координат вектора (то же самое, что функция правдоподобия):

$$f_{\mathbf{X}}(y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{X_i}(y_i) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_1^n y_i^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\|\mathbf{y}\|^2}.$$

Здесь для произвольного вектора $\mathbf{y}$ квадрат нормы $\|\mathbf{y}\|^2$ есть

$$\|\mathbf{y}\|^2 = \sum_{i=1}^{n} y_i^2 = \mathbf{y}^T \cdot \mathbf{y}.$$

Пользуясь (16), вычислим плотность распределения вектора $\mathbf{Y} = C \cdot \mathbf{X}$. Матрица C ортогональна, поэтому $C^{-1} = C^T$ и $\det C = 1$.

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\left(C^T \cdot \mathbf{y}\right) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\|C^T \cdot \mathbf{y}\|^2}.$$

Но умножение на ортогональную матрицу не меняет норму вектора. Действительно,

$$\|C^T \cdot \mathbf{y}\|^2 = (C^T \cdot \mathbf{y})^T \cdot (C^T \cdot \mathbf{y}) = (\mathbf{y}^T \cdot C) \cdot (C^T \cdot \mathbf{y}) = \mathbf{y}^T \cdot E \cdot \mathbf{y} = \|\mathbf{y}\|^2. \tag{17}$$

Окончательно имеем

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}}\, e^{-\frac{1}{2}\|\mathbf{y}\|^2} = f_{\mathbf{X}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}}\, e^{-\frac{1}{2}\sum_1^n y_i^2}.$$

Итак, вектор $\mathbf{Y}$ распределен так же, как и вектор $\mathbf{X}$, т. е. состоит из независимых случайных величин со стандартным нормальным распределением.

Упражнение. Вспомнить определение независимости случайных величин с абсолютно непрерывным распределением в терминах плотностей. Что можно сказать про независимость и про распределение координат вектора, если совместная плотность распределения координат вектора равна

$$\frac{1}{(2\pi)^{n/2}}\, e^{-\frac{1}{2}\sum_1^n y_i^2} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{y_i^2}{2}} \quad ?$$

Упражнение. Пусть ξ и η независимы и имеют стандартное нормальное распределение. Зависимы ли случайные величины $\frac{1}{\sqrt{2}}(\xi-\eta)$ и $\frac{1}{\sqrt{2}}(\xi+\eta)$? Какое распределение имеют? Является ли ортогональной матрица

$$C = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \quad ?$$

Лемма Фишера. Пусть вектор $\mathbf{X}$ состоит из независимых случайных величин со стандартным нормальным распределением, C — ортогональная матрица, и $\mathbf{Y} = C \cdot \mathbf{X}$. Тогда для любого $k = 1, \ldots, n-1$

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i^2 - Y_1^2 - \ldots - Y_k^2 \quad \text{не зависит от} \quad Y_1, \ldots, Y_k$$

и имеет χ^2-распределение H_{n-k} с $n-k$ степенями свободы.

Доказательство. Как мы видели в (17), нормы векторов $\mathbf{X}$ и $\mathbf{Y} = C \cdot \mathbf{X}$ совпадают:

$$\sum_{i=1}^{n} X_i^2 = \|\mathbf{X}\|^2 = \|C \cdot \mathbf{X}\|^2 = \sum_{i=1}^{n} Y_i^2.$$

Поэтому

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i^2 - Y_1^2 - \ldots - Y_k^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 - \ldots - Y_k^2 = Y_{k+1}^2 + \ldots + Y_n^2.$$

Случайные величины $Y_1, \ldots, Y_n$ по свойству 9 независимы и имеют стандартное нормальное распределение, поэтому $T(\mathbf{X}) = Y_{k+1}^2 + \ldots + Y_n^2$ имеет распределение H_{n-k} и не зависит от $Y_1, \ldots, Y_k$.

Второй и третий пункты следующего утверждения выглядят неправдоподобно, особенно если вспомнить обозначения:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Основное следствие леммы Фишера. Если $X_1, \ldots, X_k$ независимы и имеют нормальное распределение N_{a,σ^2}, то

1) $\sqrt{n}\,\dfrac{\overline{X}-a}{\sigma}$ имеет стандартное нормальное распределение;

2) $\dfrac{(n-1)\,S_0^2}{\sigma^2} = \sum\limits_{i=1}^{n}\dfrac{\left(X_i-\overline{X}\right)^2}{\sigma^2}$ имеет χ^2-распределение с $n-1$ степенью свободы;

3) случайные величины $\overline{X}$ и S_0^2 независимы.

Взгляните: вопреки определению 16 распределения H_{n-1}, величина $(n-1)S_0^2/\sigma^2$ — сумма не $n-1$, а n слагаемых, причем эти слагаемые зависимы из-за присутствия в каждом $\overline{X}$. К тому же они хоть и одинаково распределены, но их распределение — не $N_{0,1}$. а какое? разыскать!

Надеемся, что внимательный читатель, помня, что в невырожденном случае величины ξ и $\xi+\eta$ всегда зависимы, и видя, как в выражении для S_0^2 явным образом участвует $\overline{X}$, придет в неподдельный восторг от независимости $\overline{X}$ и S_0^2.

Отметим без доказательства, что независимость $\overline{X}$ и S_0^2 — свойство, характерное только для нормального распределения. Точно так же как и способность сохранять независимость координат после умножения на ортогональную матрицу. Подробнее об этом можно прочесть в замечательной книге В. Феллера [2, т. 2, гл. III, § 4].

Доказательство основного следствия леммы Фишера.

1. Очевидно. доказать, что очевидно!

2. Покажем сначала, что можно рассматривать выборку из стандартного нормального распределения вместо N_{a,σ^2}:

$$\frac{(n-1)S_0^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 = \sum_{i=1}^{n}\left(\frac{X_i-a}{\sigma} - \frac{\overline{X}-a}{\sigma}\right)^2 = \sum_{i=1}^{n}(z_i-\overline{z})^2,$$

где $z_i = \dfrac{X_i-a}{\sigma}$ имеют стандартное нормальное распределение, и $\overline{z} = \dfrac{\overline{X}-a}{\sigma}$.

Т. е. можно с самого начала считать, что X_i имеют стандартное нормальное распределение, и доказывать пункт 2 следствия при $a=0$, $\sigma^2=1$.

Применим лемму Фишера.

$$T(\mathbf{X}) = (n-1)S_0^2 = \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 = \sum_{i=1}^{n}X_i^2 - n(\overline{X})^2 = \sum_{i=1}^{n}X_i^2 - Y_1^2.$$

Мы обозначили через Y_1 величину

$$Y_1 = \sqrt{n}\,\overline{X} = \frac{X_1}{\sqrt{n}} + \ldots + \frac{X_n}{\sqrt{n}}.$$

Чтобы применить лемму Фишера, нужно найти ортогональную матрицу C такую, что Y_1 — первая координата вектора $\mathbf{Y} = C\cdot\mathbf{X}$. Возьмем матрицу C с первой строкой $\left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$. Так как длина (норма) этого вектора — единица, его можно дополнить до ортонормального базиса в $\mathbb{R}^n$ (иначе — этот столбец можно дополнить до ортогональной матрицы). Тогда $Y_1 = \sqrt{n}\,\overline{X}$ и будет первой координатой вектора $\mathbf{Y} = C\cdot\mathbf{X}$.

Осталось применить лемму Фишера. $\boxed{\text{непременно сделайте это!}}$

3. Из леммы Фишера сразу следует, что $T(\mathbf{X}) = (n-1)S_0^2 = \sum_{i=1}^n X_i^2 - Y_1^2$ не зависит от $Y_1 = \sqrt{n}\,\overline{X}$, то есть $\overline{X}$ и S_0^2 независимы.

Очередное следствие из леммы Фишера наконец позволит нам строить доверительные интервалы для параметров нормального распределения — то, ради чего мы и доказали уже так много утверждений. В каждом пункте указано, для какого параметра мы построим доверительный интервал с помощью данного утверждения.

Следствие основного следствия леммы Фишера.

1) $\sqrt{n}\dfrac{\overline{X}-a}{\sigma}$ имеет стандартное нормальное распределение (для a при σ известном);

2) $\displaystyle\sum_{i=1}^n \left(\frac{X_i-a}{\sigma}\right)^2$ имеет распределение H_n (для σ^2 при a известном);

3) $\displaystyle\sum_{i=1}^n \left(\frac{X_i-\overline{X}}{\sigma}\right)^2 = \frac{(n-1)S_0^2}{\sigma^2}$ имеет распределение H_{n-1} (для σ^2 при a неизвестном);

4) $\sqrt{n}\dfrac{\overline{X}-a}{\sqrt{S_0^2}} = \sqrt{n}\dfrac{\overline{X}-a}{S_0}$ имеет распределение T_{n-1} (для a при σ неизвестном).

Доказательство следствия.

1) и 3) — из леммы Фишера, 2) — из следствия 3. Осталось воспользоваться леммой Фишера и определением 17 распределения Стьюдента, чтобы доказать 4).

$$\sqrt{n}\frac{\overline{X}-a}{\sqrt{S_0^2}} = \underbrace{\sqrt{n}\frac{\overline{X}-a}{\sigma}}_{N_{0,1}} \quad\times\quad \frac{1}{\sqrt{\underbrace{\frac{(n-1)S_0^2}{\sigma^2}}_{H_{n-1}}\frac{1}{n-1}}} = \frac{\xi_0}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}.$$

$\nwarrow$ независ. $\nearrow$

По определению 17, такое отношение имеет распределение Стьюдента T_{n-1}.

6.6. Точные ДИ для параметров нормального распределения

1. Для a при известном σ^2. Этот интервал мы уже построили в примере 23:

$$\mathsf{P}_{a,\sigma^2}\left(\overline{X}-\frac{\tau_{1-\varepsilon/2}\sigma}{\sqrt{n}}<a<\overline{X}+\frac{\tau_{1-\varepsilon/2}\sigma}{\sqrt{n}}\right)=1-\varepsilon,\ \text{где}\ \Phi_{0,1}(\tau_{1-\varepsilon/2})=1-\varepsilon/2.$$

2. Для σ^2 при известном a. По п. 2 следствия,

$$\frac{nS_1^2}{\sigma^2}\quad\text{имеет распределение}\quad \mathsf{H}_n,\quad\text{где}\quad S_1^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-a)^2.$$

Пусть $g_1=\chi^2_{n,\varepsilon/2}$ и $g_2=\chi^2_{n,1-\varepsilon/2}$ — квантили распределения H_n уровня $\varepsilon/2$ и $1-\varepsilon/2$. Тогда

$$1-\varepsilon=\mathsf{P}_{a,\sigma^2}\left(g_1<\frac{nS_1^2}{\sigma^2}<g_2\right)=\mathsf{P}_{a,\sigma^2}\left(\frac{nS_1^2}{g_2}<\sigma^2<\frac{nS_1^2}{g_1}\right).$$

3. Для σ^2 при неизвестном a. По п. 3 следствия,

$$\frac{(n-1)S_0^2}{\sigma^2}\quad\text{имеет распределение}\quad \mathsf{H}_{n-1},\quad\text{где}\quad S_0^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

Пусть $g_1=\chi^2_{n-1,\varepsilon/2}$ и $g_2=\chi^2_{n-1,1-\varepsilon/2}$ — квантили распределения H_{n-1} уровня $\varepsilon/2$ и $1-\varepsilon/2$. Тогда

$$1-\varepsilon=\mathsf{P}_{a,\sigma^2}\left(g_1<\frac{(n-1)S_0^2}{\sigma^2}<g_2\right)=\mathsf{P}_{a,\sigma^2}\left(\frac{(n-1)S_0^2}{g_2}<\sigma^2<\frac{(n-1)S_0^2}{g_1}\right).$$

Упражнение. Найти 17 отличий п. 3 от п. 2.

4. Для a при неизвестном σ. По п. 4 следствия,

$$\sqrt{n}\frac{\overline{X}-a}{S_0}\quad\text{имеет распределение}\quad \mathsf{T}_{n-1},\quad\text{где}\quad S_0^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2.$$

Пусть $a_1=t_{n-1,\varepsilon/2}$ и $a_2=t_{n-1,1-\varepsilon/2}$ — квантили распределения T_{n-1} уровня $\varepsilon/2$ и $1-\varepsilon/2$. Распределение Стьюдента симметрично, поэтому $a_1=-a_2$. Тогда

$$1-\varepsilon=\mathsf{P}_{a,\sigma^2}\left(-t_{n-1,1-\varepsilon/2}<\sqrt{n}\frac{\overline{X}-a}{S_0}<t_{n-1,1-\varepsilon/2}\right)=$$
$$=\mathsf{P}_{a,\sigma^2}\left(\overline{X}-\frac{t_{n-1,1-\varepsilon/2}S_0}{\sqrt{n}}<a<\overline{X}+\frac{t_{n-1,1-\varepsilon/2}S_0}{\sqrt{n}}\right).$$

Упражнение. Сравнить п. 4 с п. 1.

Замечание 15. ДИ, полученные в п. 2 и 3, выглядят странно по сравнению с ДИ из п. 1 и 4: они содержат n в числителе, а не в знаменателе. Но если квантили нормального распределения от n не зависят вовсе, квантили распределения Стьюдента асимптотически не зависят от n по свойству $\mathsf{T}_n \Rightarrow \mathsf{N}_{0,1}$, то квантили распределения H_n зависят от n существенно. Действительно, пусть $g=g(n)$ таково, что $\mathsf{P}(\chi_n^2<g)=\delta$ при всех n, в том числе и при $n\to\infty$. Тогда последовательность $g(n)$ такова, что $\dfrac{g-n}{\sqrt{2n}}\to\tau_\delta$ при $n\to\infty$, где τ_δ — квантиль стандартного нормального распределения. В самом деле: по ЦПТ с ростом n

$$\mathsf{P}(\chi_n^2<g)=\mathsf{P}\left(\sum_{i=1}^{n}\xi_i^2<g\right)=\mathsf{P}\left(\frac{\chi_n^2-n}{\sqrt{2n}}<\frac{g-n}{\sqrt{2n}}\right)\to\Phi_{0,1}(\tau_\delta)=\delta.$$

Поэтому квантиль уровня δ распределения H_n ведет себя как $g=g(n)=n+\tau_\delta\sqrt{2n}+o(\sqrt{n})$.

Упражнение. Подставить в границы ДИ из п. 2 и 3 асимптотические выражения для квантилей и выяснить, как ведет себя длина этих ДИ с ростом n.

6.7. Вопросы и упражнения

1. Величины ξ_1 и ξ_2 независимы и имеют нормальное распределение $N(0, 16)$. Найти k, при котором величины $\xi_1 - 3\xi_2$ и $k\xi_1 + \xi_2$ независимы. Можно использовать свойство 9.

2. Доказать, что для величины χ^2_n с распределением H_n справедлива аппроксимация Фишера:

$$\sqrt{2\chi^2_n} - \sqrt{2n} \Rightarrow N_{0,1},$$

и вывести отсюда, что при больших n для вычисления квантилей распределения H_n можно пользоваться аппроксимацией

$$H_n(x) = P\left(\chi^2_n < x\right) = P\left(\sqrt{2\chi^2_n} - \sqrt{2n} < \sqrt{2x} - \sqrt{2n}\right) \approx \Phi_{0,1}\left(\sqrt{2x} - \sqrt{2n}\right).$$

Указание. Домножить и поделить на сопряженное и воспользоваться ЦПТ и ЗБЧ вместе со свойствами произведения слабо сходящейся и сходящейся по вероятности к постоянной последовательностей.

3. Изобразить квантили уровня $\varepsilon/2$ и $1-\varepsilon/2$ на графике плотности распределения H_n и T_{n-1}.

4. Вычислить, зная распределение $(n-1)S^2_0/\sigma^2$ и пользуясь известным математическим ожиданием и дисперсией распределения χ^2, математическое ожидание и дисперсию длины ДИ для дисперсии нормального распределения при неизвестном среднем.

7. Проверка гипотез

Если возможно выдвинуть несколько взаимоисключающих «гипотез» о распределении элементов выборки, то возникает задача выбора одной из этих гипотез на основании выборочных данных. Как правило, по выборке конечного объема безошибочных выводов о распределении сделано быть не может, поэтому приходится считаться с возможностью выбрать неверную гипотезу.

Пусть дана выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из распределения $\mathcal{F}$. Если не оговорено противное, считается, что все наблюдения имеют одно и то же распределение. В ряде случаев это предположение также нуждается в проверке (см., например, ниже: гипотеза об однородности или гипотеза о случайности) — в таких случаях одинаковая распределенность наблюдений не предполагается. То же касается и независимости наблюдений.

Определение 19. Гипотезой (H) называется любое предположение о распределении наблюдений:

$$H = \{\mathcal{F} = \mathcal{F}_1\} \qquad \text{или} \qquad H = \{\mathcal{F} \in \{\hat{\mathcal{F}}\}\}.$$

Гипотеза H называется простой, если она однозначно определяет распределение, т. е. $H = \{\mathcal{F} = \mathcal{F}_1\}$. Иначе H называется сложной гипотезой. Сложная гипотеза предполагает, что распределение $\mathcal{F}$ — одно из некоторого множества распределений $\{\hat{\mathcal{F}}\}$.

Если гипотез всего две, то одну из них принято называть основной, а другую — альтернативой или отклонением от основной гипотезы.

Пример 27 (типичные постановки задач).

1. Выбор из нескольких простых гипотез: $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$, ..., $H_k = \{\mathcal{F} = \mathcal{F}_k\}$ (и другие предположения невозможны).

2. Простая основная гипотеза и сложная альтернатива: $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$, $H_2 = \{\mathcal{F} \neq \mathcal{F}_1\}$. Например, $H_1 = \{\mathcal{F} = U_{0,1}\}$, $H_2 = \{\mathcal{F} \neq U_{0,1}\}$.
Еще вариант: дана выборка из семейства распределений B_p, где $0 < p \leqslant 1/2$. Простая гипотеза $H_1 = \{p = 1/2\}$. Сложная односторонняя альтернатива $H_2 = \{p < 1/2\}$. Случай $p > 1/2$ исключен априори.

3. Сложная основная гипотеза и сложная альтернатива: $H_1 = \{\mathcal{F} \in \{\hat{\mathcal{F}}\}\}$, $H_2 = \{\mathcal{F} \notin \{\hat{\mathcal{F}}\}\}$. Например, гипотеза о нормальности $H_1 = \{\mathcal{F} \in \{N_{a,\sigma^2}, a \in \mathbb{R}, \sigma > 0\}\}$, $H_2 = \{H_1 \text{ неверна}\}$.

4. Гипотеза однородности. Заданы несколько выборок:
$(X_{11}, \ldots, X_{1n_1})$ из распределения $\mathcal{F}_1$, ..., $(X_{k1}, \ldots, X_{kn_k})$ из распределения $\mathcal{F}_k$. Проверяется гипотеза $H_1 = \{\mathcal{F}_1 = \cdots = \mathcal{F}_k\}$ — сложная гипотеза — против (сложной) альтернативы $H_2 = \{H_1 \text{ неверна}\}$.

5. Гипотеза независимости. Наблюдается пара случайных величин (ξ, η).
По выборке $\big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$ из n независимых наблюдений над парой (ξ, η) проверяется гипотеза $H_1 = \{\xi \text{ и } \eta \text{ независимы}\}$ — сложная гипотеза — против (сложной) альтернативы $H_2 = \{H_1 \text{ неверна}\}$.

6. Гипотеза случайности. В эксперименте наблюдаются n случайных величин $(\xi_1, \ldots, \xi_n)$. По выборке X_1, ..., X_n, в которой каждая случайная величина представлена одним значением, проверяется гипотеза $H_1 = \{\xi_1, \ldots, \xi_n \text{ независимы и одинаково распределены}\}$ — сложная гипотеза — против (сложной) альтернативы $H_2 = \{H_1 \text{ неверна}\}$.

   Эту задачу ставят, например, если требуется проверить качество датчика случайных чисел.

Определение 20. Если имеются гипотезы H_1, ..., H_k, то критерием (нерандомизированным критерием) $\delta = \delta(X_1, \ldots, X_n)$ называется отображение

$$\delta : \mathbb{R}^n \to \{H_1, \ldots, H_k\}.$$

О рандомизированных критериях, которые предписывают принимать каждую гипотезу с некоторой (зависящей от выборки) вероятностью, мы поговорим позднее.

Определение 21. Для заданного критерия $\delta : \mathbb{R}^n \to \{H_1, \ldots, H_k\}$ будем говорить, что произошла ошибка i-го рода, если гипотеза H_i отвергнута критерием, в то время как она верна. Вероятностью ошибки i-го рода критерия δ называется

$$\alpha_i(\delta) = \mathsf{P}_{H_i}(\delta(\mathbf{X}) \neq H_i).$$

Замечание 16. Говоря «H_i верна» и вычисляя $\mathsf{P}_{H_i}(\cdot)$, мы имеем в виду, что распределение выборки именно такое, как предполагает гипотеза H_i, и вычисляем вероятность в соответствии с этим распределением. Если гипотеза H_i простая, т. е. указывает ровно на одно возможное распределение выборки, то $\alpha_i(\delta)$ — число. Если же H_i — сложная гипотеза, то $\alpha_i(\delta)$ будет зависеть от того, при каком именно из распределений $\mathcal{F}_i$, отвечающих H_i, вычисляется вероятность:

$$\alpha_i(\delta) = \alpha_i(\delta, \mathcal{F}_i) = \mathsf{P}_{\mathcal{F}_i}(\delta(\mathbf{X}) \neq H_i)$$

Пример 28 (контроль качества и ошибки).

Пусть любое изделие некоторого производства оказывается браком с вероятностью p. Контроль продукции допускает ошибки: годное изделие бракует с вероятностью γ, а бракованное пропускает (признает годным) с вероятностью ε.

Если ввести для наудачу взятого изделия две гипотезы: $H_1 = \{\text{изделие годное}\}$ и $H_2 = \{\text{изделие бракованное}\}$, и критерием выбора считать контроль продукции, то γ есть вероятность ошибки первого рода, а ε — вероятность ошибки второго рода данного критерия:

$$\gamma = \mathsf{P}_{H_1}(\delta = H_2) = \mathsf{P}_{\text{изделие годное}}(\text{контроль забраковал изделие});$$

$$\varepsilon = \mathsf{P}_{H_2}(\delta = H_1) = \mathsf{P}_{\text{изделие бракованное}}(\text{контроль пропустил изделие});$$

Упражнение. Вычислить вероятности ошибок первого и второго рода того же критерия, если гипотезы занумеровать иначе: $H_1 = \{\text{изделие бракованное}\}$ и $H_2 = \{\text{изделие годное}\}$.

Надеемся, что читатель сделал для себя следующие выводы:

1. Статистический критерий не отвечает на вопрос, верна или нет проверяемая гипотеза. Он лишь решает, противоречат или не противоречат выдвинутой гипотезе выборочные данные, можно ли принять или следует отвергнуть данную гипотезу.
2. Если есть одна основная гипотеза, а все остальное — нежелательные отклонения от нее, то вывод «данные противоречат гипотезе» всегда весомее, нежели вывод «данные не противоречат гипотезе».
3. Нам неизвестно, какая из гипотез верна в действительности, поэтому следует считаться с гипотетическими вероятностями ошибок критерия. Если много раз применять критерий к выборкам из распределения, для которого гипотеза H_i верна, то примерно доля α_i таких выборок будет признана противоречащей гипотезе H_i.

7.1. Две простые гипотезы

Рассмотрим подробно случай, когда имеются две простые гипотезы о распределении наблюдений:

$$H_1 = \{\mathcal{F} = \mathcal{F}_1\}, \quad H_2 = \{\mathcal{F} = \mathcal{F}_2\}.$$

Каков бы ни был критерий $\delta(\mathbf{X}) : \mathbb{R}^n \to \{H_1, H_2\}$, он принимает не более двух значений. То есть область $\mathbb{R}^n$ делится на две части $\mathbb{R}^n = S \cup (\mathbb{R}^n \backslash S)$ так, что

$$\delta(\mathbf{X}) = \begin{cases} H_1, & \text{если} \quad \mathbf{X} \in \mathbb{R}^n \backslash S, \\ H_2, & \text{если} \quad \mathbf{X} \in S. \end{cases}$$

Область S, в которой принимается вторая (альтернативная) гипотеза, называют критической областью.

Определение 22. Вероятность ошибки первого рода $\alpha_1 = \alpha_1(\delta)$ называют размером или уровнем значимости критерия δ:

$$\alpha_1 = \alpha_1(\delta) = \mathsf{P}_{H_1}(\delta(\mathbf{X}) \neq H_1) = \mathsf{P}_{H_1}(\delta(\mathbf{X}) = H_2) = \mathsf{P}_{H_1}(\mathbf{X} \in S).$$

Мощностью критерия δ называют величину $1 - \alpha_2$, где $\alpha_2 = \alpha_2(\delta)$ — вероятность ошибки второго рода критерия δ. Мощность критерия равна

$$1 - \alpha_2(\delta) = 1 - \mathsf{P}_{H_2}(\delta(\mathbf{X}) \neq H_2) = \mathsf{P}_{H_2}(\delta(\mathbf{X}) = H_2) = \mathsf{P}_{H_2}(\mathbf{X} \in S).$$

Заметим, что вероятности ошибок первого и второго рода вычисляются при разных предположениях о распределении (если верна H_1 и если верна H_2), так что никаких раз и навсегда фиксированных соотношений (например, $\alpha_1 = 1 - \alpha_2$, независимо от вида гипотез и вида критерия) между ними нет.

Рассмотрим крайний случай, когда критерий, независимо от выборки, всегда принимает одну и ту же гипотезу.

Пример 29. Имеются гипотезы H_1, H_2 и критерий $\delta(\mathbf{X}) \equiv H_1$, то есть S пусто. Тогда $\alpha_1 = \mathsf{P}_{H_1}(\text{принять}\ \ H_2) = 0$, $\alpha_2 = \mathsf{P}_{H_2}(\text{принять}\ \ H_1) = 1$.

Наоборот: пусть имеются гипотезы H_1, H_2 и критерий $\delta(\mathbf{X}) \equiv H_2$, то есть $S = \mathbb{R}^n$. Тогда $\alpha_1 = \mathsf{P}_{H_1}(\text{принять}\ \ H_2) = 1$, $\alpha_2 = \mathsf{P}_{H_2}(\text{принять}\ \ H_1) = 0$.

Примеры 29 и 30 показывают общую тенденцию: при попытке уменьшить одну из вероятностей ошибок критерия другая, как правило, увеличивается. Так, если уменьшать $\alpha_1 = \mathsf{P}_{H_1}(\mathbf{X} \in S)$ мы будем, сужая критическую область S, то одновременно будет уменьшаться мощность критерия $1 - \alpha_2 = \mathsf{P}_{H_2}(\mathbf{X} \in S)$.

Пример 30. Имеется выборка объема 1 из нормального распределения $\mathsf{N}_{a,1}$ и две простые гипотезы о среднем: $H_1 = \{a = 0\}$ и $H_2 = \{a = 1\}$. Рассмотрим следующий критерий (при некотором вещественном c):

$$\delta(X_1) = \begin{cases} H_1, & \text{если} \quad X_1 \leqslant c, \\ H_2, & \text{если} \quad X_1 > c. \end{cases}$$

Изобразим на графике соответствующие гипотезам плотности и вероятности ошибок первого и второго рода критерия δ:

$$\alpha_1 = \mathsf{P}_{H_1}(\delta(X_1) = H_2) = \mathsf{P}_{H_1}(X_1 > c), \quad \alpha_2 = \mathsf{P}_{H_2}(\delta(X_1) = H_1) = \mathsf{P}_{H_2}(X_1 \leqslant c).$$

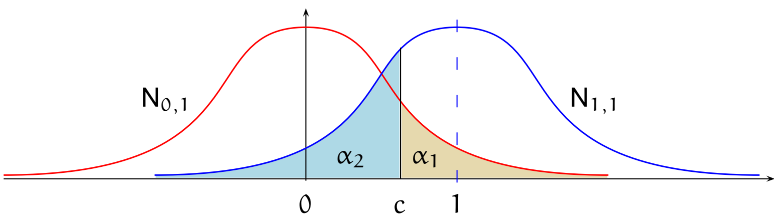


Рис. 8: Две простые гипотезы.

Видим, что с ростом c вероятность ошибки первого рода α_1 уменьшается, но вероятность ошибки второго рода α_2 растет.

Итак, критерий тем лучше, чем меньше вероятности его ошибок. Но сравнивать критерии по паре вероятностей ошибок

$$\alpha_i(\delta_1) \leqslant \alpha_i(\delta_2) \quad \text{при} \quad i = 1, 2,$$

как правило, не удается.

7.2. Подходы к сравнению критериев

Ограничимся, для простоты, задачей проверки двух простых гипотез. Пусть имеются критерии δ и ρ с вероятностями ошибок первого и второго рода $\alpha_1(\delta)$, $\alpha_2(\delta)$ и $\alpha_1(\rho)$, $\alpha_2(\rho)$.

Перечислим общепринятые подходы к сравнению критериев.

1. Минимаксный подход.

Говорят, что критерий δ не хуже критерия ρ в смысле минимаксного подхода, если

$$\max\{\alpha_1(\delta), \alpha_2(\delta)\} \leqslant \max\{\alpha_1(\rho), \alpha_2(\rho)\}.$$

Определение 23. Критерий δ называют минимаксным критерием, если он не хуже всех других критериев в смысле минимаксного подхода.

Иначе говоря, минимаксный критерий имеет самую маленькую «наибольшую ошибку» $\max\{\alpha_1(\delta), \alpha_2(\delta)\}$ среди всех прочих критериев.

Упражнение. Убедиться, что в примере 30 критерий δ является минимаксным, если $c = 1/2$.

2. Байесовский подход.

Этот подход применяют в двух случаях:

а) если известно априори, что с вероятностью r справедлива гипотеза H_1, а с вероятностью $s = 1 - r$ — гипотеза H_2,

б) если задана линейная «функция потерь»: потери от ошибочного решения равны r, если происходит ошибка первого рода, и равны s, если второго. Здесь $r + s$ уже не обязательно равно 1, но потери можно свести к единице нормировкой $r' = r/(r+s)$ и $s' = s/(r+s)$.

Пусть априорные вероятности или потери r и s заданы. Говорят, что критерий δ не хуже критерия ρ в смысле байесовского подхода, если

$$r\alpha_1(\delta) + s\alpha_2(\delta) \leqslant r\alpha_1(\rho) + s\alpha_2(\rho).$$

Определение 24. Критерий δ называют байесовским критерием, если он не хуже всех других критериев в смысле байесовского подхода.

Иначе говоря, байесовский критерий имеет самую маленькую «средневзвешенную ошибку» $r\alpha_1(\delta) + s\alpha_2(\delta)$ среди всех прочих критериев. По формуле полной вероятности это есть вероятность ошибки критерия в случае (а) или математическое ожидание потерь в случае (б).

Упражнение. Убедиться, что в примере 30 критерий δ является байесовским в случае $r = s$, если взять $c = 1/2$.

3. Выбор наиболее мощного критерия.

Ошибки первого и второго рода обычно неравноправны. Поэтому возникает желание контролировать одну из ошибок (скажем, первого рода). Например, зафиксировать ее вероятность на достаточно низком (безопасном) уровне, и рассматривать только критерии с такой или еще меньшей вероятностью ошибки первого рода. Среди них наилучшим, очевидно, следует признать критерий, обладающий наименьшей вероятностью ошибки второго рода.

Введем при $\varepsilon \in [0, 1]$ класс критериев $K_\varepsilon = \{\delta(\mathbf{X}) : \alpha_1(\delta) \leqslant \varepsilon\}$.

Определение 25. Критерий $\delta_0 \in K_\varepsilon$ называют наиболее мощным критерием (НМК) размера ε, если

$$\alpha_2(\delta_0) \leqslant \alpha_2(\delta)$$

для любого другого критерия $\delta \in K_\varepsilon$.

7.3. Построение оптимальных критериев

Следующее замечательное утверждение, по недоразумению называемое леммой, заявляет, что оптимальные во всех трех смыслах (минимаксные, байесовские, наиболее мощные) критерии могут быть построены в самом общем случае простым выбором различных констант в одном и том же критерии — критерии отношения правдоподобия.

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — выборка (набор независимых, одинаково распределенных величин), и имеются две гипотезы о распределении X_i:

$$H_1 = \{X_i \text{ имеют распределение } \mathcal{F}_1\} \quad \text{и} \quad H_2 = \{X_i \text{ имеют распределение } \mathcal{F}_2\}$$

Пусть $f_1(y)$ — плотность распределения $\mathcal{F}_1$, $f_2(y)$ — плотность распределения $\mathcal{F}_2$, а

$$f_1(\mathbf{X}) = f_1(X_1, \ldots, X_n) = \prod_{i=1}^{n} f_1(X_i) \quad \text{и} \quad f_2(\mathbf{X}) = f_2(X_1, \ldots, X_n) = \prod_{i=1}^{n} f_2(X_i)$$

— соответствующие функции правдоподобия.

Предполагается, что распределения $\mathcal{F}_1$ и $\mathcal{F}_2$ либо оба дискретны, либо оба абсолютно непрерывны.

Замечание 17. Если одно из распределений дискретно, а другое абсолютно непрерывно, то всегда существует критерий с нулевыми вероятностями ошибок. Смешанные распределения мы рассматривать не будем. Математики вместо этого могут предполагать, что оба распределения абсолютно непрерывны относительно одной и той же σ-конечной меры и имеют относительно нее плотности $f_1(y)$ и $f_2(y)$.

Мы будем выбирать гипотезу в зависимости от отношения функций правдоподобия. Напомню, что функция правдоподобия есть плотность распределения выборки.

Обратимся к примеру 30. Естественным кажется принимать вторую гипотезу, если X_1 лежит правее точки пересечения плотностей $c = 1/2$. То есть там, где вторая плотность больше, принимать вторую гипотезу, там, где первая — первую. Такой критерий сравнивает отношение $f_2(x_1, \ldots, x_n)/f_1(x_1, \ldots, x_n)$ с единицей, относя к критической области ту часть $\mathbb{R}^n$, где это отношение больше единицы. Заметим, что при этом мы получим ровно один, не обязательно оптимальный, критерий с некоторым фиксированным размером и мощностью.

Если же нужно получить критерий с заранее заданным размером $\alpha_1 = \varepsilon$, либо иметь возможность варьировать и размер, и мощность критерия, то следует рассмотреть класс похожих критериев, введя свободный параметр: там, где вторая плотность в c раз превосходит первую, выбирать вторую гипотезу, иначе — первую: сравнивать отношение плотностей $f_2(x_1, \ldots, x_n)/f_1(x_1, \ldots, x_n)$ не с единицей, а с некоторой постоянной c.

Назовем отношением правдоподобия частное

$$T(\mathbf{x}) = T(x_1, \ldots, x_n) = \frac{f_2(x_1, \ldots, x_n)}{f_1(x_1, \ldots, x_n)}, \tag{18}$$

рассматривая его лишь при таких значениях $\mathbf{x}$, когда хотя бы одна из плотностей отлична от нуля. Имеется в виду, что $0/a = 0$, $a/0 = +\infty$.

Конструкция критерия, который мы живописали выше, сильно усложнится в случае, когда распределение случайной величины $T(\mathbf{X})$ не является непрерывным, т. е. существует такое число c, что $\Delta_c = \mathsf{P}_{H_1}(f_2(\mathbf{X})/f_1(\mathbf{X}) = c)$ отлична от нуля. Это означает, что на некотором «большом» множестве значений выборки обе гипотезы «равноправны»: отношение правдоподобия постоянно. Относя это множество целиком к критическому множеству или целиком исключая из него, мы меняем вероятность ошибки первого рода (размер) критерия на положительную величину Δ_c:

$$\mathsf{P}_{H_1}(T(\mathbf{X}) \geqslant c) = \mathsf{P}_{H_1}(T(\mathbf{X}) > c) + \mathsf{P}_{H_1}(T(\mathbf{X}) = c) = \mathsf{P}_{H_1}(T(\mathbf{X}) > c) + \Delta_c.$$

И если вдруг мы захотим приравнять вероятность ошибки первого рода к заранее выбранному числу ε, может случиться так, что критерий с критическим множеством $S = \{T(\mathbf{x}) \geqslant c\}$ имеет размер больший, чем ε, а критерий с критическим множеством $S = \{T(\mathbf{x}) > c\}$ — размер меньший, чем ε.

Поэтому для математиков, не читающих [1], мы сформулируем замечательно мощное утверждение мелким шрифтом, зато в общем случае. Затем для почти математиков сформулируем и докажем частный, но наиболее частый случай, когда отношение правдоподобия $T(\mathbf{X})$ имеет при верной первой гипотезе непрерывную функцию распределения, т. е. $\Delta_c = 0$ для любого c.

7.3.1. Из человеколюбия

Определение 26. Функция $\delta(\mathbf{X})$, принимающая значения в интервале $[0, 1]$ в зависимости от значений выборки, называется рандомизированным критерием. Значение этой функции трактуют как вероятность принять вторую (так удобнее) гипотезу в некотором дополнительном эксперименте. Т. е., получив значение $\delta(\mathbf{X}) = 3/4$, мы должны дополнительно провести испытание с вероятностью успеха $3/4$, и в случае успеха принять H_2, в случае неудачи — H_1. Значение $\delta(\mathbf{X}) = 1$ предписывает принять вторую гипотезу (с вероятностью 1, т. е. обязательно), а значение $\delta(\mathbf{X}) = 0$ предписывает не принимать вторую гипотезу, а принять первую.

Напомним, что $T(\mathbf{X})$ есть отношение правдоподобия, которое мы ввели в (18).

Определение 27. Рандомизированный критерий, устроенный следующим образом:

$$\delta_{c,p}(\mathbf{X}) = \begin{cases} 0, & \text{если } T(\mathbf{X}) < c, \\ p, & \text{если } T(\mathbf{X}) = c, \\ 1, & \text{если } T(\mathbf{X}) > c, \end{cases} = \begin{cases} \text{принимается } H_1, & \text{если } T(\mathbf{X}) < c, \\ \text{с вероятностью } p \text{ принимается } H_2, & \text{если } T(\mathbf{X}) = c, \\ \text{принимается } H_2, & \text{если } T(\mathbf{X}) > c, \end{cases}$$

называется критерием отношения правдоподобия (КОП).

Размер и мощность КОП вычисляются по формуле полной вероятности. Размер равен

$$\alpha_1(\delta_{c,p}) = \mathsf{P}_{H_1}(\text{принять } H_2) = 1 \cdot \mathsf{P}_{H_1}(T(\mathbf{X}) > c) + p \cdot \mathsf{P}_{H_1}(T(\mathbf{X}) = c) = \mathsf{E}_{H_1}\delta_{c,p}(\mathbf{X}).$$

Мощность равна

$$1-\alpha_2(\delta_{c,p}) = \mathsf{P}_{H_2}(\text{принять } H_2) = 1\cdot \mathsf{P}_{H_2}(T(\mathbf{X})>c) + p\cdot \mathsf{P}_{H_2}(T(\mathbf{X})=c) = \mathsf{E}_{H_2}\delta_{c,p}(\mathbf{X}).$$

Вероятность ошибки второго рода можно найти и иначе:

$$\alpha_2(\delta_{c,p}) = \mathsf{P}_{H_2}(\text{принять } H_1) = \mathsf{P}_{H_2}(T(\mathbf{X})<c)+(1-p)\cdot \mathsf{P}_{H_2}(T(\mathbf{X})=c) = 1-\mathsf{E}_{H_2}\delta_{c,p}(\mathbf{X}).$$

Лемма Неймана — Пирсона. Существуют постоянные c и p, при которых критерий отношения правдоподобия является

1) минимаксным критерием; числа c и p следует выбрать так, чтобы вероятности ошибок первого и второго рода были одинаковы: $\alpha_1(\delta_{c,p}) = \alpha_2(\delta_{c,p})$;

2) байесовским критерием при заданных априорных вероятностях r и s; число p может быть любым, а c выбирается равным отношению r/s;

3) для любого $0<\varepsilon<1$ наиболее мощным критерием размера ε; числа c и p должны быть выбраны так, чтобы размер критерия равнялся ε:

$$\alpha_1(\delta_{c,p}) = \mathsf{P}_{H_1}(T(\mathbf{X})>c) + p\cdot \mathsf{P}_{H_1}(T(\mathbf{X})=c) = \varepsilon.$$

Мы не ограничиваем значения c областью $c>0$. Возможность брать $c=0$ (меньше бессмысленно, ибо $T(\mathbf{x})\geqslant 0$) избавляет от ограничения $\varepsilon \leqslant \mathsf{P}_{H_1}(f_2(\mathbf{X})>0)$ на возможный размер НМК. Это довольно сомнительное обобщение — ведь уже критерий $\delta(\mathbf{X}) = \mathbf{I}(f_2(\mathbf{X})>0)$ имеет единичную мощность при размере равном $\mathsf{P}_{H_1}(f_2(\mathbf{X})>0)$. Можно увеличивать размер и дальше, но мощности расти уже некуда. Такой НМК будет принимать с положительной вероятностью гипотезу H_2 там, где она верна быть не может — в области $f_2(\mathbf{x})=0$.

7.3.2. Лемма Неймана — Пирсона

Всюду далее предполагается, что

$$\boxed{\text{Функция } R(c) = \mathsf{P}_{H_1}(T(\mathbf{X})\geqslant c) \text{ непрерывна по } c \text{ при } c>0.} \tag{19}$$

Функция $R(c)$ есть просто хвост функции распределения случайной величины $T(\mathbf{X})$:

$$R(c) = 1-\mathsf{P}_{H_1}(T(\mathbf{X})<c).$$

Ее непрерывность означает, что величина $\Delta_c = \mathsf{P}_{H_1}(T(\mathbf{X})=c)$ равна нулю для любого $c>0$. Это предположение избавляет нас от необходимости рассматривать рандомизированные критерии. Итак,

Определение 28. В предположении (19) критерий

$$\delta_c(\mathbf{X}) = \begin{cases} H_1, & \text{если } T(\mathbf{X})<c,\\ H_2, & \text{если } T(\mathbf{X})\geqslant c,\end{cases} = \begin{cases} H_1, & \text{если } \dfrac{f_2(X_1,\ldots,X_n)}{f_1(X_1,\ldots,X_n)}<c,\\ H_2, & \text{если } \dfrac{f_2(X_1,\ldots,X_n)}{f_1(X_1,\ldots,X_n)}\geqslant c,\end{cases}$$

назовем критерием отношения правдоподобия (КОП).

Размер и вероятность ошибки второго рода этого критерия равны соответственно

$$\alpha_1(\delta_c) = \mathsf{P}_{H_1}(T(\mathbf{X})\geqslant c) = R(c); \qquad \alpha_2(\delta_c) = \mathsf{P}_{H_2}(T(\mathbf{X})<c).$$

Лемма Неймана — Пирсона. Пусть выполнено (19). Тогда существует постоянная c, при которой критерий отношения правдоподобия является

1) минимаксным критерием; число c следует выбрать так, чтобы вероятности ошибок первого и второго рода были одинаковы: $\alpha_1(\delta_c) = \alpha_2(\delta_c)$;

2) байесовским критерием при заданных априорных вероятностях r и s; число c выбирается равным отношению r/s;

3) для любого $0 < \varepsilon \leqslant P_{H_1}(f_2(\mathbf{X}) > 0)$ наиболее мощным критерием размера ε; число c должно быть выбрано так, чтобы размер критерия равнялся ε: $\alpha_1(\delta_c) = \varepsilon$.

Доказательство. Доказать достаточно два утверждения: существование постоянной c, удовлетворяющей первому и третьему пунктам леммы, и оптимальность соответствующего критерия. Начнем с третьего пункта.

1. Докажем, что для любого $0 < \varepsilon \leqslant P_{H_1}(f_2(\mathbf{X}) > 0)$ существует постоянная c такая, что $R(c) = \alpha_1(\delta_c) = \varepsilon$.

Функция $R(c) = P_{H_1}(T(\mathbf{X}) \geqslant c)$ не возрастает по c. Предел $R(c)$ при $c \to \infty$ равен нулю, поскольку событие $\{T(\mathbf{X}) = \infty\} = \{f_1(\mathbf{X}) = 0\}$ имеет нулевую вероятность $P_{H_1}(f_1(\mathbf{X}) = 0) = 0$. Заметим также, что

$$R(+0) = P_{H_1}(T(\mathbf{X}) > 0) = P_{H_1}(f_2(\mathbf{X}) > 0) \geqslant \varepsilon.$$

Итак, $R(c)$ непрерывно меняется от $R(+0)$ до 0, поэтому для любого $0 < \varepsilon \leqslant R(+0)$ существует c такое, что $R(c) = \alpha_1(\delta_c) = \varepsilon$.

2. Для первого пункта докажем существование такой c, что $\alpha_1(\delta_c) = \alpha_2(\delta_c)$.

Функция $\alpha_2(\delta_c) = P_{H_2}(T(\mathbf{X}) < c)$, в отличие от $R(c)$, не убывает по c. К тому же при $c \to 0$ величина $\alpha_2(\delta_c)$ стремится к $P_{H_2}(T(\mathbf{X}) = 0) = P_{H_2}(f_2(\mathbf{X}) = 0) = 0$. Она тоже, подобно $R(c)$, при $c > 0$ непрерывна по c из-за предположения (19):

$$P_{H_2}(T(\mathbf{X}) = c) = \int\limits_{\{f_2(y)=cf_1(y)\}} f_2(y)\,dy = c \int\limits_{\{f_2(y)=cf_1(y)\}} f_1(y)\,dy = cP_{H_1}(T(\mathbf{X}) = c) = 0.$$

Поэтому функции $R(c) = \alpha_1(\delta_c)$ и $\alpha_2(\delta_c)$ пересекаются хотя бы однажды.

3. Далее нам потребуется следующее красивое равенство.

Лемма 2. Введем функцию $\phi(y) = \min\{f_2(y), c f_1(y)\}$. Для нее

$$\alpha_2(\delta_c) + c\alpha_1(\delta_c) = \int\limits_{\mathbb{R}^n} \phi(y)\,dy.$$

Доказательство.

$$\alpha_2(\delta_c)+c\alpha_1(\delta_c) = P_{H_2}(T(\mathbf{X}) < c)+cP_{H_1}(T(\mathbf{X}) \geqslant c) = \int\limits_{\{T(y)<c\}} f_2(y)\,dy + c\int\limits_{\{T(y)\geqslant c\}} f_1(y)\,dy.$$

Но на множестве $\{T(y) < c\} = \{f_2(y) < cf_1(y)\}$ подынтегральная функция $f_2(y)$ совпадает с $\phi(y)$. И на множестве $\{T(y) \geqslant c\} = \{f_2(y) \geqslant cf_1(y)\}$ подынтегральная функция $cf_1(y)$ тоже совпадает с $\phi(y)$. Продолжая цепочку равенств, получим

$$\alpha_2(\delta_c) + c\alpha_1(\delta_c) = \int\limits_{\{T(y)<c\}} \phi(y)\,dy + \int\limits_{\{T(y)\geqslant c\}} \phi(y)\,dy = \int\limits_{\mathbb{R}^n} \phi(y)\,dy.$$

Пусть δ — любой другой критерий. Лемма 2 влечет странное следствие:

Следствие 4. Каково бы ни было $c > 0$, вероятности ошибок любого критерия δ связаны с вероятностями ошибок КОП δ_c неравенством

$$\alpha_2(\delta) + c\alpha_1(\delta) \geqslant \alpha_2(\delta_c) + c\alpha_1(\delta_c). \qquad (20)$$

Доказательство. Рассматривая для краткости нерандомизированный критерий δ, получим

$$\alpha_2(\delta) + c\alpha_1(\delta) = \int\limits_{\{\delta(y)=H_1\}} f_2(y)\,dy + c \int\limits_{\{\delta(y)=H_2\}} f_1(y)\,dy \geqslant$$

$$\geqslant \int\limits_{\{\delta(y)=H_1\}} \phi(y)\,dy + \int\limits_{\{\delta(y)=H_2\}} \phi(y)\,dy = \int\limits_{\mathbb{R}^n} \phi(y)\,dy = \alpha_2(\delta_c) + c\alpha_1(\delta_c).$$

4. Используя неравенство (20), докажем оптимальность КОП во всех трех смыслах.

I. Пусть c таково, что $\alpha_1(\delta_c) = \alpha_2(\delta_c) = \max\{\alpha_1(\delta_c), \alpha_2(\delta_c)\}$. Тогда правая часть неравенства (20) равна $(1+c)\max\{\alpha_1(\delta_c), \alpha_2(\delta_c)\}$, и для любого иного критерия δ

$$(1+c)\max\{\alpha_1(\delta_c), \alpha_2(\delta_c)\} \leqslant \alpha_2(\delta) + c\alpha_1(\delta) \leqslant (1+c)\max\{\alpha_1(\delta), \alpha_2(\delta)\}.$$

Итак, $\max\{\alpha_1(\delta_c), \alpha_2(\delta_c)\} \leqslant \max\{\alpha_1(\delta), \alpha_2(\delta)\}$, т. е. δ_c — минимаксный.

II. Пусть $c = r/s$ и $s \neq 0$. Тогда для любого иного критерия δ неравенство (20) превращается в определение байесовского критерия:

$$\alpha_2(\delta) + \frac{r}{s}\alpha_1(\delta) \leqslant \alpha_2(\delta_c) + \frac{r}{s}\alpha_1(\delta_c), \text{ или } s\alpha_2(\delta) + r\alpha_1(\delta) \leqslant s\alpha_2(\delta_c) + r\alpha_1(\delta_c).$$

Если же $s = 0$, то $c = \infty$. Тогда КОП имеет нулевой размер и автоматически является байесовским.

III. Пусть c таково, что $\alpha_1(\delta_c) = \varepsilon$. Любой иной критерий δ из класса K_ε имеет размер $\alpha_1(\delta) \leqslant \varepsilon$. Используя в неравенстве (20) оба этих размера, получим

$$\alpha_2(\delta_c) + c\varepsilon = \alpha_2(\delta_c) + c\alpha_1(\delta_c) \leqslant \alpha_2(\delta) + c\alpha_1(\delta) \leqslant \alpha_2(\delta) + c\varepsilon.$$

Итак, $\alpha_2(\delta_c) \leqslant \alpha_2(\delta)$, т. е. δ_c — НМК в классе K_ε.

Наконец лемма Неймана — Пирсона доказана.

Пример 31. Имеется выборка $X_1, \ldots, X_n$ из нормального распределения со средним a и единичной дисперсией. Построим минимаксный, байесовский для $r = 1/3$, $s = 2/3$ и наиболее мощный размера ε критерии для проверки гипотезы $H_1 = \{a = a_1\}$ против альтернативы $H_2 = \{a = a_2\}$, где $a_1 < a_2$.

Отношение правдоподобия имеет абсолютно непрерывное распределение при любой из гипотез, поэтому критерий отношения правдоподобия 28 будет нерандомизированным, и достаточно описать только его критическую область $\delta(\mathbf{X}) = H_2$. Она определяется неравенством

$$T(\mathbf{X}) = \frac{f_2(\mathbf{X})}{f_1(\mathbf{X})} = \exp\left\{\frac{1}{2}\sum_{i=1}^{n}(X_i - a_1)^2 - \frac{1}{2}\sum_{i=1}^{n}(X_i - a_2)^2\right\} \geqslant c. \tag{21}$$

Критерий будет байесовским при $c = r/s = 1/2$. Упростим неравенство (21). Получим

$$\delta(\mathbf{X}) = H_2 \quad \text{при} \quad \overline{X} > \frac{\frac{1}{2}(a_2^2 - a_1^2) - \frac{1}{n}\ln 2}{a_2 - a_1}.$$

Например, при $a_1 = 0$ и $a_2 = 1$ критическая область имеет вид $\overline{X} > \frac{1}{2} - \frac{1}{n}\ln 2$.

Чтобы построить минимаксный и наиболее мощный критерии, запишем неравенство (21) в эквивалентном виде $\overline{X} \geqslant c_1$, и искать будем c_1, а не c. Размер и вероятность ошибки второго рода равны соответственно

$$\begin{aligned}
\alpha_1(\delta) &= \mathsf{P}_{H_1}\left(\overline{X} \geqslant c_1\right) = \mathsf{P}_{H_1}\left(\sqrt{n}\,(\overline{X}-a_1) \geqslant \sqrt{n}\,(c_1-a_1)\right) = 1 - \Phi_{0,1}\left(\sqrt{n}\,(c_1-a_1)\right), \\
\alpha_2(\delta) &= \mathsf{P}_{H_2}\left(\overline{X} < c_1\right) = \mathsf{P}_{H_2}\left(\sqrt{n}\,(\overline{X}-a_2) < \sqrt{n}\,(c_1-a_2)\right) = \Phi_{0,1}\left(\sqrt{n}\,(c_1-a_2)\right).
\end{aligned}$$

При $\alpha_1(\delta)=\varepsilon$ получим НМК размера ε. Отсюда $\sqrt{n}(c_1-a_1) = \tau_{1-\varepsilon}$, где $\tau_{1-\varepsilon}$ — квантиль уровня $1-\varepsilon$ стандартного нормального распределения. Тогда $c_1 = a_1 + \tau_{1-\varepsilon}/\sqrt{n}$ и НМК размера ε имеет вид

$$\delta(\mathbf{X}) = H_2 \quad \text{при} \quad \overline{X} > a_1 + \frac{\tau_{1-\varepsilon}}{\sqrt{n}}.$$

При $\alpha_1(\delta) = \alpha_2(\delta)$ получим минимаксный критерий. Пользуясь свойствами функции распределения стандартного нормального закона, запишем

$$1 - \Phi_{0,1}\left(\sqrt{n}\,(c_1 - a_1)\right) = \Phi_{0,1}\left(\sqrt{n}\,(c_1 - a_2)\right) = 1 - \Phi_{0,1}\left(\sqrt{n}\,(a_2 - c_1)\right),$$

откуда $c_1 - a_1 = a_2 - c_1$ и $c_1 = (a_1 + a_2)/2$. Минимаксный критерий имеет вид

$$\delta(\mathbf{X}) = H_2 \quad \text{при} \quad \overline{X} > \frac{a_1 + a_2}{2}.$$

Пример 32. Имеется выборка $X_1, \ldots, X_n$ из нормального распределения со средним 0 и дисперсией σ^2, $\sigma > 0$. Построим наиболее мощный критерий размера ε для проверки гипотезы $H_1 = \{\sigma = \sigma_1\}$ против альтернативы $H_2 = \{\sigma = \sigma_2\}$, где $\sigma_1 < \sigma_2$.

Отношение правдоподобия снова имеет абсолютно непрерывное распределение при любой из гипотез, поэтому критерий отношения правдоподобия 28 будет нерандомизированным. Его критическая область $\delta(\mathbf{X}) = H_2$ определяется неравенством

$$T(\mathbf{X}) = \frac{\sigma_1^n}{\sigma_2^n}\exp\left\{\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)\sum_{i=1}^{n} X_i^2\right\} \geqslant c,$$

что эквивалентно неравенству $\overline{X^2} \geqslant c_1$. Найдем c_1, при котором размер критерия равен ε:

$$\alpha_1(\delta) = \mathsf{P}_{H_1}\left(\overline{X^2} \geqslant c_1\right) = \mathsf{P}_{H_1}\left(\frac{n\overline{X^2}}{\sigma_1^2} \geqslant \frac{nc_1}{\sigma_1^2}\right) = 1 - H_n\left(\frac{nc_1}{\sigma_1^2}\right) = \varepsilon.$$

Отсюда $nc_1/\sigma_1^2 = h_{1-\varepsilon}$, где $h_{1-\varepsilon}$ — квантиль χ^2-распределения с n степенями свободы уровня $1-\varepsilon$. Тогда $c_1 = h_{1-\varepsilon}\sigma_1^2/n$ и НМК размера ε имеет вид

$$\delta(\mathbf{X}) = H_2 \quad \text{при} \quad \overline{X^2} > \frac{h_{1-\varepsilon}\sigma_1^2}{n}.$$

8. Критерии согласия

Критериями согласия называют критерии, предназначенные для проверки простой гипотезы $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ при сложной альтернативе $H_2 = \{H_1 \text{ неверна}\}$. Мы рассмотрим более широкий класс основных гипотез, включающий и сложные гипотезы, а критериями согласия будем называть любые критерии, устроенные по одному и тому же принципу. А именно, пусть задана некоторая функция отклонения эмпирического распределения от теоретического, распределение которой существенно разнится в зависимости от того, верна или нет основная гипотеза. Критерии согласия принимают или отвергают основную гипотезу исходя из величины этой функции отклонения.

Итак, имеется выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из распределения $\mathcal{F}$. Мы сформулируем ряд понятий для случая простой основной гипотезы, а в дальнейшем будем их корректировать по мере изменения задачи. Проверяется простая основная гипотеза $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ при сложной альтернативе $H_2 = \{\mathcal{F} \neq \mathcal{F}_1\}$.

K1. Пусть возможно задать функцию $\rho(\mathbf{X})$, обладающую свойствами:

а) если гипотеза H_1 верна, то $\rho(\mathbf{X}) \Rightarrow G$, где G — непрерывное распределение;

б) если гипотеза H_1 неверна, то $|\rho(\mathbf{X})| \xrightarrow{p} \infty$ при $n \to \infty$.

K2. Пусть функция $\rho(\mathbf{X})$ задана. Для случайной величины η из распределения G определим постоянную C из равенства $\varepsilon = \mathsf{P}\left(|\eta| \geqslant C\right)$. Построим критерий:

$$\delta(\mathbf{X}) = \begin{cases} H_1, & \text{если} \quad |\rho(\mathbf{X})| < C, \\ H_2, & \text{если} \quad |\rho(\mathbf{X})| \geqslant C. \end{cases} \tag{22}$$

Мы построили критерий согласия. Он «работает» по принципу: если для данной выборки функция отклонения велика (по абсолютному значению), то это свидетельствует в пользу альтернативы, и наоборот. Убедимся в том, что этот критерий имеет (асимптотический) размер ε и является состоятельным.

Определение 29. Говорят, что критерий δ для проверки простой гипотезы H_1 является критерием асимптотического размера ε, если его размер приближается к ε с ростом n:

$$\alpha_1(\delta) = \mathsf{P}_{H_1}(\delta(\mathbf{X}) \neq H_1) \to \varepsilon \quad \text{при} \quad n \to \infty.$$

Поскольку альтернатива H_2 всегда является сложной, то, как мы уже отмечали в замечании 16, вероятность ошибки второго рода любого критерия δ есть функция $\alpha_2(\delta, \mathcal{F}_2)$ от конкретного распределения $\mathcal{F}_2$ из списка возможных альтернатив $\{\mathcal{F}_2 \ : \ \mathcal{F}_2 \neq \mathcal{F}_1\}$.

Определение 30. Критерий δ для проверки гипотезы H_1 против сложной альтернативы H_2 называется состоятельным, если для любого распределения $\mathcal{F}_2$, отвечающего альтернативе H_2, вероятность ошибки второго рода стремится к нулю с ростом объема выборки:

$$\alpha_2(\delta, \mathcal{F}_2) = \mathsf{P}_{\mathcal{F}_2}(\delta(\mathbf{X}) = H_1) \to 0 \quad \text{при} \quad n \to \infty.$$

Свойство 10. Для критерия δ, заданного в (22), при $n \to \infty$:

1. $\alpha_1(\delta) = \mathsf{P}_{H_1}(|\rho(\mathbf{X})| \geqslant C) \to \mathsf{P}(|\eta| \geqslant C) = \varepsilon$;
2. $\alpha_2(\delta, \mathcal{F}_2) = \mathsf{P}_{\mathcal{F}_2}(|\rho(\mathbf{X})| < C) \to 0$ для любого распределения $\mathcal{F}_2$, отвечающего H_2.

Иначе говоря, построенный критерий имеет асимптотический размер ε и состоятелен.

Упражнение. Доказать свойство 10.
Указание. По определению, запись $\xi_n \xrightarrow{p} \infty$ означает, что для любого $C > 0$

$$\mathsf{P}(\xi_n < C) \to 0 \quad \text{при} \quad n \to \infty.$$

Замечание. Если вместо «$\rho(\mathbf{X}) \Rightarrow G$» в K1(a) выполняется «$\rho(\mathbf{X})$ имеет распределение G», то критерий (22) будет иметь точный размер ε.

8.1. Критерии согласия: критерий Колмогорова

Имеется выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из распределения $\mathcal{F}$. Проверяется простая гипотеза $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ против сложной альтернативы $H_2 = \{\mathcal{F} \neq \mathcal{F}_1\}$. В том случае, когда распределение $\mathcal{F}_1$ имеет непрерывную функцию распределения F_1, можно воспользоваться критерием Колмогорова.

Пусть

$$\rho(\mathbf{X}) = \sqrt{n}\sup_y |F_n^*(y) - F_1(y)|.$$

Покажем, что $\rho(\mathbf{X})$ удовлетворяет условиям K1(а,б).

а) Если H_1 верна, то X_i имеют распределение $\mathcal{F}_1$. По теореме Колмогорова $\rho(\mathbf{X}) \Rightarrow \eta$, где η имеет распределение с функцией распределения Колмогорова.

б) Если гипотеза H_1 неверна, то X_i имеют какое-то распределение $\mathcal{F}_2$, отличное от $\mathcal{F}_1$. По теореме Гливенко — Кантелли $F_n^*(y) \xrightarrow{p} F_2(y)$ для любого y при $n \to \infty$. Поскольку $\mathcal{F}_1 \neq \mathcal{F}_2$, найдется y_0 такое, что $|F_2(y_0) - F_1(y_0)| > 0$. Но

$$\sup_y |F_n^*(y) - F_1(y)| \geqslant |F_n^*(y_0) - F_1(y_0)| \xrightarrow{p} |F_2(y_0) - F_1(y_0)| > 0.$$

Умножая на $\sqrt{n}$, получим при $n \to \infty$, что $\rho(\mathbf{X}) = \sqrt{n}\sup_y |F_n^*(y) - F_1(y)| \xrightarrow{p} \infty$.

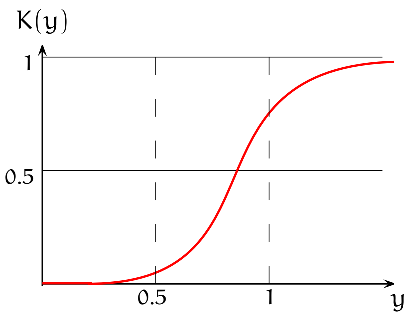


Рис. 9: График функции $K(y)$

Пусть случайная величина η имеет распределение с функцией распределения Колмогорова

$$K(y) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2y^2}, \quad y > 0.$$

Это распределение табулировано, так что по заданному ε легко найти C такое, что $\varepsilon = \mathsf{P}(\eta \geqslant C)$.

Критерий Колмогорова выглядит так:

$$\delta(\mathbf{X}) = \begin{cases} H_1, & \text{если} \quad \rho(\mathbf{X}) < C, \\ H_2, & \text{если} \quad \rho(\mathbf{X}) \geqslant C. \end{cases}$$

8.2. Критерии согласия: критерий χ^2 Пирсона

Критерий χ^2 (K.Pearson, 1903) основывается на группированных данных. Область значений предполагаемого распределения $\mathcal{F}_1$ делят на некоторое число интервалов. После чего строят функцию отклонения ρ по разностям теоретических вероятностей попадания в интервалы группировки и эмпирических частот.

Имеется выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из распределения $\mathcal{F}$. Проверяется простая гипотеза $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ против сложной альтернативы $H_2 = \{\mathcal{F} \neq \mathcal{F}_1\}$.

Пусть, как в параграфе 1.6, A_1, ..., A_k — интервалы группировки в области значений случайной величины с распределением $\mathcal{F}_1$. Обозначим для $j = 1, \ldots, k$ через ν_j число элементов выборки, попавших в интервал A_j

$$\nu_j = \{\text{число } X_i \in A_j\} = \sum_{i=1}^{n} \mathbf{I}(X_i \in A_j),$$

и через $p_j > 0$ — теоретическую вероятность $\mathsf{P}_{H_1}(X_1 \in A_j)$ попадания в интервал A_j случайной величины с распределением $\mathcal{F}_1$. С необходимостью, $p_1 + \ldots + p_k = 1$. Как правило, длины интервалов выбирают так, чтобы $p_1 = \ldots = p_k = 1/k$.

Пусть

$$\rho(\mathbf{X}) = \sum_{j=1}^{k} \frac{(\nu_j - np_j)^2}{np_j}. \tag{23}$$

Замечание 18. Свойство K1(б) выполнено далеко не для всех альтернатив. Если распределение выборки $\mathcal{F}_2 \neq \mathcal{F}_1$ имеет такие же, как у $\mathcal{F}_1$, вероятности p_j попадания в каждый из интервалов A_j, то по данной функции ρ эти распределения различить невозможно.

Поэтому на самом деле критерий, который мы построим по функции ρ из (23), решает совсем иную задачу. А именно, пусть задан набор вероятностей $p_1, \ldots, p_k$ такой, что $p_1 + \ldots + p_k = 1$. Критерий χ^2 предназначен для проверки сложной гипотезы

$$H_1' = \big\{\text{распределение } X_1 \text{ обладает свойством: } \mathsf{P}(X_1 \in A_j) = p_j \text{ для всех } j = 1, \ldots, k\big\}$$

против сложной альтернативы $H_2' = \{H_1' \text{ неверна}\}$, т. е.

$$H_2' = \big\{\text{хотя бы для одного из интервалов вероятность } \mathsf{P}(X_1 \in A_j) \text{ отличается от } p_j\big\}.$$

Покажем, что $\rho(\mathbf{X})$ удовлетворяет условию K1(а).

Теорема Пирсона. Если верна гипотеза H_1', то при фиксированном k и при $n \to \infty$

$$\rho(\mathbf{X}) = \sum_{j=1}^{k} \frac{(\nu_j - np_j)^2}{np_j} \Rightarrow \mathsf{H}_{k-1},$$

где, напомним, H_{k-1} есть χ^2-распределение с $k-1$[o]степенью свободы.

[o]Стоит остановиться и задать себе вопрос. Величина ρ есть сумма k слагаемых. Слагаемые, если вы не забыли ЦПТ или теорему Муавра — Лапласа, имеют распределения, близкие к квадратам каких-то нормальных. Куда потерялась одна степень свободы? Причина кроется, конечно, в зависимости слагаемых: $\nu_k = n - \nu_1 - \ldots - \nu_{k-1}$.

Докажем теорему Пирсона при $k = 2$.
В этом случае $\nu_2 = n - \nu_1$, $p_2 = 1 - p_1$. Посмотрим на ρ и вспомним ЦПТ:

$$\rho(\mathbf{X}) = \frac{(\nu_1 - np_1)^2}{np_1} + \frac{(\nu_2 - np_2)^2}{np_2} = \frac{(\nu_1 - np_1)^2}{np_1} + \frac{(n - \nu_1 - n(1-p_1))^2}{n(1-p_1)} =$$

$$= \frac{(\nu_1 - np_1)^2}{np_1} + \frac{(-\nu_1 + np_1)^2}{n(1-p_1)} = \frac{(\nu_1 - np_1)^2}{np_1(1-p_1)} = \left(\frac{\nu_1 - np_1}{\sqrt{np_1(1-p_1)}}\right)^2$$

Но величина ν_1 есть сумма n независимых случайных величин с распределением Бернулли B_{p_1}, и по ЦПТ

$$\frac{\nu_1 - np_1}{\sqrt{np_1(1-p_1)}} \Rightarrow \xi,$$

где ξ имеет стандартное нормальное распределение. Поэтому

$$\rho(\mathbf{X}) = \left(\frac{\nu_1 - np_1}{\sqrt{np_1(1-p_1)}}\right)^2 \Rightarrow \xi^2.$$

Величина ξ^2 имеет χ^2-распределение H_1 с одной степенью свободы. □

Для экономистов, только приступающих к знакомству с многомерным нормальным распределением, матрицами ковариаций и всевозможными квадратичными формами, составленными из (асимптотически) нормальных слагаемых, исключительно полезно познакомиться с доказательством теоремы Пирсона в общем случае. Параграф A приложения, который познакомит читателя с многомерным нормальным распределением, стоит напечатать (CTRL+P) и повесить в изголовье кровати до окончания курса эконометрики.

Функция $\rho(\mathbf{X})$ удовлетворяет условию K1(б). Действительно,

Упражнение. Вспомнить закон больших чисел и доказать, что если H_1' неверна, то найдется $j \in \{1, \ldots, k\}$ такое, что

$$\frac{(\nu_j - np_j)^2}{np_j} = \frac{n}{p_j}\left(\frac{\nu_j}{n} - p_j\right)^2 \xrightarrow{p} \infty.$$

Осталось построить критерий в соответствии с K2.

Пусть случайная величина χ^2_{k-1} имеет распределение H_{k-1}. По таблице распределения H_{k-1} найдем C равное квантили уровня $1-\varepsilon$ этого распределения. Тогда $\varepsilon = \mathsf{P}\left(\chi^2_{k-1} \geqslant C\right)$ и критерий согласия χ^2 выглядит как все критерии согласия:

$$\delta(\mathbf{X}) = \begin{cases} H_1', & \text{если } \rho(\mathbf{X}) < C, \\ H_2', & \text{если } \rho(\mathbf{X}) \geqslant C. \end{cases}$$

Замечание 19. На самом деле критерий χ^2 применяют и для решения первоначальной задачи о проверке гипотезы $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$. Необходимо только помнить, что этот критерий не состоятелен для альтернатив с теми же вероятностями попадания в интервалы разбиения, что и у $\mathcal{F}_1$. Поэтому берут большое число интервалов разбиения — чем больше, тем лучше, чтобы «уменьшить» число альтернатив, неразличимых с предполагаемым распределением.

Внимание! Опасность!

Замечание 20. Сходимость по распределению $\rho(\mathbf{X}) \Rightarrow H_{k-1}$ обеспечивается ЦПТ, поэтому разница допредельной и предельной вероятностей имеет тот же порядок, что и погрешность нормального приближения

$$|\mathsf{P}\left(\rho(\mathbf{X}) \geqslant C\right) - \mathsf{P}\left(\chi^2_{k-1} \geqslant C\right)| \leqslant \text{ примерно! } \max\left\{\frac{b}{\sqrt{np_j(1-p_j)}}\right\}$$

(см. неравенство Берри — Эссеена для погрешности в ЦПТ), где b – некоторая постоянная. Маленькие значения np_j в знаменателе приведут к тому, что распределение $\rho(\mathbf{X})$ будет существенно отличаться от H_{k-1}. Тогда и реальная вероятность $\mathsf{P}\left(\rho \geqslant C\right)$ — точный размер полученного критерия — будет сильно отличаться от ε. Поэтому для выборки объема n число интервалов разбиения выбирают так, чтобы обеспечить нужную точность при замене распределения $\rho(\mathbf{X})$ на H_{k-1}.

Обычно требуют, чтобы $np_1 = \ldots = np_k$ были не менее 5-6.

8.3. Критерий χ^2 Пирсона для проверки параметрической гипотезы

Критерий χ^2 часто применяют для проверки гипотезы о виде распределения, т. е. о принадлежности распределения выборки некоторому параметрическому семейству.

Имеется выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из неизвестного распределения $\mathcal{F}$. Проверяется сложная гипотеза

$$H_1 = \left\{\mathcal{F} \in \{\mathcal{F}_\theta\}\right\},$$

где $\theta \in \Theta \subseteq \mathbb{R}^l$ — неизвестный параметр (скалярный или векторный), l — его размерность.

Пусть $\mathbb{R}$ разбито на $k > l$ интервалов группировки $A_1 \cup \cdots \cup A_k$, и ν_j — число элементов выборки, попавших в A_j. Но вероятность $p_j = \mathsf{P}_{H_1}(X_1 \in A_j) = p_j(\theta)$ теперь зависит от неизвестного параметра θ.

Функция отклонения (23) также зависит от неизвестного параметра θ, и использовать ее в критерии Пирсона нельзя — мы не можем вычислить ее значение:

$$\rho(\mathbf{X}, \theta) = \sum_{j=1}^{k} \frac{(\nu_j - np_j(\theta))^2}{np_j(\theta)}. \tag{24}$$

Пусть $\hat\theta = \hat\theta(\mathbf{X})$ — значение параметра θ, доставляющее минимум функции $\rho(\mathbf{X}, \theta)$ при данной выборке $\mathbf{X}$. Подставив вместо истинных вероятностей p_j их оценки $p_j(\hat\theta)$, получим функцию отклонения

$$\rho(\mathbf{X}, \hat\theta) = \sum_{j=1}^{k} \frac{(\nu_j - np_j(\hat\theta))^2}{np_j(\hat\theta)}. \tag{25}$$

Условие K1(а) (при выполнении некоторых условий относительно гладкости $p_j(\theta)$[o]) обеспечивается теоремой (R. Fisher, 1924), которую мы доказывать не будем:

Теорема 8. Если верна гипотеза H_1, и $\dim(\theta) = l$ — размерность параметра (вектора) θ, то при фиксированном k и при $n \to \infty$

$$\rho(\mathbf{X}, \hat{\theta}) = \sum_{j=1}^{k} \sum_{j=1}^{k} \frac{(\nu_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})} \Rightarrow \mathsf{H}_{k-1-l},$$

где H_{k-1-l} есть χ^2-распределение с $k-1-l$ степенями свободы.

Условие (K1(б)) выполнено, если, скажем, рассматривать альтернативные распределения $\mathcal{F}_2$ такие, что ни при каких θ набор вероятностей $\mathsf{P}_{\mathcal{F}_2}(X_1 \in A_1), \ldots, \mathsf{P}_{\mathcal{F}_2}(X_1 \in A_k)$ не совпадает с $p_1(\theta), \ldots, p_k(\theta)$.

Построим критерий χ^2.

Пусть случайная величина χ^2_{k-1-l} имеет распределение H_{k-1-l}. По заданному ε найдем C такое, что $\varepsilon = \mathsf{P}\left(\chi^2_{k-1-l} \geqslant C\right)$.

Критерий согласия χ^2 имеет такой же вид, как все критерии согласия:

$$\delta(\mathbf{X}) = \begin{cases} H_1, & \text{если } \rho(\mathbf{X}, \hat{\theta}) < C, \\ H_2, & \text{если } \rho(\mathbf{X}, \hat{\theta}) \geqslant C. \end{cases}$$

[o] Все $\partial^2 p_j(\theta)/\partial\theta_i \partial\theta_l$ непрерывны по θ; ранг матрицы $\|\partial p_j(\theta)/\partial\theta_i\|$ равен l.

Замечания 19, 20 о количестве интервалов разбиения остаются в силе.

Замечание 21. Оценку $\hat{\theta}$, минимизирующую функцию $\rho(\mathbf{X}, \theta)$, нельзя заменить на оценку максимального правдоподобия для θ, построенную по выборке $X_1, \ldots, X_n$. При такой замене предельное распределение величины $\rho(\mathbf{X}, \theta)$

а) уже не равно H_{k-1-l}, а совпадает с распределением величины

$$\xi_1^2 + \ldots + \xi_{k-1-l}^2 + a_1(\theta)\xi_{k-l}^2 + \ldots + a_l(\theta)\xi_{k-1}^2,$$

где все ξ_i независимы и имеют распределение $\mathsf{N}_{0,1}$, а коэффициенты $a_i(\theta)$, вообще говоря, отличны от 0 и 1 (H.Chernoff, E.Lehmann, 1954);

б) зависит от θ.

Почувствуйте разницу:

Замечание 22. Оценку $\hat{\theta}$, минимизирующую функцию $\rho(\mathbf{X}, \theta)$, можно получить как оценку максимального правдоподобия для θ, построенную по вектору $\nu_1, \ldots, \nu_k$ из полиномиального распределения. Функция правдоподобия имеет вид

$$f(\boldsymbol{\nu}; \theta) = \frac{n!}{\nu_1! \ldots \nu_k!} (p_1(\theta))^{\nu_1} \cdot \ldots \cdot (p_k(\theta))^{\nu_k}, \text{ где } \sum_{i=1}^{k} p_i(\theta) = 1 \text{ и } \sum_{i=1}^{k} \nu_i = n.$$

Вычисление точки максимума по θ такой функции в общем случае возможно лишь численно, равно как и вычисление точки минимума функции $\rho(\mathbf{X}, \theta)$.

8.4. Проверка гипотезы однородности: критерий Колмогорова — Смирнова

Даны две выборки $\mathbf{X} = (X_1, \ldots, X_n)$ и $\mathbf{Y} = (Y_1, \ldots, Y_m)$ из неизвестных распределений $\mathcal{F}$ и $\mathcal{G}$ соответственно. Проверяется сложная гипотеза $H_1 = \{\mathcal{F} = \mathcal{G}\}$ против (еще более сложной) альтернативы $H_2 = \{H_1$ неверна$\}$.

Критерий Колмогорова — Смирнова используют, если $\mathcal{F}$ и $\mathcal{G}$ имеют непрерывные функции распределения.

Пусть $F_n^*(y)$ и $G_m^*(y)$ — эмпирические функции распределения, построенные по выборкам $\mathbf{X}$ и $\mathbf{Y}$,

$$\rho(\mathbf{X}, \mathbf{Y}) = \sqrt{\frac{mn}{m+n}} \sup_y \left|F_n^*(y) - G_m^*(y)\right|.$$

Теорема 9. Если гипотеза H_1 верна, то $\rho(\mathbf{X}, \mathbf{Y}) \Rightarrow \eta$ при $n, m \to \infty$, где η имеет распределение с функцией распределения Колмогорова.

Упражнение. Доказать, что $\rho(\mathbf{X}, \mathbf{Y}) \xrightarrow{p} \infty$ при $n, m \to \infty$, если H_2 верна.

И снова: в таблице распределения Колмогорова по заданному ε найдем C такое, что $\varepsilon = \mathsf{P}(\eta \geqslant C)$, и построим критерий согласия Колмогорова — Смирнова:

$$\delta(\mathbf{X}) = \begin{cases} H_1, & \text{если } \rho(\mathbf{X}) < C, \\ H_2, & \text{если } \rho(\mathbf{X}) \geqslant C. \end{cases}$$

Замечание 23. Если есть более двух выборок, и требуется проверить гипотезу однородности, часто пользуются одним из вариантов критерия χ^2 Пирсона. Этот критерий (и ряд других критериев) рекомендую посмотреть в §3.4.2, с. 124 книги Г. И. Ивченко, Ю. И. Медведев. Математическая статистика. Москва, 1984, 248 с.

8.5. Проверка гипотезы независимости: критерий χ^2 Пирсона

Есть выборка $(\mathbf{X}, \mathbf{Y}) = ((X_1, Y_1), \ldots, (X_n, Y_n))$ значений двух наблюдаемых совместно случайных величин X и Y в n независимых экспериментах. Проверяется гипотеза $H_1 = \{X$ и Y независимы$\}$.

Введем k интервалов группировки $\Delta_1, \ldots, \Delta_k$ для значений X и m интервалов группировки $\nabla_1, \ldots, \nabla_m$ для значений Y.

X \ Y	∇_1	∇_2	$\ldots$	∇_m	$\sum_{j=1}^{m}$
Δ_1	$\nu_{1,1}$	$\nu_{1,2}$	$\ldots$	$\nu_{1,m}$	$\nu_{1,\cdot}$
$\vdots$			$\ldots$		
Δ_k	$\nu_{k,1}$	$\nu_{k,2}$	$\ldots$	$\nu_{k,m}$	$\nu_{k,\cdot}$
$\sum_{i=1}^{k}$	$\nu_{\cdot,1}$	$\nu_{\cdot,2}$	$\ldots$	$\nu_{\cdot,m}$	n

Посчитаем эмпирические частоты:

$\nu_{i,j} = \{$число пар (X_l, Y_l), попавших в $\Delta_i \times \nabla_j\}$,

$\nu_{\cdot,j} = \{$число Y_l, попавших в $\nabla_j\}$,

$\nu_{i,\cdot} = \{$число X_l, попавших в $\Delta_i\}$.

Если гипотеза H_1 верна, то теоретические вероятности попадания пары (X, Y) в любую из областей $\Delta_i \times \nabla_j$ равны произведению вероятностей: для всех i и j

$$p_{i,j} = \mathsf{P}\left((X, Y) \in \Delta_i \times \nabla_j\right) = \mathsf{P}\left(X \in \Delta_i\right) \cdot \mathsf{P}\left(Y \in \nabla_j\right) = p_i^x \cdot p_j^y.$$

Именно эту гипотезу (назовем ее H_1') мы в действительности и проверяем. По ЗБЧ

$$\frac{\nu_{i,\cdot}}{n} \xrightarrow{p} p_i^x, \qquad \frac{\nu_{\cdot,j}}{n} \xrightarrow{p} p_i^y, \qquad \frac{\nu_{i,j}}{n} \xrightarrow{p} p_{i,j}.^{\circ}$$

[°]Поэтому значительная разница между $\frac{\nu_{i,j}}{n}$ и $\frac{\nu_{i,\cdot}}{n}\frac{\nu_{\cdot,j}}{n}$ (или между $\nu_{i,j}$ и $\frac{\nu_{i,\cdot}\nu_{\cdot,j}}{n}$) может служить основанием для отклонения гипотезы независимости.

Пусть

$$\rho(\mathbf{X},\mathbf{Y}) = n\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\left(\nu_{i,j}-(\nu_{i,\cdot}\nu_{\cdot,j})/n\right)^2}{\nu_{i,\cdot}\nu_{\cdot,j}}. \tag{26}$$

Теорема 10. Если гипотеза H_1 верна, то $\rho(\mathbf{X},\mathbf{Y}) \Rightarrow H_{(k-1)(m-1)}$ при $n\to\infty$.

Критерий согласия асимптотического уровня ε строится обычным образом.

Упражнение. Чтобы функция ρ и теорема 10 не падали с неба, убедитесь, что гипотеза H_1' есть гипотеза о принадлежности распределения выборки параметрическому семейству распределений с вектором неизвестных параметров $(p_1^x,\ldots,p_{k-1}^x,p_1^y,\ldots,p_{m-1}^y)$ размерности $l{=}k{+}m{-}2$. Подставив ОМП $\nu_{i,\cdot}/n$ для p_i^x и $\nu_{\cdot,j}/n$ для p_j^y в функцию

$$\rho = \sum_{i,j}\frac{\left(\nu_{i,j}-np_i^xp_j^y\right)^2}{np_i^xp_j^y} \qquad \text{(см. (24))},$$

получим (26). Всего есть $k{\cdot}m$ интервалов, и по теореме 8 при верной H_1' предельное χ^2-распределение имеет $k{\cdot}m{-}1{-}(k{+}m{-}2)=(k{-}1)(m{-}1)$ степеней свободы.

Замечания 19 и 20 по поводу числа $k\cdot m$ интервалов группировки остаются в силе.

8.6. Совпадение дисперсий двух нормальных выборок

Есть две независимые выборки из нормальных распределений: $\mathbf{X}=(X_1,\ldots,X_n)$ из $\mathrm{N}_{a_1,\sigma_1^2}$ и $\mathbf{Y}=(Y_1,\ldots,Y_m)$ из $\mathrm{N}_{a_2,\sigma_2^2}$, средние которых, вообще говоря, неизвестны. Критерий Фишера предназначен для проверки гипотезы $H_1=\{\sigma_1^2=\sigma_2^2\}$.

Обозначим через $S_0^2(\mathbf{X})$ и $S_0^2(\mathbf{Y})$ несмещенные выборочные дисперсии:

$$S_0^2(\mathbf{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2,\quad S_0^2(\mathbf{Y}) = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i-\overline{Y})^2$$

и зададим функцию отклонения $\rho(\mathbf{X},\mathbf{Y})$ как их отношение $\rho(\mathbf{X},\mathbf{Y}) = \dfrac{S_0^2(\mathbf{X})}{S_0^2(\mathbf{Y})}$.

Теорема 11. Если гипотеза H_1 верна, то случайная величина $\rho(\mathbf{X},\mathbf{Y})$ имеет распределение Фишера $F_{n-1,m-1}$ с $n{-}1$, $m{-}1$ степенями свободы.

Доказательство. По лемме Фишера, независимые случайные величины

$$\xi_{n-1}^2 = \frac{(n{-}1)\,S_0^2(\mathbf{X})}{\sigma_1^2} \quad\text{и}\quad \xi_{m-1}^2 = \frac{(m{-}1)\,S_0^2(\mathbf{Y})}{\sigma_2^2}$$

имеют распределения H_{m-1} и H_{n-1} соответственно. При $\sigma_1^2=\sigma_2^2$ отношение

$$\frac{\xi_{n-1}^2/(n{-}1)}{\xi_{m-1}^2/(m{-}1)} = \frac{S_0^2(\mathbf{X})}{\cancel{\sigma_1^2}}\cdot\frac{\cancel{\sigma_2^2}}{S_0^2(\mathbf{Y})} = \rho(\mathbf{X},\mathbf{Y})$$

имеет распределение Фишера с $n{-}1$, $m{-}1$ степенями свободы по определению 18 и совпадает с $\rho(\mathbf{X},\mathbf{Y})$. □

С условием K1(б) дело обстоит сложнее.

Упражнение. Доказать, что для любой альтернативы $\sigma_1^2 \neq \sigma_2^2$

$$\rho(\mathbf{X}, \mathbf{Y}) \xrightarrow{p} \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \quad \text{при} \quad n, m \to \infty. \tag{27}$$

Построим критерий Фишера и убедимся, что (27) обеспечивает его состоятельность. Возьмем квантили $f_{\varepsilon/2}$ и $f_{1-\varepsilon/2}$ распределения Фишера $F_{n-1,m-1}$. Критерием Фишера называют критерий

$$\delta(\mathbf{X}, \mathbf{Y}) = \begin{cases} H_1, & \text{если} \quad f_{\varepsilon/2} \leqslant \rho(\mathbf{X}, \mathbf{Y}) \leqslant f_{1-\varepsilon/2}, \\ H_2 & \text{иначе.} \end{cases}$$

Доказательство состоятельности критерия Фишера.

Покажем, что последовательность квантилей $f_\delta = f_\delta(n, m)$ любого уровня $0 < \delta < 1$ распределения $F_{n,m}$ сходится к 1 при $n, m \to \infty$. Возьмем величину $f_{n,m}$ с этим распределением. По определению, $\mathsf{P}\left(f_{n,m} < f_\delta\right) = \delta$, $\mathsf{P}\left(f_{n,m} > f_\delta\right) = 1 - \delta$ при всех n, m. По свойству 2 распределения Фишера, $f_{n,m} \xrightarrow{p} 1$. Поэтому для любого $\epsilon > 0$ обе вероятности $\mathsf{P}\left(f_{n,m} < 1 - \epsilon\right)$ и $\mathsf{P}\left(f_{n,m} > 1 + \epsilon\right)$ стремятся к нулю при $n, m \to \infty$, становясь рано или поздно меньше как δ, так и $1 - \delta$. Следовательно, при достаточно больших n, m выполнено $1 - \epsilon < f_\delta < 1 + \epsilon$.

Для доказательства состоятельности осталось предположить, что гипотеза H_1 не верна, взять ϵ равное, например, половине расстояния от 1 до σ_1^2/σ_2^2 и использовать сходимость (27). Пусть, скажем, при достаточно больших n и m

$$\frac{\sigma_1^2}{\sigma_2^2} < \frac{\sigma_1^2}{\sigma_2^2} + \epsilon = 1 - \epsilon < f_{\varepsilon/2}.$$

Тогда вероятность ошибки второго рода удовлетворяет неравенствам

$$\alpha_2(\delta) = \mathsf{P}_{H_2}(f_{\varepsilon/2} \leqslant \rho \leqslant f_{1-\varepsilon/2}) \leqslant \mathsf{P}_{H_2}(1 - \epsilon < \rho) = \mathsf{P}_{H_2}\left(\rho > \frac{\sigma_1^2}{\sigma_2^2} + \epsilon\right) \to 0.$$

Аналогично рассматривается случай, когда (при достаточно больших n и m)

$$f_{1-\varepsilon/2} < 1 + \epsilon = \frac{\sigma_1^2}{\sigma_2^2} + \epsilon < \frac{\sigma_1^2}{\sigma_2^2}.$$

Упражнение. Сформулировать критерий Фишера в случае, когда средние известны. Какой статистикой вы воспользуетесь теперь?

Критерий Фишера используют в качестве первого шага в задаче проверки однородности двух независимых нормальных выборок. Особенно часто возникает необходимость проверить равенство средних двух нормальных совокупностей — например, в медицине или биологии для выяснения наличия или отсутствия действия препарата. Эта задача решается с помощью критерия Стьюдента (с ним мы познакомимся на следующей странице), но только в случае, когда неизвестные дисперсии равны. Для проверки этого предположения пользуются сначала критерием Фишера. Самое печальное, если гипотеза равенства дисперсий отвергается критерием Фишера, либо если сразу заведомо известно, что неизвестные дисперсии различны. Задачу о проверке равенства средних в этих условиях называют проблемой Беренса — Фишера. Ее решение возможно лишь в частных случаях, и больше о ней мы ничего говорить не будем.

8.7. Совпадение средних двух нормальных выборок с равными дисперсиями

Есть две независимые выборки: $\mathbf{X} = (X_1, \ldots, X_n)$ из $\mathsf{N}_{a_1,\sigma^2}$ и $\mathbf{Y} = (Y_1, \ldots, Y_m)$ из $\mathsf{N}_{a_2,\sigma^2}$, причем дисперсия σ^2 одинакова для обоих распределений, но, вообще говоря, неизвестна. Проверяется сложная гипотеза $H_1 = \{a_1 = a_2\}$.

Эта задача есть частный случай задачи об однородности. Для ее решения построим критерий Стьюдента точного размера ε.

Из леммы Фишера вытекает следующее утверждение.

Теорема 12. Случайная величина t_{n+m-2}, равная

$$t_{n+m-2} = \sqrt{\frac{nm}{n+m}} \cdot \frac{(\overline{X} - a_1) - (\overline{Y} - a_2)}{\sqrt{\dfrac{(n-1)S_0^2(\mathbf{X}) + (m-1)S_0^2(\mathbf{Y})}{n+m-2}}}$$

имеет распределение Стьюдента T_{n+m-2} с $n+m-2$ степенями свободы.

Доказательство теоремы 12.

1. Легко видеть, убедиться, что легко! что $\overline{X} - a_1$ имеет распределение $\mathsf{N}_{0,\sigma^2/n}$, а $\overline{Y} - a_2$ имеет распределение $\mathsf{N}_{0,\sigma^2/m}$. Тогда их разность распределена тоже нормально с нулевым средним и дисперсией равной

$$\mathsf{D}\left((\overline{X} - a_1) - (\overline{Y} - a_2)\right) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2 \cdot \frac{n+m}{nm}.$$

Нормируем эту разность. Величина

$$\xi_0 = \frac{1}{\sigma}\sqrt{\frac{nm}{n+m}}\left((\overline{X} - a_1) - (\overline{Y} - a_2)\right)$$

имеет стандартное нормальное распределение.

2. Из леммы Фишера следует, что независимые случайные величины $(n-1)\,S_0^2(\mathbf{X})/\sigma^2$ и $(m-1)\,S_0^2(\mathbf{Y})/\sigma^2$ имеют распределения H_{m-1} и H_{n-1} соответственно, а их сумма

$$S^2 = \frac{1}{\sigma^2}\left((n-1)S_0^2(\mathbf{X}) + (m-1)S_0^2(\mathbf{Y})\right)$$

имеет χ^2-распределение H_{n+m-2} с $n+m-2$ степенями свободы и не зависит от $\overline{X}$ и от $\overline{Y}$.

3. По определению 17, отношение $\dfrac{\xi_0}{\sqrt{S^2/(n+m-2)}}$ как раз имеет распределение Стьюдента T_{n+m-2}. Осталось подставить в эту дробь ξ_0 и S^2 и убедиться, что σ сократится и получится в точности t_{n+m-2} из теоремы 12.

Введем функцию $\rho(\mathbf{X}, \mathbf{Y}) = \sqrt{\dfrac{nm}{n+m}} \cdot \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{(n-1)S_0^2(\mathbf{X}) + (m-1)S_0^2(\mathbf{Y})}{n+m-2}}}.$

Из теоремы 12 следует свойство K1(a): если H_1 верна, т. е. если $a_1 = a_2$, то величина $\rho = t_{n+m-2}$ имеет распределение Стьюдента T_{n+m-2}.

Упражнение. Доказать свойство K1(б): для любой альтернативы к основной гипотезе (т. е. как только $a_1 \neq a_2$) величина $|\rho|$ неограниченно возрастает по вероятности с ростом n и m.

> **Указание.** Воспользовавшись ЗБЧ или свойствами 2–4 из 1-й лекции, доказать, что числитель и знаменатель сходятся к постоянным:
>
> $$\overline{X} - \overline{Y} \xrightarrow{p} \mathrm{const} \neq 0, \qquad \frac{(n-1)S_0^2(\mathbf{X}) + (m-1)S_0^2(\mathbf{Y})}{n+m-2} \xrightarrow{p} \mathrm{const} \neq 0,$$
>
> тогда как корень перед дробью неограниченно возрастает.

Поэтому остается по ε найти $C = \tau_{1-\varepsilon/2}$ — квантиль распределения T_{n+m-2}. Для такого C величина t_{n+m-2} из распределения T_{n+m-2} удовлетворяет равенству

$$\mathsf{P}\left(|t_{n+m-2}| > C\right) = 2\mathsf{P}\left(t_{n+m-2} > C\right) = \varepsilon.$$

И критерий Стьюдента выглядит как все критерии согласия:

$$\delta(\mathbf{X}, \mathbf{Y}) = \begin{cases} H_1, & \text{если } |\rho(\mathbf{X}, \mathbf{Y})| < C, \\ H_2, & \text{если } |\rho(\mathbf{X}, \mathbf{Y})| \geqslant C. \end{cases}$$

Упражнение. Доказать, что этот критерий имеет точный размер ε.

Упражнение. Построить критерий для проверки гипотезы о равенстве средних двух независимых нормальных выборок с произвольными известными дисперсиями.

8.8. Гипотеза о среднем нормальной совокупности с известной дисперсией

Имеется выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из нормального распределения N_{a,σ^2} с известной дисперсией σ^2. Проверяется простая гипотеза $H_1 = \{a = a_0\}$ против сложной альтернативы $H_2 = \{a \neq a_0\}$.

Построим критерий точного размера ε с помощью функции отклонения $\rho(\mathbf{X})$

$$\rho(\mathbf{X}) = \sqrt{n}\frac{\overline{X} - a_0}{\sigma}.$$

Очевидно свойство K1(а): если H_1 верна, то $\rho(\mathbf{X})$ имеет стандартное нормальное распределение.

Упражнение. Доказать свойство K1(б): если $a \neq a_0$, то $|\rho(\mathbf{X})| \xrightarrow{p} \infty$.

По ε выберем $C = \tau_{1-\varepsilon/2}$ — квантиль стандартного нормального распределения. Тогда

$$\varepsilon = \mathsf{P}_{H_1}(|\rho(\mathbf{X})| \geqslant C).$$

Критерий выглядит как все критерии согласия:

$$\delta(\mathbf{X}) = \begin{cases} H_1, & \text{если } |\rho(\mathbf{X})| < C, \\ H_2, & \text{если } |\rho(\mathbf{X})| \geqslant C. \end{cases} \tag{28}$$

Упражнение. Доказать, что этот критерий имеет точный размер ε и является состоятельным.

Упражнение. Построить критерий для различения трех гипотез: $H_1 = \{a = a_0\}$, $H_2 = \{a < a_0\}$ и $H_3 = \{a > a_0\}$.

8.9. Гипотеза о среднем нормальной совокупности с неизвестной дисперсией

Проверяется та же гипотеза, что и в предыдущем разделе, но в случае, когда дисперсия σ^2 неизвестна. Критерий, который мы построим, тоже называют критерием Стьюдента, только одновыборочным.

Введем функцию отклонения $\rho(\mathbf{X})$ равную

$$\rho(\mathbf{X}) = \sqrt{n}\frac{\overline{X} - a_0}{\sqrt{S_0^2}}, \qquad \text{где } S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Сразу по п. 4 следствия леммы Фишера имеем K1(a): если $a = a_0$, то ρ имеет распределение Стьюдента T_{n-1}.

Упражнение. Доказать свойство K1(б).

Критерий строится в точности как в (28), но в качестве C следует брать квантиль распределения Стьюдента, а не стандартного нормального распределения. $\boxed{\text{почему?}}$

Упражнение. Нарисовать критерий и доказать, что этот критерий имеет точный размер ε и является состоятельным.

Упражнение. В самом ли деле три последних критерия состоятельны?

Напоминание. А вы доказали выполнение свойства K1(б) для функций отклонения этих критериев, чтобы говорить о состоятельности?

Примечание. А что такое «состоятельность» критерия?

8.10. Критерии, основанные на доверительных интервалах

Имеется выборка $\mathbf{X} = (X_1, \ldots, X_n)$ из семейства распределений $\mathcal{F}_\theta$. Проверяется простая гипотеза $H_1 = \{\theta = \theta_0\}$ против сложной альтернативы $H_2 = \{\theta \neq \theta_0\}$.

Пусть имеется точный (асимптотически точный) доверительный интервал (θ^-, θ^+) для параметра θ уровня доверия $1 - \varepsilon$. Взяв произвольное θ', для выборки из распределения $\mathcal{F}_{\theta'}$ имеем

$$\mathsf{P}_{\theta'}(\theta^- < \theta' < \theta^+) = 1 - \varepsilon \quad (\to 1 - \varepsilon).$$

Тогда критерий

$$\delta(\mathbf{X}) = \begin{cases} H_1, & \text{если } \; \theta_0 \in (\theta^-, \theta^+), \\ H_2, & \text{если } \; \theta_0 \notin (\theta^-, \theta^+) \end{cases}$$

имеет точный (асимптотический) размер ε. Действительно,

$$\alpha_1(\delta) = \mathsf{P}_{H_1}(\delta{=}H_2) = \mathsf{P}_{\theta_0}(\theta_0 \notin (\theta^-, \theta^+)) = 1 - \mathsf{P}_{\theta_0}(\theta^- {<}\, \theta_0 \,{<}\, \theta^+) = \varepsilon \ (\to \varepsilon).$$

Если доверительный интервал строится с помощью «функции отклонения» $G(\mathbf{X}, \theta)$, то эта же функция годится и в качестве «функции отклонения» $\rho(\mathbf{X})$ для построения критерия согласия.

Пример 33. Посмотрим на критерий (28). Основная гипотеза H_1 принимается, только если $|\rho(\mathbf{X})| < C = \tau_{1-\varepsilon/2}$, что равносильно неравенству

$$\left|\sqrt{n}\,\frac{\overline{X} - a_0}{\sigma}\right| < \tau_{1-\varepsilon/2}, \qquad \text{или} \qquad \overline{X} - \frac{\tau_{1-\varepsilon/2}\sigma}{\sqrt{n}} < a_0 < \overline{X} + \frac{\tau_{1-\varepsilon/2}\sigma}{\sqrt{n}}.$$

Сравните то, что получилось, с точным доверительным интервалом (13) для параметра a нормального распределения с известной дисперсией.

9. Исследование статистической зависимости

Часто требуется определить, как зависит наблюдаемая случайная величина от одной или нескольких других величин. Самый общий случай такой зависимости — зависимость статистическая: например, $X = \xi + \eta$ и $Z = \xi + \phi$ зависимы, но эта зависимость не функциональная.

Для зависимых случайных величин имеет смысл рассмотреть математическое ожидание одной из них при фиксированном значении другой (других). Такое условное математическое ожидание показывает, как влияет на среднее значение первой величины изменение значений второй. Скажем, стоимость квартиры зависит от площади, этажа, района и других параметров, но не является функцией от них. Зато в широких предположениях можно считать ее математическое ожидание функцией от этих величин. Разумеется, наблюдать это среднее значение мы не можем — в нашей власти лишь наблюдать значения первой случайной величины при разных значениях остальных. Эту зависимость можно воображать как вход и выход некоторой машины — «ящика с шуршавчиком». Входные данные, или «факторы», как правило, известны. На выходе мы наблюдаем результат преобразования входных данных в ящике по каким-либо правилам.

9.1. Математическая модель регрессии

Пусть наблюдаемая случайная величина X зависит от случайной величины или случайного вектора Z. Значения Z мы либо задаем, либо наблюдаем. Обозначим через $f(t)$ функцию, отражающую зависимость среднего значения X от значений Z:

$$\mathsf{E}(X \mid Z = t) = f(t). \tag{29}$$

Функция $f(t)$ называется линией регрессии X на Z, а уравнение $x = f(t)$ — регрессионным уравнением. После n экспериментов, в которых Z последовательно принимает значения $Z = t_1, \ldots, Z = t_n$, получим значения наблюдаемой величины X, равные $X_1, \ldots, X_n$. Обозначим через ε_i разницу $X_i - \mathsf{E}(X \mid Z = t_i) = X_i - f(t_i)$ между наблюдаемой в i-м эксперименте случайной величиной и ее математическим ожиданием.

Итак, $X_i = f(t_i) + \varepsilon_i$, $i = 1, \ldots, n$, где ε_i — ошибки наблюдения, равные в точности разнице между реальным и усредненным значением случайной величины X при значении $Z = t_i$. Про совместное распределение $\varepsilon_1, \ldots, \varepsilon_n$ обычно что-либо известно или предполагается: например, что вектор ошибок $\boldsymbol{\varepsilon}$ состоит из независимых и одинаково нормально распределенных случайных величин с нулевым средним. Нулевое среднее тут необходимо:

$$\mathsf{E}\,\varepsilon_i = \mathsf{E}\,X_i - f(t_i) = \mathsf{E}(X \mid Z = t_i) - \mathsf{E}(X \mid Z = t_i) = 0.$$

Требуется по значениям $t_1, \ldots, t_n$ и $X_1, \ldots, X_n$ оценить как можно точнее функцию $f(t)$. Величины t_i не являются случайными, так что вся случайность сосредоточена в неизвестных ошибках ε_i и в наблюдаемых X_i. Но пытаться в классе всех возможных функций восстанавливать $f(t)$ по «наилучшим оценкам» для $f(t_i)$ довольно глупо — наиболее точными приближениями к $f(t_i)$ оказываются X_i, и функция $f(t)$ будет просто ломаной, построенной по точкам (t_i, X_i). Поэтому сначала заранее определяют вид функции $f(t)$. Часто предполагают, что $f(t)$ есть полином (редко больше третьей или четвертой степени) с неизвестными коэффициентами. Будем пока предполагать, что функция $f(t)$ полностью определяется неизвестными параметрами $\theta_1, \ldots, \theta_k$.

9.2. Метод максимального правдоподобия

Оценки неизвестных параметров находят с помощью метода максимального правдоподобия. Он предписывает выбирать неизвестные параметры так, чтобы максимизировать функцию правдоподобия случайного вектора $X_1, \ldots, X_n$.

Будем, для простоты, предполагать, что вектор ошибок $\boldsymbol{\varepsilon}$ состоит из независимых и одинаково распределенных случайных величин с плотностью распределения $h(x)$ из некоторого семейства распределений с нулевым средним и, вообще говоря, неизвестной дисперсией. Очень часто полагают, что ε_i имеют симметричное распределение — нормальное N_{0,σ^2}, Стьюдента, Лапласа, логистическое и т.п. Поскольку X_i от ε_i зависят линейно, то распределение X_i окажется таким же, как у ε_i, но с центром уже не в нуле, а в точке $f(t_i)$. Поэтому X_i имеет плотность $h\big(x - f(t_i)\big)$, и функция правдоподобия вектора $X_1, \ldots, X_n$ равна, в силу независимости координат,

$$f(X_1, \ldots, X_n; \theta_1, \ldots, \theta_k) = h\big(X_1 - f(t_1)\big) \cdot \ldots \cdot h\big(X_n - f(t_n)\big) = h(\varepsilon_1) \cdot \ldots \cdot h(\varepsilon_n). \quad (30)$$

Если величины ε_i имеют разные распределения, то h следует заменить на соответствующие h_i. В отсутствие независимости произведение плотностей в (30) заменится плотностью совместного распределения координат вектора $\boldsymbol{\varepsilon}$.

Метод максимального правдоподобия предписывает находить оценки неизвестных параметров θ_i функции $f(t)$ и оценки неизвестной дисперсии (или дисперсий) $\mathsf{D}\,\varepsilon_i$, максимизируя по этим параметрам функцию правдоподобия (30). Рассмотрим, во что превращается метод максимального правдоподобия в наиболее частых на практике предположениях.

9.3. Метод наименьших квадратов

Предположим, что вектор ошибок $\boldsymbol{\varepsilon}$ состоит из независимых случайных величин с нормальным распределением N_{0,σ^2}. Функция правдоподобия (30) имеет вид

$$f(\mathbf{X}; \boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(X_1 - f(t_1))^2}{2\sigma^2}\right\} \cdot \ldots \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(X_1 - f(t_n))^2}{2\sigma^2}\right\} = \\ = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - f(t_i))^2\right\}.$$

Очевидно, что при любом фиксированном σ^2 максимум функции правдоподобия достигается при наименьшем значении суммы квадратов ошибок $\sum (X_i - f(t_i))^2 = \sum \varepsilon_i^2$.

Определение 31. Оценкой метода наименьших квадратов (ОМНК) для неизвестных параметров $\theta_1, \ldots, \theta_k$ уравнения регрессии называется набор значений параметров, доставляющий минимум сумме квадратов отклонений

$$\sum_{i=1}^{n} (X_i - f(t_i))^2 = \sum_{i=1}^{n} \varepsilon_i^2.$$

Найдя оценки для θ_i, найдем тем самым оценку $\hat{f}(t)$ для $f(t)$. Обозначим через $\hat{f}(t_i)$ значения этой функции, и через $\hat{\varepsilon}_i = X_i - \hat{f}(t_i)$ соответствующие оценки ошибок. Оценка максимального правдоподобия для σ^2, она же точка максимума по σ^2 функции правдоподобия, равна вычислить!

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{f}(t_i))^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2. \quad (31)$$

Мудрый читатель понял, что основная цель рассмотренного выше примера — показать, что метод наименьших квадратов не падает с неба, а есть в точности метод максимального правдоподобия в случае, когда вектор ошибок, а вместе с ним и вектор наблюдаемых откликов регрессии, состоит из независимых и одинаково распределенных случайных величин с нормальным распределением.

Пример 34. Пусть плотность независимых случайных величин ε_i имеет вид

$$h(x) = \frac{1}{2\sigma}\,\exp\{-|x|/\sigma\}, \ \text{т. е. } \varepsilon_i \text{ имеют распределение Лапласа.}$$

Тогда при любом фиксированном σ^2 максимум функции правдоподобия достигается при наименьшем значении суммы $\sum|X_i - f(t_i)|$ абсолютных отклонений. Оценка максимального правдоподобия (ОМП) для набора $\theta_1, \ldots, \theta_k$ уже не есть ОМНК.

9.4. Примеры

Найдем ОМНК для функций $f(t)$ в ряде частных случаев. Напомним, что ОМП совпадает с ОМНК почти исключительно в случае нормальности вектора ошибок.

Пример 35. Пусть функция $f(t) = \theta$ — постоянная, θ — неизвестный параметр. Тогда наблюдения равны $X_i = \theta + \varepsilon_i$, $i = 1, \ldots, n$. Легко узнать задачу оценивания неизвестного математического ожидания θ по выборке из независимых и одинаково распределенных случайных величин $X_1, \ldots, X_n$. Найдем ОМНК $\hat\theta$ для параметра θ:

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n}(X_i-\theta)^2 = -2\sum_{i=1}^{n}(X_i-\theta)\bigg|_{\theta=\hat\theta} = 0 \ \text{ при } \ \hat\theta = \overline{X}.$$

Трудно назвать этот ответ неожиданным. Соответственно, $\hat\sigma^2 = S^2$.

Упражнение. Покажите, что в условиях примера 34 ОМП для θ, минимизирующая $\sum|X_i - \theta|$, есть выборочная медиана

$$\hat\theta = \begin{cases} X_{(m)}, & \text{если } n = 2m-1 \text{ (нечётно)},\\ \frac12\left(X_{(m)}+X_{(m+1)}\right), & \text{если } n = 2m \text{ (чётно)}, \end{cases}$$

а ОМП для дисперсии равна $\hat\sigma^2 = \frac{1}{n}\sum\limits_{i=1}^{n}|X_i - \hat\theta|$. Вместо полусуммы можно брать любую точку отрезка $\left[X_{(m)},\, X_{(m+1)}\right]$.

Пример 36. Линейная регрессия.
Рассмотрим линейную регрессию $X_i = \theta_1 + t_i\theta_2 + \varepsilon_i$, $i = 1, \ldots, n$, где θ_1 и θ_2 — неизвестные параметры. Здесь $f(t) = \theta_1 + t\theta_2$ — прямая.

Найдем оценку метода наименьших квадратов $\hat\theta_1, \hat\theta_2$, на которой достигается минимум величины $\sum \varepsilon_i^2 = \sum(X_i - \theta_1 - t_i\theta_2)^2$. Приравняв к нулю частные производные этой суммы по параметрам, найдем точку экстремума.

Упражнение. Убедиться, что решением системы уравнений

$$\frac{\partial}{\partial\theta_1}\sum_{i=1}^{n}\varepsilon_i^2 = 0\,, \qquad \frac{\partial}{\partial\theta_2}\sum_{i=1}^{n}\varepsilon_i^2 = 0$$

является пара

$$\hat\theta_2 = \frac{\frac1n\sum t_iX_i - \overline{X}\cdot\overline{t}}{\frac1n\sum(t_i-\overline{t})^2}\,, \qquad \hat\theta_1 = \overline{X} - \overline{t}\hat\theta_2.$$

Определение 32. Величина

$$\rho^* = \frac{\frac{1}{n}\sum t_i X_i - \overline{X}\cdot\overline{t}}{\sqrt{\frac{1}{n}\sum(t_i-\overline{t})^2\cdot\frac{1}{n}\sum(X_i-\overline{X})^2}}$$

называется выборочным коэффициентом корреляции и характеризует степень линейной зависимости между наборами чисел $X_1,\ldots,X_n$ и $t_1,\ldots,t_n$.

Пример 37. Термин «регрессия» появился впервые в работе Francis Galton, “Regression towards mediocrity in hereditary stature” (Journal of the Anthropological Institute V. 15, p. 246–265, 1886).

Гальтон исследовал, в частности, рост детей высоких родителей и установил, что он «регрессирует» в среднем, т. е. в среднем дети высоких родителей не так высоки, как их родители. Пусть X — рост сына (дочери), а Z_1 и Z_2 — рост отца и матери. Для линейной модели регрессии $\mathsf{E}\,(X \mid Z_1 = t, Z_2 = u) = f(t,u) = \theta_1 t + \theta_2 u + c$ Гальтон нашел оценки параметров:

$$\mathsf{E}\,(\text{роста сына} \mid Z_1 = t, Z_2 = u) = 0,27t + 0,2u + \mathrm{const},$$

а средний рост дочери еще в 1,08 раз меньше.

9.5. Общая модель линейной регрессии

Введем два вектора: $\mathbf{Z} = (Z_1,\ldots,Z_k)$ — факторы регрессии и $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_k)$ — неизвестные параметры регрессии. Каждый вектор есть вектор-столбец, а изображен по горизонтали для удобства. Обозначать вектора мы, как и ранее, будем жирным шрифтом.

Рассматривается модель регрессии, которая в курсе «Эконометрика» называется простой (линейной) регрессией:

$$\mathsf{E}\,(X \mid \mathbf{Z} = \mathbf{t}) = f(\mathbf{t}) = \beta_1\,t_1 + \ldots + \beta_k\,t_k, \quad \text{или} \quad \mathsf{E}\,(X \mid \mathbf{Z}) = f(\mathbf{Z}) = \beta_1\,Z_1 + \ldots + \beta_k\,Z_k.$$

Пусть в i-м эксперименте факторы регрессии принимают заранее заданные значения $\mathbf{Z}^{(i)} = (Z_1^{(i)},\ldots,Z_k^{(i)})$, где $i = 1,\ldots,n$.

После $n \geqslant k$ экспериментов получен набор откликов $\mathbf{X} = (X_1,\ldots,X_n)$, где

$$\begin{cases} X_1 = \beta_1\,Z_1^{(1)} + \ldots + \beta_k\,Z_k^{(1)} + \varepsilon_1\\ X_2 = \beta_1\,Z_1^{(2)} + \ldots + \beta_k\,Z_k^{(2)} + \varepsilon_2\\ \ldots\\ X_n = \beta_1\,Z_1^{(n)} + \ldots + \beta_k\,Z_k^{(n)} + \varepsilon_n, \end{cases}$$

или, в матричной форме, $\mathbf{X} = Z^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, где матрица $Z(k \times n)$ (матрица плана) равна

$$Z = \begin{pmatrix} Z_1^{(1)} & \ldots & Z_1^{(n)}\\ \vdots & \ldots & \vdots\\ Z_k^{(1)} & \ldots & Z_k^{(n)} \end{pmatrix} = (\mathbf{Z}^{(1)} \;\ldots\; \mathbf{Z}^{(n)}).$$

Вектор $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)$ состоит из случайных ошибок в данных экспериментах.

Требуется по данным матрице плана Z и вектору результатов $\mathbf{X}$ найти оценки для параметров регрессии β и параметров распределения вектора ошибок ε.

9.6. Метод наименьших квадратов. Нормальное уравнение

Предположение 1. Матрица Z имеет ранг k, т. е. все k ее строк линейно независимы.

Лемма 3. Предположение 1 означает, что матрица $A = Z{\cdot}Z^T$ положительно определена.

Напоминание 1. Матрица $A(k \times k)$ положительно определена, если $\mathbf{t}^T A \mathbf{t} \geqslant 0$ для любого $\mathbf{t} = (t_1, \ldots, t_k)$, причем $\mathbf{t}^T A \mathbf{t} = 0$, если и только если $\mathbf{t} = \mathbf{0}$.

Напоминание 2. Квадрат нормы вектора $\mathbf{u}$ равен $\|\mathbf{u}\|^2 = \mathbf{u}^T\mathbf{u} = \sum u_i^2 \geqslant 0$. Норма равна нулю, если и только если $\mathbf{u} = \mathbf{0}$.

Доказательство леммы 3. Благодаря напоминанию 2,

$$\mathbf{t}^T A \mathbf{t} = \mathbf{t}^T Z{\cdot}Z^T \mathbf{t} = (Z^T\mathbf{t}\ )^T \cdot (Z^T\mathbf{t}\ ) = \|Z^T\mathbf{t}\|^2 \geqslant 0,$$

причем $\|Z^T\mathbf{t}\| = 0$, если и только если $Z^T\mathbf{t} = \mathbf{0}$. Но «ранг Z равен k» как раз и означает, по определению, что $Z^T\mathbf{t} = \mathbf{0}$ тогда и только тогда, когда $\mathbf{t} = \mathbf{0}$. □

Скоро нам пригодится корень из матрицы A, существование которого гарантирует

Лемма 4. Положительная определенность и симметричность матрицы A влекут существование вещественной симметричной матрицы $\sqrt{A}$ такой, что $\sqrt{A}\sqrt{A} = A$.

Действительно, матрица A симметрична, поскольку $A = ZZ^T$ и $A^T = A$. Существование[o] матрицы $\sqrt{A}$ с нужными свойствами следует из возможности привести A ортогональными преобразованиями $A = Q^T D Q$ к диагональному виду с положительными, в силу положительной определенности, собственными значениями A на диагонали D. Тогда $\sqrt{A} = Q^T\sqrt{D}Q$. □

Найдем ОМНК $\hat{\beta}$ для вектора β, доставляющий минимум функции $S(\beta)$, равной

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \|\varepsilon\|^2 = \|\mathbf{X} - Z^T\beta\|^2 = (\mathbf{X} - Z^T\beta)^T \cdot (\mathbf{X} - Z^T\beta).$$

Вместо того, чтобы искать точку экстремума функции $S(\beta)$ дифференцированием по β_i, заметим следующее. Величина $S(\beta)$ есть квадрат расстояния от точки $\mathbf{X} \in \mathbb{R}^n$ до точки $Z^T\beta$ — одной из точек линейного подпространства (гиперплоскости) в $\mathbb{R}^n$ с координатами вида $Z^T\mathbf{t}$, где $\mathbf{t} \in \mathbb{R}^k$.

Минимальное расстояние $S(\hat{\beta})$ мы получим, когда вектор $\mathbf{X} - Z^T\hat{\beta}$ будет ортогонален всем векторам этого подпространства, т. е. когда для любого $\mathbf{t} \in \mathbb{R}^k$ скалярное произведение векторов $Z^T\mathbf{t}$ и $\mathbf{X} - Z^T\hat{\beta}$ обратится в ноль. Запишем это скалярное произведение в матричном виде

$$\left(Z^T\mathbf{t}, \mathbf{X} - \hat{\beta}\right) = \left(Z^T\mathbf{t}\right)^T\left(\mathbf{X} - Z^T\hat{\beta}\right) = \mathbf{t}^T \cdot \left(Z\mathbf{X} - ZZ^T\hat{\beta}\right) = 0.$$

[o]См., например, А. И. Мальцев «Основы линейной алгебры», раздел «Унитарные и евклидовы пространства», параграф «Унитарные и симметрические преобразования», теорема 7.

Подставляя в качестве $\mathbf{t}$ базисные вектора в $\mathbb{R}^k$ вида $(0, \ldots, 0, 1, 0, \ldots, 0)$, сразу же получим, что все координаты вектора $Z\mathbf{X} - ZZ^T\widehat{\boldsymbol{\beta}}$ равны нулю. Итак, оценка метода наименьших квадратов $\widehat{\boldsymbol{\beta}}$ есть любое решение уравнения

$$ZZ^T\widehat{\boldsymbol{\beta}} = Z\mathbf{X} \quad \text{или} \quad A\widehat{\boldsymbol{\beta}} = Z\mathbf{X}. \tag{32}$$

По лемме 3, уравнение (32) имеет единственное решение

$$\widehat{\boldsymbol{\beta}} = A^{-1}Z\mathbf{X} \tag{33}$$

в том и только в том случае, когда матрица $Z(k \times n)$ имеет полный ранг k, где $k \leqslant n$. Уравнение (32) называется нормальным уравнением.

В предположении, что вектор ошибок $\boldsymbol{\varepsilon}$ состоит из независимых случайных величин с нормальным распределением N_{0,σ^2} с одной и той же дисперсией, ОМНК совпадает с оценкой максимального правдоподобия, а ОМП для σ^2, согласно (31), равна

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}\|\mathbf{X} - Z^T\widehat{\boldsymbol{\beta}}\|^2 = \frac{1}{n}S(\widehat{\boldsymbol{\beta}}). \tag{34}$$

9.7. Свойства ОМНК

Отметим несколько свойств, которые, возможно, нам понадобятся в дальнейшем.

1. Разница $\widehat{\boldsymbol{\beta}}$ и $\boldsymbol{\beta}$ равна $A^{-1}Z\boldsymbol{\varepsilon}$:

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = A^{-1}Z\mathbf{X} - \boldsymbol{\beta} = A^{-1}Z(Z^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} = A^{-1}A\boldsymbol{\beta} + A^{-1}Z\boldsymbol{\varepsilon} - \boldsymbol{\beta} = A^{-1}Z\boldsymbol{\varepsilon}.$$

2. Если $\mathsf{E}\,\boldsymbol{\varepsilon} = 0$, то $\widehat{\boldsymbol{\beta}}$ — несмещенная оценка для $\boldsymbol{\beta}$: $\mathsf{E}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + A^{-1}Z\mathsf{E}\,\boldsymbol{\varepsilon} \equiv \boldsymbol{\beta}$.

Пусть выполнены предположения **1** и **2**:

Предположение 2. Вектор ошибок $\boldsymbol{\varepsilon}$ состоит из независимых случайных величин с нормальным распределением N_{0,σ^2} с одной и той же дисперсией.

Напоминание 3. Для произвольного случайного вектора $\boldsymbol{x}$, координаты которого имеют вторые моменты, матрица ковариаций $\mathsf{D}\,\boldsymbol{x} = \mathsf{E}\left((\boldsymbol{x} - \mathsf{E}\,\boldsymbol{x})(\boldsymbol{x} - \mathsf{E}\,\boldsymbol{x})^T\right)$ — это матрица, чей (i,j)-й элемент равен

$$\mathrm{cov}(x_i, x_j) = \mathsf{E}\,(x_i - \mathsf{E}\,x_i)(x_j - \mathsf{E}\,x_j).$$

В частности, $\mathsf{D}\,\boldsymbol{\varepsilon} = \sigma^2 \cdot E_n$, где E_n — единичная $(n \times n)$-матрица.

3. Матрица ковариаций вектора $\sqrt{A}\widehat{\boldsymbol{\beta}}$ равна σ^2E_k:

$$\mathsf{D}\sqrt{A}\widehat{\boldsymbol{\beta}} = \mathsf{E}\left(\sqrt{A}\widehat{\boldsymbol{\beta}} - \mathsf{E}\sqrt{A}\widehat{\boldsymbol{\beta}}\right)\left(\sqrt{A}\widehat{\boldsymbol{\beta}} - \mathsf{E}\sqrt{A}\widehat{\boldsymbol{\beta}}\right)^T = \mathsf{E}\left(\sqrt{A}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)\left(\sqrt{A}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)^T =$$

$$= \mathsf{E}\left(\sqrt{A}A^{-1}Z\boldsymbol{\varepsilon}\right)\left(\sqrt{A}A^{-1}Z\boldsymbol{\varepsilon}\right)^T = \sqrt{A}A^{-1}Z\,\mathsf{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right)Z^T{A^{-1}}^T\sqrt{A}^T.$$

И так как $A^T = A$, $\mathsf{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T = \sigma^2E_n$, то $\mathsf{D}\sqrt{A}\widehat{\boldsymbol{\beta}} = \sigma^2 \cdot \sqrt{A}A^{-1}ZZ^TA^{-1}\sqrt{A} = \sigma^2E_k$.

Свойство **3** означает, что координаты вектора $\sqrt{A}\widehat{\boldsymbol{\beta}}$ некоррелированы. Сформулируем дальнейшее следствие этого свойства первым пунктом следующей теоремы. С утверждениями второго и третьего пунктов читатель встретится в следующем семестре многократно.

Теорема 13.

1. Вектор $\dfrac{1}{\sigma}\sqrt{A}(\hat\beta-\beta)$ имеет k-мерное стандартное нормальное распределение, т. е. состоит из k независимых случайных величин с распределением $\mathsf{N}_{0,1}$.

2. Величина $\dfrac{n\hat\sigma^2}{\sigma^2}=\dfrac{1}{\sigma^2}\|\mathbf{X}-Z^T\hat\beta\|^2$ имеет распределение χ^2 с $n-k$ степенями свободы и не зависит от $\hat\beta$.

3. Оценка $(\sigma^2)^*=\dfrac{n\hat\sigma^2}{n-k}=\dfrac{1}{n-k}\|\mathbf{X}-Z^T\hat\beta\|^2$ является несмещенной оценкой для σ^2.

Доказательство теоремы 13.

1. Вектор $\sqrt{A}(\hat\beta-\beta)=\sqrt{A}A^{-1}Z\varepsilon=(\sqrt{A})^{-1}Z\varepsilon$ есть линейное преобразование нормального вектора ε и поэтому имеет нормальное совместное распределение. По свойству **3**, матрица ковариаций этого вектора есть σ^2E_k, поэтому матрица ковариаций нормированного вектора $\sqrt{A}(\hat\beta-\beta)/\sigma$ есть просто E_k, а математическое ожидание равно нулю по свойству **2**. Напомним, что координаты многомерного нормального вектора независимы тогда и только тогда, когда они некоррелированы — см. теорему 14. Первое утверждение теоремы доказано.

2. По построению ОМНК, вектор $X-Z^T\hat\beta$ ортогонален любому вектору вида Z^Tt. В частности, он ортогонален вектору $Z^T(\hat\beta-\beta)$. По теореме Пифагора, для треугольника с такими катетами сумма квадратов их длин равна квадрату длины гипотенузы:
$$\|X-Z^T\hat\beta\|^2+\|Z^T(\hat\beta-\beta)\|^2=\|X-Z^T\hat\beta+Z^T(\hat\beta-\beta)\|^2=\|X-Z^T\beta\|^2.$$
Поэтому
$$\|X-Z^T\hat\beta\|^2=\|X-Z^T\beta\|^2-\|Z^T(\hat\beta-\beta)\|^2=\|\varepsilon\|^2-\|Z^T(\hat\beta-\beta)\|^2. \tag{35}$$
Но квадрат нормы $\|Z^T(\hat\beta-\beta)\|^2$ равен квадрату нормы $\|\sqrt{A}(\hat\beta-\beta)\|^2$:
$$\|Z^T(\hat\beta-\beta)\|^2=(\hat\beta-\beta)^TZZ^T(\hat\beta-\beta)=(\hat\beta-\beta)^T\sqrt{A}^T\sqrt{A}(\hat\beta-\beta)=$$
$$=\|\sqrt{A}(\hat\beta-\beta)\|^2=\|(\sqrt{A})^{-1}Z\varepsilon\|^2.$$
Осталось заметить, что строки $(k\times n)$-матрицы $(\sqrt{A})^{-1}Z$ ортогональны:
$$\left((\sqrt{A})^{-1}Z\right)\left((\sqrt{A})^{-1}Z\right)^T=(\sqrt{A})^{-1}ZZ^T(\sqrt{A})^{-1}=E_k,$$
поэтому k её строк можно дополнить до некоторой ортогональной $(n\times n)$-матрицы C. Первые k координат n-мерного вектора $\mathbf{Y}=C\varepsilon/\sigma$ совпадают с вектором $(\sqrt{A})^{-1}Z\varepsilon/\sigma$. В результате из (35) получим
$$\frac{n\hat\sigma^2}{\sigma^2}=\frac{1}{\sigma^2}\|X-Z^T\hat\beta\|^2=\|\varepsilon/\sigma\|^2-\|(\sqrt{A})^{-1}Z\varepsilon/\sigma\|^2=\sum_{i=1}^n\left(\frac{\varepsilon_i}{\sigma}\right)^2-Y_1^2-\ldots-Y_k^2. \tag{36}$$
Не забудьте, что вектор ε/σ имеет n-мерное стандартное нормальное распределение. Тогда вся разность (36) по лемме Фишера имеет распределение χ^2 с $n-k$ степенями свободы и не зависит от вычитаемого, т. е. от случайного вектора ε (и от $\hat\beta$ тоже, поскольку $\hat\beta$ есть функция ε).

3. Напомним, что $\mathsf{E}\,\chi^2_{n-k}=n-k$. Отсюда и из второго утверждения теоремы получим
$$\mathsf{E}\,(\sigma^2)^*=\mathsf{E}\left(\frac{n\hat\sigma^2}{n-k}\right)=\frac{\sigma^2}{n-k}\,\mathsf{E}\left(\frac{n\hat\sigma^2}{\sigma^2}\right)=\frac{\sigma^2}{n-k}\cdot(n-k)=\sigma^2.$$

A. Многомерное нормальное распределение

Определение 33. Пусть случайный вектор $\xi = (\xi_1, \ldots, \xi_m)$ имеет вектор средних $\mathbf{a} = \mathsf{E}\,\xi$ и невырожденную матрицу ковариаций Σ, составленную из элементов $\Sigma_{ij} = \mathrm{cov}(\xi_i, \xi_j)$. Говорят, что вектор ξ имеет нормальное распределение $\mathsf{N}_{\mathbf{a},\Sigma}$ в $\mathbb{R}^m$, если плотность этого вектора равна

$$f_\xi(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^m \sqrt{|\det \Sigma|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{a})^T \Sigma^{-1}(\mathbf{x}-\mathbf{a})\right\}, \quad \text{где} \quad \mathbf{x} \in \mathbb{R}^m$$

Квадратичная форма $(\mathbf{x}-\mathbf{a})^T \Sigma^{-1}(\mathbf{x}-\mathbf{a})$ в показателе экспоненты равна

$$(\mathbf{x}-\mathbf{a})^T \Sigma^{-1}(\mathbf{x}-\mathbf{a}) = \sum_{i,j} (x_i - a_i) \cdot (\Sigma^{-1})_{ij} \cdot (x_j - a_j). \tag{37}$$

Замечание 24. Вектор, составленный из нормальных случайных величин, не обязательно имеет многомерное нормальное распределение. Так, вектор $(\xi, c\xi)$ имеет вырожденную матрицу ковариаций $\left(\begin{smallmatrix} 1 & c \\ c & c^2 \end{smallmatrix}\right)$, если $\mathsf{D}\,\xi = 1$, и не имеет плотности в $\mathbb{R}^2$.

Поэтому в условиях следующей теоремы существование совместной нормальной плотности координат вектора обязательно.

Теорема 14. Пусть вектор ξ имеет многомерное нормальное распределение $\mathsf{N}_{\mathbf{a},\Sigma}$. Координаты этого вектора независимы тогда и только тогда, когда они некоррелированы, т. е. когда матрица ковариаций Σ диагональна.

Замечание 25. Следовало бы крикнуть «ура!»: свойство, о котором мы давно мечтали, — чтобы независимость следовала из некоррелированности, — имеет все же место. Но только для наборов случайных величин с нормальным совместным распределением, и это — очередное изумительное качество нормального распределения.

Доказательство. Только в случае диагональной матрицы Σ с элементами $\Sigma_{ii} = \sigma_i^2 = \mathsf{D}\,\xi_i$ квадратичная форма (37) превращается в сумму квадратов

$$\sum_{i,j} (x_i - a_i) \cdot (\Sigma^{-1})_{ij} \cdot (x_j - a_j) = \sum_i \frac{(x_i - a_i)^2}{\sigma_i^2},$$

и многомерная плотность распадается в произведение плотностей координат. □

Надеюсь, читатель поверит следующей теореме без доказательства, будучи в состоянии доказать ее, как и в одномерном случае, с помощью многомерных характеристических функций.

Многомерная центральная предельная теорема.
Пусть $\xi^{(1)}, \xi^{(2)}, \ldots$ — последовательность независимых и одинаково распределенных случайных векторов, каждый из которых имеет среднее $\mathsf{E}\,\xi^{(1)} = \mathbf{a}$ и невырожденную матрицу ковариаций Σ. Обозначим через $S_n = \xi^{(1)} + \ldots + \xi^{(n)}$ вектор частичных сумм.
Тогда при $n \to \infty$ имеет место слабая сходимость распределений векторов

$$\eta^{(n)} = \frac{S_n - n\mathbf{a}}{\sqrt{n}} \Rightarrow \eta, \quad \text{где} \quad \eta \text{ имеет распределение } \mathsf{N}_{0,\Sigma}.$$

В условиях многомерной ЦПТ распределение любых непрерывных функций $g\left(\eta^{(n)}\right)$ слабо сходится к распределению $g(\eta)$. В качестве $g(\mathbf{x})$ нам будет нужна только $g(\mathbf{x}) = \sum x_i^2 = \|\mathbf{x}\|^2$.

Следствие 5. В условиях многомерной ЦПТ имеет место сходимость $\|\eta^{(n)}\|^2 \Rightarrow \|\eta\|^2$.

Осталось доказать теорему Пирсона.

B. Доказательство теоремы Пирсона

План действий:

1. Сначала покажем, что величина $\rho = \sum_{j=1}^{k} (\nu_j - np_j)^2/np_j$ есть квадрат нормы некоторого вектора $\boldsymbol{\eta}^{(n)} = (S_n - n\boldsymbol{a})/\sqrt{n}$ в $\mathbb{R}^k$. Затем убедимся в том, что матрица ковариаций типичного слагаемого $\boldsymbol{\xi}^{(1)}$ в сумме S_n вырождена, что мешает использовать ЦПТ.

2. Найдем ортогональное преобразование C, приводящее $\boldsymbol{\xi}^{(1)}$ к виду $C \cdot \boldsymbol{\xi}^{(1)} = (\widehat{\boldsymbol{\xi}}^{(1)}, 0)$, где вектор $\widehat{\boldsymbol{\xi}}^{(1)} \in \mathbb{R}^{k-1}$ уже имеет невырожденную единичную матрицу ковариаций. В силу линейности умножения, вектор $\boldsymbol{\eta}^{(n)}$ тоже перейдет в вектор $C \cdot \boldsymbol{\eta}^{(n)} = (\widehat{\boldsymbol{\eta}}^{(n)}, 0)$ с нулевой последней координатой. Но его норма не изменится из-за ортогональности матрицы C.

3. К вектору сумм $\widehat{\boldsymbol{\eta}}^{(n)}$ применим многомерную ЦПТ. В пределе получим $(k-1)$-мерный нормальный вектор с нулевым средним и единичной матрицей ковариаций, т. е. составленный из независимых величин со стандартным нормальным распределением. Воспользуемся следствием 5 и тем, что квадрат нормы этого вектора имеет χ^2-распределение H_{k-1}.

Реализация:

1. С каждым элементом выборки X_i свяжем вектор-столбец $\boldsymbol{\xi}^{(i)}$:

$$\boldsymbol{\xi}^{(i)} = (\xi_1, \ldots, \xi_k) = \left(\frac{\mathbf{I}(X_i \in A_1) - p_1}{\sqrt{p_1}}, \ldots, \frac{\mathbf{I}(X_i \in A_k) - p_k}{\sqrt{p_k}}\right), \quad i = 1, 2, \ldots, n.$$

Получим n независимых и одинаково распределенных векторов. Среднее $\boldsymbol{a} = \mathsf{E}\,\boldsymbol{\xi}^{(1)}$ равно нулю, поскольку $\mathsf{E}\,\mathbf{I}(X_1 \in A_j) = p_j$ для любого $j = 1, \ldots, k$. Далее, $\nu_j = \sum_{i=1}^{n} \mathbf{I}(X_i \in A_j)$, поэтому

$$\left(\frac{\nu_1 - np_1}{\sqrt{np_1}}, \ldots, \frac{\nu_k - np_k}{\sqrt{np_k}}\right) = \frac{\sum_{i=1}^{n} \boldsymbol{\xi}^{(i)}}{\sqrt{n}} = \frac{S_n}{\sqrt{n}} = \frac{S_n - n\boldsymbol{a}}{\sqrt{n}}.$$

Найдем матрицу ковариаций вектора $\boldsymbol{\xi}^{(1)}$, составленную из элементов

$$\sigma_{ij} = \operatorname{cov}\left(\frac{\mathbf{I}(X_1 \in A_i) - p_i}{\sqrt{p_i}}, \frac{\mathbf{I}(X_1 \in A_j) - p_j}{\sqrt{p_j}}\right) = \frac{1}{\sqrt{p_i p_j}} \cdot \big((\mathsf{E}\,\mathbf{I}(X_1 \in A_i) \cdot \mathbf{I}(X_1 \in A_j) - p_i p_j\big) =$$

$$= \frac{1}{\sqrt{p_i p_j}} \cdot \begin{cases} p_i - p_i p_j, & \text{если } i = j, \\ 0 - p_i p_j, & \text{если } i \neq j \end{cases} = \begin{cases} 1 - p_i, & \text{если } i = j, \\ -\sqrt{p_i p_j}, & \text{если } i \neq j. \end{cases}$$

Вырождена эта матрица хотя бы оттого, что координаты вектора $\boldsymbol{\xi}^{(1)}$ линейно связаны:

$$\sum_{j=1}^{k} \sqrt{p_j}\,\xi_j = \sum_{j=1}^{k} \mathbf{I}(X_1 \in A_j) - \sum_{j=1}^{k} p_j = 1 - 1 = 0. \tag{38}$$

2. Из (38) мораль: если последняя строка ортогональной матрицы C будет иметь вид $(\sqrt{p_1}, \ldots, \sqrt{p_k})$ (что вполне возможно — норма такой строки равна единице), то после умножения C на $\boldsymbol{\xi}^{(1)}$ получим вектор с нулевой последней координатой — в точности (38).

При умножении вектора $\boldsymbol{\xi}$ на матрицу C слева его матрица ковариаций $\Sigma = \mathsf{E}\,\boldsymbol{\xi}\boldsymbol{\xi}^T$ перейдет в $B = C\Sigma C^T$. Убедимся, что, какой бы ни была ортогональная матрица C, в результате получим диагональную матрицу из нулей с единицами на главной диагонали, кроме элемента $b_{kk} = 0$. Ортогональность C означает, что для любых $m \neq k$ и $l \neq m$ имеют место равенства

$$\sum_{j=1}^{k} c_{mj} c_{kj} = \sum_{j=1}^{k} c_{mj} \sqrt{p_j} = 0, \quad \sum_{j=1}^{k} c_{mj}^2 = 1, \quad \sum_{j=1}^{k} c_{mj} c_{lj} = 0.$$

Учитывая, что il-й элемент матрицы C^T есть c_{li}, получим

$$b_{ml} = \sum_{i=1}^{k} \left(\sum_{j=1}^{k} c_{mj} \sigma_{ji}\right) c_{li} = \sum_{i=1}^{k} \left(\sum_{j \neq i} -c_{mj} \sqrt{p_i p_j} + c_{mi}(1 - p_i)\right) c_{li} =$$

$$= \sum_{i=1}^{k} \left(\sqrt{p_i}\big(\sum_{j \neq i} -c_{mj}\sqrt{p_j} - c_{mi}\sqrt{p_i}\big) + c_{mi}\right) c_{li} = \sum_{i=1}^{k} c_{li} \cdot \begin{cases} c_{mi}, & m \neq k \\ 0, & m = k \end{cases} =$$

$$= \begin{cases} 1, & m \neq k,\ m = l \\ 0, & m = k \text{ или } m \neq l \end{cases} = \begin{pmatrix} E_{k-1} & 0 \\ 0 & 0 \end{pmatrix}. \tag{39}$$

3. Осталось повторить то, что мы уже описали в плане: умножение $C \cdot \xi^{(1)} = (\hat{\xi}^{(1)}, 0)$ приводит к вектору с нулевой последней координатой по (38). Равенствами (39) мы показали, что вектор $\hat{\xi}^{(1)} \in \mathbb{R}^{k-1}$ имеет невырожденную единичную матрицу ковариаций E_{k-1}. Вектора $\hat{\xi}^{(1)}, \hat{\xi}^{(2)}, \ldots$ независимы, одинаково распределены, имеют нулевое среднее $C \cdot \mathsf{E}\,\xi^{(1)} = \mathbf{0}$. Все условия многомерной ЦПТ выполнены, поэтому

$$\hat{\eta}^{(n)} = \frac{\hat{\xi}^{(1)} + \ldots + \hat{\xi}^{(n)}}{\sqrt{n}} \Rightarrow \eta, \quad \text{где } \eta \text{ имеет распределение } \mathrm{N}_{\mathbf{0}, E_{k-1}}.$$

По следствию 5, норма вектора $\hat{\eta}^{(n)}$ слабо сходится к норме вектора η, состоящего, согласно теореме 14, из независимых случайных величин со стандартным нормальным распределением:

$$\|\hat{\eta}^{(n)}\|^2 \Rightarrow \|\eta\|^2 = \sum_{i=1}^{k-1} \eta_i^2 = \chi^2_{k-1}, \text{ где } \chi^2_{k-1} \text{ имеет распределение } \mathrm{H}_{k-1}. \qquad (40)$$

Распределение H_{k-1} возникло здесь по определению 16. Осталось заметить, что у векторов $\boldsymbol{\eta}^{(n)}$, $C \cdot \boldsymbol{\eta}^{(n)}$, $\hat{\boldsymbol{\eta}}^{(n)}$, связанных равенствами

$$C \cdot \boldsymbol{\eta}^{(n)} = \frac{C \cdot \xi^{(1)} + \ldots + C \cdot \xi^{(n)}}{\sqrt{n}} = (\hat{\boldsymbol{\eta}}^{(n)}, 0),$$

нормы одинаковы в силу (17): $\|\boldsymbol{\eta}^{(n)}\|^2 = \|C \cdot \boldsymbol{\eta}^{(n)}\|^2 = \|(\hat{\boldsymbol{\eta}}^{(n)}, 0)\|^2 = \|\hat{\boldsymbol{\eta}}^{(n)}\|^2$. И все эти нормы ведут себя так же как и (40).

Упражнение. Найти среди этих норм величину ρ из теоремы Пирсона.

Указатель терминов

Литература

[1] Боровков А.А. Математическая статистика. М.: Наука, 1984.

[2] Феллер В. Введение в теорию вероятностей и ее приложения. М.: Мир, Т.2, 1984.

[3] Коршунов Д.А., Чернова Н.И. Сборник задач и упражнений по математической статистике. Новосибирск: Изд-во Института математики им. С.Л.Соболева СО РАН, 2001.

[4] Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М.: Наука, 1965.