Задачи и упражнения по математической статистике

Широбоков М.Г.
ФПМИ МФТИ

21 ноября 2022 г.

Оглавление

Введение

В данном пособии содержатся задачи, упражнения и вопросы по «Математической статистике», которые автор предлагает студентам 4-го курса физтех-школы прикладной математики и информатики МФТИ на основе читаемого много лет курса кафедры математических основ управления. Задачи посвящены ключевым разделам математической статистики: проверке гипотез о согласии распределений, независимости случайных величин, однородности выборок, двух простых гипотез и сложных гипотез, а также свойствам и методам изучения точечных и интервальных оценок. По традиции этот курс делится на две части: первые несколько лекций и семинаров составляют содержание так называемого *первого задания* и посвящены в основном задачам проверки гипотез, а последующие лекции и семинары посвящаются в основном задачам построения оценок параметров распределений и относятся ко *второму заданию*. В конце пособия читатель может найти списки вопросов на понимание, которые автор предлагает студентам во время приема заданий и экзаменов. Кроме того, в конце приводятся вопросы для проведения диктантов: преподаватель называет пункт из списка, а студенты, не пользуясь материалами, самостоятельно пишут определения и формулировки теорем. Приводятся задачи повышенной сложности, предлагаемые автором студентам на сдачах заданий и экзаменах, при этом задачи, отмеченные звездочкой, сложнее не отмеченных. Замечания к тексту можно направлять на электронный адрес автора shirobokov@phystech.edu.

Список обозначений и сокращений

$(\Omega, \mathcal{F}, \mathbb{P})$ – вероятностное пространство (Ω – множество исходов, $\mathcal{F}$ – сигма-алгебра, $\mathbb{P}$ – вероятностная мера).
$\mathbb{E}X$ – математическое ожидание случайной величины X.
$\mathbb{D}X$ – дисперсия случайной величины X.
$\mathrm{cov}(X, Y)$ – корреляционный момент (ковариация) случайных величин X и Y.
$\overset{\circ}{X}$ – «центрированная» случайная величина X, то есть $\overset{\circ}{X} = X - \mathbb{E}X$.
$\mathrm{Be}(p)$ – распределение Бернулли.
$\mathrm{Bi}(n, p)$ – биномиальное распределение, $\mathrm{Bi}(1, p) = \mathrm{Be}(p)$.
$\mathrm{Beta}(a, b)$ – бета-распределение.
$\mathrm{Po}(\lambda)$ – распределение Пуассона.
$\mathrm{M}(n, p_1, \ldots, p_N)$ – полиномиальное распределение.
$\Gamma(\lambda, n)$ – гамма-распределение, $\Gamma(\lambda, 1) = \mathrm{Exp}$.
$\mathrm{U}(a, b)$ – равномерное непрерывное распределение на отрезке $[a, b]$.
$\mathrm{N}(\mu, \sigma^2)$ – нормальное (гауссовское) распределение с математическим ожиданием μ и дисперсией σ^2.
$\mathrm{Exp}(\lambda)$ – показательное распределение с параметром λ, плотность распределения $f(x) = \lambda \exp(-\lambda x)$, $x \geq 0$.
$\xrightarrow{\text{с.к.}}$ – сходимость в среднем квадратичном.
$\xrightarrow{\text{п.н.}}$ – сходимость почти наверное.
$\xrightarrow{\mathbb{P}}$ – сходимость по вероятности.
$\xrightarrow{d}$ – сходимость по распределению.
l.i.m. – предел в среднем квадратичном (limit in mean).
$\overset{d}{=}$ – равенство по распределению.
$\overset{\text{п.н.}}{=}$ – равенство почти наверное.
$\overline{X}$ – среднее выборочное, посчитанное по выборке X.
$I(A)$ – индикаторная функция события A: $I(A) = 1$, если A верно, и $I(A) = 0$ иначе.
$\mathrm{N}_p(\mu, \sigma^2)$, $\chi^2_p(n)$ и пр. – p-квантили соответствующих распределений.
с.к. – в среднем квадратичном.
п.н. – почти наверное.
х.ф. – характеристическая функция.
ЗБЧ – закон больших чисел.
ЦПТ – центральная предельная теорема.

1. Введение в математическую статистику

1.1. Введение

Математическая статистика занимается восстановлением численных характеристик случайных величин по их измерениям. Измерение понимается не в том смысле, какое имеет место в функциональном анализе, а в смысле получения конкретной реализации случайной величины на ее множестве значений. Например, каждый раз, когда мы бросаем кубик, мы получаем конкретные числа (1, 2, 3, 4, 5, 6) в какой-то последовательности. Это и есть измерения некоторой случайной величины, принимающей шесть значений. Математическая статистика начинается там, где мы задаем вопрос: с какими вероятностями выпадают грани вот этого реального, лежащего перед нами, кубика?

Сразу надо оговорить, что абсолютно точное восстановление распределения случайной величины теоретически невозможно[1]. Однако можно попробовать это распределение *оценить*, аппроксимировать, как аппроксимируют функции по известным ее значениям в отдельных точках. Поэтому в математической статистике не ставится задача определения параметров распределения, но только их оценки. Вообще оценок тоже можно придумать бесконечное множество. Какая из них лучше другой тоже изучается в математической статистике.

Предложить какую-нибудь оценку распределения легко. Но обосновать, почему эта оценка заслуживает употребления, – совсем иное дело. В этом нам помогают сведения из теории вероятностей: в первую очередь закон больших чисел, центральная предельная теорема, неравенства Чебышева и Маркова, а также свойства сходимости *почти наверное*, *по вероятности*, *распределению* и *в среднем квадратичном*. Рассмотрим показательный пример подкидывания монеты и оценки вероятности выпадения решки.

1.2. Пример с монеткой

Рассмотрим монету, которая может падать только на решку или орел. Чтобы найти вероятность выпадения решки, рассмотрим случайную величину $\xi \in \mathrm{Be}(p)$, которая равна 1 при выпадении решки и 0 при выпадении орла. Параметр p в распределении Бернулли – это

[1] Ну, может, только в вырожденных случаях: например, когда известно, что случайная величина может принимать только одно значение и его достаточно определить по одному измерению. Но вряд ли такую величину можно назвать «случайной». На практике такие случаи неинтересны.

вероятность события $\xi = 1$, то есть вероятность выпадения решки. По условию, p неизвестно, и его требуется.

Попробуем оценить p просто поделив количество выпадений решки на общее количество подбрасываний монеты. Это очень естественный подход, и как мы потом увидим, он лежит в основе чуть ли не всей математической статистики!

Обозначим за X_1 результат первого броска монеты. До опыта этот результат неизвестен. Поэтому X_1 – это случайная величина, которая имеет то же распределение, как и ξ. Обозначим за X_2 результат второго броска монеты. Опять же, до опыта результат неизвестен, поэтому это случайная величина, распределенная так же, как ξ. Введем аналогично $X_3, \ldots, X_n$ – результаты бросков с 3-го по n-й. Будем считать броски независимыми, и на языке теории вероятностей эту независимость будем понимать как стохастическую независимость. Тогда мы будем иметь множество независимых одинаково распределенных случайных величин $X_1, \ldots, X_n$. Наша оценка параметра p запишется так:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Заметим, что $\hat{p}$ – это случайная величина. Ее значение зависит от реализаций случайных величин $X_1, \ldots, X_n$, то есть от результата всех бросков.

Кто-то в качестве оценки параметра p может предложить использовать $1/2$ даже без всяких бросков. Выясним, почему оценка выше лучше оценки $1/2$. Во-первых, при повторном подкидывании n раз монеты оценка $\hat{p}$ в среднем будет принимать истинное значение p:

$$\mathbb{E}\hat{p} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}X_i = \frac{1}{n} \cdot np = p.$$

Это уже какой-то результат, потому что оценка $1/2$ может быть далека от истинного значения p, а при использовании оценки $\hat{p}$ мы уверены, что $\hat{p}$ будет лежать где-то в окрестности истинного, но неизвестного параметра p. Насколько велика эта окрестность? Вычислим дисперсию оценки $\hat{p}$:

$$\mathbb{D}\hat{p} = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{D}X_i = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}.$$

Здесь мы воспользовались независимостью измерений X_i, поэтому дисперсия суммы равна сумме дисперсий. Видно, что $\mathbb{D}\hat{p}$ стремится к нулю

с ростом числа измерений. Это значит, что чем больше измерений, тем меньше разброс $\hat{p}$ вокруг истинного значения p. Про оценку $1/2$ ничего подобного не скажешь: либо она точно попала в истинное значение (вряд ли подобное возможно в реальности), либо не попала, и никакие измерения этого не изменят. А в случае оценки $\hat{p}$ мы хотя бы уверены, что наша оценка не будет далека от истинного и даже стремится к истинному с ростом числа измерений n. Сходимость к истинному значению подтверждается законом больших чисел по Хинчину:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mathbb{E}\hat{p} = p.$$

Значит, какое бы малое ε мы ни взяли, вероятность отклониться от p на величину ε стремится к нулю. Согласно усиленному закону больших чисел:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{п.н.}} \mathbb{E}\hat{p} = p,$$

какова бы ни была полученная в результате измерений последовательность[2] нулей и единиц, с ростом числа подбрасываний оценка сойдется к p. Это значит, оценка $\hat{p}$ приближает неизвестное значение вероятности p, то есть приближает распределение случайной величины ξ.

1.3. Выборка и оценка функции распределения

Измерению может подвергаться не только случайная величина, но и случайный вектор. Набор $X_1, \ldots, X_n$ нам тоже будет встречаться на протяжении всего курса. Введем поэтому несколько определений общего характера.

Определение. Пусть $\xi \in \mathbb{R}^k$, $k \geq 1$, – случайный вектор. Тогда набор случайных векторов $X = (X_1, ..., X_n)$, распределение которых совпадает с распределением вектора ξ, называется *выборкой* из распределения случайного вектора ξ, а количество n векторов в наборе – *объемом выборки*.

Определение. Любая измеримая функция выборки (то есть та, которая является случайной величиной) называется *статистикой*.

Если функция зависит не только от выборки, но и от неизвестных параметров распределения, то статистикой она не является. Например, функция $\hat{p}(X)$, рассмотренная в примере с монетной, является статистикой, а функция $\varphi(X, p) = \hat{p}(X) - p$ статистикой не является.

[2]Кроме, быть может, тех последовательностей, что составляют множество меры 0.

Функцию распределения вектора ξ будем обозначать $F_\xi(x)$, по определению она совпадает с функциями распределения $F_{X_1}(x)$, ..., $F_{X_n}(x)$ векторов $X_1, \ldots, X_n$, то есть $F_\xi(x) = F_{X_1}(x) = \cdots = F_{X_n}(x)$. Реализации компонент $X_1, \ldots, X_n$ будем обозначать $x_1, \ldots, x_n$ соответственно.

Определение. Выборка $X = (X_1, ..., X_n)$ называется *простой*, если случайные векторы X_i, $i = 1, \ldots, n$, независимы в совокупности.

В дальнейшем, если не оговорено обратное, мы будет предполагать, что

1) размерность ξ равна 1, то есть ξ – это случайная величина,
2) под выборкой понимается простая выборка.

Итак, пусть даны произвольная случайная величина ξ и выборка X_1, ..., X_n из ее распределения. Пусть необходимо оценить функцию распределения $F_\xi(x)$ случайной величины ξ. Вспомним, что функция распределения в точке $x \in \mathbb{R}$ это вероятность

$$F_\xi(x) = \mathbb{P}(\xi < x),$$

так что наша задача состоит в оценке этой вероятности. Как и прежде, попробуем оценить эту вероятность через отношение благоприятных исходов к общему числу измерений:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i < x),$$

где $I(A)$ – индикаторная функция события A, она равна 1, если A выполнено, и равна 0, иначе. При вычислении этой оценки мы просто посчитали, сколько измерений $X_1, \ldots, X_n$ оказались меньше x, и поделили это число на объем всей выборки n. Эта оценка функции распределения имеет центральное место в математической статистике, и поэтому мы введем новое определение.

Определение. *Эмпирической функцией распределения* (далее – ЭФР) выборки $X = (X_1, ..., X_n)$ называется случайная функция

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i < x), \ x \in \mathbb{R},$$

где $I(A)$ – индикаторная функция события A, принимающая значение 1, если событие A имеет место, и 0, иначе.

ЭФР обладает рядом свойств:

1. $n\hat{F}_n(x) \in \mathrm{Bi}(n, F_\xi(x))$.
2. $\mathbb{E}\hat{F}_n(x) = F_\xi(x)$.
3. $\mathbb{D}\hat{F}_n(x) = \dfrac{1}{n}F_\xi(x)(1 - F_\xi(x))$.
4. Из ЗБЧ по Хинчину (или из неравенства Чебышева) следует сходимость по вероятности $\hat{F}_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} F_\xi(x)$, $n \to \infty$, в каждой точке $x \in \mathbb{R}$.
5. Согласно УЗБЧ по Колмогорову, верно более сильное утверждение: сходимость почти наверное $\hat{F}_n(x) \xrightarrow[n\to\infty]{\text{п.н.}} F_\xi(x)$, $n \to \infty$, в каждой точке $x \in \mathbb{R}$.
6. Из ЦПТ получаем сходимость по распределению:

$$\frac{n\hat{F}_n(x) - nF_\xi(x)}{\sqrt{nF_\xi(x)(1 - F_\xi(x))}} \xrightarrow[n\to\infty]{\mathrm{d}} \xi_{N(0,1)} \in \mathrm{N}(0,1),\ n \to \infty,$$

в каждой точке $x \in \mathbb{R}$.

Первое свойство следует из того, что каждый из индикаторов $I(X_i < x)$ обладает распределением Бернулли с параметром

$$p = \mathbb{P}(I(X_i < x) = 1) = \mathbb{P}(X_i < x) = \mathbb{P}(\xi < x) = F_\xi(x),$$

и так как X_i независимы в совокупности, то и $I(X_i < x)$ независимы в совокупности. Отсюда и получаем, что сумма n независимых случайных величин, распределенных как $\mathrm{Be}(F_\xi(x))$, имеет биномиальное распределение $\mathrm{Bi}(n, F_\xi(x))$.

Эти свойства говорят о поточечной (относительно x) сходимости оценки $\hat{F}_n(x)$ к истинному и неизвестному значению $F_\xi(x)$. Но оказывается, справедливы два гораздо более сильных свойства: теорема Гливенко и теорема Колмогорова.

Введем величину

$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_\xi(x)|,$$

отражающую величину отклонения эмпирической функции распределения от истинной на всей числовой оси.

Теорема Гливенко. $D_n \xrightarrow[n\to\infty]{\text{п.н.}} 0$.

Теорема Колмогорова. *Если функция распределения $F_\xi(x)$ непрерывна, то при любом $t > 0$:*

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}D_n \le t) = \mathrm{K}(t) = \sum_{j=-\infty}^{\infty} (-1)^j \exp\left(-2j^2t^2\right).$$

При $t \leq 0$ *по определению считаем* $\mathrm{K}(t) = 0$.

Эти теоремы говорят уже не просто о поточечной сходимости, а о сходимости в равномерной метрике. Функция $\mathrm{K}(t)$ является функцией так называемого *распределения Колмогорова*. Все это говорит о том, что ЭФР является очень хорошей аппроксимацией истинного распределения. Введем еще одно понятие, которое нам понадобится в дальнейшем.

1.4. Порядковые статистики

Эмпирическая функция распределения – это случайный процесс, а более конкретно – это случайная функция. Действительно, она зависит от исхода (через случайные величины $X_1, \ldots, X_n$) и от детерминированного параметра $x \in \mathbb{R}$. Реализациями этого процесса являются кусочно-постоянные функции: если зафиксировать исход (а значит и выборку) и увеличивать x от $-\infty$ до $+\infty$, то мы увидим, что ЭФР равна нулю на участке $(-\infty, \min\{X_1, \ldots, X_n\}]$; далее, она испытывает скачок, так как x переходит через один из элементов выборки. Если все элементы выборки различны, то это будет скачок на величину $1/n$, n – объем выборки. В момент, когда x перейдет через второй по возрастанию элемент выборки, ЭФР увеличится еще на величину $1/n$, и так далее. Итак, для того чтобы нарисовать графики реализаций ЭФР, требуется упорядочить элементы выборки по возрастанию (для фиксированного исхода) и найти значения, которые принимает ЭФР между этими точками. Так мы приходим к очень важному в математической статистике понятию – порядковой статистике. Дам формальное определение.

Определение. Пусть дано вероятностное пространство $(\Omega, \mathfrak{S}, \mathbb{P})$, на котором определены элементы выборки $X_1, \ldots, X_n$, и $x_i = X_i(\omega)$, $i = 1, \ldots, n$, $\omega \in \Omega$. Перенумеруем последовательность $\{x_i\}_{i=1}^n$ в порядке неубывания так, что

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Тогда функция $X_{(k)}(\omega) = x_{(k)}$ называется *k-й порядковой статистикой выборки* $X = (X_1, \ldots, X_n)$.

Есть два частных случая, которые будут нам встречаться наиболее часто: $X_{(1)}$ и $X_{(n)}$. Это соответственно наименьшее и наибольшее значения выборки, то есть

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

Определение. Последовательность $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ называют *вариационным рядом* выборки X.

Именно в точках $X_{(k)}$ происходят скачки ЭФР, причем

$$\hat{F}_n(x) = k/n \text{ при } X_{(k)} < x \leq X_{(k+1)},$$

где для удобства принято $X_{(0)} = -\infty$ и $X_{(n+1)} = +\infty$. Эти равенства справедливы в случаях как совпадающих между собой порядковых статистик, так и для различающихся, то есть в самом общем случае.

1.5. Распределение порядковых статистик

Итак, *порядковыми статистиками* $X_{(k)}$, $k = 1, \ldots, n$, называются упорядоченные (для каждого исхода) элементы выборки $X_1, \ldots, X_n$, так что почти наверное

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

Элементы выборки – независимые одинаково распределенные случайные величины. Порядковые статистики – зависимые и неодинаково распределенные случайные величины.

Задача 1. Найти распределение k-й порядковой статистики.

Решение. Найдем распределение $X_{(k)}$, $k = 1, \ldots, n$. Для этого рассмотрим событие $X_{(k)} < x$. Заметим, что k-е по величине значение из $X_1, \ldots, X_n$ оказалось меньше x *тогда и только тогда*, когда как минимум k случайных величин из $X_1, \ldots, X_n$ оказались меньше x. Меньше x могло оказаться k значений выборки, $k + 1$ значение и так далее до n значений. Вероятность того, что какой-то X_k оказался меньше x равна

$$\mathbb{P}(X_k < x) = \mathbb{P}(\xi < x) = F_\xi(x),$$

а вероятность того, что какой-то X_k оказался больше или равен x, равна

$$\mathbb{P}(X_k \geq x) = \mathbb{P}(\xi \geq x) = 1 - F_\xi(x).$$

Так как элементы выборки одинаково распределены и независимы, то функция распределения статистики $X_{(k)}$ выражается по формуле

$$F_{X_{(k)}}(x) = \sum_{m=k}^{n} C_n^m F_\xi^m(x) \left(1 - F_\xi(x)\right)^{n-m}.$$

Теперь допустим, что ξ имеет плотность распределения, в таком случае и порядковые статистики имеют плотность распределения. Плотность

статистики $X_{(k)}$ можно искать как производную функции $F_{X_{(k)}}(x)$, но это очень неудобно. Гораздо проще искать плотность как предел:

$$f_{X_{(k)}}(x) = \lim_{\varepsilon \to 0+} \frac{\mathbb{P}(X_{(k)} \in [x, x+\varepsilon))}{\varepsilon}.$$

При подсчете числителя достаточно выделить слагаемое, пропорциональное ε, потому как остальные слагаемые, деленные на ε, в результате вычисления предела дадут 0. Теперь заметим, что k-е по величине значение из $X_1, \ldots, X_n$ попадает в интервал $[x, x+\varepsilon)$ тогда и только тогда, когда выполнено одно из следующих несовместимых событий:

1) $k-1$ значений меньше x, 1 значение лежит в интервале $[x, x+\varepsilon)$, и $n-k$ значений превышают $x+\varepsilon$,
2) $k-2$ значений меньше x, 2 значения лежат в интервале $[x, x+\varepsilon)$, и $n-k$ значений превышают $x+\varepsilon$,
3) $k-3$ значений меньше x, 3 значения лежат в интервале $[x, x+\varepsilon)$, и $n-k$ значений превышают $x+\varepsilon$

   ...

Если m случайных величин из $X_1, \ldots, X_n$ попали в интервал $[x, x+\varepsilon)$, то

$$\mathbb{P}(X_1 \in [x, x+\varepsilon), \ldots, X_m \in [x, x+\varepsilon)) =$$

$$= \mathbb{P}(X_1 \in [x, x+\varepsilon)) \cdots \mathbb{P}(X_m \in [x, x+\varepsilon)) \sim \varepsilon^m, \ \varepsilon \to 0.$$

Значит достаточно рассмотреть только событие, при котором лишь одна из величин $X_1, \ldots, X_n$ попадает в интервал $[x, x+\varepsilon)$. Итак,

$$f_{X_{(k)}}(x) = \lim_{\varepsilon \to 0} \frac{\mathbb{P}(X_{(k)} \in [x, x+\varepsilon))}{\varepsilon} =$$

$$= \lim_{\varepsilon \to 0} \frac{C_n^1 \, \mathbb{P}^1(\xi \in [x, x+\varepsilon)) \cdot C_{n-1}^{k-1} \, \mathbb{P}^{k-1}(\xi < x) \cdot C_{n-k}^{n-k} \, \mathbb{P}^{n-k}(\xi \geq x+\varepsilon)}{\varepsilon},$$

откуда окончательно получаем

$$f_{X_{(k)}}(x) = n C_{n-1}^{k-1} F_\xi^{k-1}(x) \left(1 - F_\xi(x)\right)^{n-k} f_\xi(x).$$

Полезно помнить, что если выборка взята из равномерного распределения $\mathrm{U}(0,1)$, то $X_{(k)} \in \mathrm{Beta}(k, n-k+1)$.

2. Критерии согласия

2.1. Введение

Математическую статистику можно условно разделить на две части: проверку статистических гипотез и теорию оценивания. Мы начнем с проверки гипотез. *Гипотезой* называется любое утверждение о распределении случайной величины. Например, гипотезой может быть предположение, что данная нам для измерений случайная величина имеет распределение $\mathrm{N}(0,1)$. Бывают задачи, в которых требуется проверить лишь принадлежность распределения к какому-либо классу распределений. В этом случае, например, гипотезой может быть предположение о том, что случайная величина имеет нормальное распределение, а вот с какими параметрами не важно. Можно выдвинуть несколько гипотез: например, у нас могут иметься все основания считать, что случайная величина распределена либо по закону $\mathrm{N}(0,1)$, либо по закону $\mathrm{N}(0,2)$, а третьего не дано. С этого момента и далее мы начнем выяснять, как проверять такие гипотезы.

В первую очередь отмечу, что за исключением непрактичных, вырожденных случаев, установить верность той или иной гипотезы невозможно, ошибки установления верности гипотезы неизбежны. Задача теория проверки гипотез и состоит в том, чтобы в том или ином смысле минимизировать эти ошибки.

С формальной точки зрения, проверка гипотез состоит в поиске некоторого отображения, которое бы однозначно сопоставляло реализации выборки какую-то гипотезу. Перейдем теперь к точным определениям.

Определение. Множество $\Omega \subseteq \mathbb{R}^n$ всех значений выборки называется *выборочным пространством.*

Пусть функция распределения F_ξ измеряемой случайной величины ξ неизвестна и принадлежит некоторому множеству априори допустимых распределений $\mathcal{F}$. Например, $\mathcal{F}$ может быть множеством всех возможных распределений.

Определение. Любое утверждение о принадлежности F_ξ какому-либо подмножеству $\mathcal{F}' \subset \mathcal{F}$ называется *гипотезой* и обозначается, например, так:

$$H : F_\xi \in \mathcal{F}' \subset \mathcal{F}.$$

Определение. Если $\mathcal{F}'$ в гипотезе $H : F_\xi \in \mathcal{F}'$ состоит из одного элемента, то гипотеза H называется *простой*, иначе – *сложной.* Другими словами, *простой гипотезой* будем называть любое предположение, однозначно определяющее распределение выборки.

На проверку может быть выдвинуто несколько гипотез. Некоторые из них могут быть простыми, некоторые – сложными. Бывает, что выдвинута одна гипотеза. В этом случае на самом деле предполагается, что гипотез две, просто вторая гипотеза не пишется, но по умолчанию дополняет множество из основной гипотезы до множества всех априори допустимых гипотез $\mathcal{F}$.

Определение. Пусть выдвинуто r гипотез $H_1, \ldots, H_r$. *Статистическим критерием* называется измеримая функция $\delta : \Omega \to \to \{H_1, \ldots, H_r\}$, сопоставляющая выборке какую-либо гипотезу.

Задание функции δ равносильно разбиению выборочного пространства Ω на r непересекающихся подмножеств $\Omega_1, \ldots, \Omega_r$ таких, что при попадании выборки x в область Ω_1 принимается гипотеза H_1, при попадании в область Ω_2 принимается гипотеза H_2 и так далее.

Определение. Пусть дано r простых гипотез $H_1, \ldots, H_r$. Вероятность

$$\alpha_i(\delta) = \mathbb{P}_i(X \notin \Omega_i) = \mathbb{P}_i(\delta(X) \neq H_i)$$

называется *вероятностью ошибки i-го рода*. Индекс i под символом вероятности означает, что вероятность подсчитывается в случае, когда выборка X распределена по закону гипотезы H_i. Другими словами, вероятность ошибки i-го рода – это вероятность отклонить i-ю гипотезу, если на самом деле она верна.

Иногда используют и запись с вертикальной чертой:

$$\mathbb{P}(\delta(X) \neq H_i | H_i),$$

но она несколько неудачна, так как ее можно спутать с условной вероятностью. Об условной вероятности можно говорить лишь в том случае, если на множестве гипотез $\{H_1, \ldots, H_r\}$ задано вероятностное распределение. Такое бывает, и мы с этим столкнемся, когда речь пойдет о байесовских решающих правилах. Пока же никакого распределения на множестве гипотез нет. Для определенности будем пользоваться обозначением с индексом.

Определение. В случае двух гипотез H_1 и H_2 множество Ω_2 называют *критической областью* гипотезы H_1. Критическую область обычно задают с помощью вспомогательной статистики – меры отклонения эмпирических данных от гипотетических. Тогда критическая область – это просто область достаточно больших значений вспомогательной статистики. Эту вспомогательную статистику еще называют *статистикой критерия*.

В случае двух простых гипотез H_1 и H_2 вероятность ошибки 1-го рода обычно обозначается символом α, а вероятность ошибки 2-го

рода обозначается символом β. В этом же случае

$$\alpha = \mathbb{P}_1(X \in \Omega_2) = \mathbb{P}_1(\delta(X) = H_2),$$

$$\beta = \mathbb{P}_2(X \in \Omega_1) = \mathbb{P}_2(\delta(X) = H_1).$$

Определение. Если $\alpha(\delta) \leq \alpha_0$, то говорят, что критерий δ имеет *уровень значимости* α_0.

Определение. Пусть $F(x)$ – некоторая функция распределения. Тогда любое решение уравнения $F(x) = p \in (0,1)$, если оно существует, называется *p-квантúлем*. Если решение не существует, то p-квантилем для непрерывной слева функции $F(x)$ называется $x = \sup\{y : F(y) \leq p\}$. Обозначать квантиль будем по имени соответствующего распределения, вынося в индекс значение p. Например, для функции распределения $\xi \in \mathrm{N}(0,1)$ p-квантиль будем обозначать так: $x = \mathrm{N}_p(0,1)$.

2.2. Критерий согласия Колмогорова

Условия. Даны выборка $x_1, \ldots, x_n$ и непрерывная функция $F(x)$. Выдвинута простая гипотеза

$$H_1 : F_\xi(x) = F(x).$$

Требуется составить критерий проверки гипотезы H_1 на заданном уровне значимости α.

Алгоритм

1. Составить эмпирическую функцию распределения на данных $x_1, \ldots, x_n$ и вычислить реализацию *статистики Колмогорова–Смирнова*:

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|.$$

2. Выбрать в качестве критической области $\Omega_2 = \{x \in \Omega : D_n(x) \geq t_\alpha\}$, и найти t_α из условия на уровень значимости:

$$\mathbb{P}_1(X \in \Omega_2) = \mathbb{P}_1(D_n \geq t_\alpha) = \alpha,$$

то есть найти $(1-\alpha)$-квантиль распределения случайной величины D_n.

3. Принять решение по следующей схеме:

$$H_1 \text{ отвергается} \Leftrightarrow D_n \geq t_\alpha,\ \mathbb{P}_1(D_n \geq t_\alpha) = \alpha.$$

Замечание 1. Статистика Колмогорова–Смирнова равна максимальному отклонению эмпирической функции распределения $\hat{F}_n(x)$, построенной по выборке X, от гипотетической $F(x)$. Согласно теореме

Гливенко, если гипотеза H_1 верна (то есть $F(x) = F_\xi(x)$), то это отклонение почти наверное стремится к нулю с ростом объема выборки. Если гипотеза верна, мы *ожидаем*, что отклонение D_n будет небольшим. Поэтому в качестве критической области выбираются достаточно большие значения этой статистики, а граница для этих значений определяется из условия на уровень значимости.

Замечание 2. Функция $F(x)$ используется только при расчете D_n. В случае верности гипотезы H_1 (то есть при $F(x) = F_\xi(x)$) распределение D_n от $F(x)$ не зависит, поэтому можно пользоваться одной и той же таблицей квантилей вне зависимости от функции $F(x)$.

Замечание 3. Критерий используется только для непрерывных функций $F(x)$.

Замечание 4. Если значения n достаточно большие ($n \geq 20$), то благодаря теореме Колмогорова можно воспользоваться приближением

$$\mathbb{P}_1(D_n \geq t_\alpha) = \mathbb{P}_1(\sqrt{n}D_n \geq \sqrt{n}t_\alpha) \approx 1 - \mathrm{K}(\sqrt{n}t_\alpha) = \alpha$$

и находить $t_\alpha = \mathrm{K}_{1-\alpha}/\sqrt{n}$, то есть отыскивать границу критической области через квантиль распределения Колмогорова.

Замечание 5. Если мы отклонили гипотезу H_1, то мы можем гарантировать, что вероятность нашей ошибки не превышает уровень значимости α. Если же гипотезу мы не отклоняем, то утверждать, что она верная, нельзя, т.к. мы не знаем вероятность ошибки такого утверждения. Поэтому в случае неотклонения гипотезы H_1 мы скромно говорим: «данные гипотезе не противоречат».

Замечание 6. На практике статистику D_n рассчитывают по формулам

$$D_n = \max\{D_n^+, D_n^-\},$$

$$D_n^+ = \max_{1\leq k\leq n}\left(\frac{k}{n} - F(X_{(k)})\right),\ D_n^- = \max_{1\leq k\leq n}\left(F(X_{(k)}) - \frac{k-1}{n}\right).$$

Задача 1. Дана выборка $x = (0.1, 0.9, 0.3, 0.4, 0.7)$ из непрерывного распределения. На уровне значимости $\alpha = 0.05$ проверить гипотезу о равномерном на отрезке $(0, 1)$ распределении измеряемой случайной величины:

$$H_1 : \mathrm{U}(0, 1).$$

Решение. Функция распределения $\mathrm{U}(0, 1)$ есть

$$F(x) = x,\ x \in (0, 1).$$

Вычислим реализацию статистики D_n. Для этого воспользуемся формулами

$$D_n = \max\{D_n^+, D_n^-\},$$

$$D_n^+ = \max_{1\le k\le n}\left(\frac{k}{n} - F(x_{(k)})\right),\ D_n^- = \max_{1\le k\le n}\left(F(x_{(k)}) - \frac{k-1}{n}\right).$$

В нашем случае

$$(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}) = (0.1, 0.3, 0.4, 0.7, 0.9),$$

$$D_n^+ = \max\left\{\frac{1}{5} - \frac{1}{10},\ \frac{2}{5} - \frac{3}{10},\ \frac{3}{5} - \frac{4}{10},\ \frac{4}{5} - \frac{7}{10},\ 1 - \frac{9}{10}\right\} = \frac{1}{5},$$

$$D_n^- = \max\left\{\frac{1}{10} - 0,\ \frac{3}{10} - \frac{1}{5},\ \frac{4}{10} - \frac{2}{5},\ \frac{7}{10} - \frac{3}{5},\ \frac{9}{10} - \frac{4}{5}\right\} = \frac{1}{10},$$

$$D_n = \max\left\{\frac{1}{5},\ \frac{1}{10}\right\} = \frac{1}{5}.$$

Теперь заглянем в таблицу квантилей[3] и найдем квантиль $t_\alpha = 0.563$. Мы видим, что $D_n < t_\alpha$, поэтому ответ: «данные гипотезе не противоречат».

2.3. Критерий согласия хи-квадрат Пирсона

Условия. Дана дискретная случайная величина ξ, принимающая значения $1, 2, \ldots, N$ с некоторыми неизвестными вероятностями $p_1, \ldots, p_N$, которые образуют вектор $p = (p_1, \ldots, p_N)$; $p_1 + \cdots + p_N = 1$. Имеется выборка $X = (X_1, \ldots, X_n)$ и вектор вероятностей $p^\circ = (p_1^\circ, \ldots, p_N^\circ)$, причем $0 < p_j^\circ < 1$ для всех $j = 1, \ldots, N$. Выдвинута простая гипотеза:

$$H_1 : p = p^\circ.$$

Составить критерий проверки гипотезы H_1 на заданном уровне значимости α.

Алгоритм

1. Вычислить частоты исходов

$$\nu_j = \sum_{i=1}^{n} I(X_i = j),\ j = 1, \ldots, N.$$

2. Вычислить статистику критерия

$$T_{\chi^2} = \sum_{j=1}^{N} \frac{(\nu_j - np_j^\circ)^2}{np_j^\circ}.$$

[3] Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику : учебник. Москва : Издательство ЛКИ, 2010. С. 579. См. случай $n = 5$ и $p = 0.05$.

3. В качестве критической области выбрать $\Omega_{\text{кр}} = \{x \in \Omega : T_{\chi^2}(x) > t_\alpha\}$, где t_α соответствует условию на уровень значимости: $\mathbb{P}_1(T_{\chi^2} > t_\alpha) = \alpha$.

Известно, что при истинности гипотезы H_1 статистика $T_{\chi^2} \xrightarrow{d} \xrightarrow{d} \chi^2(N-1)$ при $n \to \infty$, поэтому границу t_α можно вычислять как $(1-\alpha)$-квантиль распределения $\chi^2(N-1)$. Итак, критерий согласия хи-квадрат Пирсона формулируется следующим образом:

$$H_1 \text{ отвергается} \Leftrightarrow T_{\chi^2} > t_\alpha,\ t_\alpha = \chi^2_{1-\alpha}(N-1).$$

Замечание 1. Для того чтобы воспользоваться фактом сходимости распределения статистики критерия к распределению хи-квадрат, критерий рекомендуется применять при $n \geq 50$ и $\nu_j \geq 5$ для всех $j = 1, \ldots, N$.

Замечание 2. Вектор $\nu = (\nu_1, \ldots, \nu_N)$ имеет полиномиальное распределение $\mathrm{M}(n, p_1, \ldots, p_N)$ с функцией вероятности

$$\mathbb{P}(\nu_1 = k_1, \ldots, \nu_N = k_N) = \frac{N!}{k_1! \ldots k_N!}\, p_1^{k_1} \cdots p_N^{k_N}, \quad k_1 + \cdots + k_N = n.$$

Про это распределение известно, что

$$\nu_j \in \mathrm{Bi}(n, p_j),\ \mathbb{E}\nu_j = np_j,\ \mathbb{D}\nu_j = np_j(1-p_j),$$

$$\forall i, j,\ i \neq j \quad \operatorname{cov}(\nu_i, \nu_j) = -np_ip_j.$$

Замечание 3. Статистика T_{χ^2} представляет собой *меру хи-квадрат* отклонения эмпирических данных от гипотетических. Чтобы лучше себе представить содержимое этого выражения, надо вспомнить, что согласно закону больших чисел, $\nu_j/n \xrightarrow{\mathbb{P}} p_j$ с ростом объема выборки $n \to \infty$. Поэтому при достаточно больших n и при условии истинности гипотезы H_1 мы ожидаем, что разница $(\nu_j - np_j^\circ)^2$ будет небольшой. Весовой коэффициент $1/np_j^\circ$ позволяет получить сходимость $T_{\chi^2} \xrightarrow{d} \chi^2(N-1)$ при $n \to \infty$, избавить нас от расчета распределения статистики T_{χ^2} и дать возможность воспользоваться известными таблицами распределений хи-квадрат.

Замечание 4. Если исходные данные представляют собой выборку из некоторого непрерывного распределения, то, чтобы воспользоваться критерием хи-квадрат, можно предварительно применить *метод группировки наблюдений*: разбить пространство значений случайной величины на N непересекающихся интервалов, задать гипотетические вероятности p_j° попадания в них, подсчитать частоты попадания данных

в эти интервалы и т.д. Недостатком такого подхода является то, что при группировке данных происходит некоторая потеря информации. Кроме того, возникает вопрос о выборе числа интервалов N и их виде.

Замечание 5. Преимущества критерия хи-квадрат состоят в том, что при его применении нет необходимости учитывать точные значения наблюдений. Так же следует отметить его универсальность и наглядность.

Задача 1. При 4040 бросаниях монеты решка выпала 2048 раз, а орел выпал 1992 раза. Проверить на уровне значимости 5% гипотезу о симметричности монеты.

Решение. По условию задачи нам даны частоты исходов $\nu_1 = 2048$ и $\nu_2 = 1992$, $N = 2$, объем выборки $n = \nu_1 + \nu_2 = 4040$, уровень значимости $\alpha = 0.05$ и гипотеза о симметричности монеты

$$H_1 : p = p^\circ,$$

где вектор вероятностей $p^\circ = [1/2, 1/2]$. Первым делом вычисляем реализацию статистики критерия:

$$T_{\chi^2} = \sum_{j=1}^{N} \frac{(\nu_j - np_j^\circ)^2}{np_j^\circ} = \frac{(\nu_1 - np_1^\circ)^2}{np_1^\circ} + \frac{(\nu_2 - np_2^\circ)^2}{np_2^\circ} =$$

$$= 0.3881 + 0.3881 = 0.7762.$$

Затем по таблице смотрим границу критической области

$$t_\alpha = \chi^2_{1-\alpha}(N-1) = \chi^2_{0.95}(1) = 3.841.$$

Так как $T_{\chi^2} = 0.7762 < 3.841 = \chi^2_{1-\alpha}(N-1)$, то наш ответ: «данные гипотезе не противоречат».

2.4. Критерий хи-квадрат для сложной гипотезы

Условия. Дана дискретная случайная величина ξ, принимающая значения $1, 2, \ldots, N$ с некоторыми неизвестными вероятностями $p_1, \ldots, p_N$, которые образуют вектор $p = (p_1, \ldots, p_N)$; $p_1 + \cdots + p_N = 1$. Имеется выборка $X = (X_1, \ldots, X_n)$ и гладкая вектор-функция $p^\circ(\theta) = (p_1^\circ(\theta), \ldots, p_N^\circ(\theta))$, $\theta = (\theta_1, \ldots, \theta_r) \in \Theta$, где $r < N - 1$. Выдвинута сложная гипотеза:

$$H_1 : p = p^\circ(\theta) \text{ для некоторого } \theta \in \Theta.$$

Требуется составить критерий проверки гипотезы H_1 на заданном уровне значимости α.

Алгоритм

1. Вычислить частоты исходов

$$\nu_j = \sum_{i=1}^{n} I(X_i = j),\ j = 1, \ldots, N.$$

2. Оценить неизвестный параметр θ. Для этого использовать оценку максимального правдоподобия для вектора с полиномиальным распределением:

$$\hat{\theta} = \arg\max_{\theta} \prod_{j=1}^{N} (p_j^\circ(\theta))^{\nu_j},$$

которая находится из системы r уравнений

$$\sum_{j=1}^{N} \frac{\nu_j}{p_j^\circ(\theta)} \cdot \frac{\partial p_j^\circ(\theta)}{\partial \theta_k} = 0,\ k = 1, \ldots, r.$$

Решение обозначить символом $\hat{\theta}$.

3. Рассчитать статистику критерия

$$\hat{T}_{\chi^2} = \sum_{j=1}^{N} \frac{(\nu_j - np_j^\circ(\hat{\theta}))^2}{np_j^\circ(\hat{\theta})}.$$

4. В качестве критической области выбрать $\Omega_{\text{кр}} = \{x \in \Omega :\ \hat{T}_{\chi^2}(x) > t_\alpha\}$, где t_α ищется из условия на уровень значимости $\mathbb{P}_1(T_{\chi^2} > t_\alpha) = \alpha$.

Известно[4], что при истинности гипотезы H_1 и выполнении условий

1) $\sum\limits_{j=1}^{N} p_j^\circ(\theta) = 1,\ \forall \theta \in \Theta$,

2) $p_j^\circ(\theta) \geq c > 0, \forall j$, и существуют непрерывные производные

$$\frac{\partial p_j^\circ(\theta)}{\partial \theta_k},\ \frac{\partial^2 p_j^\circ(\theta)}{\partial \theta_k \partial \theta_l},\ k = 1, \ldots, r,\ l = 1, \ldots, r,$$

3) $(N \times r)$-матрица $\left\| \partial p_j^\circ(\theta)/\partial \theta_k \right\|$ имеет ранг r для всех $\theta \in \Theta$,

[4]Крамер Г. Математические методы статистики. Москва : Мир, 1975. С. 460.

статистика $\hat{T}_{\chi^2} \xrightarrow{d} \chi^2(N-1-r)$ при $n \to \infty$. Поэтому граница t_α критической области вычисляется как $(1-\alpha)$-квантиль распределения $\chi^2(N-1-r)$. В критерий получаем следующий критерий:

$$H_1 \text{ отвергается} \Leftrightarrow \hat{T}_{\chi^2} > t_\alpha,\ t_\alpha = \chi^2_{1-\alpha}(N-1-r).$$

Замечание 1. Для того, чтобы воспользоваться фактом сходимости распределения статистики критерия к распределению хи-квадрат, критерий рекомендуется применять при $n \geq 50$ и $\nu_j \geq 5$ для всех $j = 1, \ldots, N$.

Замечание 2. Оценка максимального правдоподобия – это такое значение параметра, при котором вероятность получения имеющихся данных (в нашем случае набора частот $\nu_1, \ldots, \nu_N$) максимальна.

Замечание 3. Этот критерий применяется и в случае непрерывных распределений в основной гипотезе, если применить метод группировки наблюдений.

Задача 2. Среди 2020 семей, имеющих двух детей, 527 семей, в которых два мальчика, 476 имеют двух девочек (в остальных 1017 семьях дети разного пола). Можно ли с уровнем значимости $\alpha = 0.05$ считать, что количество мальчиков в семье с двумя детьми – биномиальная величина?

Решение. По условию задачи измеряемая случайная величина ξ (число мальчиков в семье) – случайная величина, которая принимает $N = 3$ значения: 0, 1 и 2. По условию нам даны частоты исходов $\nu_0 = 476$, $\nu_1 = 1017$, $\nu_2 = 527$, уровень значимости $\alpha = 0.05$ и гипотеза

$$H_1 : \xi \in \mathrm{Bi}(2, \theta).$$

Сначала найдем вероятности $p_j^\circ(\theta)$:

$$p_0^\circ(\theta) = \mathbb{P}_0(\xi = 0) = C_2^0 \theta^0 (1-\theta)^2 = (1-\theta)^2,$$
$$p_1^\circ(\theta) = \mathbb{P}_0(\xi = 1) = C_2^1 \theta^1 (1-\theta)^1 = 2\theta(1-\theta),$$
$$p_2^\circ(\theta) = \mathbb{P}_0(\xi = 2) = C_2^2 \theta^2 (1-\theta)^0 = \theta^2.$$

Теперь оценим неизвестный параметр θ по методу максимального правдоподобия. Так как параметр один, то $r = 1$, и требуется решить одно уравнение

$$\sum_{j=0}^{2} \frac{\nu_j}{p_j^\circ(\theta)} \cdot \frac{\partial p_j^\circ(\theta)}{\partial \theta} = 0.$$

Подставим сюда выражения для $p_j^\circ(\theta)$ и получим

$$\sum_{j=0}^{2} \frac{\nu_j}{p_j^\circ(\theta)} \cdot \frac{\partial p_j^\circ(\theta)}{\partial \theta} = \frac{-2\nu_0}{1-\theta} + \frac{1-2\theta}{\theta(1-\theta)}\nu_1 + \frac{2\nu_2}{\theta} = 0.$$

Решая это уравнение относительно θ, мы получаем его оценку

$$\hat{\theta} = \frac{\nu_1 + 2\nu_2}{2n} = 0.5126$$

и оценку гипотетических вероятностей

$$p_0^\circ(\hat{\theta}) = 0.2375,\ p_1^\circ(\hat{\theta}) = 0.4997,\ p_2^\circ(\hat{\theta}) = 0.2628.$$

Теперь можно вычислить реализацию статистики критерия

$$\hat{T}_{\chi^2} = \sum_{j=0}^{2} \frac{(\nu_j - np_j^\circ(\hat{\theta}))^2}{np_j^\circ(\hat{\theta})} = 0.1158.$$

Наконец по таблице смотрим границу критической области

$$t_\alpha = \chi^2_{1-\alpha}(N-1-r) = \chi^2_{0.95}(1) = 3.84.$$

Так как

$$\hat{T}_{\chi^2} = 0.1158 < 3.84 = \chi^2_{1-\alpha}(N-1-r),$$

то данные гипотезе не противоречат.

Замечание. Рождение мальчика в семье можно представить как случайную величину η с распределением Бернулли $\mathrm{Be}(\theta)$. Если $\eta = 1$, то считаем, что родился мальчик. А если $\eta = 0$, то считаем, что родилась девочка.

1) Допустим, что $\eta_1 \in \mathrm{Be}(\theta)$ есть результат рождения мальчика первым ребенком в семье и $\eta_2 \in \mathrm{Be}(\theta)$ есть результат рождения мальчика вторым ребенком в семье. Заметьте, что в обоих случаях распределение одно и то же, пусть это будет нашим априори верным предположением. Из теории вероятности и из приложения следует, что независимость η_1 и η_2 равносильна $\xi = \eta_1 + \eta_2 \in \mathrm{Bi}(2, \theta)$. Таким образом, в предположении однородности (то есть неизменности) распределения рождения мальчиков от одного ребенка к следующему, гипотеза о биномиальном распределении мальчиков в семье с двумя детьми равносильна гипотезе о независимости рождения мальчиков от одного ребенка к следующему.

2) Теперь допустим, что $\eta_1 \in \mathrm{Be}(\theta_1)$ есть результат рождения мальчика первым ребенком в семье, а $\eta_2 \in \mathrm{Be}(\theta_2)$ есть результат рождения мальчика вторым ребенком в семье. Пусть в этом случае нам неизвестно, есть ли равенство $\theta_1 = \theta_2$. Но пусть нам априори известно, что результат рождения второго ребенка не зависит от результата рождения первого ребенка. Из теории вероятностей и приложения следует, что $\theta_1 = \theta_2 = \theta$ равносильно $\eta_1 + \eta_2 \in \mathrm{Bi}(2, \theta)$. Таким образом, в

предположении независимости рождения мальчиков от одного ребенка к следующему, гипотеза о биномиальном распределении мальчиков в семье с двумя детьми равносильна гипотезе об однородности распределения рождения мальчиков от одного ребенка к следующему.

2.5. Приложение

Теорема. *Пусть* $\xi \in \mathrm{Be}(\theta_1)$ *и* $\eta \in \mathrm{Be}(\theta_2)$. *Пусть* $\xi + \eta \in \mathrm{Bi}(2, \varphi)$ *для некоторого* φ. *Тогда*
1) *из равенства* $\theta_1 = \theta_2 = \theta$ *следует независимость* ξ *и* η *и* $\varphi = \theta$;
2) *из независимости* ξ *и* η *будет следовать равенство* $\theta_1 = \theta_2 = \varphi$.

Доказательство. Введем совместную функцию вероятности

$$p(i,j) = \mathbb{P}(\xi = i, \eta = j),\ i,j = 0,1.$$

Эта функция задана в четырех точках $(0,0)$, $(1,0)$, $(0,1)$ и $(1,1)$. Так как $\xi + \eta \in \mathrm{Bi}(2,\varphi)$, то

$$\begin{aligned}
&\mathbb{P}(\xi + \eta = 0) = p(0,0) = (1-\varphi)^2,\\
&\mathbb{P}(\xi + \eta = 1) = p(1,0) + p(0,1) = 2\varphi(1-\varphi),\\
&\mathbb{P}(\xi + \eta = 2) = p(1,1) = \varphi^2.
\end{aligned}$$

Отсюда сразу получаем семейство решений:

$$\begin{aligned}
p(0,0) &= (1-\varphi)^2,\\
p(1,0) &= 2\varphi(1-\varphi) - c,\\
p(0,1) &= c,\\
p(1,1) &= \varphi^2,
\end{aligned}$$

параметризованное числом $c \in [0, 2\varphi(1-\varphi)]$.

1. Если ξ и η имеют одинаковое распределение $\mathrm{Be}(\theta)$, то необходимо, чтобы

$$\mathbb{P}(\xi = 0) = \mathbb{P}(\eta = 0),$$

что равносильно равенству

$$p(0,0) + p(0,1) = p(0,0) + p(1,0),$$

откуда $p(0,1) = p(1,0)$, а значит $c = \varphi(1-\varphi)$ и совместное распределение будет следующим:

$$\begin{aligned}
p(0,0) &= (1-\varphi)^2,\\
p(1,0) &= \varphi(1-\varphi),\\
p(0,1) &= \varphi(1-\varphi),\\
p(1,1) &= \varphi^2.
\end{aligned}$$

Это распределение соответствует только независимым ξ и η.

2. Если $\xi \in \mathrm{Be}(\theta_1)$ и $\eta \in \mathrm{Be}(\theta_2)$ независимы, то

$$\begin{aligned}(1-\theta_1)(1-\theta_2) &= (1-\varphi)^2,\\ \theta_1(1-\theta_2) &= 2\varphi(1-\varphi) - c,\\ (1-\theta_1)\theta_2 &= c,\\ \theta_1\theta_2 &= \varphi^2.\end{aligned}$$

Складывая второе и третье уравнения, получаем $\varphi = (\theta_1+\theta_2)/2$, тогда четвертое уравнение

$$\theta_1\theta_2 = \left(\frac{\theta_1+\theta_2}{2}\right)^2$$

равносильно равенству между средним геометрическим и средним арифметическим между неотрицательными θ_1 и θ_2, что возможно лишь в случае $\theta_1 = \theta_2$. Тогда $\varphi = \theta_1 = \theta_2$. Теорема доказана.

2.6. p-value для проверки простой гипотезы

Пусть дана выборка $X = (X_1, \ldots, X_n)$, выдвинута некоторая гипотеза H_1, выбрана тестовая статистика $T = T(X)$ и вид критической области $\Omega_{\text{кр}} = \{x \in \Omega : T(x) > t\}$. Пусть требуется составить критерий проверки гипотезы H_1 на заданном уровне значимости α. В этом случае мы обычно поступали так: находили границу критической области t_α такую, что $\mathbb{P}_1(T(X) > t_\alpha) = \alpha$, и если на конкретной реализации выборки $X = x$ получилось $T(x) > t_\alpha$, то гипотезу мы отклоняли, иначе не отклоняли. Оказывается, что решение об отклонении или неотклонении можно принимать, не вычисляя границу t_α.

Действительно, достаточно вычислить вероятность $\mathbb{P}_1(T(X) > T(x))$ и сравнить ее с α. Если оказывается, что $\mathbb{P}_1(T(X) > T(x)) > \alpha$, то (в силу монотонности функции распределения $T(X)$) отсюда следует $T(x) < t_\alpha$, и гипотеза не отклоняется. Если же оказалось, что $\mathbb{P}_1(T(X) > T(x)) \le \alpha$, то отсюда следует $T(x) \ge t_\alpha$, и гипотеза отклоняется. Итак, критерий можно записать таким образом:

$$H_1 \text{ отвергается} \quad \Leftrightarrow \quad \mathbb{P}_1(T(X) > t) \le \alpha,\ t = T(x).$$

Вероятность $\mathbb{P}_1(T(X) > T(x))$ называется *p-value* или *p-значение* (p – от слова probability). Это вероятность того, что статистика критерия превышает значение, которое только что было получено в эксперименте.

Использование p-value не является чем-то принципиально отличным от использования квантилей распределения. Действительно,

если раньше для статистики критерия мы находили границу, удовлетворяющую уровню значимости (то есть находили аргумент функции распределения, квантиль), то в случае p-value мы сравниваем сами значения функций (вероятности, распределения) со значением уровня значимости. В отличие от квантилей, когда известны заранее подготовленные таблицы для наиболее распространенных уровней критерия, p-value не затабулируешь, поэтому их приходится вычислять с использованием не таблиц, а математических пакетов.

Задача 1. При снятии показаний измерительного прибора десятые доли деления шкалы прибора оцениваются «на глаз» наблюдателем. Количества цифр 0, 1, 2, ..., 9, записанных наблюдателем в качестве десятых долей при 100 независимых измерениях, равны 5, 8, 6, 12, 14, 18, 11, 6, 13, 7 соответственно. Проверить гипотезы о согласии данных с законом равномерного распределения и с законом нормального распределения. Для ответа на вопрос можно сравнить значения p-value для обеих гипотез.

Замечание. Имеется в виду дискретное равномерное распределение.

Решение. Пусть ξ обозначает истинное значение десятой доли измеряемой величины, тогда она может принимать 10 значений $j = 0, \ldots, 9$. Обозначим ν_j число событий $\{\xi = j\}$, по условию задачи

$$\{\nu_j\} = \{5, 8, 6, 12, 14, 18, 11, 6, 13, 7\}. \tag{1}$$

Истинные неизвестные вероятности обозначим $p_j = \mathbb{P}(\xi = j)$, $0 \le j \le 9$. По условию, число измерений равно $n = 100$. Уровень значимости выберем равным $\alpha = 0.05$.

а) Гипотеза о равномерном дискретном распределении. Здесь $N = 10$, $p_j^\circ = 1/10$ для всех $j = 0, \ldots, 9$, и для статистики критерия получаем

$$T_{\chi^2} = \sum_{j=0}^{9} \frac{\left(\nu_j - np_j^\circ\right)^2}{np_j^\circ} = 16.4.$$

Граница критической области $t_\alpha = \chi^2_{0.95,9} = 16.9$. Отсюда делаем вывод, что данные гипотезе не противоречат. Тот же результат получаем с использованием p-value:

$$p = \mathbb{P}_1\left(T_{\chi^2} > 16.4\right) = 0.0590 > \alpha,$$

откуда следует, что данные гипотезе не противоречат. В отличие от квантилей, наиболее часто встречающиеся значения которых можно найти в таблицах, значение p-value рассчитывается с помощью подходящего математического пакета.

б) Гипотеза о нормальном распределении. Пусть H_1 – гипотеза, состоящая в том, что выборка получена из нормального распределения $\mathrm{N}(\theta_1, \theta_2^2)$, вектор параметров $\theta = (\theta_1, \theta_2^2)$. Так как математическое ожидание и дисперсия не заданы, то гипотеза сложная. Более того, гипотеза состоит из непрерывных распределений. Поэтому чтобы использовать критерий хи-квадрат, применим сначала метод группировки данных. Для этого разобьем числовую ось на $N = 10$ интервалов:

$$\Delta_0 = (-\infty, 0.5]\,,\ \Delta_1 = (0.5, 1.5]\,,\ \ldots,\ \Delta_8 = (7.5, 8.5]\,,\ \Delta_9 = (8.5, +\infty)\,.$$

Пусть ν_j обозначает количество измерений, попавших в интервал Δ_j, для каждого $j = 0, \ldots, N-1$; они принимают те же значения, что и в выражении (1). Введем функцию плотности нормального распределения:

$$f\,(x;\theta) = \frac{1}{\sqrt{2\pi\theta_2^2}} \exp\left(-\frac{(x-\theta_1)^2}{2\theta_2^2}\right)$$

и вероятности попадания в каждый из интервалов Δ_j:

$$p_j^\circ\,(\theta) = \int\limits_{\Delta_j} f\,(x;\theta)\,dx,\ j = 0, \ldots, N-1.$$

Теперь нужно оценить неизвестные параметры. Оценки максимального правдоподобия находятся из системы уравнений:

$$\sum_{j=0}^{N-1} \frac{\nu_j}{p_j^\circ\,(\theta)} \frac{\partial p_j^\circ}{\partial \theta_1} = 0,\ \sum_{j=0}^{N-1} \frac{\nu_j}{p_j^\circ\,(\theta)} \frac{\partial p_j^\circ}{\partial \theta_2^2} = 0,$$

откуда получаем

$$\theta_1 = \frac{1}{n} \sum_{j=0}^{N-1} \nu_j \frac{\int_{\Delta_j} x f\,(x;\theta)\,dx}{\int_{\Delta_j} f\,(x;\theta)\,dx},\ \theta_2^2 = \frac{1}{n} \sum_{j=0}^{N-1} \nu_j \frac{\int_{\Delta_j} (x-\theta_1)^2 f\,(x;\theta)\,dx}{\int_{\Delta_j} f\,(x;\theta)\,dx}.$$

Полученная система уравнений успешно решается численно, если в качестве начального приближения взять

$$\hat\theta_1^{(0)} = \frac{1}{n} \sum_{j=0}^{N-1} \nu_j z_j,\ \hat\theta_2^{2(0)} = \frac{1}{n} \sum_{j=0}^{N-1} \nu_j \Big(z_j - \hat\theta_1^{(0)}\Big)^2,$$

где z_j – любая точка из интервала Δ_j при $j = 0, \ldots, N-1$. Возьмем для определенности $z_j = j$. Тогда

$$\hat\theta_1^{(0)} = 4.7700,\ \hat\theta_2^{2(0)} = 6.2771.$$

Процедура численного решения исходных уравнений завершается уже через три итерации при допустимой относительной точности 10^{-15}:

$$\hat{\theta}_1 = 4.7897,\ \hat{\theta}_2^2 = 7.1798.$$

Статистика критерия, отвечающая точному решению, равна

$$\hat{T}_{\chi^2} = \sum_{j=0}^{N-1} \frac{\left(\nu_j - n\hat{p}_j^{\circ}\right)^2}{n\hat{p}_j^{\circ}} = 9.8026,$$

в то время как та же статистика для начального приближения равна

$$\hat{T}_{\chi^2}^{(0)} = \sum_{j=0}^{N-1} \frac{\left(\nu_j - n\hat{p}_j^{\circ(0)}\right)^2}{n\hat{p}_j^{\circ(0)}} = 10.7990.$$

Пусть уровень значимости равен $\alpha = 0.05$. Так как $N = 10$ и число параметров равно 2, то $t_\alpha = \chi^2_{0.95,7} = 14.1$. В этом случае гипотезу H_1 не отклоняем. Что касается p-value, то для точного решения

$$p = \mathbb{P}_0\left(\hat{T}_{\chi^2} > 9.8026\right) = 0.2000 > \alpha,$$

а для начального приближения

$$p^{(0)} = \mathbb{P}_0\left(\hat{T}_{\chi^2}^{(0)} > 10.7990\right) = 0.1476 > \alpha.$$

Заметим, что p-value при переходе от начального приближения к точному меняется достаточно сильно, на 35.5%, то есть $|p - p^{(0)}|/p^{(0)} = 0.355$. Вместе с тем численное решение системы уравнений быстро дает точный ответ. Пользоваться приближенными значениями вместо точных смысла нет.

Замечание. Для уменьшения расчетов мы могли бы сначала оценить параметры нормального распределения, например, получив

$$\hat{\theta}_1 \approx \frac{1}{n}\sum_{j=0}^{N-1} j\nu_j = 4.7700,\ \hat{\theta}_2^2 \approx \frac{1}{n}\sum_{j=0}^{N-1} \nu_j\left(j - \hat{\theta}_1\right)^2 = 6.2771,$$

а затем в качестве интервалов выбрать

$$\Delta_0 = (-\infty, 0.77]\,,\ \Delta_1 = (0.77, 1.77]\,,\ \Delta_8 = (7.77, 8.77]\,,\ \Delta_9 = (8.77, +\infty)\,.$$

В силу симметрии нормального распределения достаточно было бы только вычислить p_0, p_1, и p_2, откуда $p_3 = 0.5 - (p_0 + p_1 + p_2)$. Вероятности же на остальных участках будут равны соответствующим

вероятностям на этих участках. Стоит отметить, впрочем, что в данном случае интервалы группировки оказываются случайными, ведь если провести измерения вновь, то результаты будут другими и интервалы, которые мы подбираем для симметрии, тоже окажутся другими. В классическом методе хи-квадрат интервалы группировки заданы априори и от данных не зависят. Поэтому классический метод хи-квадрат для апостериори подобранных интервалов применять нельзя.

Задача 2. За первый час счетчиком зарегистрировано 150 событий пуассоновского потока, за следующие два часа – 250 событий. Была ли постоянной интенсивность наступления событий в единицу времени в течение всех трех часов наблюдения (уровень значимости α принять равным 0.05)?

Замечание. Гипотеза H_1 состоит в том, что дан пуассоновский поток с постоянной интенсивностью λ. Пусть ν_1 обозначает число событий за первый час, а ν_2 – число событий за последующие два часа. В силу H_1 случайные величины ν_1 и ν_2 независимы. Тогда

$$\mathbb{P}_0\left(\nu_1 = k_1, \nu_2 = k_2\right) = e^{-\lambda}\frac{\lambda^{k_1}}{k_1!}e^{-2\lambda}\frac{(2\lambda)^{k_2}}{k_2!} = e^{-3\lambda}\frac{\lambda^{k_1}(2\lambda)^{k_2}}{k_1!k_2!}.$$

Параметр λ неизвестен, а групп всего две: первый час и два последующих. Значит $N = 2$, $r = 1$ и $N - r - 1 = 0$ и критерием хи-квадрат воспользоваться не удастся, не хватает групп.

Решение. Однако распределение Пуассона обладает одним важным и интересным свойством, которое поможет нам избавиться от неизвестного параметра. Заметим, что $\nu_1 + \nu_2$ – тоже пуассоновская случайная величина с интенсивностью $\lambda + 2\lambda = 3\lambda$, поэтому

$$\mathbb{P}_0\left(\nu_1 + \nu_2 = k\right) = e^{-3\lambda}(3\lambda)^k/k!.$$

Теперь вычислим условную вероятность

$$\mathbb{P}_0\left(\nu_1 = k_1 | \nu_1 + \nu_2 = k\right) = \frac{\mathbb{P}_0\left(\nu_1 = k_1, \nu_1 + \nu_2 = k\right)}{\mathbb{P}_0\left(\nu_1 + \nu_2 = k\right)} =$$

$$= \frac{\mathbb{P}_0\left(\nu_1 = k_1, \nu_2 = k - k_1\right)}{\mathbb{P}_0\left(\nu_1 + \nu_2 = k\right)} = \frac{\dfrac{e^{-\lambda}\lambda^{k_1}}{k_1!} \cdot \dfrac{e^{-2\lambda}(2\lambda)^{k-k_1}}{(k-k_1)!}}{\dfrac{e^{-3\lambda}(3\lambda)^k}{k!}} =$$

$$= \frac{k!}{k_1!\,(k-k_1)!}\left(\frac{2}{3}\right)^{k-k_1}\left(\frac{1}{3}\right)^{k_1}.$$

Видно, что условное распределение представляет собой биномиальное распределение $\mathrm{Bi}(k, 1/3)$. Здесь $k = 400$ – число испытаний, $k_1 = 150$ – число успехов, $1/3$ – вероятность успеха. Условное распределение не зависит от неизвестного параметра λ. Также можно заметить, что с учетом $k = 400$ распределение хорошо приближается нормальным распределением с параметрами $\mu = k/3 = 133.33$ и $\sigma^2 = k \cdot 1/3 \cdot 2/3 = = 88.89$.

Теперь, в качестве тестовой статистики будем рассматривать величину ν_1, а вероятности каких-либо событий будем подсчитывать не только при условии H_1, но и при условии $\nu_1 + \nu_2 = k$, получая таким образом условные вероятности. Формально это значит, что вместо исходного вероятностного пространства рассматривается его сужение на событие $\{\nu_1+\nu_2 = k\}$. Далее, будем считать, что критическая область представляет собой все достаточно далекие от среднего μ значения ν_1. Действительно, если ν_1 слишком маленькое, то это говорит в пользу маленькой интенсивности; наоборот, при большом значении ν_1 мы ожидаем высокую интенсивность процесса. Формально критическая область записывается в виде

$$\Omega_{\text{кр}} = \{x \in \mathbb{Z} : x \geq 0,\ |x - \mu| \geq t_\alpha\},$$

где t_α находится из условия на уровень значимости:

$$\mathbb{P}_1(|\nu_1 - \mu| \geq t_\alpha \,|\, \nu_1 + \nu_2 = k) = \alpha.$$

Так как $k = 400$ достаточно велико, то можно приближенно считать симметричным распределение ν_1 относительно математического ожидания μ. Поэтому условие на уровень значимости можно переписать так:

$$\mathbb{P}_1(\nu_1 \geq \mu + t_\alpha \,|\, \nu_1 + \nu_2 = k) = \alpha/2. \tag{2}$$

Теперь вычислим сумму $(\mu + t_\alpha)$ как $(1 - \alpha/2)$-квантиль распределения $\mathrm{Bi}(k, \frac{1}{3})$ или распределения $\mathrm{N}(\mu, \sigma^2)$. Получим $(\mu + t_\alpha) = 152$ для биномиального распределения и $(\mu + t_\alpha) = 151.8120$ для нормального распределения. Далее, $(\mu + t_\alpha)$ сравним со значением $k_1 = 150$. Так как $k_1 < (\mu + t_\alpha)$, то отсюда следует вывод, что опытные данные гипотезе не противоречат. Заметим также, что в силу дискретности биномиального распределения подобрать целое значение $(\mu + t_\alpha)$, чтобы в уравнении (2) достигалось равенство, не удастся. Поэтому нужно взять минимальное значение $(\mu + t_\alpha)$, при котором выполняется неравенство $\mathbb{P}_1(\nu_1 \geq \mu + t_\alpha | \nu_1 + \nu_2 = k) \leq \alpha/2$. Фактически, ошибка первого рода будет немного меньше α. Нормальное распределение непрерывно,

поэтому всегда удастся подобрать $(\mu + t_\alpha)$, чтобы было достигнуто равенство в выражении (2).

Ответ можно было бы получить с помощью расчета значения p-value. Для биномиального распределения имеем

$$\mathbb{P}_0\left(\nu_1 \geq k_1 \,|\, \nu_1 + \nu_2 = k\right) = 0.0353,$$

а для нормального получаем

$$\mathbb{P}_0\left(\nu_1 \geq k_1 \,|\, \nu_1 + \nu_2 = k\right) = 1 - \Phi\left(\frac{k_1 - \mu}{\sigma}\right) = 0.0385.$$

Так как уровень значимости $\alpha/2 = 0.025 < 0.0353$, то делаем тот же вывод: *опытные данные гипотезе не противоречат.*

3. Критерии независимости, однородности, случайности

3.1. Критерий независимости хи-квадрат

Условия. Дан случайный вектор $\xi = (\xi_1, \xi_2)$, где $\xi_1 \in \{1, \ldots, k\}$, а $\xi_2 \in \{1, \ldots, s\}$. Распределение и степень зависимости ξ_1 и ξ_2 неизвестны. Число измерений вектора ξ равно n. Обозначим $p_{ij} = \mathbb{P}(\xi_1 = i, \xi_2 = j)$, $i = 1, \ldots, k$, $j = 1, \ldots, s$. Выдвинута гипотеза (*гипотеза независимости*):

$$H_1: \quad \mathbb{P}(\xi_1 = i, \xi_2 = j) = \mathbb{P}(\xi_1 = i)\mathbb{P}(\xi_2 = j) \quad \forall i, j,$$

или, что то же самое,

$$H_1: \quad p_{ij} = \sum_{m=1}^{s} p_{im} \sum_{l=1}^{k} p_{lj} \equiv p_{i\bullet} p_{\bullet j} \quad \forall i, j.$$

Составить критерий проверки гипотезы H_1 на заданном уровне значимости α.

Алгоритм

1. Рассчитать ν_{ij} – число наблюдений пары (i, j).

2. Оценить неизвестные параметры $p_{i\bullet}$ и $p_{\bullet j}$ методом максимального правдоподобия в предположении H_1:

$$(\hat{p}_{i\bullet}, \hat{p}_{\bullet j}) = \arg \max_{p_{i\bullet}, p_{\bullet j}} \prod_{i,j} (p_{i\bullet} p_{\bullet j})^{\nu_{ij}}.$$

Решение этой задачи известно:

$$\hat{p}_{i\bullet} = \frac{\nu_{i\bullet}}{n},\ \hat{p}_{\bullet j} = \frac{\nu_{\bullet j}}{n}.$$

3. Вычислить статистику критерия

$$\hat{T}_{\chi^2} = \sum_{i=1}^{k}\sum_{j=1}^{s}\frac{(\nu_{ij} - n\hat{p}_{i\bullet}\hat{p}_{\bullet j})^2}{n\hat{p}_{i\bullet}\hat{p}_{\bullet j}},$$

а в качестве критической области выбрать $\Omega_{\text{кр}} = \{x \in \Omega : \hat{T}_{\chi^2}(x) > t_\alpha\}$, где t_α найти из условия на уровень значимости. Известно, что

$$\hat{T}_{\chi^2} \xrightarrow{d} \chi^2((s-1)(k-1))$$

при $n \to \infty$. Таким образом, при большом количестве измерений можно приближенно вычислять t_α как $(1-\alpha)$-квантиль распределения $\chi^2((s-1)(k-1))$. Итак, критерий независимости хи-квадрат выглядит следующим образом:

$$H_1 \text{ отвергается} \Leftrightarrow \hat{T}_{\chi^2} > \chi^2_{1-\alpha}((s-1)(k-1)).$$

Замечание 1. Известно, что по ЗБЧ $\nu_{ij}/n \xrightarrow{\mathbb{P}} p_{ij}$, $\nu_{i\bullet}/n \xrightarrow{\mathbb{P}} p_{i\bullet}$ и $\nu_{\bullet j}/n \xrightarrow{\mathbb{P}} p_{\bullet j}$ при $n \to \infty$. При этом, если гипотеза H_1 верна и $p_{ij} = p_{i\bullet}p_{\bullet j}$, то следует ожидать, что отклонение эмпирических данных от гипотетических $(\nu_{ij} - n\hat{p}_{i\bullet}\hat{p}_{\bullet j})^2$ будет малой величиной.

Замечание 2. Гипотеза является сложной, так как конкретное распределение не фиксировано. Поэтому и прибегают к оценкам неизвестных параметров распределения – в данном случае $p_{i\bullet}$ и $p_{\bullet j}$. Кроме того, эта задача сводится к задаче проверки сложной параметрической гипотезы. Этим объясняется вид функции правдоподобия и количество степеней свободы в распределении хи-квадрат: $p_{i\bullet}$ и $p_{\bullet j}$ образуют $r = s + k - 2$ неизвестных параметров, а число возможных значений вектора ξ равно $N = sk$, следовательно число степеней равно $N - 1 - r = sk - 1 - s - k + 2 = (s-1)(k-1)$.

Замечание 3. Так как данный критерий сводится к критерию хи-квадрат проверки сложной параметрической гипотезы, то и рекомендации к частотам и объему выборки для пользования асимптотическим результатом распределения статистики хи-квадрат остаются теми же: $n \geq 50$, $\nu_{ij} \geq 5\ \forall i, j$.

Задача 1. В таблице приведены статистические данные о приеме в ВУЗ (здесь A – принят, B – мужчина). Проверить гипотезу о

	B	$\bar{B}$	Σ
A	97	40	137
$\bar{A}$	263	42	305
Σ	360	82	442

независимости признаков «результат» и «пол» на уровне значимости $\alpha = 0.001$, то есть 0.1%.

Решение. Формально нам даны две дискретные случайные величины, которые могут принимать два значения, поэтому $s = k = 2$. По условию задачи нам уже даны частоты исходов $\nu_{11} = 97$, $\nu_{12} = 40$, $\nu_{21} = 263$, $\nu_{22} = 42$, а также суммы $\nu_{1\bullet} = 137$, $\nu_{2\bullet} = 305$, $\nu_{\bullet 1} = 360$, $\nu_{\bullet 2} = 82$. Объем выборки

$$n = 442 = \nu_{\bullet\bullet} = \nu_{1\bullet} + \nu_{2\bullet} = \nu_{\bullet 1} + \nu_{\bullet 2}.$$

Сначала оцениваем вероятности:

$$\hat{p}_{1\bullet} = \frac{\nu_{1\bullet}}{n} = 0.3099,\ \hat{p}_{2\bullet} = \frac{\nu_{2\bullet}}{n} = 0.6900,$$

$$\hat{p}_{\bullet 1} = \frac{\nu_{\bullet 1}}{n} = 0.8145,\ \hat{p}_{\bullet 2} = \frac{\nu_{\bullet 2}}{n} = 0.1855.$$

Теперь вычисляем статистику критерия:

$$\hat{T}_{\chi^2} = \frac{(\nu_{11} - n\hat{p}_{1\bullet}\hat{p}_{\bullet 1})^2}{n\hat{p}_{1\bullet}\hat{p}_{\bullet 1}} + \frac{(\nu_{12} - n\hat{p}_{1\bullet}\hat{p}_{\bullet 2})^2}{n\hat{p}_{1\bullet}\hat{p}_{\bullet 2}} + \\ + \frac{(\nu_{21} - n\hat{p}_{2\bullet}\hat{p}_{\bullet 1})^2}{n\hat{p}_{2\bullet}\hat{p}_{\bullet 1}} + \frac{(\nu_{22} - n\hat{p}_{2\bullet}\hat{p}_{\bullet 2})^2}{n\hat{p}_{2\bullet}\hat{p}_{\bullet 2}} =$$

$$= 1.9061 + 8.3680 + 0.8562 + 3.7588 = 14.8890.$$

Далее, это значение сравниваем с границей критической области:

$$t_\alpha = \chi^2_{1-\alpha}((s-1)(k-1)) = \chi^2_{0.999}(1) = 10.8.$$

Так как $14.8890 > 10.8$, то гипотезу о независимости следует отвергнуть. Утверждая, что данные зависимы, мы совершаем возможную ошибку с вероятностью α, то есть 0.1%.

Замечание. Для $s = k = 2$ можно доказать, что $\hat{T}_{\chi^2} = Z_n^2$, где

$$Z_n = \left(\frac{\nu_{11}}{\nu_{1\bullet}} - \frac{\nu_{21}}{\nu_{2\bullet}}\right)\sqrt{\frac{n\nu_{1\bullet}\nu_{2\bullet}}{\nu_{\bullet 1}\nu_{\bullet 2}}} \equiv \left(\frac{\nu_{11}}{\nu_{\bullet 1}} - \frac{\nu_{12}}{\nu_{\bullet 2}}\right)\sqrt{\frac{n\nu_{\bullet 1}\nu_{\bullet 2}}{\nu_{1\bullet}\nu_{2\bullet}}}.$$

Статистика Z_n является знакопеременной, ее знак может нести дополнительную информацию о направлении отклонения от основной гипотезы. А именно, можно доказать, что

$$\frac{Z_n}{\sqrt{n}} \xrightarrow{\mathbb{P}} \rho = \operatorname{cov}(\xi_1, \xi_2),\ n \to \infty.$$

Таким образом, Z_n может быть использована для оценки степени корреляции данных. Например, в нашей задаче получаем

$$Z_n = \left(\frac{97}{137} - \frac{263}{305}\right)\sqrt{\frac{442 \cdot 137 \cdot 305}{360 \cdot 82}} = -3.8586,$$

а оценка корреляционного момента $\rho = -3.8586/\sqrt{442} = -0.1835$. Можно проверить, что $(-3.8586)^2 = 14.8890$. Так как $\rho < 0$, то корреляция между A и B отрицательная, а корреляция между A и $\bar{B}$ положительная. Это можно интерпретировать так: чем больше приходит мужчин, тем больше суммарно отказов. Чем больше приходит женщин, тем больше суммарно принятых. В таблице видна и другая «зависимость»: если ты мужчина, тебе скорее откажут, чем примут. Если ты женщина, тебя скорее примут, чем откажут. Несмотря на кажущуюся простоту корреляционного анализа, нужно всегда помнить: статистические взаимосвязи не означают прямые причинно-следственные. Например, ущерб от пожара обычно положительно коррелирует с числом пожарных, участвовавших в тушении пожара. Но глупо было бы сокращать число пожарных, чтобы уменьшить ущерб от пожара. Другими словами, управлять системой на основе статистической зависимости не получится.

3.2. Критерий однородности хи-квадрат

Условия. Даны k независимых случайных величин $\xi_1, \ldots, \xi_k$, каждая из которых может принимать целые значения от 1 до N. Пусть проведено n_1 измерений ξ_1, n_2 измерений ξ_2, ... и n_k измерений ξ_k. Обозначим $p_i = (p_{i1}, ..., p_{iN})$ векторы вероятностей $(i = 1, \ldots, k)$, где p_{ij} – вероятность того, что i-я случайная величина принимает значение j. Выдвинута гипотеза (*гипотеза однородности*):

$$H_1 : p_1 = p_2 = ... = p_k = P = (P_1, ..., P_N),$$

где P – некоторый (неизвестный) вектор вероятностей, $P_1 + ... + P_N = 1$. Составить критерий проверки гипотезы H_1 на уровне значимости α.

Алгоритм

1. Рассчитать ν_{ij} – частоты того, что i-я случайная величина приняла j-е значение.

2. Оценить неизвестный вектор параметров P методом максимального правдоподобия:

$$\hat{P} = \arg\max_P \prod_{i,j} P_j^{\nu_{ij}} = \arg\max_P \prod_j P_j^{\nu_{\bullet j}}, \text{ где } \nu_{\bullet j} = \sum_{i=1}^{k} \nu_{ij}.$$

Решение этой задачи известно и выражается формулой $\hat{P}_j = \nu_{\bullet j}/n$, $j = 1, ..., N$, $n = n_1 + ... + n_k$.

3. Вычислить тестовую статистику:

$$\hat{T}_{\chi^2} = \sum_{i=1}^{k} T_{\chi_i^2} = \sum_{i=1}^{k} \sum_{j=1}^{N} \frac{\left(\nu_{ij} - n_i \hat{P}_j\right)^2}{n_i \hat{P}_j},$$

а в качестве критической области выбрать $\Omega_{\text{кр}} = \{x \in \Omega : \hat{T}_{\chi^2}(x) > t_\alpha\}$, где граница t_α будем искать исходя из условия на уровень значимости. Так как известно, что при всех $n_i \to \infty$, $i = 1, \ldots, k$:

$$\hat{T}_{\chi^2} \xrightarrow{d} \chi^2((k-1)(N-1)),$$

то границу t_α вычисляют как $(1-\alpha)$-квантиль распределения $\chi^2((k-1)(N-1))$. Итак, критерий однородности хи-квадрат выглядит следующим образом:

$$H_1 \text{ отвергается} \Leftrightarrow \hat{T}_{\chi^2} > \chi^2_{1-\alpha}((k-1)(N-1)).$$

Замечание 1. Появление n как суммы объемов всех выборок и $\nu_{\bullet j}$ как суммы частот исхода j можно интерпретировать как объединение выборок в одну выборку объема n с общим распределением, которое определяется вектором вероятности P. Это объединение возникает естественным образом, ведь в рамках гипотезы H_1 все выборки получены из одного и того же распределения.

Замечание 2. Величины

$$T_{\chi_i^2} = \frac{\left(\nu_{ij} - n_i \hat{P}_j\right)^2}{n_i \hat{P}_j}$$

можно интерпретировать как меры хи-квадрат для каждой из случайных величин в отдельности, а итоговую меру хи-квадрат $\hat{T}_{\chi^2}$ – как сумму таких мер.

Замечание 3. Чтобы воспользоваться асимптотическим результатом распределения статистики хи-квадрат рекомендуется $n_i \geq 50$, $\nu_{\bullet j} \geq 5\ \forall i, j$.

Задача 2. Поступающие в ВУЗ абитуриенты разбиты на два потока по 300 человек в каждом. Итоги экзамена по одному и тому же предмету на каждом потоке оказались следующими: на 1-м потоке баллы 2, 3, 4 и 5 получили соответственно 33, 43, 80 и 144 человека; соответствующие же данные для второго потока – 39, 35, 72 и 154. Можно ли при уровне значимости $\alpha = 0.05$ считать оба потока однородными?

Решение. В данной задаче имеются две выборки, они отвечают разным потокам. Случайные величины в каждом потоке могут принимать четыре значения. Формально имеем $k = 2$, $N = 4$, $n_1 = n_2 = 300$, $n = n_1 + n_2 = 600$. Для удобства сразу сформируем таблицу данных:

	«2»	«3»	«4»	«5»
1 поток	33	43	80	144
2 поток	39	35	72	154
Σ	72	78	152	298

Здесь $\nu_{\bullet 1} = 72$, $\nu_{\bullet 2} = 78$, $\nu_{\bullet 3} = 152$, $\nu_{\bullet 4} = 298$. Оценим вероятности исходов:

$$\hat{P}_1 = 0.12,\ \hat{P}_2 = 0.13,\ \hat{P}_3 = 0.2533,\ \hat{P}_4 = 0.4967.$$

Теперь вычисляем реализацию статистики:

$$\hat{T}_{\chi^2} = \frac{(\nu_{11} - n_1\hat{P}_1)^2}{n_1\hat{P}_1} + \frac{(\nu_{12} - n_1\hat{P}_2)^2}{n_1\hat{P}_2} + \cdots + \frac{(\nu_{24} - n_2\hat{P}_4)^2}{n_2\hat{P}_4} = 2.0771.$$

Теперь это значение сравниваем с границей:

$$t_\alpha = \chi^2_{1-\alpha}((k-1)(N-1)) = \chi^2_{0.95}(3) = 7.82.$$

Так как $2.0771 < 7.82$, то наш ответ: «данные гипотезе не противоречат».

Замечание. Однородность потоков означает независимость (в бытовом понимании этого слова) номера потока от получения абитуриентами тех или иных оценок. Именно в бытовом понимании, потому что номер потока в этой задаче неслучайный, нам не нужно устанавливать зависимость случайности попадания на поток от случайности получить ту или иную оценку. Даже если бы в условии задачи было явно сказано, что распределение по потокам случайно, то ставить задачу проверки гипотезы о независимости было бы непрактично, ибо в таком

случае мы бы проверяли стохастическую зависимость распределяющей процедуры от полученных оценок (ясно и так, что они независимы во всех смыслах).

3.3. Критерий случайности (инверсий)

Условия. Дана выборка (не обязательно простая) $X = (X_1, \ldots, X_n)$ с неизвестным непрерывным распределением. Выдвинута гипотеза

$$H_1 : F_X(x_1, \ldots, x_n) = F(x_1) \ldots F(x_n),$$

где $F(x)$ – некоторая одномерная и неизвестная функция непрерывного распределения. Требуется составить критерий проверки гипотезы H_1 на уровне значимости α.

Алгоритм

1. Сначала для выборки $(X_1, ..., X_n)$ вычисляется так называемое *число инверсий*. Говорят, что компоненты X_i и X_j образуют инверсию, если $i < j$, но $X_i > X_j$. Введем обозначение

$$\eta_i = \sum_{j=i+1}^{n} I(X_j < X_i),\ i = 1, \ldots, n-1,$$

тогда число инверсий в выборке определяется по формуле

$$T(X) = \sum_{i=1}^{n-1} \eta_i.$$

2. Статистика $T = T(X)$ рассматривается как тестовая статистика, а в качестве критической области выбирается

$$\Omega_{\text{кр}} = \left\{ x \in \Omega : \left| T(x) - \frac{n(n-1)}{4} \right| > t_\alpha \right\}.$$

Критическую границу t_α выбирают из условия на уровень значимости

$$\mathbb{P}_1 \left(\left| T(X) - \frac{n(n-1)}{4} \right| > t_\alpha \right) \leq \alpha.$$

При достаточно больших n, можно пользоваться нормальным приближением:

$$\frac{T - \mathbb{E}T}{\sqrt{\mathbb{D}T}} \xrightarrow{d} \xi_{N(0,1)} \in N(0,1),\ n \to \infty,$$

где

$$\mathbb{E}T = \frac{n(n-1)}{4},\ \mathbb{D}T = \frac{n(n-1)(2n+5)}{72}.$$

В таких случаях можно приближенно считать

$$t_\alpha = \frac{t_{\alpha/2} n^{3/2}}{6},\ \Phi(-t_{\alpha/2}) = \frac{\alpha}{2}.$$

Итак, критерий инверсий в случае больших n выглядит следующим образом:

$$H_1 \text{ отвергается} \Leftrightarrow \left| T - \frac{n(n-1)}{4} \right| > \frac{t_{\alpha/2} n^{3/2}}{6},\ t_{\alpha/2} = -\mathrm{N}_{\alpha/2}(0,1).$$

Замечание 1. Идея метода исходит из того, что в случае верности гипотезы H_1 ожидается, что компоненты X не будут как-либо упорядочены, потому что будут стохастически независимы в совокупности (то есть здесь «случайность» понимается как стохастическая независимость). Статистика T используется как мера беспорядка данных. При этом крайние случаи, то есть когда $X_1 < ... < X_n$ или $X_n < ... < X_1$, естественно рассматривать как свидетельства «полного отсутствия беспорядка». Поэтому-то значения T, достаточно далекие от среднего значения, записывают в критическую область.

Замечание 2. Утверждение гипотезы о независимости компонент выборки используется при расчете математического ожидания и дисперсии статистики T, а также в теореме о предельном нормальном распределении.

Замечание 3. Нормальное приближение рекомендуется применять при $n > 10$.

Задача 3. Измерения одной непрерывной случайной величины дали следующую выборку: 0.8147, 0.9058, 0.1270, 0.9134, 0.6324, 0.0975, 0.2785, 0.5469, 0.9575, 0.9649, 0.1576, 0.9706, 0.9572, 0.4854, 0.8003, 0.1419, 0.4218, 0.9157, 0.7922, 0.9595. Объем данной выборки $n = 20$. Проверить гипотезу о случайности на уровне значимости $\alpha = 0.01$.

Решение. Подсчитать количество инверсий можно так. Берем первое число 0.8147 и смотрим, сколько раз это число оказывается больше по величине со следующими за ним в списке, то есть со 2-го по 20-е. Это число $\eta_1 = 11$. Затем выбираем второе число в выборке 0.9058, сравниваем его с элементами выборки с 3-го по 20-е: $\eta_2 = 11$. Аналогично можно получить число инверсий для каждого элемента выборки, кроме последнего: $\eta_3 = 1$, $\eta_4 = 10$, $\eta_5 = 7$, $\eta_6 = 0$, $\eta_7 = 2$, $\eta_8 = 4$, $\eta_9 = 8$, $\eta_{10} = 9$, $\eta_{11} = 1$, $\eta_{12} = 8$, $\eta_{13} = 6$, $\eta_{14} = 2$, $\eta_{15} = 3$,

$\eta_{16} = \eta_{17} = 0$, $\eta_{18} = 1$, $\eta_{19} = 0$. Тогда суммарное число инверсий равно $T = 84$, отсюда

$$\left| T - \frac{n(n-1)}{4} \right| = 11.$$

Так как объем выборки достаточно большой, то границу t_α вычислим по приближенной формуле

$$t_\alpha = \frac{t_{\alpha/2} n^{3/2}}{6} = 38.3982,\ t_{\alpha/2} = -\mathrm{N}_{\alpha/2}(0,1) = 2.5758.$$

Так как $11 < 38.3982$, то данные гипотезе не противоречат.

4. Наиболее мощные критерии

4.1. Критерий Неймана–Пирсона

С этого момента нас будут интересовать задачи с двумя и более гипотезами. Ранее мы ввели понятие статистического критерия – отображения из выборочного пространства в множество гипотез. Отображения рассматривались однозначные: каждой реализации выборки соответствует одна и только одна гипотеза. Однако, как мы увидим впоследствии, не все оптимизационные задачи имеют решение в классе подобных отображений. Нам понадобится более общее понятие статистического критерия.

Определение. *Рандомизированным статистическим критерием* проверки r гипотез называется измеримое отображение $\delta : \Omega \to \mathbb{R}^{(r)}$, где $\mathbb{R}^{(r)}$ – пространство векторов вероятностей $(\pi_1, \ldots, \pi_r)$, где $\pi_j \geq 0$ $\forall j$ и $\pi_1 + \cdots + \pi_r = 1$.

Теперь каждому значению выборки будем сопоставлять не гипотезу, а распределение на множестве гипотез. Например, если гипотез две ($r = 2$), то каждой реализации выборки будет соответствовать какой-то вектор вероятностей (π_1, π_2), и для того чтобы указать на гипотезу, требуется провести дополнительный эксперимент, *рандомизацию*: подкинуть монету, которая выпадает на решку с вероятностью π_1 и на орел с вероятностью π_2. Если выпадет «решка», следует принять первую гипотезу, а если «орел» – вторую. Мы увидим, что такой, казалось бы странный, подход не только может увеличить вероятность принятия неверной гипотезы, но и увеличить вероятность принятия верной гипотезы!

Замечание. Обычный статистический критерий является частным случаем рандомизированного критерия. В этом случае π_j могут

принимать только значения 0 и 1. Вообще, когда говорят о рандомизированных критериях, то имеют ввиду все расширенное множество критериев, которое включает в себя и обычные, *нерандомизированные*, критерии. А когда говорят о нерандомизированных критериях, то имеют ввиду вот это подмножество рандомизированных критериев, в которых выбор гипотезы осуществляется без дополнительных процедур рандомизации.

Здесь речь пойдет о проверке двух простых гипотез H_1 и H_2. Гипотезу H_1 будем называть *основной* гипотезой, а H_2 – *альтернативной* гипотезой (или просто *альтернативой*). Последующие ниже определения вероятности ошибки 1-го рода, вероятности ошибки 2-го рода и мощности критерия, будут даны именно для случая двух простых гипотез. Конечно эти понятия можно обобщить на случай сложных гипотез, это будет сделано в следующих разделах.

Будем предполагать, что распределения в обеих гипотезах являются либо абсолютно непрерывными, либо дискретными.

В случае двух гипотез рандомизированный критерий

$$\delta(x) = (\pi_1(x), \pi_2(x))$$

однозначно определяется функцией $\pi_2(x)$, поскольку $\pi_1(x) = 1 - \pi_2(x)$. Функцию $\pi_2(x)$ мы будем обозначать как $\pi(x)$.

Определение. Функция $\pi(x)$ называется *критической функцией*. Соответствующий критической функции $\pi(x)$ критерий будем также обозначать буквой π.

Критическая функция $0 \le \pi(x) \le 1$ для любого $x \in \Omega$. Если она для каждого $x \in \Omega$ принимает лишь одно из двух значений $\{0, 1\}$, то соответствующий критерий мы называем нерандомизированным. В этом случае выборочное пространство разбивается на два непересекающихся подмножества $\Omega = \Omega_1 \cup \Omega_2$, и если реализация выборки $x \in \Omega_1$, значит следует принять гипотезу H_1, а если $x \in \Omega_2$, то принимается H_2. Множество Ω_2 называется *критической областью*, а критическая функция есть ее индикаторная функция: $\pi(x) = I(x \in \Omega_2)$. Если же критическая функция может принимать значения, отличные от 0 и 1, то говорить о какой-либо критической области не принято.

Если для некоторого x выполнено $\pi(x) \in (0, 1)$, то в случае реализации выборки $X = x$ проводится процедура рандомизации: моделируется случайная величина с распределением Бернулли $\xi_{\text{Be}} \in \text{Be}(\pi(x))$, и если $\xi_{\text{Be}} = 1$, то принимается альтернативная гипотеза H_2, иначе принимается основная гипотеза H_1.

Качество критерия чаще всего характеризуется вероятностями ошибочного принятия гипотез.

Определение. *Вероятностью ошибки 1-го рода* критерия π называется

$$\alpha(\pi) = \mathbb{E}_1\pi(X).$$

Таким образом, вероятность ошибки 1-го рода определяет вероятность отклонения гипотезы H_1 при условии, что на самом деле она верна. В случае нерандомизированного критерия π с критической областью Ω_2 приходим к более простой формуле

$$\alpha(\pi) = \mathbb{P}_1(X \in \Omega_2).$$

Определение. *Вероятностью ошибки 2-го рода* называется

$$\beta(\pi) = 1 - \mathbb{E}_2\pi(X).$$

Вероятность ошибки 2-го рода определяет вероятность принятия гипотезы H_1 при условии, что на самом деле верна альтернативная гипотеза. В случае нерандомизированного критерия π с критической областью Ω_2 приходим к формуле

$$\beta(\pi) = 1 - \mathbb{P}_2(X \in \Omega_2).$$

Определение. *Мощностью критерия* π (при альтернативе) называется

$$W(\pi) = \mathbb{E}_2\pi(X) \equiv 1 - \beta(\pi).$$

Мощность критерия определяет вероятность принятия альтернативы в случае, когда она справедлива. В случае нерандомизированного критерия π с критической областью Ω_2:

$$W(\pi) = \mathbb{P}_2(X \in \Omega_2).$$

Определение. Если $\alpha(\pi) \leq \varepsilon$, то говорят, что критерий π имеет *уровень значимости* ε.

Определение. Плотность выборки (или функция вероятности выборки) называется также *функцией правдоподобия* и обозначается $L_j(x)$, если она соответствует гипотезе H_j, $j = 1, 2$.

Определение. Функция

$$l(x) = \frac{L_2(x)}{L_1(x)},$$

определенная на множестве $\{x : L_1(x) + L_2(x) > 0\}$, называется *функцией отношения правдоподобия*. Если $L_1(x) = 0$ и $L_2(x) > 0$, то считается $l(x) = +\infty$.

Если $L_1(x) = L_2(x) = 0$, то можно принимать любое решение – H_1 или H_2, так как вероятности ошибок 1-го и 2-го рода от этого зависеть не будут.

Определение. *Наиболее мощным критерием* проверки двух простых гипотез на уровне значимости ε называется рандомизированный критерий π, который является решением оптимизационной задачи:

$$W(\pi) \to \max, \quad \alpha(\pi) = \varepsilon.$$

Решение этой оптимизационной задачи не обязано существовать для любых ε, а в случае существование не обязано быть единственным. Легко построить примеры, когда существует бесконечное множество наиболее мощных критериев. У этих критериев равны мощности и вероятности ошибки первого рода.

Оказывается, существует простой алгоритм построения наиболее мощного критерия. Введем статистику $Z = l(X)$.

Лемма Неймана–Пирсона (неполный вариант)[5]. *Для любого $\varepsilon > 0$ такого, что $\mathbb{P}_1(Z > 0) \geq \varepsilon$, существуют числа $c > 0$ и $p \in [0, 1]$ такие, что критерий π с критической функцией*

$$\pi_{c,p}(x) = \begin{cases} 1 & \text{при } \ l(x) > c \\ p & \text{при } \ l(x) = c \\ 0 & \text{при } \ l(x) < c \end{cases}$$

имеет вероятность ошибки 1-го рода, равную ε, то есть $\pi_{c,p} \in K_\varepsilon \equiv \equiv \{\pi : \alpha(\pi) = \varepsilon\}$ и является наиболее мощным критерием в K_ε. Эти числа c и p удовлетворяют уравнениям

$$\alpha(\pi_{c,p}) \equiv \mathbb{E}_1 \pi_{c,p}(X) \equiv \mathbb{P}_1(Z > c) + p\mathbb{P}_1(Z = c) = \varepsilon.$$

Более того, любые числа c и p, удовлетворяющие этому уравнению, дают наиболее мощный критерий $\pi_{c,p}$ в K_ε.

Определение. Наиболее мощный критерий $\pi_{c,p}$, о котором идет речь в лемме, называется *критерием Неймана–Пирсона.*

Критерий Неймана–Пирсона уровня ε является наиболее мощным критерием уровня ε. Но не всякий наиболее мощный критерий является критерием Неймана–Пирсона, то есть имеет вид функции $\pi_{c,p}$ из леммы. Если в условии задачи требуется найти наиболее мощный критерий (любой), то можно воспользоваться для этого леммой Неймана–Пирсона или же вывести другой наиболее мощный критерий,

[5] Боровков А.А. Математическая статистика : учебник. 4-е изд., стер., 2010. С. 311.

доказав, что он действительно наиболее мощный. Если же спрашивается найти именно критерий Неймана–Пирсона, то нужно искать критерий из леммы.

Задача 1. Построить наиболее мощный критерий для проверки гипотез

$$H_1 : \ \mathrm{N}(0,1)$$
$$H_2 : \ \mathrm{N}(0,4)$$

на уровне значимости $0 < \varepsilon < 1$. Объем выборки считать равным $n \geq 1$. Также найти мощность полученного критерия и вероятность ошибки второго рода.

Решение. Сначала запишем функции правдоподобия для обеих гипотез:

$$L_1(x) = f_{\mathrm{N}(0,1)}(x_1) \dots f_{\mathrm{N}(0,1)}(x_n) = \left(\frac{1}{\sqrt{2\pi \cdot 1}}\right)^n \exp\left(-\sum_{j=1}^{n} \frac{(x_j - 0)^2}{2 \cdot 1}\right),$$

$$L_2(x) = f_{\mathrm{N}(0,4)}(x_1) \dots f_{\mathrm{N}(0,4)}(x_n) = \left(\frac{1}{\sqrt{2\pi \cdot 4}}\right)^n \exp\left(-\sum_{j=1}^{n} \frac{(x_j - 0)^2}{2 \cdot 4}\right).$$

Теперь составим функцию отношения правдоподобия:

$$l(x) = \frac{L_2(x)}{L_1(x)} = \frac{1}{2^n} \exp\left(-\sum_{j=1}^{n} \left[\frac{x_j^2}{8} - \frac{x_j^2}{2}\right]\right) = \frac{1}{2^n} \exp\left(\sum_{j=1}^{n} \frac{3x_j^2}{8}\right).$$

Лемма Неймана–Пирсона утверждает, что наиболее мощный критерий с уровнем значимости $\varepsilon \leq \mathbb{P}_1(Z > 0) \equiv 1$ можно искать среди критериев вида

$$\pi_{c,p}(x) = \begin{cases} 1 & \text{при} \ \ l(x) > c, \\ p & \text{при} \ \ l(x) = c, \\ 0 & \text{при} \ \ l(x) < c. \end{cases}$$

Конкретные значения $c > 0$ и $p \in [0, 1]$ находятся из условия на уровень значимости

$$\mathbb{P}_1(Z > c) + p\mathbb{P}_1(Z = c) = \varepsilon,$$

где $Z = l(X)$. В нашем случае случайная величина Z является непрерывной при первой гипотезе, поэтому $\mathbb{P}_1(Z = c) = 0$ для любого $c > 0$. Отсюда получаем одно уравнение на c: $\mathbb{P}_1(Z > c) = \varepsilon$, таким образом, c является $(1-\varepsilon)$-квантилем распределения случайной величины Z. Что

же касается p, то его можно выбирать любым. Пусть будет, например, $p = 0$.

Условие на c бывает удобно заменить на эквивалентное, которое бы опиралось на квантиль какого-нибудь известного распределения. Заметим, что неравенство $l(x) > c$ равносильно неравенству $\sum_{j=1}^{n} x_j^2 > > \bar{c}$ с некоторым новым параметром $\bar{c}$, который как-то (не важно как) связан с параметром c. Таким образом,

$$\mathbb{P}_1(Z > c) = \mathbb{P}_1\left(\sum_{j=1}^{n} X_j^2 > \bar{c}\right) = \varepsilon.$$

Так как в силу гипотезы H_1 случайные величины X_j распределены по стандартному нормальному закону $\mathrm{N}(0,1)$ и независимы, то

$$\sum_{j=1}^{n} X_j^2 \in \chi^2(n),$$

и значение $\bar{c}$ находится как $(1-\varepsilon)$-квантиль распределения $\chi^2(n)$. Окончательно, критерий Неймана–Пирсона можно записать в виде

$$\pi(x) = \begin{cases} 1 \ \text{ при } \ \displaystyle\sum_{j=1}^{n} x_j^2 > \chi^2_{1-\varepsilon}(n), \\ 0 \ \text{ при } \ \displaystyle\sum_{j=1}^{n} x_j^2 \le \chi^2_{1-\varepsilon}(n). \end{cases}$$

Согласно лемме Неймана–Пирсона, такой критерий будет иметь уровень значимости, равный ε, и будет иметь наибольшую мощность из всех возможных. Найдем эту мощность:

$$W = \mathbb{E}_2\pi(X) = \mathbb{P}_2\left(\sum_{j=1}^{n} X_j^2 > \chi^2_{1-\varepsilon}(n)\right).$$

Вероятность здесь подсчитывается в рамках гипотезы H_2, значит X_j распределены по закону $\mathrm{N}(0,4)$. Заметим, что

$$\sum_{j=1}^{n} \left(\frac{X_j}{2}\right)^2 \in \chi^2(n)$$

как сумма квадратов независимых случайных величин, распределенных по закону $\mathrm{N}(0,1)$. Значит

$$W = \mathbb{P}_2\left(\sum_{j=1}^{n}\left(\frac{X_j}{2}\right)^2 > \frac{\chi^2_{1-\varepsilon}(n)}{4}\right) = 1 - F_{\chi^2(n)}\left(\frac{\chi^2_{1-\varepsilon}(n)}{4}\right).$$

Задача 2. Построить наиболее мощный критерий для проверки гипотез

$$H_1:\ \mathrm{U}(0,1)$$
$$H_2:\ \mathrm{U}(0,2)$$

на уровне значимости $0 < \varepsilon < 1$. Объем выборки считать равным $n \geq 1$. Также найти мощность полученного критерия и вероятность ошибки второго рода.

Решение. Сначала запишем функции правдоподобия для обеих гипотез:

$$L_1(x) = f_{\mathrm{U}(0,1)}(x_1)\dots f_{\mathrm{U}(0,1)}(x_n) =$$

$$= I(0 \leq x_1 \leq 1)\dots I(0 \leq x_n \leq 1) = I(x \in [0,1]^n),\ x \in \mathbb{R}^n,$$

$$L_2(x) = f_{\mathrm{U}(0,2)}(x_1)\dots f_{\mathrm{U}(0,2)}(x_n) =$$

$$= \frac{1}{2}I(0 \leq x_1 \leq 2)\dots\frac{1}{2}I(0 \leq x_n \leq 2) = \frac{1}{2^n}I(x \in [0,2]^n),\ x \in \mathbb{R}^n.$$

Теперь составим функцию отношения правдоподобия:

$$l(x) = \frac{L_2(x)}{L_1(x)} = \frac{1}{2^n}\frac{I(x \in [0,2]^n)}{I(x \in [0,1]^n)}.$$

Прежде чем двигаться дальше, выясним, какие значения может принимать функция $l(x)$. Если $x \notin [0,2]^n$, то $L_1(x) = L_2(x) = 0$, в этом случае можно принимать любое решение – H_1 или H_2, так как вклад в значения вероятностей ошибок первого и второго рода эта область не дает. Ограничимся поэтому только областью $x \in [0,2]^n$. В этом случае числитель $l(x)$ всегда равен 1, а знаменатель может принимать как значение 0 при $x \notin [0,1]^n$, так и значение 1 при $x \in [0,1]^n$. Поэтому

$$l(x) = \begin{cases} 2^{-n} & \text{при}\ \ x \in [0,1]^n, \\ +\infty & \text{при}\ \ x \notin [0,1]^n. \end{cases}$$

Лемма Неймана–Пирсона утверждает, что наиболее мощный критерий с уровнем значимости $\varepsilon \leq \mathbb{P}_1(Z > 0) \equiv 1$ можно искать среди критериев вида

$$\pi_{c,p}(x) = \begin{cases} 1 & \text{при } l(x) > c, \\ p & \text{при } l(x) = c, \\ 0 & \text{при } l(x) < c. \end{cases}$$

Конкретные значения $c > 0$ и $p \in [0, 1]$ находятся из условия на уровень значимости

$$\mathbb{P}_1(Z > c) + p\mathbb{P}_1(Z = c) = \varepsilon,$$

где $Z = l(X)$. Теперь будем просто перебирать все значения $c > 0$ и решать вышеуказанное уравнение.

1. Если c достаточно большое, а именно $c > 1/2^n$, то

$$\mathbb{P}_1(Z = c) = 0.$$

Поэтому получаем уравнение $\mathbb{P}_1(Z > c) = \varepsilon$. Теперь выясняем, что $Z > c$ только, когда Z принимает значение $+\infty$, то есть для $X \notin [0, 1]^n$. А так как в силу первой гипотезы $\mathrm{U}(0, 1)$ вероятность попадания в эту область равна нулю, то имеем $\mathbb{P}_1(Z > c) = 0 \neq \varepsilon$. Противоречие, продолжаем поиски c.

2. Пусть теперь $c = 1/2^n$. Сразу получаем

$$\mathbb{P}_1(Z = c) = \mathbb{P}_1(X \in [0, 1]^n) = 1.$$

$$\mathbb{P}_1(Z > c) = \mathbb{P}_1(X \notin [0, 1]^n) = 0.$$

Условие на уровень значимости в этом случае такое:

$$0 + p \cdot 1 = \varepsilon,$$

что по сути сразу же дает $p = \varepsilon \in [0, 1]$. Итак, пара $c > 0$, $p \in [0, 1]$, удовлетворяющая условию на уровень значимости, найдена.

3. Пусть наконец $0 < c < 1/2^n$. В этом случае

$$\mathbb{P}_1(Z = c) = 0, \ \mathbb{P}_1(Z > c) = 1,$$

и условие на уровень значимости $1 + p \cdot 0 = \varepsilon$ содержит противоречие.

Итак, наш критерий выглядит следующим образом:

$$\pi(x) = \begin{cases} 1 & \text{при } l(x) > 1/2^n \Leftrightarrow x \notin [0, 1]^n, \\ \varepsilon & \text{при } l(x) = 1/2^n \Leftrightarrow x \in [0, 1]^n. \end{cases}$$

Что означает полученный ответ? Если $x \notin [0,1]^n$, то есть выпало значение случайной величины от 1 до 2, то следует принять вторую гипотезу $\mathrm{U}(0,2)$, вместо $\mathrm{U}(0,1)$. Этот результат можно было бы получить из «здравого смысла», но критерий Неймана–Пирсона здравому смыслу не противоречит. Если же $x \in [0,1]^n$, то есть все элементы выборки попали на интервал $[0,1]$, то необходимо провести процедуру рандомизации: реализовать случайную величину $\xi_{\mathrm{Be}} \in \mathrm{Be}(\varepsilon)$ и посмотреть на результат. Если $\xi_{\mathrm{Be}} = 1$, то следует принять H_2, иначе – H_1. Процедура рандомизации привносит дополнительную случайность в систему ради того, чтобы было в точности выполнено условие на уровень значимости. Может показаться, что внесение дополнительной случайности в систему повышает ошибку. Так и есть, вероятность ошибки 1-го рода увеличивается, но до безопасного значения, равного уровню значимости. А вот вероятность ошибки 2-го рода уменьшается, потому что благодаря рандомизации повышается шанс правильного решения.

Рассчитаем теперь мощность критерия. Она равна

$$W = \mathbb{E}_2\pi(X) = \mathbb{P}_2(X \notin [0,1]^n) + \varepsilon\mathbb{P}_2(X \in [0,1]^n)$$

$$= \mathbb{P}_2(X \in [0,2]^n) - \mathbb{P}_2(X \in [0,1]^n) + \varepsilon\mathbb{P}_2(X \in [0,1]^n)$$

$$= 1 - \frac{1}{2^n} + \varepsilon \cdot \frac{1}{2^n} = 1 - \frac{1-\varepsilon}{2^n}.$$

Вероятность ошибки 2-го рода равна

$$\beta = 1 - W = \frac{1-\varepsilon}{2^n}.$$

Важное замечание. Наиболее мощный критерий не обязан быть единственным. Лемма Неймана–Пирсона гарантированно дает лишь один из наиболее мощных критериев. Например, в данной задаче, можно было бы предложить такой нерандомизированный критерий:

$$\pi(x) = \begin{cases} 1 & \text{при} \quad x \notin [0,c]^n, \\ 0 & \text{при} \quad x \in [0,c]^n \end{cases}$$

и найти c из условия на уровень значимости

$$\varepsilon = \mathbb{E}_1\pi(X) = \mathbb{P}_1(X \notin [0,c]^n) = 1 - c^n.$$

Видно, что решение $c \in (0,1)$ всегда существует. Мощность такого критерия

$$W = \mathbb{E}_2\pi(X) = \mathbb{P}_2(X \notin [0,c]^n) = 1 - \frac{c^n}{2^n} = 1 - \frac{1-\varepsilon}{2^n}$$

совпадает с мощностью критерия Неймана–Пирсона. И таких критериев в данной задаче можно построить бесконечное множество, просто выбирая произвольное подмножество отрезка $[0,1]$ с мерой $1-\varepsilon$ при первой гипотезе. Эти подмножества не обладают никакими преимуществами друг перед другом.

4.2. Свойства критерия Неймана–Пирсона

Напомню, что *критерием Неймана–Пирсона* мы называем критерий, о котором идет речь в лемме Неймана–Пирсона. Этот критерий обладает несколькими важными свойствами, которые следует знать.

Пусть, как и прежде, $l(x)$ означает функцию отношения правдоподобия, случайная величина $Z = l(X)$.

Лемма Неймана–Пирсона (неполный вариант). *Для любого $\varepsilon > 0$ такого, что $\mathbb{P}_1(Z > 0) \geq \varepsilon$, существуют числа $c > 0$ и $p \in [0,1]$ такие, что критерий π с критической функцией*

$$\pi^*(x) = \begin{cases} 1 & \text{при } \ l(x) > c, \\ p & \text{при } \ l(x) = c, \\ 0 & \text{при } \ l(x) < c \end{cases}$$

имеет вероятность ошибки 1-го рода, равную ε, то есть $\pi^* \in K_\varepsilon \equiv \equiv \{\pi : \alpha(\pi) = \varepsilon\}$, и является наиболее мощным критерием в K_ε. При этом числа c и p находятся из уравнения

$$\alpha(\pi^*) \equiv \mathbb{P}_1(Z > c) + p\mathbb{P}_1(Z = c) = \varepsilon.$$

И наоборот: любые $c > 0$, $p \in [0,1]$, удовлетворяющие этому уравнению, дают наиболее мощный в K_ε критерий.

Свойства критерия Неймана–Пирсона.

1. Мощность $W(\pi^*)$ критерия Неймана–Пирсона π^* с уровнем значимости ε удовлетворяет неравенству $W(\pi^*) \geq \varepsilon$. Действительно, рассмотрим критерий π с критической функцией $\pi(x) = \varepsilon \ \forall x \in \Omega$. Легко видеть, что $\alpha(\pi) = \mathbb{E}_1\pi(X) = \varepsilon$ (то есть $\pi \in K_\varepsilon$) и $W(\pi) = \mathbb{E}_2\pi(X) = \varepsilon$, а так как критерий Неймана–Пирсона является наиболее мощным критерием в классе K_ε, то его мощность должна быть не меньше, чем мощность критерия π, значит $W(\pi^*) \geq W(\pi) = \varepsilon$.

2. Согласно лемме, критерий минимизирует вероятность ошибки 2-го рода при фиксированной вероятности ошибки 1-го рода ε. Однако можно доказать, что этот же критерий π^* минимизирует вероятность ошибки 1-го рода в классе критериев K с фиксированной вероятностью ошибки 2-го рода:

$$K = \{\pi : \ \beta(\pi) = \beta(\pi^*)\}.$$

Что это значит? Вот мы построили критерий Неймана–Пирсона на уровне значимости ε, и он имеет вероятность ошибки второго рода β. Если мы теперь рассмотрим все возможные критерии с вероятностью ошибки 2-го рода, равной β (той самой, какая у критерия Неймана–Пирсона), то критерий с более низкой вероятностью ошибки 1-го рода (ниже ε) найти не удастся.

3. Для любого $c \geq 0$ выполнено

$$\mathbb{P}_2(l(X) = c) = \mathbb{P}_2\left(\frac{L_2(X)}{L_1(X)} = c\right) = \int\limits_{x:\ \frac{L_2(x)}{L_1(x)}=c} L_2(x)\,dx =$$

$$= \int\limits_{x:\ \frac{L_2(x)}{L_1(x)}=c} cL_1(x)\,dx = c\mathbb{P}_1(l(X) = c),$$

откуда окончательно получаем

$$\mathbb{P}_2(l(X) = c) = c\mathbb{P}_1(l(X) = c) \quad \forall c \geq 0.$$

Поэтому если найдено $c > 0$, для которого $\mathbb{P}_1(l(X) = c) = 0$, то p можно выбирать произвольно, т.к. на вероятностях ошибок 1-го и 2-го рода это никак не отражается:

$$\alpha = \mathbb{P}_1(l(X) > c) + p \cdot \underbrace{\mathbb{P}_1\left(l\left(X\right) = c\right)}_{=0},$$

$$\beta = \mathbb{P}_2(l(X) < c) + (1-p) \cdot \underbrace{\mathbb{P}_2\left(l\left(X\right) = c\right)}_{=c\cdot\mathbb{P}_1(l(X)=c)=0}.$$

Чтобы не вводить процедуру рандомизации, выбирают просто $p = 0$ или $p = 1$.

4. Если $\varepsilon > 0$ и $\mathbb{P}_1(Z > 0) \equiv \gamma < \varepsilon$, то рассмотрим критерий

$$\pi(x) = \begin{cases} 1 & \text{при } \ l(x) > 0, \\ 0 & \text{при } \ l(x) = 0. \end{cases}$$

Вероятность ошибки первого рода для такого критерия

$$\alpha(\pi) = \mathbb{P}_1(Z > 0) = \gamma < \varepsilon,$$

то есть меньше требуемого значения. Мощность этого критерия

$$W(\pi) = \mathbb{P}_2(Z > 0) = 1 - \mathbb{P}_2(Z = 0) = 1 - \mathbb{P}_2(L(X|H_2) = 0) = 1 - 0 = 1$$

достигает своего теоретически максимального значения. Поэтому критерий π является наиболее мощным критерием с меньшим уровнем значимости и неулучшаемым значением мощности.

5. Если носители функций $L_1(x)$ и $L_2(x)$ не пересекаются, то $l(x) \in \{0, +\infty\}$ и $\mathbb{P}_1(Z > 0) = \mathbb{P}_1(L_1(X) = 0) = 0 < \varepsilon$, и тогда из п. 4 следует, что критерий $\pi(x) = I(l(x) > 0)$ имеет вероятности ошибок 1-го и 2-го рода, равные нулю. Отмечу также, что если x принадлежит носителю распределения только одной гипотезы, то критерий Неймана–Пирсона выбирает с вероятностью единица эту гипотезу.

4.3. Равномерно наиболее мощные критерии

Рассмотрим задачу проверки простой гипотезы H_1 и сложной гипотезы H_2. Обозначим за $\mathcal{F}_2$ семейство распределений, отвечающих гипотезе H_2. Обобщим понятия ошибок 1-го и 2-го рода, а также понятие мощности критерия.

Как и в случае двух простых гипотез, критерии будем рассматривать рандомизированные, а задавать их будем посредством критических функций $\pi(x)$, которые каждому значению выборки x сопоставляют вероятность принятия альтернативной гипотезы H_2. Понятие вероятности ошибки 1-го рода остается без изменений:

$$\alpha(\pi) = \mathbb{E}_1 \pi(X).$$

Вероятность ошибки 2-го рода (как и мощность) теперь становится функцией альтернативы – конкретного распределения из $\mathcal{F}_2$:

$$\beta(\pi, F) = 1 - \mathbb{E}_F \pi(X), \quad F \in \mathcal{F}_2,$$

$$W(\pi, F) = \mathbb{E}_F \pi(X), \quad F \in \mathcal{F}_2.$$

Как и раньше, введем в рассмотрение множество

$$K_\varepsilon = \{\pi : \mathbb{E}_1 \pi(X) = \varepsilon\}.$$

Определение. Критерий π_u называется *равномерно наиболее мощным критерием* (РНМК) на уровне значимости ε, если $\pi_u \in K_\varepsilon$ и

$$\forall F \in \mathcal{F}_2 \;\; \forall \pi \in K_\varepsilon \quad \mathbb{E}_F \pi_u \geq \mathbb{E}_F \pi.$$

Присутствие квантора всеобщности $\forall F \in \mathcal{F}_2$ и означает равномерно наибольшую мощность по всем альтернативам. Обратите внимание, что $\pi_u = \pi_u(x)$ является функцией только выборки и не зависит от $F \in \mathcal{F}_2$, от $F \in \mathcal{F}_2$ зависит только математическое ожидание при альтернативе.

Займемся теперь вопросом построения РНМК. Иногда такие критерии удается построить, воспользовавшись леммой Неймана–Пирсона.

Пример 1. Рассмотрим задачу проверки двух простых гипотез

$$H_1 : \ \mathrm{U}(0, \theta_1),$$
$$H_2 : \ \mathrm{U}(0, \theta_2),$$

где известно, что $\theta_2 > \theta_1$. Пусть для простоты объем выборки $n = 1$. Критерий Неймана–Пирсона на уровне значимости ε в такой задаче выглядит следующим образом:

$$\pi^*(x) = \begin{cases} 1 & \text{при} \quad x \geq \theta_1, \\ \varepsilon & \text{при} \quad x < \theta_1. \end{cases}$$

Согласно лемме Неймана–Пирсона, это наиболее мощный критерий с уровнем значимости ε, значит

$$\forall \pi \in K_\varepsilon \ \ \mathbb{E}_{\theta_2} \pi^*(X) \geq \mathbb{E}_{\theta_2} \pi(X).$$

Построенный критерий не зависит от θ_2, стало быть мы можем даже написать

$$\forall \theta > \theta_1 \ \ \forall \pi \in K_\varepsilon \ \ \mathbb{E}_\theta \pi^*(X) \geq \mathbb{E}_\theta \pi(X).$$

По определению π^* – это равномерно наиболее мощный критерий в задаче проверки гипотез:

$$H_1 : \ \theta = \theta_1,$$
$$H_2 : \ \theta > \theta_1.$$

Пример 2. Теперь рассмотрим задачу проверки двух простых гипотез:

$$H_1 : \ \mathrm{U}(0, \theta_1),$$
$$H_2 : \ \mathrm{U}(0, \theta_2).$$

но уже с $\theta_2 < \theta_1$. Здесь критерий Неймана–Пирсона выглядит так:

$$\pi^*(x) = \begin{cases} \min\left(\dfrac{\varepsilon\theta_1}{\theta_2}, 1\right) & \text{при} \quad x \in (0, \theta_2), \\ 0 & \text{при} \quad x \geq \theta_2. \end{cases}$$

Данный критерий зависит от альтернативы θ_2, его мощность равна

$$W(\pi^*) = \min\left(\varepsilon\theta_1/\theta_2, 1\right).$$

Теперь рассмотрим критерий

$$\pi'(x) = \begin{cases} 1 & \text{при } x \in (0, \varepsilon\theta_1), \\ 0 & \text{при } x \geq \varepsilon\theta_1. \end{cases}$$

Легко видеть, что вероятность ошибки 1-го рода $\alpha(\pi') = \varepsilon$, а мощность

$$W(\pi') = \min(\varepsilon\theta_1/\theta_2, 1)$$

совпадает с мощностью критерия Неймана–Пирсона. Получается, что $\pi'(x)$ тоже является наиболее мощным критерием и

$$\forall \pi \in K_\varepsilon \;\; \mathbb{E}_{\theta_2}\pi'(X) \geq \mathbb{E}_{\theta_2}\pi(X),$$

но в отличие от критерия Неймана–Пирсона π' не зависит от альтернативы θ_2. Поэтому для π' мы можем написать

$$\forall \theta < \theta_1 \;\; \forall \pi \in K_\varepsilon \;\; \mathbb{E}_\theta\pi'(X) \geq \mathbb{E}_\theta\pi(X).$$

По определению критерий π' является равномерно наиболее мощным критерием в задаче проверки гипотез

$$H_1 : \; \theta = \theta_1,$$
$$H_2 : \; \theta < \theta_1.$$

В общем случае можно предложить следующий «алгоритм» построения РНМК:

1. Зафиксировать одно из распределений F в альтернативе H_2.
2. Построить критерий Неймана–Пирсона $\pi^*(x, F)$ для двух простых гипотез H_1 и $H_2 : F$ на требуемом уровне значимости ε.
3. Если вдруг критерий Неймана–Пирсона не зависит от F, то он же и есть РНМК. Если же он зависит от F, то можно попробовать найти другой критерий, не зависящий от F, но на том же уровне значимости ε и с точно такой же мощностью как у критерия Неймана–Пирсона. Полученный таким образом критерий является РНМК.

Ясно, что если нам «не повезло» и критерий Неймана–Пирсона зависит от F, то придется заниматься творчеством и искать независящий от F критерий. К счастью, есть целый класс моделей (экспоненциальное семейство распределений) и задач (с двусторонней альтернативой), для которых критическая функция РНМК имеет вполне определенный вид.

Определение. Говорят, что модель с функцией плотности (или функцией вероятности) $f(x,\theta)$ принадлежит *экспоненциальному семейству распределений*, если

$$f(x,\theta) = h(x)\exp\left(a(\theta)U(x) + V(\theta)\right). \tag{3}$$

Например, экспоненциальному семейству принадлежат модели

$$\mathrm{N}(\theta,\sigma^2),\ \mathrm{N}(\mu,\theta^2),\ \mathrm{Be}(\theta),\ \mathrm{Bi}(n,\theta),\ \mathrm{Po}(\theta),\ \Gamma(\theta,\lambda).$$

Определение. Критерий называется *несмещенным*, если

$$\forall F \in \mathcal{F}_2 \quad \mathbb{E}_F\pi(X) \geq \mathbb{E}_1\pi(X).$$

Определение. Критерий называется *состоятельным*, если

$$\forall F \in \mathcal{F}_2 \quad \mathbb{E}_F\pi(X) \to 1, \quad n \to \infty.$$

Теорема о существовании несмещенного РНМК. *Пусть дана экспоненциальная модель*

$$f(x,\theta) = h(x)\exp\left(a(\theta)U(x) + V(\theta)\right)$$

со строго монотонной функцией $a(\theta)$, и требуется проверить гипотезы

$$H_1:\ \theta = \theta_1,$$
$$H_2:\ \theta \neq \theta_1.$$

Тогда в классе несмещенных критериев уровня ε существует РНМК вида

$$\pi_u(x) = \begin{cases} 1 & \text{при } T(x) \notin [c_1, c_2], \\ p_1 & \text{при } T(x) = c_1, \\ p_2 & \text{при } T(x) = c_2, \\ 0 & \text{при } c_1 < T(x) < c_2, \end{cases}$$

где $T(x) = \sum\limits_{i=1}^{n} U(x_i)$, а параметры c_1, c_2, p_1 и p_2 ищутся из уравнений

$$\mathbb{E}_1\pi_u(X) = \varepsilon,\ \mathbb{E}_1(\pi_u(X) - \varepsilon)T(X) = 0.$$

И наоборот, любые числа c_1, c_2, p_1, p_2, удовлетворяющие этим уравнениям, дают РНМК уровня ε.

Важнейшее замечание. В этой теореме утверждается существование РНМК лишь в классе несмещенных критериев. Если же мы

расширим область поиска еще и на смещенные критерии, то РНМК для экспоненциального семейства существовать не будет! В лемме Неймана–Пирсона такого условия нет, там ограничение было лишь на уровень значимости. Несмещенным критерий Неймана–Пирсона получался уже по факту.

Задача 1. Построить критерий для проверки гипотез

$$H_1 : \ \theta = 1/2,$$
$$H_2 : \ \theta \neq 1/2$$

о распределении Бернулли $\mathrm{Be}(\theta)$ по результатам восьми измерений на уровне значимости $\varepsilon = 0.05$.

Решение. Будем искать РНМК в классе несмещенных критериев. Распределение Бернулли принадлежит экспоненциальной модели, ее функция вероятности

$$f(x, \theta) = \theta^x (1-\theta)^{1-x}, \ x \in \{0, 1\}.$$

Приведем ее к виду (3):

$$f(x,\theta) = \exp\left(x \ln\theta + (1-x)\ln(1-\theta)\right) = \exp\left(\left[\ln\frac{\theta}{1-\theta}\right] \cdot x + \ln(1-\theta)\right).$$

Отсюда видно, что

$$h(x) = 1, \ a(\theta) = \ln\left(\frac{\theta}{1-\theta}\right), \ U(x) = x, \ V(\theta) = \ln(1-\theta)$$

и функция $a(\theta)$ является строго монотонной. Теперь составим статистику

$$T(x) = \sum_{i=1}^{n} U(x_i) = \sum_{i=1}^{n} x_i$$

и уравнения на параметры p_1, p_2, c_1, c_2:

$$\mathbb{E}_1 \pi_u(X) = 1 \cdot \mathbb{P}_1\left(T \notin [c_1, c_2]\right) + p_1 \cdot \mathbb{P}_1\left(T = c_1\right) + p_2 \cdot \mathbb{P}_1\left(T = c_2\right) = \varepsilon,$$

$$\mathbb{E}_1\left(\pi_u(X) - \varepsilon\right) T(X) = \mathbb{E}_1\left(\pi_u(X) \cdot T(X)\right) - \varepsilon \cdot \frac{n}{2} = 0.$$

Вычислим теперь $\mathbb{E}_1(\pi_u T)$:

$$\mathbb{E}_1(\pi_u T) = \sum_{x \in \Omega} \pi_u(x) T(x) \mathbb{P}_1(X = x) =$$

$$= \sum_{T(x)\notin[c_1,c_2]} 1\cdot T(x)\mathbb{P}_1(X=x) + \sum_{T(x)=c_1} p_1 T(x)\mathbb{P}_1(X=x) + \\ + \sum_{T(x)=c_2} p_2 T(x)\mathbb{P}_1(X=x) =$$

$$= \sum_{k\notin[c_1,c_2]} k\mathbb{P}_1(T=k) + p_1c_1\mathbb{P}_1(T=c_1) + p_2c_2\mathbb{P}_1(T=c_2).$$

В результате получаем два уравнения:

$$\sum_{k\notin[c_1,c_2]} \mathbb{P}_1(T=k) + p_1\mathbb{P}_1(T=c_1) + p_2\mathbb{P}_1(T=c_2) = \varepsilon,$$

$$\sum_{k\notin[c_1,c_2]} k\mathbb{P}_1(T=k) + p_1c_1\mathbb{P}_1(T=c_1) + p_2c_2\mathbb{P}_1(T=c_2) = \frac{\varepsilon n}{2}.$$

Нам достаточно, чтобы подобрались хоть какие-нибудь параметры c_1, c_2, p_1, p_2, удовлетворяющие уравнениям, мощность будет максимизирована в любом случае. Попытаемся решить уравнение при $c_1 = n - c_2 = c$ и $p_1 = p_2 = p$. В этом случае уравнения упрощаются:

$$2\mathbb{P}_1(T<c) + 2p\mathbb{P}_1(T=c) = \varepsilon,$$

$$\sum_{k<c} k\mathbb{P}_1(T=k) + \sum_{k>n-c} k\mathbb{P}_1(T=k) + \\ + pc\mathbb{P}_1(T=c) + p(n-c)\mathbb{P}_1(T=n-c) = \frac{\varepsilon n}{2}.$$

Заметим, что

$$\sum_{k<c} k\mathbb{P}_1(T=k) + \sum_{k>n-c} k\mathbb{P}_1(T=k) = n\sum_{k=0}^{c-1}\mathbb{P}_1(T=k) = n\mathbb{P}_1(T<c).$$

Тогда наши два уравнения упростятся еще сильнее:

$$\mathbb{P}_1(T<c) + p\mathbb{P}_1(T=c) = \frac{\varepsilon}{2},$$

$$n\mathbb{P}_1(T<c) + pn\mathbb{P}_1(T=c) = \frac{\varepsilon n}{2}.$$

Очевидно, эти уравнения равносильны, и достаточно решить одно из них. Для этого будем перебирать целые $c \in [0, n]$ и решать получающиеся уравнения относительно $p \in [0, 1]$. Будем учитывать, что

$\mathbb{P}_1(T=k) = C_n^k/2^n$ для всех $k = 0, 1, \ldots, n$. Если $c = 0$, то $\mathbb{P}_1(T < c) = \mathbb{P}_1(T < 0) = 0$,

$$\mathbb{P}_1(T=c) = \mathbb{P}_1(T=0) = C_8^0/2^8 < \varepsilon/2 = 0.05/2,$$

поэтому нужный $p \in [0,1]$ не найдется. Если $c = 1$, то

$$\mathbb{P}_1(T < c) = \mathbb{P}_1(T=0) = C_8^0/2^8,$$

$$\mathbb{P}_1(T=c) = \mathbb{P}_1(T=1) = C_8^1/2^8,$$

и тогда

$$p = \frac{\varepsilon/2 - \mathbb{P}_1(T=0)}{\mathbb{P}_1(T=1)} = \frac{0.025 - C_8^0/2^8}{C_8^1/2^8} = \frac{27}{40} \in [0,1].$$

Окончательно получаем искомый критерий:

$$\pi_u(x) = \begin{cases} 1 & \text{при } T(x) \in \{0, 8\}, \\ 27/40 & \text{при } T(x) \in \{1, 7\}, \\ 0 & \text{при } T(x) \in (1, 7). \end{cases}$$

5. Байесовские и минимаксные решающие правила

В прошлый раз мы рассмотрели один из популярных методов проверки гипотез: поиск наиболее мощного критерия на заданном уровне значимости. По сути это была оптимизационная задача с функционалом (мощность) и ограничениями (на уровень значимости). Можно ставить, однако, и другие оптимизационные задачи, например, вводя понятие штрафа за неверное решение и понятие риска.

Будем рассматривать случай, когда дано $k \geq 2$ простых гипотез:

$$\begin{aligned} H_1 &: f_1(x), \\ H_2 &: f_2(x), \\ &\ldots \\ H_k &: f_k(x), \end{aligned}$$

где $f_j(x)$ – функция плотности (или функция вероятности) выборки, соответствующая гипотезе H_j, $j = 1, \ldots, k$. В терминах наших предыдущих обозначений $f_j(x) = L_j(x)$. Предположим, что распределение, из

которого извлечена выборка, само было выбрано случайно. Формально это значит, что для каждой гипотезы H_j определена вероятность q_j ее использования для построения выборки X. Числа $q_1, \ldots, q_k$ могут быть как известны в конкретной задаче, так и неизвестны. Называются они *априорными* вероятностями.

Предположим, что задана т.н. *матрица штрафов* $C = \|c_{ij}\|_{i,j=1}^{k}$, где c_{ij} – штраф за принятие гипотезы H_i, когда верна H_j. Если $i \neq j$, то обычно $c_{ij} \geq 0$, а если $i = j$, то $c_{ij} \leq 0$. Отрицательные значения штрафа можно интерпретировать как «вознаграждение» за правильно принятое решение.

Как и раньше, символом δ будем обозначать рандомизированные статистические критерии, которые в теории байесовского оценивания и проверки гипотез еще носят название *решающих правил*:

$$\delta : \Omega \to \mathbb{R}^{(k)}, \quad \mathbb{R}^{(k)} = \{(\pi_1, \ldots, \pi_k) : \pi_i \geq 0, \sum \pi_i = 1\}.$$

Определение. *Функцией риска* $R(\delta) = (R_1(\delta), \ldots, R_k(\delta))$ называется k-мерная функция с компонентами

$$R_j(\delta) = \sum_{i=1}^{k} c_{ij} p_{ij}(\delta), \ j = 1, \ldots, k,$$

где $p_{ij} = \mathbb{P}_j(H_i)$ – вероятность принять гипотезу H_i, если верна H_j. Число $R_j(\delta)$ представляет собой средние потери от использования решающего правила δ в случае, когда справедлива гипотеза H_j. Заметим также, что вероятности $p_{ij} = p_{ij}(\delta)$ зависят от решающего правила δ.

Определение. *Байесовским риском* называется функция

$$r(\delta) = \sum_{j=1}^{k} R_j(\delta) q_j.$$

Это полная потеря от использования решающего правила δ.

Определение. *Байесовским* решающим правилом называется

$$\delta^* = \arg\min_{\delta} r(\delta),$$

то есть это решающее правило, минимизирующее полную среднюю потерю.

Оказывается, строить байесовские решающие правила легко и просто. Пусть даны k простых гипотез $H_1, \ldots, H_k$ с функциями правдоподобия $f_1(x), \ldots, f_k(x)$ и априорными вероятностями $q_1, \ldots, q_k$

соответственно. Пусть дана матрица штрафов $C = \|c_{ij}\|_{i,j=1}^{k}$. Введем вспомогательные функции

$$h_i(x) = \sum_{j=1}^{k} c_{ij} q_j f_j(x),\ i = 1, \ldots, k.$$

Теорема. *Байесовское решающее правило δ^* определяется так:*

$$\delta^*(x) = H_j \Leftrightarrow h_j(x) = \min_{1 \le i \le k} h_i(x)$$

для каждого $x \in \Omega$. Если минимум достигается на нескольких индексах $i = 1, \ldots, k$, то мы вправе выбрать любую из соответствующих им гипотез, полный риск от этого не зависит.

Задача 1. Пусть требуется проверить простые гипотезы

$$H_1 : \mathrm{Be}(1/2),\ q_1 = 1/4,$$
$$H_2 : \mathrm{Be}(1/3),\ q_2 = 3/4$$

с соответствующими априорными вероятностями q_1 и q_2. Пусть дана матрица штрафов $C = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}$. Требуется построить байесовское решающее правило.

Решение. Сначала выпишем функции правдоподобия:

$$f_1(x) = \mathbb{P}_1(X = x) = \begin{bmatrix} 1/2, & x = 0, \\ 1/2, & x = 1; \end{bmatrix}$$
$$f_2(x) = \mathbb{P}_2(X = x) = \begin{bmatrix} 2/3, & x = 0, \\ 1/3, & x = 1. \end{bmatrix}$$

Теперь сформируем вспомогательные функции:

$$h_1(x) = c_{11} q_1 f_1(x) + c_{12} q_2 f_2(x) = \frac{3}{4} f_2(x) = \begin{bmatrix} 1/2, & x = 0, \\ 1/4, & x = 1; \end{bmatrix}$$
$$h_2(x) = c_{21} q_1 f_1(x) + c_{22} q_2 f_2(x) = \frac{1}{2} f_1(x) = \begin{bmatrix} 1/4, & x = 0, \\ 1/4, & x = 1. \end{bmatrix}$$

Теперь мы готовы построить байесовское решающее правило. Если $x = 0$, то $h_2(x) \le h_1(x)$, поэтому следует принять гипотезу H_2. Если же $x = 1$, то $h_1(x) = h_2(x)$, и поэтому можно принять как гипотезу H_1, так и H_2. Пусть это будет H_2. Получается, байесовским решающим правилом в данном примере будет такое решающее правило, которое для любого входа x принимает гипотезу H_2:

$$x = 0 \ \rightarrow \ H_2,$$
$$x = 1 \ \rightarrow \ H_2.$$

Другое решающее правило:

$$\begin{aligned} x = 0 \quad &\to \quad H_2, \\ x = 1 \quad &\to \quad H_1 \end{aligned}$$

также будет являться байесовским.

Замечание. В случае двух гипотез, гипотеза H_2 принимается в случае

$$c_{11}q_1f_1(x) + c_{12}q_2f_2(x) \geq c_{21}q_1f_1(x) + c_{22}q_2f_2(x),$$

что равносильно неравенству

$$\frac{f_2(x)}{f_1(x)} \geq \frac{c_{21} - c_{11}}{c_{12} - c_{22}} \cdot \frac{q_1}{q_2}.$$

Это значит, что критическая область имеет вид: $\{x : l(x) \geq c\}$, где $l(x)$ – функция отношения правдоподобия, а c – постоянная, однозначно определяемая матрицей штрафов и априорными вероятностями.

Определение. *Минимаксным* решающим правилом называется

$$\tilde{\delta} = \arg\min_{\delta} \max_{j} R_j(\delta).$$

Такое решающее правило можно искать в случаях, когда априорные вероятности неизвестны. В отличие от предыдущего случая, строить на практике $\tilde{\delta}$ очень сложно. Простого алгоритма на случаи любого числа гипотез и произвольных распределений, по-видимому, не существует. Однако полезно знать связь между байесовским и минимаксным решающими правилами.

Теорема. *Если существуют такие априорные вероятности, при которых соответствующее байесовское решающее правило δ^* дает равные средние риски*

$$R_1(\delta^*) = R_2(\delta^*) = \cdots = R_k(\delta^*),$$

то это байесовское решающее правило является минимаксным, то есть $\tilde{\delta} = \delta^*$.

Таким образом, минимаксное решающее правило ищется как байесовское, которое бы уравнивало средние риски. Перебираемыми переменными в этом случае являются априорные вероятности. Конечно, не всегда такие априорные вероятности легко находятся или хотя бы существуют. Вернемся к задаче 1. Попытаемся найти априорные вероятности q_1 и $q_2 = 1 - q_1$ для того, чтобы средние риски на соответствующем байесовском решающем правиле сравнялись:

$$R_1(\delta^*) = R_2(\delta^*).$$

Из замечания выше следует, что в этой задаче байесовское решающее правило для произвольного q_1 имеет вид

$$\delta^*(x) = \left[\begin{array}{ll} H_1, & l\,(x) < \dfrac{2q_1}{1-q_1}, \\ H_2, & l\,(x) \geq \dfrac{2q_1}{1-q_1}, \end{array} \right.$$

где функция отношения правдоподобия

$$l(x) = \frac{f_2(x)}{f_1(x)} = \left[\begin{array}{ll} 4/3, & x = 0, \\ 2/3, & x = 1. \end{array} \right.$$

Введем обозначение

$$g = g(q_1) = \frac{2q_1}{1-q_1} \in [0, +\infty),$$

и тогда поиски q_1 будут равносильны поискам величины g. Средние риски выражаются формулам

$$R_1(\delta^*) = c_{11}p_{11} + c_{21}p_{21} = 2p_{21},$$
$$R_2(\delta^*) = c_{12}p_{12} + c_{22}p_{22} = p_{12},$$

а их равенство означает $2p_{21} = p_{12}$, или конкретно для нашего решающего правила δ^*:

$$2\mathbb{P}_1\left(l(X) \geq g\right) = \mathbb{P}_2\left(l(X) < g\right).$$

Рассмотрим несколько случаев.

1. Если $g \in [0, 2/3]$, то получим равенства

$$\mathbb{P}_1(l(X) \geq g) = \mathbb{P}_1(X \in \{0, 1\}) = 1,$$
$$\mathbb{P}_2(l(X) < g) = \mathbb{P}_2(X \in \varnothing) = 0,$$

с учетом которых наше уравнение перепишется в виде $2 \cdot 1 = 0$ и будет означать противоречие.

2. Если $g \in (2/3, 4/3]$, то получим равенства

$$\mathbb{P}_1(l(X) \geq g) = \mathbb{P}_1(X = 0) = 1/2,$$
$$\mathbb{P}_2(l(X) < g) = \mathbb{P}_2(X = 1) = 1/3.$$

В этом случае также получаем противоречие.

3. Если $g > 4/3$, то получим равенства

$$\mathbb{P}_1(l(X) \geq g) = \mathbb{P}_1(X \in \varnothing) = 0,$$
$$\mathbb{P}_2(l(X) < g) = \mathbb{P}_2(X \in \{0,1\}) = 1.$$

И в этом случае также получаем противоречие.

Были рассмотрены все возможные значения $g(q_1)$, а значит и все возможные значения q_1, и ни в каких случаях не удалось сравнять средние риски. Конечно это не значит, что минимаксного решающего правила не существует и его нельзя найти каким-либо другим способом. Оказывается, что для случая двух простых гипотез алгоритм построения минимаксного решающего правила есть и формулируется в одном из пунктов следующей (уже полной) леммы Неймана–Пирсона.

Лемма Неймана–Пирсона. *Пусть даны две простые гипотезы H_1 и H_2 с соответствующими априорными вероятностями q_1 и q_2 и матрицей штрафов $C = \|c_{ij}\|_{i,j=1}^2$. Определим класс критериев вида*

$$\pi_{c,p}(x) = \begin{cases} 1 & \text{при } \ l(x) > c, \\ p & \text{при } \ l(x) = c, \\ 0 & \text{при } \ l(x) < c, \end{cases}$$

где $l(x)$ – функция отношения правдоподобия. Тогда верны следующие утверждения:

1. *При $c = \frac{c_{21}-c_{11}}{c_{12}-c_{22}} \frac{q_1}{q_2}$ критерий $\pi_{c,p}$ для любого $p \in [0,1]$ определяет байесовское решающее правило.*
2. *Существуют $c > 0$ и $p \in [0,1]$ такие, что критерий $\pi_{c,p}$ будет определять минимаксное решающее правило. Это любые числа $c > 0$ и $p \in [0,1]$, удовлетворяющие уравнению*

$$R_1(\pi_{c,p}) = R_2(\pi_{c,p}).$$

3. *Для любого $\varepsilon > 0$ такого, что $\mathbb{P}_1(l(X) > 0) \geq \varepsilon$, существуют константы $c > 0$ и $p \in [0,1]$ такие, что критерий $\pi_{c,p}$ имеет уровень значимости ε и является наиболее мощным критерием в классе критериев с уровнем значимости ε. Это любые числа $c > 0$ и $p \in [0,1]$, удовлетворяющие уравнению*

$$\mathbb{P}_1(l(X) > c) + p\mathbb{P}_1(l(X) = c) = \varepsilon.$$

Замечания. Если матрица штрафов имеет вид $C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, то условие

$$R_1(\pi_{c,p}) = R_2(\pi_{c,p})$$

равносильно равенству вероятностей ошибок первого и второго рода:

$$\alpha(\pi_{c,p}) = \beta(\pi_{c,p}).$$

Заметим также, что хотя рандомизированные байесовские критерии в принципе возможны, рассматривать их не имеет смысла, так как они не дают никаких преимуществ (в плане оптимизации полного среднего риска) по сравнению с нерандомизированными критериями.

Задача 2. Построить минимаксное решающее правило для проверки двух простых гипотез

$$H_1 : \mathrm{Be}(1/2),$$
$$H_2 : \mathrm{Be}(1/3).$$

Матрицу штрафов взять равной $C = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}$.

Решение. Ранее мы выяснили, что байесовского решающего правила, которое бы уравнивало средние риски, в этой задаче не существует. Теперь рассмотрим критерии более общего вида:

$$\pi_{c,p}(x) = \begin{cases} 1 & \text{при } \ l(x) > c, \\ p & \text{при } \ l(x) = c, \\ 0 & \text{при } \ l(x) < c, \end{cases}$$

и будем искать константы $c > 0$ и $p \in [0, 1]$ из уравнения

$$R_1(\pi_{c,p}) = R_2(\pi_{c,p}).$$

Это уравнение равносильно $2p_{21} = p_{12}$, или, в терминах критерия $\pi_{c,p}$,

$$2\mathbb{P}_1(l(X) > c) + 2p\mathbb{P}_1(l(X) = c) = \mathbb{P}_2(l(X) < c) + (1 - p)\mathbb{P}_2(l(X) = c).$$

Будем перебирать все $c > 0$, пока не найдем ответ. Для этого нам пригодится уже вычисленное выражение для функции отношения правдоподобия:

$$l(x) = \frac{f_2(x)}{f_1(x)} = \left[\begin{array}{ll} 4/3, & x = 0, \\ 2/3, & x = 1. \end{array} \right.$$

1. Если $c \in (0, 2/3)$, то получаем равенства

$$\mathbb{P}_1(l(X) > c) = \mathbb{P}_1(X \in \{0, 1\}) = 1,$$
$$\mathbb{P}_1(l(X) = c) = \mathbb{P}_1(X \in \varnothing) = 0,$$
$$\mathbb{P}_2(l(X) < c) = \mathbb{P}_2(X \in \varnothing) = 0,$$
$$\mathbb{P}_2(l(X) = c) = \mathbb{P}_2(X \in \varnothing) = 0,$$

и наше уравнение переписывается в виде

$$2 \cdot 1 + 2p \cdot 0 = 0 + (1 - p) \cdot 0,$$

что означает противоречие.

2. Если $c = 2/3$, то получаем равенства

$$\mathbb{P}_1(l(X) > c) = \mathbb{P}_1(X = 0) = 1/2,$$
$$\mathbb{P}_1(l(X) = c) = \mathbb{P}_1(X = 1) = 1/2,$$
$$\mathbb{P}_2(l(X) < c) = \mathbb{P}_2(X \in \varnothing) = 0,$$
$$\mathbb{P}_2(l(X) = c) = \mathbb{P}_2(X = 1) = 1/3$$

и наше уравнение, записанное в виде

$$2 \cdot 1/2 + 2p \cdot 1/2 = 0 + (1 - p) \cdot 1/3$$

не имеет решений при $p \in [0, 1]$.

3. Если $c \in (2/3, 4/3)$, то получаем равенства

$$\mathbb{P}_1(l(X) > c) = \mathbb{P}_1(X = 0) = 1/2,$$
$$\mathbb{P}_1(l(X) = c) = \mathbb{P}_1(X \in \varnothing) = 0,$$
$$\mathbb{P}_2(l(X) < c) = \mathbb{P}_2(X = 1) = 1/3,$$
$$\mathbb{P}_2(l(X) = c) = \mathbb{P}_2(X \in \varnothing) = 0,$$

и наше уравнение преобразуется к виду

$$2 \cdot 1/2 + 2p \cdot 0 = 1/3 + (1 - p) \cdot 0,$$

что дает противоречие.

4) Если $c = 4/3$, то получаем равенства

$$\mathbb{P}_1(l(X) > c) = \mathbb{P}_1(X \in \varnothing) = 0,$$
$$\mathbb{P}_1(l(X) = c) = \mathbb{P}_1(X = 0) = 1/2,$$
$$\mathbb{P}_2(l(X) < c) = \mathbb{P}_2(X = 1) = 1/3,$$
$$\mathbb{P}_2(l(X) = c) = \mathbb{P}_2(X = 0) = 2/3,$$

и наше уравнение преобразуется к виду

$$2 \cdot 0 + 2p \cdot 1/2 = 1/3 + (1-p) \cdot 2/3,$$

откуда получаем $p = 3/5 \in [0,1]$. Решение найдено, и согласно лемме Неймана–Пирсона критерий

$$\pi_{c,p}(x) = \begin{cases} 3/5 & \text{при } x = 0, \\ 0 & \text{при } x = 1 \end{cases}$$

определяет минимаксное решающее правило.

6. Другие критерии

6.1. Критерий Стьюдента для проверки гипотезы о равенстве средних нормальной модели

Условия (дисперсии известны). Пусть имеются две выборки: $X = (X_1, \ldots, X_n)$ из распределения $\mathrm{N}(\mu_1, \sigma_1^2)$ и $Y = (Y_1, \ldots, Y_m)$ из распределения $\mathrm{N}(\mu_2, \sigma_2^2)$, причем параметры μ_1, μ_2 неизвестны, а σ_1, σ_2 известны. Выдвинута гипотеза

$$H_1 : \mu_1 = \mu_2.$$

Требуется составить критерий проверки гипотезы H_1 на заданном уровне значимости α.

Решение. Введем выборочные средние

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \ \overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

и предложим статистику критерия

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}.$$

В условиях гипотезы H_1 статистика T имеет стандартное нормальное распределение $\mathrm{N}(0,1)$. Решение принимается в зависимости от рассматриваемой альтернативы:

1. Для альтернативы $\mu_1 \neq \mu_2$:

$$H_1 \text{ отвергается} \Leftrightarrow |T| > \mathrm{N}_{1-\alpha/2}(0,1).$$

2. Для альтернативы $\mu_1 > \mu_2$:

$$H_1 \text{ отвергается} \Leftrightarrow T > \mathrm{N}_{1-\alpha}(0,1).$$

3. Для альтернативы $\mu_1 < \mu_2$:

$$H_1 \text{ отвергается} \Leftrightarrow T < \mathrm{N}_{\alpha}(0,1).$$

Условия (дисперсии неизвестны, но равны). Пусть имеются две выборки: $X = (X_1, \ldots, X_n)$ из распределения $\mathrm{N}(\mu_1, \sigma^2)$ и $Y = (Y_1, \ldots, Y_m)$ из распределения $\mathrm{N}(\mu_2, \sigma^2)$, причем параметры μ_1, μ_2, σ неизвестны. Выдвинута гипотеза

$$H_1 : \mu_1 = \mu_2.$$

Требуется составить критерий проверки гипотезы H_1 на заданном уровне значимости α.

Решение. Введем выборочные средние $\overline{X}, \overline{Y}$, выборочные дисперсии S^2_{0X}, S^2_{0Y} и вспомогательную статистику S^2:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i,$$

$$S^2_{0X} = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2,\ S^2_{0Y} = \frac{1}{m-1}\sum_{i=1}^{m} (Y_i - \overline{Y})^2,$$

$$S^2 = \frac{(n-1)S^2_{0X} + (m-1)S^2_{0Y}}{n+m-2}.$$

Далее, предлагается тестовая статистика

$$T = \frac{\overline{X} - \overline{Y}}{S}\sqrt{\frac{nm}{n+m}}.$$

При условии H_1 статистика T имеет распределение Стьюдента $\mathrm{St}(n+m-2)$. Решение принимается в зависимости от рассматриваемой альтернативы:

1. Для альтернативы $\mu_1 \neq \mu_2$:

$$H_1 \text{ отвергается} \Leftrightarrow |T| > \mathrm{St}_{1-\alpha/2}(n+m-2).$$

2. Для альтернативы $\mu_1 > \mu_2$:

$$H_1 \text{ отвергается} \Leftrightarrow T > \mathrm{St}_{1-\alpha}(n+m-2).$$

3. Для альтернативы $\mu_1 < \mu_2$:

$$H_1 \text{ отвергается} \Leftrightarrow T < \mathrm{St}_{\alpha}(n+m-2).$$

6.2. Критерий Фишера для проверки гипотезы о равенстве дисперсий нормальной модели

Условия. Пусть имеются выборки $X = (X_1, \ldots, X_n)$ из распределения $\mathrm{N}(\mu_1, \sigma_1^2)$ и $Y = (Y_1, \ldots, Y_m)$ из $\mathrm{N}(\mu_2, \sigma_2^2)$, причем параметры $\mu_1, \mu_2, \sigma_1, \sigma_2$ неизвестны. Выдвинута гипотеза

$$H_1 : \sigma_1 = \sigma_2.$$

Требуется составить критерий проверки гипотезы H_1 на заданном уровне значимости α.

Решение. Запишем выборочные средние двух выборок

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i,$$

а также выборочные дисперсии

$$S_{0X}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,\ S_{0Y}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2.$$

В качестве тестовой статистики выберем $T = S_{0X}^2/S_{0Y}^2$, которая в условиях H_1 имеет распределение Фишера: $\mathrm{F}(n-1, m-1)$. Решение принимается в зависимости от рассматриваемой альтернативы:

1. Для альтернативы $\sigma_1 \neq \sigma_2$:

$$H_1 \text{ отвергается} \Leftrightarrow \{T < t_{\alpha/2}\} \cup \{T > t_{1-\alpha/2}\},$$
$$t_{\alpha/2} = \mathrm{F}_{\alpha/2}(n-1, m-1),\ t_{1-\alpha/2} = \mathrm{F}_{1-\alpha/2}(n-1, m-1).$$

2. Для альтернативы $\sigma_1 > \sigma_2$:

$$H_1 \text{ отвергается} \Leftrightarrow T > \mathrm{F}_{1-\alpha}(n-1, m-1).$$

3. Для альтернативы $\sigma_1 < \sigma_2$:

$$H_1 \text{ отвергается} \Leftrightarrow T < \mathrm{F}_{\alpha}(n-1, m-1).$$

Задача 44. При изучении некоторого физического явления в термостате получены данные (в градусах Цельсия): 21.2, 21.8, 21.3, 21.0, 21.4, 21.3. Результаты измерений суть значения, принимаемые нормальными случайными величинами. К термостату применено некоторое усовершенствование, после чего при другом режиме получены данные: 37.7, 37.6, 37.6, 37.4. Можно ли при уровне значимости $\alpha = 0.05$ усовершенствование признать эффективным?

Решение. По условию задачи имеются две выборки:

$$X = (21.2, 21.8, 21.3, 21.0, 21.4, 21.3)$$

из нормального распределения $\mathrm{N}(\mu_1, \sigma_1)$ и

$$Y = (37.7, 37.6, 37.6, 37.4)$$

из нормального распределения $\mathrm{N}(\mu_2, \sigma_2)$. Будем считать усовершенствование эффективным, если $\sigma_2 < \sigma_1$. Чтобы это выяснить, выдвинем гипотезу

$$H_1 : \ \sigma_1 = \sigma_2$$

и будем рассматривать ее против альтернативы

$$H_2 : \ \sigma_1 > \sigma_2.$$

Для проверки H_1 воспользуемся критерием Фишера. Вычислим выборочные средние и выборочные дисперсии:

$$\overline{X} = 21.333, \ \overline{Y} = 37.575,$$
$$S^2_{0X} = 0.071, \ S^2_{0Y} = 0.016,$$

тогда статистика критерия Фишера примет значение $T = S^2_{0X}/S^2_{0Y} = = 4.463$. Так как рассматривается альтернатива $\sigma_1 > \sigma_2$, то граница критической области $t_{1-\alpha} = \mathrm{F}_{1-\alpha}(n-1, m-1) = 9.013$, где $n = 6$ и $m = 4$. Гипотеза H_1 не отклоняется. Следовательно, данных недостаточно для того, чтобы с уверенностью 95% утверждать, что усовершенствование было успешным. Этот вывод мы делаем несмотря на то, что оценка S^2_{0Y} дисперсии после усовершенствования меньше оценки S^2_{0X} дисперсии до усовершенствования.

6.3. Последовательный критерий Вальда

Вернемся к задаче проверке двух простых гипотез. До сих пор мы предполагали, что число испытаний n (объем выборки) является известной, детерминированной величиной. Но существуют критерии, в которых решение о приеме той или иной гипотезы или о продолжении измерений формируется на каждом шаге испытания в зависимости от значений полученной выборки и условий критерия. В таких случаях, объем выборки является случайной величиной ν, а подобные критерии называются *последовательными*. Из последовательных критериев широкое распространение получил последовательный критерий отношения правдоподобия, разработанный А. Вальдом.

Условия. Даны две простые гипотезы

$$H_1 : f_1(x),$$
$$H_2 : f_2(x)$$

с функциями плотности (или функциями вероятности) $f_1(x)$ и $f_2(x)$, отличными на множестве ненулевой меры. Требуемые значения вероятностей ошибок первого и второго рода равны α и β соответственно. Построить последовательный критерий отношения правдоподобия.

Решение. Сначала вычисляем величины

$$\Gamma_1' = \ln \frac{\beta}{1-\alpha}, \quad \Gamma_2' = \ln \frac{1-\beta}{\alpha}.$$

Далее вычисляем последовательно при $n = 1, 2, 3, \ldots$ логарифмическую функцию правдоподобия

$$l_n(x) = \ln \frac{L_2(x)}{L_1(x)} \equiv \sum_{k=1}^{n} \ln \frac{f_2(x_k)}{f_1(x_k)}$$

до тех пор, пока не нарушится одно из неравенств

$$\Gamma_1' < l_n(x) < \Gamma_2'.$$

Если оказалось $l_n(x) \le \Gamma_1'$, то следует принять H_1, а если $l_n(x) \ge \Gamma_2'$, то следует принять H_2.

Комментарии

1. В результате получится критерий силы (α', β'), где истинные вероятности α' и β' ошибок первого и второго рода связаны с запрашиваемыми вероятностями ошибок α и β соотношениями

$$\alpha' + \beta' \le \alpha + \beta,$$
$$\alpha' \le \frac{\alpha}{1-\beta}, \ \beta' \le \frac{\beta}{1-\alpha}.$$

Точные значения α' и β' определяются численными методами, хотя при стремлении α и β к нулю абсолютные погрешности $|\alpha' - \alpha|$ и $|\beta' - \beta|$ стремятся к нулю. Более того, фактически вероятности ошибок α' и β' могут оказаться либо меньше величин α и β (что хорошо), либо превзойдут их на малую, известную величину.

2. Если все же требуется получить критерий силы (α, β), и никакой другой, то можно воспользоваться численными методами подбора границ Γ_1 и Γ_2 логарифмического отношения правдоподобия $l_n(x)$:

$\Gamma_1 < l_n(x) < \Gamma_2$. Эти границы будут, вообще говоря, отличаться от Γ_1' и Γ_2'.

3. Справедливы также следующие утверждения:

$$\lim_{n\to\infty} \mathbb{P}_1(\nu > n) = 0, \quad \lim_{n\to\infty} \mathbb{P}_2(\nu > n) = 0.$$

которые означают, что вышеизложенный критерий с вероятностью единица останавливается за конечное число шагов.

4. Более того, все моменты случайной величины ν конечны в рамках обеих гипотез. Оценить среднее число шагов последовательного критерия можно по формулам

$$\mathbb{E}_1\nu \approx \frac{(1-\alpha)\Gamma_1' + \alpha\Gamma_2'}{\mathbb{E}_1 Z}, \quad \mathbb{E}_2\nu \approx \frac{\beta\Gamma_1' + (1-\beta)\Gamma_2'}{\mathbb{E}_2 Z},$$

где $Z = \ln\left(f_2(X_k)/f_1(X_k)\right)$. Приближенные равенства здесь стоят потому, что границы Γ_1' и Γ_2', вообще говоря, не соответствуют истинным границам Γ_1 и Γ_2, отвечающим вероятностям ошибок α и β, хотя при достаточно малых α и β близки к ним.

5. Отметим наконец, что описанный выше критерий одновременно минимизирует $\mathbb{E}_1\nu$ и $\mathbb{E}_2\nu$ среди всех критериев, у которых вероятность ошибки первого рода не превышает α, а вероятность ошибки второго рода не превышает β.

Пример. Пусть $\xi \in \mathrm{Be}(p)$ и поставлены две простые гипотезы:

$$H_1 : p = p_1,$$
$$H_2 : p = p_2,$$

где $p_1 = 1/2$, $p_2 = 2/3$, а запрашиваемые вероятности ошибок первого и второго рода равны соответственно $\alpha = 0.05$ и $\beta = 0.05$. Построить последовательный критерий отношения правдоподобия.

Решение. Вычисляем границы

$$\Gamma_1' = \ln\frac{\beta}{1-\alpha} = -2.9444, \quad \Gamma_2' = \ln\frac{1-\beta}{\alpha} = +2.9444.$$

Теперь запишем выражение для логарифмического отношения правдоподобия:

$$l_n(x) = \sum_{k=1}^{n} \ln\frac{f_2(x_k)}{f_1(x_k)} = \sum_{k=1}^{n} \ln\frac{p_2^{x_k}(1-p_2)^{1-x_k}}{p_1^{x_k}(1-p_1)^{1-x_k}} = \ln\frac{p_2^{r_n}(1-p_2)^{n-r_n}}{p_1^{r_n}(1-p_1)^{n-r_n}},$$

где $r_n = \sum\limits_{k=1}^{n} x_k$. Логарифмируя частное, получим

$$l_n(x) = r_n \ln\left(\frac{p_2}{p_1}\right) + (n - r_n)\ln\left(\frac{1-p_2}{1-p_1}\right) =$$

$$= r_n \ln\left(\frac{p_2(1-p_1)}{p_1(1-p_2)}\right) + n\ln\left(\frac{1-p_2}{1-p_1}\right),$$

откуда получаем условия на r_n:

$$\frac{\Gamma_1' - n\ln\left(\frac{1-p_2}{1-p_1}\right)}{\ln\left(\frac{p_2(1-p_1)}{p_1(1-p_2)}\right)} < \sum_{k=1}^{n} x_k < \frac{\Gamma_2' - n\ln\left(\frac{1-p_2}{1-p_1}\right)}{\ln\left(\frac{p_2(1-p_1)}{p_1(1-p_2)}\right)},$$

или для наших конкретных данных

$$\frac{-2.9444 + n\cdot 0.4055}{0.6931} < \sum_{k=1}^{n} x_k < \frac{+2.9444 + n\cdot 0.4055}{0.6931}.$$

Измерения проводятся последовательно при $n = 1, 2, \ldots$ до тех пор, пока не будет нарушено одно из этих неравенств. Если нарушилось левое неравенство, то следует принять H_1, а если нарушилось правое неравенство, то следует принять H_2. Оценим количество измерений до остановки процедуры. Для этого сначала вычислим $\mathbb{E}_1 Z$ и $\mathbb{E}_2 Z$, где $Z = \ln\left(f_2(X_k)/f_1(X_k)\right)$:

$$\mathbb{E}_1 Z = \ln\frac{p_2^X(1-p_2)^{1-X}}{p_1^X(1-p_1)^{1-X}} =$$

$$= \ln\frac{p_2^0(1-p_2)^{1-0}}{p_1^0(1-p_1)^{1-0}}\cdot\mathbb{P}_1(X=0) + \ln\frac{p_2^1(1-p_2)^{1-1}}{p_1^1(1-p_1)^{1-1}}\cdot\mathbb{P}_1(X=1) =$$

$$= \ln\left(\frac{1-p_2}{1-p_1}\right)\cdot\frac{1}{2} + \ln\left(\frac{p_2}{p_1}\right)\cdot\frac{1}{2} = -0.0589.$$

$$\mathbb{E}_2 Z = \ln\left(\frac{1-p_2}{1-p_1}\right)\cdot\mathbb{P}_2(X=0) + \ln\left(\frac{p_2}{p_1}\right)\cdot\mathbb{P}_2(X=1) =$$

$$= \ln\left(\frac{1-p_2}{1-p_1}\right)\cdot\frac{1}{3} + \ln\left(\frac{p_2}{p_1}\right)\cdot\frac{2}{3} = 0.0566.$$

Отсюда сразу получаем оценки среднего числа измерений:

$$\mathbb{E}_1\nu \approx \frac{(1-\alpha)\Gamma_1' + \alpha\Gamma_2'}{\mathbb{E}_1 Z} = 44.9914.$$

$$\mathbb{E}_2\nu \approx \frac{\beta\Gamma_1' + (1-\beta)\Gamma_2'}{\mathbb{E}_2 Z} = 46.8197.$$

7. Точечное оценивание

7.1. Введение в точечное оценивание

Определение. Точечная оценка $T(X)$ параметрической функции $\tau(\theta)$ называется *несмещенной*, если

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

Пример. Рассмотрим модель $\mathrm{U}(0,\theta)$, $\theta>0$. Статистика $T_1(X)=$ $=2\overline{X}$ является несмещенной оценкой параметра θ, т.к.

$$\forall\theta>0\ \ \mathbb{E}_\theta T_1(X)=\mathbb{E}_\theta\left(\frac{2}{n}\sum_{k=1}^{n}X_k\right)=\frac{2}{n}\sum_{k=1}^{n}\mathbb{E}_\theta X_k=\frac{2}{n}\sum_{k=1}^{n}\frac{\theta}{2}=\theta.$$

Несмещенной оценкой параметра θ является и статистика $T_2(X)=$ $=X_{(1)}+X_{(n)}$. Чтобы это показать, заметим, что X_i/θ имеет равномерное распределение $\mathrm{U}(0,1)$, стало быть $X_{(1)}/\theta$ и $X_{(n)}/\theta$ являются соответственно первой и максимальной порядковыми статистиками распределения $\mathrm{U}(0,1)$. Известно, что такие порядковые статистики имеют бета-распределение: $\mathrm{Beta}(1,n)$ для первой порядковой статистики и $\mathrm{Beta}(n,1)$ для максимальной порядковой статистики. Напомню, что если $\xi\in\mathrm{Beta}(\alpha,\beta)$, то $\mathbb{E}\xi=\alpha/(\alpha+\beta)$. Значит $\forall\theta>0$:

$$\mathbb{E}_\theta T_2(X)=\theta\cdot\mathbb{E}_\theta\left(\frac{X_{(1)}}{\theta}\right)+\theta\cdot\mathbb{E}_\theta\left(\frac{X_{(n)}}{\theta}\right)=\theta\cdot\frac{1}{n+1}+\theta\cdot\frac{n}{n+1}=\theta.$$

Определение. Параметрическая функция $b(\theta)=\mathbb{E}_\theta T(X)-\tau(\theta)$ называется *смещением* оценки $T(X)$ параметрической функции $\tau(\theta)$. Для несмещенных оценок смещение тождественно равно нулю.

Определение. Оценка называется *асимптотически несмещенной*, если для любого значения параметра ее смещение относительно оцениваемой функции стремится к нулю с ростом объема выборки, то есть

$$\forall\theta\in\Theta\ \ b_n(\theta)\to 0,\ n\to\infty.$$

Здесь индекс n явно показывает зависимость смещения от объема выборки, хотя его часто опускают для краткости обозначений.

Пример. В модели $\mathrm{U}(0,\theta)$, $\theta>0$, статистика $T(X)=X_{(n)}$ является асимптотически несмещенной оценкой параметра θ:

$$\forall\theta>0\ \ b_n(\theta)=\mathbb{E}_\theta T(X)-\theta=-\frac{\theta}{n+1}\to 0,\ n\to\infty.$$

Несмещенные оценки не всегда существуют. Для примера рассмотрим модель $\mathrm{Be}(\theta)$, $\theta \in (0,1)$. Пусть $T(X)$ – произвольная статистика, тогда в общем виде можно записать

$$\mathbb{E}_\theta T(X) = \sum_{x=(x_1,\ldots,x_n)} T(x) f(x_1;\theta)\ldots f(x_n;\theta),$$

где функция вероятности $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, $x \in \{0,1\}$, $\theta \in (0,1)$, а суммирование производится по всем наборам $(x_1,\ldots,x_n)$ при $x_k \in \{0,1\}$, $k = 1,\ldots,n$. Это выражение можно упростить:

$$\mathbb{E}_\theta T(X) = \sum_{x=(x_1,\ldots,x_n)} T(x)\theta^{S_n}(1-\theta)^{n-S_n},\ S_n = \sum_{k=1}^n x_k.$$

Теперь хорошо видно, что математическое ожидание от *произвольной* статистики $T(X)$ является полиномом степени не выше n относительно параметра θ. Поэтому, если данная $\tau(\theta)$ не является полиномом θ, или является полиномом степени выше n, то несмещенных статистик для $\tau(\theta)$ не существует. Пусть $n = 1$, тогда для произвольной статистики $T(X)$ получаем

$$\mathbb{E}_\theta T(X) = T(0)\cdot(1-\theta) + T(1)\cdot\theta,$$

и поэтому несмещенные оценки могут существовать (и существуют) только для функций вида $\tau(\theta) = a + b\theta$ с произвольными постоянными a и b. Если же $n = 2$, то получаем

$$\mathbb{E}_\theta T(X) = T(0,0)\cdot(1-\theta)^2 + T(0,1)\cdot(1-\theta)\theta + T(1,0)\cdot\theta(1-\theta) + T(1,1)\cdot\theta^2,$$

и поэтому несмещенные оценки могут существовать (и существуют) только для функций вида $\tau(\theta) = a + b\theta + c\theta^2$ с произвольными постоянными a, b и c.

Нужно знать, что выборочные моменты

$$\hat{\alpha}_m = \frac{1}{n}\sum_{k=1}^n X_k^m$$

для каждого $m > 0$ в любой модели являются несмещенными оценками истинных моментов $\mathbb{E}_\theta X_1^m$. Что же касается центральных моментов

$$\hat{\mu}_m = \frac{1}{n}\sum_{k=1}^n (X_k - \overline{X})^m,$$

то такие оценки, вообще говоря, являются смещенными. Например, выборочная дисперсия

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

имеет смещение $b_n(\theta) = \mathbb{E}_\theta S^2 - \mathbb{D}_\theta X_1 = -\dfrac{1}{n}\mathbb{D}_\theta X_1 \to 0$, $n \to \infty$, то есть S^2 является асимптотически несмещенной оценкой истинной дисперсии. Часто рассматривают исправленную выборочную дисперсию

$$S_0^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2.$$

Она является несмещенной оценкой дисперсии $\mathbb{D}_\theta X_1$ в любой модели (в любой, в которой существует дисперсия).

Определение. Оценка $T_n(X)$ называется *состоятельной* оценкой параметрической функции $\tau(\theta)$, если

$$\forall\theta \in \Theta \;\; T_n(X) \xrightarrow{\mathbb{P}_\theta} \tau(\theta), \;\; n \to \infty,$$

то есть если для каждого значения параметра θ с ростом объема выборки эта оценка сходится по вероятности к оцениваемой функции.

Для исследования оценок на состоятельность нужно помнить несколько сведений из теории вероятности. Перечислим их по порядку.

1. Неравенство Чебышева. Для любой случайной величины ξ с конечными математическим ожиданием $\mathbb{E}\xi$ и дисперсией $\mathbb{D}\xi$ верно неравенство

$$\forall\varepsilon > 0 \;\; \mathbb{P}\left(|\xi - \mathbb{E}\xi| > \varepsilon\right) \le \frac{\mathbb{D}\xi}{\varepsilon^2}.$$

Например, пусть статистика $T_n(X)$ является несмещенной оценкой функции $\tau(\theta)$. Тогда можно записать

$$\forall\theta \in \Theta \;\; \forall\varepsilon > 0 \;\; \mathbb{P}_\theta\left(|T_n(X) - \tau(\theta)| > \varepsilon\right) \le \frac{\mathbb{D}_\theta T_n(X)}{\varepsilon^2},$$

где было учтено для несмещенной оценки $\mathbb{E}_\theta T(X) = \tau(\theta)$. Теперь если для каждого θ дисперсия $\mathbb{D}_\theta T_n(X) \to 0$, $n \to \infty$, то вероятность в левой части выражения для каждого θ также стремится к нулю. По определению это будет означать сходимость по вероятности $T_n(X) \xrightarrow{\mathbb{P}_\theta} \tau(\theta)$, а значит и состоятельность оценки $T_n(X)$.

Заметим, что статистика $T_n(X)$ необязательно должна являться несмещенной для того, чтобы воспользоваться неравенством Чебышева

для проверки оценки на состоятельность. Достаточно, чтобы она являлась асимптотически несмещенной. А именно, если $T_n(X)$ является асимптотически несмещенной оценкой функции $\tau(\theta)$ и для каждого θ дисперсия $\mathbb{D}_\theta T_n(X) \to 0$, $n \to \infty$, то $T_n(X)$ является состоятельной оценкой функции $\tau(\theta)$.

2. Закон больших чисел Хинчина. Пусть $\xi_1, \xi_2, \ldots$ – последовательность независимых одинаково распределенных случайных величин с конечным математическим ожиданием $\mathbb{E}\xi_k = \mu$, $\forall k \geq 1$. Тогда

$$\frac{1}{n}\sum_{k=1}^{n} \xi_k \xrightarrow{\mathbb{P}} \mu, \quad n \to \infty.$$

Например, для простой выборки $X = (X_1, X_2, \ldots, X_n)$ отсюда прямо следует, что в любой модели среднее выборочное сходится к математическому ожиданию $\mathbb{E}_\theta X_1$. Значит для $\tau(\theta) = \mathbb{E}_\theta X_1$:

$$\forall \theta \in \Theta \quad \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{\mathbb{P}_\theta} \tau(\theta), \quad n \to \infty.$$

По определению это значит, что выборочное среднее является состоятельной оценкой истинного математического ожидания. То же самое можно сказать и про произвольные выборочные моменты: они являются состоятельными оценками истинных моментов:

$$\forall \theta \in \Theta \quad \frac{1}{n}\sum_{k=1}^{n} X_k^m \xrightarrow{\mathbb{P}_\theta} \mathbb{E}_\theta X_1^m, \quad n \to \infty.$$

3. Свойства сходимости по вероятности. Очень важно знать и применять следующие свойства сходимости по вероятности (всюду $n \to \infty$):

1) если $\xi_n \xrightarrow{\mathbb{P}} \xi$ и $\eta_n \xrightarrow{\mathbb{P}} \eta$, то $\xi_n \pm \eta_n \xrightarrow{\mathbb{P}} \xi \pm \eta$ и $\xi_n \eta_n \xrightarrow{\mathbb{P}} \xi\eta$,

2) если $\xi_n \xrightarrow{\mathbb{P}} \xi$, то $c\xi_n \xrightarrow{\mathbb{P}} c\xi$ для любого $c \in \mathbb{R}$,

3) если $\xi_n \xrightarrow{\mathbb{P}} \xi$ и функция φ непрерывна, то $\varphi(\xi_n) \xrightarrow{\mathbb{P}} \varphi(\xi)$.

Например, нам известно, что выборочное среднее $\overline{X}$ в любой модели является состоятельной оценкой математического ожидания $\mathbb{E}_\theta X_1$. Тогда согласно свойству 3.3 квадрат выборочного среднего $\overline{X}^2$ является состоятельной оценкой квадрата математического ожидания $(\mathbb{E}_\theta X_1)^2$. Далее, слагаемое $\overline{X^2}$ является состоятельной оценкой второго момента $\mathbb{E}_\theta X_1^2$ согласно закону больших чисел Хинчина. Из свойства 3.2 следует, что разность $\overline{X^2} - \overline{X}^2$ является состоятельной оценкой

разности $\mathbb{E}_\theta X_1^2 - (\mathbb{E}_\theta X_1)^2$, то есть дисперсии $\mathbb{D}_\theta X_1$. Таким образом, выборочная дисперсия

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

является состоятельной оценкой истинной дисперсии $\mathbb{D}_\theta X_1$. То же самое касается и выборочной дисперсии S_0^2, так как

$$S_0^2 = \frac{n}{n-1}S^2,$$

множитель $n/(n-1)$ сходится к 1 (во всех смыслах, в том числе и по вероятности), а S^2 сходится к дисперсии.

Задача. Найти несмещенную состоятельную оценку θ^2 в модели $\mathrm{N}(\theta, \sigma^2)$.

Решение. В первую очередь вспоминаем, что выборочное среднее $\overline{X}$ является несмещенной и состоятельной оценкой математического ожидания, то есть θ в нашем случае. Далее, согласно свойству 3.3 статистика $\overline{X}^2$ является состоятельной оценкой θ^2. Математическое ожидание от $\overline{X}^2$ равно

$$\mathbb{E}_\theta \overline{X}^2 = \mathbb{D}_\theta \overline{X} + \left(\mathbb{E}_\theta \overline{X}\right)^2 = \frac{\sigma^2}{n} + \theta^2.$$

Отсюда видно, что смещение равно σ^2/n и убывает к нулю с ростом выборки, значит $\overline{X}^2$ является асимптотически несмещенной оценкой θ^2. Так как σ^2 по условию задачи нам известна, то рассмотрим статистику

$$T(X) = \overline{X}^2 - \frac{\sigma^2}{n}.$$

Так как σ^2/n не зависит от выборки, то она сходится к нулю во всех смыслах сходимости, в том числе и по вероятности. Тогда по свойству 3.1 заключаем, что статистика $T(X)$ сходится по вероятности к θ^2 и поэтому является несмещенной состоятельной оценкой θ^2.

Определение. *Оценкой максимального правдоподобия* (ОМП) параметра $\theta \in \Theta$ в точке $x \in \Omega$ называется такая точка параметрического множества $\hat{\theta}(x) \in \Theta$, в которой функция правдоподобия $L(x;\theta)$ достигает максимума:

$$\forall \theta \in \Theta \quad L(x;\hat{\theta}(x)) \geq L(x;\theta),$$

если такая точка $\hat{\theta}(x) \in \Theta$ существует.

Комментарии. Оценка максимального правдоподобия не обязана существовать для любого $x \in \Omega$ и не обязана быть единственной для каждого x, в котором существует. Чтобы найти ОМП в точке $x \in \Omega$, нужно зафиксировать этот x и решить оптимизационную задачу

$$L(x;\theta) \to \max_{\theta\in\Theta}.$$

Следует помнить, что этот максимум может достигаться как во внутренних точках Θ (и тогда его удобно искать среди корней уравнения $\partial L/\partial\theta = 0$), так и на границе множества Θ. Функция $L(x;\theta)$ относительно θ может иметь минимумы и точки перегиба, поэтому корни уравнения $\partial L/\partial\theta = 0$ необязательно являются точками максимума этой функции. Кроме того, функция $L(x;\theta)$ вообще не обязана быть дифференцируемой по θ. С каждой ситуацией нужно разбираться отдельно.

Задача 1. Найти ОМП параметра θ в модели $\mathrm{Be}(\theta)$, $\theta \in [0,1]$.

Решение. Функция правдоподобия в этой модели записывается так:

$$L(x;\theta) = \theta^s(1-\theta)^{n-s},\ \theta \in (0,1),\ s \in [0,n],$$

$$L(x;0) = \begin{cases} 1, & s = 0, \\ 0, & s > 0; \end{cases}$$

$$L(x;1) = \begin{cases} 1, & s = n, \\ 0, & s < n, \end{cases}$$

где $s = s(x)$ – количество единиц в векторе x, то есть $s = \sum_{k=1}^n x_k$.

Чтобы найти ОМП, мы будем поочередно фиксировать возможные x (и соответственно s) и максимизировать полученное выражение относительно θ. Если $s = 0$, то очевидно, что максимум функции $L(x;\theta)$ достигается при $\theta = 0$. Если $s = n$, то аналогично получаем, что максимум достигается при $\theta = 1$.

Теперь разберем случай, когда $s \in (0,n)$. В этом случае $L(x;0) = L(x;1) = 0$, а для $\theta \in (0,1)$ функция $L(x;\theta) > 0$, поэтому если максимум достигается, то только в точках $\theta \in (0,1)$.

Итак, пусть $s \in (0,n)$ и $\theta \in (0,1)$. В этом случае функция $L(x;\theta)$ представляет собой гладкую функцию параметра θ, заданную на открытом множестве $\theta \in (0,1)$. Заметим, что максимум функции $L(x;\theta)$ достигается там же, где максимум функции

$$\ln L(x;\theta) = s\ln\theta + (n-s)\ln(1-\theta).$$

Это тоже гладкая функция на открытом множестве, поэтому ее максимум (если существует) необходимо удовлетворяет уравнению

$$\frac{\partial \ln L}{\partial\theta} = \frac{s}{\theta} - \frac{n-s}{1-\theta} = 0,$$

откуда $\theta = s/n \in (0,1)$. Так как вторая производная

$$\frac{\partial^2 \ln L}{\partial \theta^2} = -\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2} = -\frac{n^2}{s} - \frac{n^2}{n-s} < 0$$

в точке $\theta = s/n$, то она является точкой максимума.

Мы рассмотрели все возможные случаи $s \in [0,n]$ и готовы заключить, что

$$\hat{\theta}(s) = \left[\begin{array}{ll} 0, & s = 0, \\ s/n, & s/n \in (0,1), \\ 1, & s = n. \end{array} \right.$$

Три случая можно объединить в один:

$$\hat{\theta}(x) = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \forall x \in \{0,1\}^n.$$

Задача 2. Найти ОМП параметра θ в модели $\mathrm{U}(0,\theta)$, $\theta > 0$.

Решение. Функция правдоподобия

$$L(x;\theta) = \theta^{-n} I(x_{(n)} \le \theta) \cdot I(x_{(1)} \ge 0).$$

Если $\theta < x_{(n)}$, то $L(x;\theta) = 0$. Если же $\theta \ge x_{(n)}$, то $L(x;\theta) > 0$ и убывает с ростом θ. Поэтому максимум достигается при $\theta = x_{(n)}$. Значит оценка максимального правдоподобия есть $\hat{\theta} = X_{(n)}$.

Замечание. Как уже было замечено, в произвольной параметрической модели может оказаться, что для некоторых x из выборочного пространства максимум не достигается ни при каком $\theta \in \Theta$. Так обстоит дело, например, в модели $\mathrm{Be}(\theta)$, $\theta \in (0,1)$. А именно, если $\overline{x} \in (0,1)$, то максимум функции правдоподобия достигается в некоторой точке $\hat{\theta}(x) \in (0,1)$. Однако при $\overline{x} = 0$ и $\overline{x} = 1$ максимум достигается лишь на границе множества $(0,1)$, то есть в точках $\theta = 0$ и $\theta = 1$ соответственно. Эта известная неприятность для ограниченных множеств Θ решается замыканием параметрического множества Θ, то есть переходом к $\Theta = [0,1]$. В случае же неограниченных множеств Θ можно ограничиваться лишь компактными подмножествами, но только в тех случаях, когда вероятностная мера подмножества $\tilde{\Omega} \subset \Omega$, на котором максимум не достигается, стремится к нулю с ростом объема выборки[6, 7]. Для практики и развития асимптотической теории оценок максимального правдоподобия этого оказывается достаточно.

[6] Боровков А.А. Математическая статистика : учебник. 4-е изд., стер., 2010. С. 84.

[7] Dudley R.M. Mathematical Statistics. Chapter 3. Maximum Likelihood and M-estimation. 2003.

7.2. Еще об оценке максимального правдоподобия

Напоминаю, что в случаях, когда выборка X распределена дискретно, функция правдоподобия $L(x, \theta)$ представляет собой функцию вероятности выборки, то есть $L(x, \theta) = \mathbb{P}_\theta(X = x)$. Если выборка имеет абсолютно непрерывное распределение, то $L(x, \theta)$ – это просто плотность ее распределения. Оценка максимального правдоподобия $\hat{\theta}(x)$ – это точка пространства параметров Θ, в которой достигается максимум функции $L(x, \theta)$. Можно сказать, это то значение параметра, при котором выпадение наблюдаемого сейчас вектора x наиболее вероятно. Ясно, что в качестве оценки параметра разумно не брать те значения, для которых наблюдение x наименее вероятно, но почему именно наиболее вероятно?

Во-первых, оказывается, что при определенных условиях регулярности на функцию $L(x, \theta)$ и множество параметров Θ оценка максимального правдоподобия оказывается состоятельной, асимптотически несмещенной и асимптотически нормальной[8]. Во-вторых, справедливо более глубокое свойство оценки максимального правдоподобия. Рассмотрим все множество распределений и выделим в нем множество $\{L(x, \theta)\}_{\theta \in \Theta}$, то есть множество моделей, среди которых мы ищем истинное распределение выборки. В пространстве всех распределений также выделим распределение, соответствующее эмпирической функции распределения $\hat{F}_n(x)$. Оказывается[9], что в пространстве распределений можно ввести такую метрику, расстояние между распределениями, что оценка максимального правдоподобия минимизирует это расстояние между множеством $\{L(x, \theta)\}_{\theta \in \Theta}$ и эмпирическим распределением $\hat{F}_n(x)$. Кроме того, как мы уже знаем, эмпирическая функция распределения сходится к истинной функции распределения с ростом объема выборки. Из этого следует, что оценка максимального правдоподобия с ростом объема выборки приближается к истинному параметру распределения.

7.3. Эффективные и оптимальные оценки

Пусть имеется параметрическая функция $\tau(\theta)$, $\theta \in \Theta$. Произвольный класс оценок $T(X)$ этой функции будем обозначать символом K, а класс несмещенных оценок обозначим символом K_τ:

$$K_\tau = \{T(X) : \forall \theta \in \Theta \;\; \mathbb{E}_\theta T(X) = \tau(\theta)\}$$

[8] См., например, Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику : учебник. Москва : Издательство ЛКИ. 2010. С. 230.

[9] См., например, Боровков А.А. Математическая статистика : учебник. 4-е изд., стер., 2010. С. 103.

и, в частности,

$$K_\theta = \{T(X) : \forall \theta \in \Theta \;\; \mathbb{E}_\theta T(X) = \theta\}.$$

Теорема Рао–Крамера. *Пусть выполнены условия:*

1) *выборочное пространство* Ω *не зависит от оцениваемого параметра* θ*;*

2) *для всех* $\theta = (\theta_1, \ldots, \theta_r) \in \Theta$ *и всех* $T(X) \in K_\theta$ *допускается дифференцирование по компонентам параметра* θ *под знаком интегралов:*

$$\forall j = 1, \ldots, r \quad \frac{\partial}{\partial \theta_j} \int\limits_\Omega L(x;\theta)dx = \int\limits_\Omega \frac{\partial}{\partial \theta_j} L(x;\theta)dx,$$

$$\forall j = 1, \ldots, r \quad \frac{\partial}{\partial \theta_j} \int\limits_\Omega T(x)L(x;\theta)dx = \int\limits_\Omega T(x)\frac{\partial}{\partial \theta_j} L(x;\theta)dx$$

для непрерывного распределения выборки X *и под знаком рядов:*

$$\forall j = 1, \ldots, r \quad \frac{\partial}{\partial \theta_j} \sum_{x\in\Omega} L(x;\theta) = \sum_{x\in\Omega} \frac{\partial}{\partial \theta_j} L(x;\theta),$$

$$\forall j = 1, \ldots, r \quad \frac{\partial}{\partial \theta_j} \sum_{x\in\Omega} T(x)L(x;\theta) = \sum_{x\in\Omega} T(x)\frac{\partial}{\partial \theta_j} L(x;\theta)$$

в случае дискретного распределения выборки X*,*

3) для каждого $\theta \in \Theta$ *существует невырожденная матрица* $J = \|J_{ij}\|_{i,j=1}^r$ *с компонентами*

$$J_{ij} = \mathbb{E}_\theta \left(\frac{\partial \ln L}{\partial \theta_i} \frac{\partial \ln L}{\partial \theta_j} \right)$$

.

Тогда $\forall \theta \in \Theta \;\; \forall T \in K_\theta \;\; \forall e \in \mathbb{R}^r$

$$e^T R_T e \geq e^T J^{-1} e,$$

где R_T *– ковариационная матрица вектора* $T(X)$*. Это неравенство называется неравенством Рао–Крамера.*

Так как параметр θ в общем случае предполагается векторным и имеет некоторую размерность $r \geq 1$, то и оценивающая его статистика $T(X)$ является вектором размерности r. Если $r = 1$, то $R_T = \mathbb{D}_\theta T$, и тогда неравенство Рао–Крамера перепишется в виде

$$\forall T \in K_\theta \;\; \forall \theta \in \Theta \;\; \mathbb{D}_\theta T \geq \frac{1}{\mathbb{D}_\theta \left(\frac{\partial \ln L}{\partial \theta} \right)}.$$

Если $r = 1$ и оценивается скалярная функция $\tau(\theta)$, то неравенство Рао–Крамера записывается в виде

$$\forall T \in K_\tau \quad \forall \theta \in \Theta \quad \mathbb{D}_\theta T \geq \frac{[\tau'(\theta)]^2}{\mathbb{D}_\theta \left(\frac{\partial \ln L}{\partial \theta}\right)},$$

где $\tau'(\theta) = d\tau/d\theta$. Всюду далее будем полагать $r = 1$.

Определение. Если для несмещенной оценки T выполняется равенство в неравенстве Рао–Крамера, то оценка T параметра θ называется *эффективной* (или *R-эффективной*) оценкой параметра θ.

Определение. Оценка $T^*(X)$ параметрической функции $\tau(\theta)$ называется *оптимальной (в среднеквадратичном смысле) в классе оценок* K, если

$$\forall \theta \in \Theta \quad \forall T \in K \quad \mathbb{E}_\theta (T^*(X) - \tau(\theta))^2 \leq \mathbb{E}_\theta (T(X) - \tau(\theta))^2.$$

Определение. Оценки, оптимальные в классе несмещенных оценок, называются просто *оптимальными*[10]. Для оптимальных оценок $T^*(X)$ выполнено

$$\forall \theta \in \Theta \quad \forall T \in K_\tau \quad \mathbb{D}_\theta T^*(X) \leq \mathbb{D}_\theta T(X).$$

Свойства эффективных оценок

1. Если эффективная оценка параметра θ существует, то она является оптимальной и потому единственной эффективной оценкой параметра θ. Единственность понимается в том смысле, что если T_1 и T_2 являются оптимальными оценками параметра θ, то $T_1 = T_2$ почти всюду.

2. Если $T(X)$ является эффективной оценкой функции $\tau(\theta)$, то для любых постоянных $a, b \in \mathbb{R}$, $a \neq 0$, оценка $aT(X) + b$ является эффективной оценкой функции $a\tau(\theta) + b$.

3. Если $T(X)$ является эффективной оценкой функции $\tau(\theta)$, то

$$T(X) - \tau(\theta) = a(\theta) V(X; \theta),$$

где $a(\theta) = \mathbb{D}_\theta T / \tau'(\theta)$ и *вклад выборки*

$$V(X; \theta) = \frac{\partial \ln L(X; \theta)}{\partial \theta}.$$

[10] В учебниках Боровкова А.А. и Черновой Н.И., а также во многих других источниках, то, что мы называем оптимальными (в каком-то классе) оценками, эти авторы называют их *эффективными* (в каком-то классе) оценками, а то, что мы называем эффективными оценками, они называют *R-эффективными* оценками. Наша терминология совпадает с терминологией в учебнике Ивченко Г.И. и Медведева Ю.И., а также с терминологией в учебнике Натана А.А., Горбачева О.Г. и Гуза С.А.

4. Если $T_1(X)$ является эффективной оценкой функции $\tau_1(\theta)$, а $T_2(X)$ является эффективной оценкой функции $\tau_2(\theta)$, то существуют постоянные a, b, c, не равные одновременно нулю, такие, что $aT_1(X) + bT_2(X) = c$ почти всюду и $a\tau_1(\theta) + b\tau_2(\theta) = c$ всюду.

5. Эффективные оценки существуют только для моделей экспоненциального семейства распределений. А именно, если функция правдоподобия имеет вид

$$L(x;\theta) = h(x)\exp\left(a(\theta)U(x) + b(\theta)\right),$$

то существует эффективная оценка $T(X) = U(X)$ параметрической функции $\tau(\theta) = -b'(\theta)/a'(\theta)$.

6. Если эффективная оценка скалярной функции $\tau(\theta)$ существует, то она является также и оценкой максимального правдоподобия функции $\tau(\theta)$.

Задача 1. Дана модель $\mathrm{Be}(\theta)$, $\theta \in (0,1)$. Доказать, что статистика $T(X) = \overline{X}$ является эффективной оценкой параметра θ.

Решение 1 (по определению, самое тяжелое). Сначала убеждаемся в том, что для данной модели выполнены все условия теоремы о неравенстве Рао–Крамера. Выборочное пространство $\Omega = \{0,1\}^n$ не зависит от θ. Функция правдоподобия

$$L(x;\theta) = \theta^S(1-\theta)^{n-S}, \quad S = \sum_{i=1}^{n} x_i,$$

дифференцируема в каждой точке $\theta \in (0,1)$. Так как выборочное пространство Ω состоит из конечного числа точек, а распределение выборки X является дискретным, то знак производной можно вносить внутрь сумм. Далее, квадрат производной

$$\left(\frac{\partial \ln L}{\partial \theta}\right)^2 = \left(\frac{S - n\theta}{\theta(1-\theta)}\right)^2$$

имеет конечное и ненулевое математическое ожидание. Статистика $T(X)$ является несмещенной оценкой параметра θ:

$$\forall \theta \in (0,1) \;\; \mathbb{E}_\theta T(X) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\theta X_i = \theta.$$

Теперь проверим, что оценка $T(X)$ доставляет равенство в неравенстве Рао–Крамера. Для этого сначала вычислим дисперсию статистики

$$\mathbb{D}_\theta T(X) = \frac{\theta(1-\theta)}{n},$$

затем найдем вклад выборки

$$V(X;\theta) = \frac{\partial \ln L(X;\theta)}{\partial \theta} = \frac{S - n\theta}{\theta(1-\theta)}$$

и ее дисперсию

$$\mathbb{D}_\theta V(X;\theta) = \frac{1}{\theta^2(1-\theta)^2} \sum_{i=1}^{n} \theta(1-\theta) = \frac{n}{\theta(1-\theta)}.$$

Так как $\mathbb{D}_\theta T(X) = 1/\mathbb{D}_\theta V(X;\theta)$ верно для всех θ, то по определению статистика $T(X)$ является эффективной оценкой параметра θ.

Решение 2 (через специальное представление, полегче). Проверяем регулярность модели, как в предыдущем решении. Далее достаточно проверить равенство

$$T(X) - \theta = a(\theta) V(X;\theta),\ a(\theta) = \mathbb{D}_\theta T.$$

Равенство выше равносильно

$$\frac{1}{n} S - \theta = \frac{\theta(1-\theta)}{n} \frac{S - n\theta}{\theta(1-\theta)},$$

что представляет собой верное при всех $\theta \in (0,1)$ равенство. Заметим, что это решение проще предыдущего в том смысле, что здесь не нужно вычислять дисперсию вклада выборки.

Решение 3 (самое простое). Даже без проверки модели на регулярность сразу замечаем, что модель $\mathrm{Be}(\theta)$, $\theta \in (0,1)$, принадлежит экспоненциальному семейству распределений, причем

$$L(x;\theta) = \theta^S (1-\theta)^{n-S} = \exp\left(S \ln\left(\frac{\theta}{1-\theta} \right) + n \ln(1-\theta) \right),\ \ S = \sum_{i=1}^{n} X_i.$$

Следовательно, статистика S является эффективной оценкой функции

$$\tau(\theta) = -n \frac{\partial \ln(1-\theta)/\partial\theta}{\partial \ln\left(\frac{\theta}{1-\theta}\right)/\partial\theta} = n\theta,$$

а значит $\overline{X} = S/n$ является эффективной оценкой $\tau(\theta)/n = \theta$. В данном варианте решения никаких дисперсий рассчитывать не нужно.

Определение. Статистика $S = S(X)$ называется *достаточной относительно параметра* θ, если условная функция правдоподобия $L(x;\theta|S(x) = s)$ не зависит от θ ни при каких s.

Смысл достаточных статистик состоит в следующем. Как мы вскоре увидим, сколько-нибудь содержательные оценки параметров и параметрических функций не просто зависят от выборки, они зависят от выборки через некоторую функцию. Например, эффективная оценка параметра θ в модели $\mathrm{Be}(\theta)$ зависит от суммы X_i. Оказывается, что в теории оценивания это обычное дело, когда для оценки параметра достаточно знать какую-то комбинацию $S(X)$ элементов выборки (часто скалярную функцию), а не каждый элемент выборки по отдельности. Другими словами, для оценки параметра нужно куда меньше информации, чем имеется. Достаточные статистики – это способ формализовать это наблюдение.

Обосновать это определение можно так. Если мы возьмем поверхность $S(x) = s$ для фиксированного s и рассмотрим вероятностное подпространство на этой поверхности (то есть условное распределение $L(x;\theta|S(x) = s)$), то для достаточной статистики S окажется, что распределение выборки на этой поверхности от θ не зависит, поэтому мы никаким образом не сможем воспользоваться этой информацией для оценки θ. Вот если бы выпадение тех или иных x на этой поверхности зависело бы от θ, мы бы смогли этим воспользоваться, например, найдя оценку максимального правдоподобия, то есть найти значение параметра, которое наиболее вероятно соответствует выборке. Но для достаточных статистик распределение выборки на этой поверхности от параметра не зависит. Выходит так, что оценка параметра зависит не от вектора x, а скорее от s – некоторой комбинации элементов вектора x. На самом деле, выборка X – это пример достаточной статистики, но неэкономный, в ряде случаев удается найти достаточные статистики малой размерности, например размерности 1.

Приведу пример, как можно доказать, что некоторая статистика является достаточной, воспользовавшись при этом только определением.

Пример. Рассмотрим модель Пуассона $\mathrm{Po}(\theta)$, $\theta > 0$. Покажем, что статистика $S = \sum_{k=1}^{n} X_k$ является достаточной для параметра θ. Рассмотрим вероятность

$$L(x;\theta|S(x) = s) = \mathbb{P}_\theta(X = x|S = s) =$$

$$= \frac{\mathbb{P}_\theta\left(X_1 = x_1, \dots, X_n = x_n, \sum\limits_{k=1}^{n} X_k = s\right)}{\mathbb{P}_\theta\left(\sum\limits_{k=1}^{n} X_k = s\right)}.$$

Если $\sum_{k=1}^{n} x_k \neq s$, то это выражение равно нулю, и поэтому не зависит

от θ. Рассмотрим случай, когда $\sum_{k=1}^n x_k = s$. В этом случае выражение $\sum_{k=1}^n X_k = s$ в числителе можно убрать. Более того, $\sum_{k=1}^n X_k$ имеет распределение $\mathrm{Po}(n\theta)$. Отсюда следует, что искомая вероятность равна

$$\left(e^{-n\theta}\frac{(n\theta)^s}{s!}\right)^{-1}\cdot\prod_{k=1}^n\frac{\theta^{x_k}}{x_k!}\,e^{-\theta}=\frac{s!}{n^s\prod\limits_{k=1}^n x_k!}$$

и от θ не зависит. Поэтому статистика S является достаточной для θ.

Это было доказательство достаточности статистики исходя из определения. Но на практике гораздо удобнее пользоваться следующей теоремой.

Критерий факторизации. *Для того чтобы статистика $S(X)$ была достаточной для параметра θ, необходимо и достаточно, чтобы функция правдоподобия $L(x;\theta)$ имела вид*

$$L(x;\theta)=g(S(x);\theta)h(x),$$

где g и h – неотрицательные функции, $g(s;\theta)$ измерима относительно s, а $h(x)$ измерима относительно x, и $h(x)$ не зависит от θ.

Пример. Вернемся к предыдущему примеру. Функция правдоподобия

$$L(x;\theta)=\prod_{k=1}^n\frac{\theta^{x_k}}{x_k!}\,e^{-\theta}=e^{-n\theta}\theta^{S(x)}\cdot\prod_{k=1}^n\frac{1}{x_k!}.$$

Видно, что мы можем взять $g(s;\theta)=e^{-n\theta}\theta^s$, а $h(x)=\prod_{k=1}^n\frac{1}{x_k!}$. Согласно критерию факторизации, отсюда следует, что $S(X)=\sum_{k=1}^n X_k$ является достаточной статистикой для θ в модели $\mathrm{Po}(\theta)$.

Пример. Рассмотрим модель $\mathrm{U}(0,\theta)$, $\theta>0$. Запишем функцию правдоподобия

$$L(x;\theta)=\theta^{-1}I(x_1\in[0,\theta])\ldots\theta^{-1}I(x_n\in[0,\theta])=\theta^{-n}I(x_{(n)}\le\theta)\cdot I(x_{(1)}\ge 0).$$

В данном случае статистика $X_{(n)}$ является достаточной, так как для нее можно взять $g(s;\theta)=\theta^{-n}I(s\le\theta)$ и $h(x)=I(x_{(1)}\ge 0)$.

Теперь о свойствах достаточных статистик.

Теорема. *Пусть S – достаточная статистика для параметра θ, а $u=\varphi(v)$ – взаимно однозначное отображение. Тогда $S_1=\varphi(S)$ – тоже достаточная статистика для параметра θ. Если $\tau(\theta)$ осуществляет взаимно однозначное отображение, то S является достаточной статистикой и для $\tau(\theta)$.*

Теорема. *Пусть $S(X)$ – достаточная статистика для параметра θ. Тогда если оценка максимального правдоподобия параметра*

θ существует и единственна, то она является функцией этой достаточной статистики.

Определение. Статистика $T(X)$ называется *полной достаточной статистикой* для параметра θ, если она является достаточной и если из утверждения

$$\forall \theta \in \Theta \;\; \mathbb{E}_\theta \varphi(T) = 0$$

следует $\varphi = 0$ п.н. относительно распределения T.

Другими словами, математическое ожидание только тривиальной функции полной достаточной статистики тождественно равно нулю.

Теорема Рао–Блекуэлла–Колмогорова. *Оптимальная оценка, если она существует, является функцией от достаточной статистики.*

Теорема о полных достаточных статистиках. *Любая функция полной достаточной статистики является оптимальной оценкой своего математического ожидания.*

Значит это ровно следующее. Пусть $S(X)$ – полная достаточная статистика. Возьмем произвольную измеримую функцию $\varphi(s)$. Тогда $T(X) = \varphi(S(X))$ является оптимальной оценкой параметрической функции $\tau(\theta) = \mathbb{E}_\theta \varphi(S(X))$. Значит, если мы хотим найти оптимальную оценку для какой-либо параметрической функции $\tau(\theta)$, то достаточно найти полную достаточную статистику S и найти такую функцию $\varphi(S)$, чтобы

$$\forall \theta \in \Theta \;\; \tau(\theta) = \mathbb{E}_\theta \varphi(S).$$

Это уравнение называется *уравнением несмещенности*. Неизвестной здесь является функция φ.

Ранее мы нашли оптимальную оценку параметра θ в модели $\mathrm{Be}(\theta)$, просто найдя ее эффективную оценку. Теперь найдем оптимальную оценку через полную достаточную статистику.

Задача 2. Дана модель $\mathrm{Be}(\theta)$, $\theta \in (0, 1)$. Найти оптимальную оценку параметра θ.

Решение. Сначала ищем достаточную статистику. Для этого запишем функцию. Для этого записываем функцию правдоподобия

$$L(x; \theta) = \theta^S (1 - \theta)^{n-S}, \quad S = \sum_{i=1}^{n} x_i.$$

Эта функция зависит от x только через S. По критерию факторизации заключаем, что статистика S является достаточной статистикой для параметра θ. Теперь исследуем ее на полную достаточность. Возьмем произвольную функцию $\varphi(\cdot)$ подадим ей на вход статистику S

и предположим, что для любого $\theta \in (0,1)$ выполнено равенство

$$\mathbb{E}_\theta \varphi(S) = \sum_{k=0}^{n} \varphi(k) \cdot C_n^k \theta^k (1-\theta)^{n-k} = 0.$$

Мы воспользовались тем, что $S \in \mathrm{Bi}(n,\theta)$. Нам надо доказать, что отсюда последует $\varphi \equiv 0$, то есть что $\varphi(k) = 0$ для каждого $k = 0, \ldots, n$. Чтобы это показать, введем сначала переменную

$$z = \frac{\theta}{1-\theta} \in (0, +\infty),$$

и тогда наше уравнение запишется в виде

$$\forall z \in (0,+\infty) \quad \sum_{k=0}^{n} \varphi(k) \cdot C_n^k z^k = 0.$$

Полином равен тождественно нулю тогда и только тогда, когда все его коэффициенты равны нулю. Следовательно, $\varphi(k) = 0$ для каждого $k = 0, \ldots, n$. По определению это значит, что статистика S является полной достаточной статистикой. Попробуем теперь подобрать такую функцию $\varphi(S)$, чтобы

$$\forall \theta \in (0,1) \quad \mathbb{E}_\theta \varphi(S) = \theta.$$

Так как $\mathbb{E}_\theta S = n\theta$, то

$$\mathbb{E}_\theta \left(\frac{1}{n} S \right) = \theta.$$

Мы нашли подходящую $\varphi(S) = S/n = \overline{X}$. Согласно теореме о полных достаточных статистиках, статистика $\overline{X}$ является оптимальной оценкой своего математического ожидания, то есть θ.

Задача 3. Найти оптимальную оценку параметра θ в модели $\mathrm{Exp}(\theta)$, $\theta > 0$.

Решение. Из вида функции правдоподобия

$$L(x,\theta) = \prod_{i=1}^{n} \theta \exp(-\theta x_i) I(x_i \geq 0) = \theta^n \exp\left(-\theta \sum_{i=1}^{n} x_i \right) I(x_{(1)} \geq 0)$$

следует, что статистика $S = \sum_{i=1}^{n} x_i$ является достаточной. Докажем, что она полная. Статистика $S \in \Gamma(n,\theta)$, и плотность этой статистики равна

$$f_S(s) = \theta \exp(-\theta s) \frac{(s\theta)^{n-1}}{(n-1)!} I(s \geq 0).$$

Возьмем произвольную функцию φ и приравняем к нулю

$$\mathbb{E}\varphi(S) = \int\limits_{-\infty}^{+\infty} \varphi(s) f_S(s)\, ds = \int\limits_{0}^{+\infty} \varphi(s)\theta \exp(-\theta s) \frac{(s\theta)^{n-1}}{(n-1)!}\, ds = 0,$$

что равносильно

$$\int\limits_{0}^{+\infty} \exp(-\theta s)\varphi(s) s^{n-1}\, ds = 0.$$

Это равенство означает, что преобразование Лапласа функции $\varphi(s)s^{n-1}$ тождественно равно нулю, а это возможно только если (почти всюду) $\varphi(s)s^{n-1} = 0$, откуда следует $\varphi(s) = 0$ почти всюду. Полнота статистики S доказана.

Подберем теперь функцию статистики S такую, чтобы ее математическое ожидание было равно θ. Математическое ожидание от самой статистики равно $\mathbb{E}S = n/\theta$, а нам нужна θ. Отсюда конечно же не следует, что $\mathbb{E}(1/S) = \theta/n$, но попробовать вычислить $\mathbb{E}(1/S)$ стоит. Сделаем это:

$$\mathbb{E}\frac{1}{S} = \int\limits_{0}^{+\infty} \frac{1}{s}\theta \exp(-\theta s) \frac{(s\theta)^{n-1}}{(n-1)!}\, ds = \frac{\theta}{n-1}.$$

Остается взять $\varphi(S) = (n-1)/S$, тогда $\mathbb{E}\varphi(S) = \theta$. По теореме о полных достаточных статистиках получается, что $T(X) = (n-1)/\sum_{i=1}^{n} X_i$ – оптимальная оценка θ в модели $\mathrm{Exp}(\theta)$, $\theta > 0$.

7.4. Байесовское оценивание

Условия. Дана функция правдоподобия модели $f(x;\theta)$, $\theta \in \Theta$, априорное распределение $\pi(\theta)$ на множестве Θ и функция потерь $L(\theta, d)$. Требуется найти байесовскую оценку δ^* параметра θ, то есть

$$\delta^* = \arg\min_{\delta} \int R(\theta, \delta)\pi(\theta) d\theta,$$

где функция риска $R(\theta, \delta) = \mathbb{E}_\theta L(\theta, \delta(X))$.

Решение. 1. Найти апостериорное распределение

$$\pi(\theta|x) = \frac{f(x;\theta)\pi(\theta)}{\int f(x;\theta)\pi(\theta)\, d\theta}.$$

2. Вычислить средние потери относительно апостериорного распределения

$$\mathbb{E}(L(\theta,d)|x)=\int L(\theta,d)\pi(\theta|x)\,d\theta.$$

3. В качестве искомого решения δ^* взять такое d, которое минимизирует эти средние потери.

Замечания. Вычислять знаменатель в выражении для $\pi(\theta|x)$ необязательно, так как эта величина не зависит ни от θ, ни от d, и поэтому никак не скажется на оптимальном значении d. Также обратите внимание, что интегралы всюду берутся только по θ. По выборке интегралы будут браться, например, если мы захотим вычислить функцию риска $R(\theta,\delta)$, но в решении ее расчет не требуется.

Задача. Дана модель $\mathrm{Be}(\theta)$, $\theta\in[0,1]$. Априорное распределение на Θ считается непрерывным равномерным, то есть $\theta\in\mathrm{U}(0,1)$. Функция потерь задана в виде

$$L(\theta,d)=(d-\theta)^2.$$

Найти байесовскую оценку параметра θ.

Решение. Апостериорное распределение, с точностью до знаменателя

$$\pi(\theta|x)\cong\theta^S(1-\theta)^{n-S},\quad S=\sum_{i=1}^n x_i,\quad \theta\in[0,1].$$

Средние потери относительно этого апостериорного распределения

$$\mathbb{E}(L(\theta,d)|x)=\int_0^1(d-\theta)^2\cdot\theta^S(1-\theta)^{n-S}\,d\theta=$$

$$=d^2\cdot\mathrm{B}(S+1,n-S+1)-2d\cdot\mathrm{B}(S+2,n-S+1)+\mathrm{B}(S+3,n-S+1).$$

Здесь $\mathrm{B}(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$ – это бета-функция, а $\Gamma(m)$ – гамма-функция. Напомним, что гамма-функция при целых m равна $\Gamma(m)=(m-1)!$. Отсюда легко получить оптимальное значение d:

$$d^*=\frac{\mathrm{B}(S+2,n-S+1)}{\mathrm{B}(S+1,n-S+1)}=\frac{(S+1)!(n-S)!(n+1)!}{(n+2)!S!(n-S)!}=\frac{S+1}{n+2}\in[0,1].$$

Значит статистика

$$T(X)=\frac{\sum\limits_{i=1}^n X_i+1}{n+2}$$

является искомой байесовской оценкой параметра θ.

8. Интервальное оценивание

8.1. Введение

Определение. *γ-доверительным интервалом* скалярного параметра $\theta \in \Theta$ называется любой случайный интервал $(T_1(X), T_2(X))$, для которого

$$\forall \theta \in \Theta \;\; \mathbb{P}_\theta(T_1(X) < \theta < T_2(X)) \geq \gamma.$$

Величина γ называется *уровнем доверия*, или *коэффициентом доверия*. Статистики $T_1(X)$ и $T_2(X)$ называются *границами* доверительного интервала.

Есть несколько распространенных способов построения доверительных интервалов. Мы рассмотрим два из них: через центральную статистику и с помощью критериев отношения правдоподобия.

8.2. Построение доверительных интервалов с использованием центральной статистики

Определение. Пусть функция $G(X, \theta)$ обладает двумя свойствами: 1) ее распределение не зависит от θ; 2) при каждом значении $x \in \Omega$ функция $G(x, \theta)$ непрерывна и строго монотонна относительно θ, причем тип монотонности общий для всех $x \in \Omega$. Тогда такую функцию $G(X, \theta)$ называют *центральной статистикой*.

Пусть дана центральная статистика $G(X, \theta)$ и она обладает плотностью распределения $f_G(g)$. Тогда с одной стороны можно записать

$$\mathbb{P}_\theta(g_1 < G(X, \theta) < g_2) = \int_{g_1}^{g_2} f_G(g) dg = \gamma$$

для любых g_1 и g_2, для которых интеграл справа равен γ. С другой стороны, в силу свойств центральной статистики, неравенства

$$g_1 < G(X, \theta) < g_2$$

можно разрешить относительно θ и получить

$$T_1(g_1, g_2, X) < \theta < T_2(g_1, g_2, X).$$

Поэтому

$$\mathbb{P}_\theta(g_1 < G(X, \theta) < g_2) = \mathbb{P}_\theta(T_1(g_1, g_2, X) < \theta < T_2(g_1, g_2, X)) = \gamma.$$

Значит $(T_1,\, T_2)$ является γ-доверительным интервалом для θ. Степень свободы в выборе g_1 и g_2 может быть устранена по-разному. Часто ищут кратчайший γ-доверительный интервал, то есть для каждого $x \in \Omega$ решают оптимизационную задачу

$$|T_2(g_1, g_2, x) - T_1(g_1, g_2, x)| \to \min_{g_1, g_2}, \quad \int\limits_{g_1}^{g_2} f_G(g)dg = \gamma,$$

если решения не зависят от x. Другой возможный подход – оптимизировать математическое ожидание разности $T_2 - T_1$. В тех случаях, когда кратчайший интервал найти трудно (нет аналитического выражения или сложно с численной точки зрения), а плотность распределения $f_G(g)$ функции $G(X, \theta)$ является почти симметричной, то ищут *центральный интервал*, то есть такой интервал $(T_1(g_1, g_2, X), T_2(g_1, g_2, X))$, в котором g_1 и g_2 определяются из уравнений

$$\int\limits_{-\infty}^{g_1} f_G(g)dg = \int\limits_{g_2}^{+\infty} f_G(g)dg = \frac{1-\gamma}{2}.$$

Доверительный интервал, построенный для таких значений g_1 и g_2, не обязан быть кратчайшим, хотя в некоторых случаях, например, когда $G(X, \theta) \in \chi^2(m)$ для какого-либо m, при большом объеме выборки от кратчайшего интервала он отличается несильно, как следует из центральной предельной теоремы и «симметризации» распределения хи-квадрат с ростом объема выборки.

Итак, чтобы построить доверительный интервал, достаточно обзавестись какой-нибудь центральной статистикой. В нормальных моделях центральные статистики необходимо запомнить.

В модели $\mathrm{N}(\theta, \sigma^2)$ с известной дисперсией σ^2 центральная статистика для оценивания среднего θ выглядит так:

$$G(X, \theta) = \frac{\overline{X} - \theta}{\sqrt{\sigma^2/n}} \in \mathrm{N}(0, 1).$$

В модели $\mathrm{N}(\mu, \theta^2)$ с известным средним μ центральная статистика для оценивания дисперсии θ^2 выглядит так:

$$G(X, \theta) = \frac{1}{\theta^2} \sum_{k=1}^{n} (X_k - \mu)^2 \in \chi^2(n).$$

В модели $\mathrm{N}(\theta_1, \theta_2^2)$ с неизвестной дисперсией θ_2^2 центральная статистика для оценивания среднего θ_1 выглядит так:

$$G(X,\theta) = \frac{\overline{X} - \theta_1}{\sqrt{S_0^2/n}} \in \mathrm{St}(n-1), \quad S_0^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2.$$

Здесь, так как дисперсия неизвестна, пользуются ее несмещенной оценкой S_0^2.

В модели $\mathrm{N}(\theta_1, \theta_2^2)$ с неизвестным средним θ_1 центральная статистика для оценивания дисперсии θ_2^2 выглядит так:

$$G(X,\theta) = \frac{1}{\theta_2^2}\sum_{k=1}^{n}(X_k - \overline{X})^2 \in \chi^2(n-1).$$

Здесь, так как среднее неизвестно, пользуются его несмещенной оценкой $\overline{X}$.

В модели $\mathrm{U}(0,\theta)$ функция $G(X,\theta) = X_{(n)}/\theta \in \mathrm{Beta}(n,1)$ является центральной статистикой. В модели $\mathrm{Exp}(\theta)$ функция $G(X,\theta) = \theta\sum_{i=1}^{n} X_i \in \Gamma(n,1)$ является центральной статистикой.

Задача 1. Построить кратчайший γ-доверительный интервал для параметра θ модели $\mathrm{U}(0,\theta)$.

Решение. Вспоминаем, что функция $G(X,\theta) = X_{(n)}/\theta$ является центральной статистикой и ее распределение $\mathrm{Beta}(n,1)$ имеет плотность распределения $f_G(g) = ng^{n-1}$, $g \in [0,1]$. Тогда, с одной стороны,

$$\mathbb{P}_\theta(g_1 < G(X,\theta) < g_2) = \int_{g_1}^{g_2} f_G(g)dg = g_2^n - g_1^n = \gamma, \quad g_1, g_2 \in [0,1],$$

а, с другой стороны, $g_1 < G(X,\theta) < g_2$ равносильно $X_{(n)}/g_2 < \theta < X_{(n)}/g_1$, и поэтому

$$\mathbb{P}_\theta(X_{(n)}/g_2 < \theta < X_{(n)}/g_1) = g_2^n - g_1^n = \gamma.$$

Значит, в нашем случае $T_1(g_1,g_2,X) = X_{(n)}/g_2$ и $T_2(g_1,g_2,X) = X_{(n)}/g_1$. По условию необходимо найти кратчайший интервал (T_1,T_2), поэтому будем решать оптимизационную задачу

$$1/g_1 - 1/g_2 \to \min,$$

$$g_2^n - g_1^n = \gamma,\ g_1 \in [0,1],\ g_2 \in [0,1].$$

Сначала попробуем найти решение в открытой области $(g_1,g_2) \in (0,1)^2$. Для этого составим функцию Лагранжа:

$$\mathcal{L}(g_1,g_2,\lambda) = 1/g_1 - 1/g_2 + \lambda(g_2^n - g_1^n - \gamma)$$

и запишем необходимые условия оптимальности:

$$\frac{\partial \mathcal{L}}{\partial g_1} = -\frac{1}{g_1^2} - \lambda \cdot n \cdot g_1^{n-1} = 0,$$

$$\frac{\partial \mathcal{L}}{\partial g_2} = \frac{1}{g_2^2} + \lambda \cdot n \cdot g_2^{n-1} = 0,$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = g_2^n - g_1^n - \gamma = 0,$$

что равносильно

$$\lambda \cdot n \cdot g_1^{n+1} = -1,$$

$$\lambda \cdot n \cdot g_2^{n+1} = -1,$$

$$g_2^n - g_1^n - \gamma = 0.$$

При четных n из первых двух уравнений следует $g_1 = g_2$, что противоречит третьему уравнению. А если n нечетно, то либо снова получаем $g_1 = g_2$ (противоречие), либо $g_1 = -g_2$ (это противоречит $g_1, g_2 \in (0,1)$). Значит, ни при каких $g_1, g_2 \in (0,1)$ решения этой оптимизационной задачи не существует.

Будем искать решение на границе множества $[0,1]^2$. Множество

$$\{(g_1, g_2) : g_2^n - g_1^n = \gamma\}$$

пересекает множество $[0,1]^2$ всего в двух точках: $(0, \gamma^{1/n})$ и $((1 - -\gamma)^{1/n}, 1)$, причем в первой точке функционал принимает бесконечное значение, а во второй точке – конечное значение. Следовательно, пара $g_1 = (1-\gamma)^{1/n}$, $g_2 = 1$ образует решение задачи, и кратчайший γ-доверительный интервал для оценивания θ в модели $\mathrm{U}(0,\theta)$ выглядит следующим образом:

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

Задача 2. Пятикратное измерение некоторой физической величины W одним и тем же прибором дало результаты: 1.78, 1.81, 1.94, 1.86, 2.00. Тем же прибором было произведено пятикратное измерение

эталона, истинная величина которого равна одной единице измерения прибора. Результаты измерения эталона есть: 0.92, 0.78, 0.89, 0.82, 0.92. Предполагая, что ошибки измерений независимы и имеют одно и то же нормальное распределение, построить доверительный интервал для значений величины W при доверительной вероятности 0.95 (систематическая ошибка в обеих сериях измерений одинакова).

Решение. Пусть a – систематическая ошибка измерений. Тогда компоненты выборки X_i при измерении величины W распределены по нормальному закону $\mathrm{N}(W+a, \sigma^2)$, где σ^2 – дисперсия распределения, неизвестная по условию. Компоненты выборки Y_i при измерении эталона распределены по нормальному закону $\mathrm{N}(1+a, \sigma^2)$. Так как по условию задачи систематическая ошибка измерений неизвестна, избавимся от нее вычитанием одних измерений из других:

$$Z_i = X_i - Y_i \in \mathrm{N}(W-1, 2\sigma^2),\ i = 1, 2, \ldots, 5.$$

Теперь построим γ-доверительный интервал для величины $W-1$. Дисперсия $2\sigma^2$ неизвестна, значит воспользуемся центральной статистикой

$$G(Z;\theta) = \frac{\overline{Z} - \theta}{\sqrt{S_0^2/n}} \in \mathrm{St}(n-1),\ S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2,$$

где $\theta = W - 1$, $n = 5$. С одной стороны,

$$\mathbb{P}_\theta(g_1 < G(Z;\theta) < g_2) = \int_{g_1}^{g_2} f_G(g)dg = \gamma,\quad g_1, g_2 \in (-\infty, +\infty),$$

а, с другой стороны, $g_1 < G(Z;\theta) < g_2$ равносильно

$$\overline{Z} - g_2\sqrt{S_0^2/n} < \theta < \overline{Z} - g_1\sqrt{S_0^2/n}.$$

Получаем отсюда

$$\mathbb{P}_\theta(\overline{Z} - g_2\sqrt{S_0^2/n} < \theta < \overline{Z} - g_1\sqrt{S_0^2/n}) = \int_{g_1}^{g_2} f_G(g)dg = \gamma$$

и границы интервала $T_1 = \overline{Z} - g_2\sqrt{S_0^2/n}$ и $T_2 = \overline{Z} - g_1\sqrt{S_0^2/n}$. Будем искать кратчайший интервал (T_1, T_2), для этого поставим оптимизационную задачу

$$g_2 - g_1 \to \min,\quad \int_{g_1}^{g_2} f_G(g)dg = \gamma.$$

Решение этой задачи хорошо известно (см. приложение) и выражается квантилями распределения: $g_1 = \mathrm{St}_{\frac{1-\gamma}{2}}(n-1)$ и $g_2 = \mathrm{St}_{\frac{1+\gamma}{2}}(n-1)$. Значит искомый γ-доверительный интервал для θ:

$$\overline{Z} - \mathrm{St}_{\frac{1+\gamma}{2}}(n-1)\sqrt{S_0^2/n} < \theta < \overline{Z} - \mathrm{St}_{\frac{1-\gamma}{2}}(n-1)\sqrt{S_0^2/n},$$

а γ-доверительный интервал для W:

$$\overline{Z} + 1 - \mathrm{St}_{\frac{1+\gamma}{2}}(n-1)\sqrt{S_0^2/n} < W < \overline{Z} + 1 - \mathrm{St}_{\frac{1-\gamma}{2}}(n-1)\sqrt{S_0^2/n}.$$

В нашем случае

$$\begin{aligned}
&\overline{Z} = 1.0120,\\
&S_0^2 = 0.0076,\\
&\mathrm{St}_{\frac{1-\gamma}{2}}(n-1) = \mathrm{St}_{0.025}(4) = -2.7764,\\
&\mathrm{St}_{\frac{1+\gamma}{2}}(n-1) = \mathrm{St}_{0.975}(4) = 2.7764,
\end{aligned}$$

и значит $1.9040 < W < 2.1200$.

Утверждая это, мы ошибаемся с вероятностью 5%.

8.3. Построение доверительных интервалов с использованием критерия отношения правдоподобия

Пусть дана модель $F(x,\theta)$, $\theta \in \Theta$. Рассмотрим задачу проверки гипотез

$$\begin{aligned}
&H_1 : \theta = \theta_0,\\
&H_2 : \theta \neq \theta_0.
\end{aligned}$$

Допустим, что на уровне значимости α нам удалось построить равномерно наиболее мощный критерий, причем возможно не обязательно среди всех критериев, а только среди несмещенных критериев или критериев с какими-то другими ограничениями. Предположим даже, что эту задачу мы можем решить для произвольных θ_0 и все соответствующие критерии являются нерандомизированными. Получается, что для каждого фиксированного θ_0:

$$\mathbb{P}_{\theta_0}(X \in \Omega_1(\theta_0)) = 1 - \alpha,$$

где $\Omega_1(\theta_0)$ – область принятия первой гипотезы (вообще говоря, зависящая от θ_0), а α – уровень значимости. Логическое выражение под

вероятностью часто можно переписать относительно θ_0 эквивалентным образом и получить для каждого θ_0:

$$\mathbb{P}_{\theta_0}(X \in \Omega_1(\theta_0)) = \mathbb{P}_{\theta_0}(\theta_0 \in \Theta_1(X)) = 1 - \alpha,$$

где $\Theta_1(X)$ – некоторое случайное множество, которое, по определению, является $(1-\alpha)$-доверительным множеством для параметра θ.

Рассмотрим пример. Допустим требуется найти γ-доверительный интервал для математического ожидания θ в модели $\mathrm{N}(\theta, \sigma^2)$ с известной дисперсией σ^2. Для этого мы ставим задачу проверки гипотез

$$H_1 : \theta = \theta_0,$$
$$H_2 : \theta \neq \theta_0$$

с произвольным θ_0, и для каждого θ_0 ищем равномерно наиболее мощный критерий с уровнем значимости α. Так как модель $\mathrm{N}(\theta, \sigma^2)$ является моделью экспоненциального семейства, то такого критерия в классе всех критериев не существует, но он существует в классе несмещенных критериев. Можно получить (см. теорему о виде критерия из раздела 4.3), что в классе несмещенных критериев равномерно наиболее мощный критерий выглядит следующим образом: H_1 отклоняет тогда и только тогда, когда

$$\frac{|\overline{X} - \theta_0|}{\sigma}\sqrt{n} \geq \mathrm{N}_{\alpha/2}(0,1).$$

Гипотеза H_1 принимается тогда и только тогда, когда

$$\frac{|\overline{X} - \theta_0|}{\sigma}\sqrt{n} \leq \mathrm{N}_{\alpha/2}(0,1).$$

Итак, в обозначениях выше получается, что

$$\Omega_1(\theta_0) = \left\{ x : \ \frac{|\overline{x} - \theta_0|}{\sigma}\sqrt{n} \leq \mathrm{N}_{\alpha/2}(0,1) \right\},$$

и гипотеза принимается, если $x \in \Omega_1(\theta_0)$. Теперь в неравенстве выше выразим равносильным образом θ_0. Получится, что

$$\mathbb{P}_{\theta_0}\left(\frac{|\overline{X} - \theta_0|}{\sigma}\sqrt{n} \leq \mathrm{N}_{\alpha/2}(0,1) \right) =$$

$$= \mathbb{P}_{\theta_0}\left(\underbrace{\overline{x} - \frac{\sigma \mathrm{N}_{\alpha/2}(0,1)}{\sqrt{n}} \leq \theta_0 \leq \overline{x} + \frac{\sigma \mathrm{N}_{\alpha/2}(0,1)}{\sqrt{n}}}_{\theta_0 \in \Theta_1(x)} \right).$$

То, что мы получили в скобках, есть некоторый доверительный интервал. Мы знаем, что эта вероятность для любых θ_0 равна $1-\alpha$, поэтому это $(1-\alpha)$-доверительный интервал.

Итак, алгоритм для построения γ-доверительных оценок параметра получается следующий:

1) поставить и решить задачу поиска равномерного наиболее мощного критерия в задаче проверки простой гипотезы с двусторонней альтернативой на уровне значимости $\alpha = 1-\gamma$;

2) выразить условие принятия основной гипотезы относительно параметра основной гипотезы;

3) то, что получится – это $(1-\alpha)$-доверительный интервал (или множество, если θ – многомерный вектор) для искомого параметра.

Построенные таким образом γ-доверительные множества являются *наиболее точными доверительными множествами* и в некотором смысле оптимизируют размер множества[11,12].

Существуют и другие методы построения доверительных множеств, например методы построения асимптотических доверительных интервалов и множеств, методы построения доверительных интервалов с помощью точечных статистик и байесовский подход[13].

8.4. Построение доверительных интервалов с использованием точечных оценок

Пусть $T(X)$ – точечная оценка параметра θ, а $F_T(t;\theta)$ – ее функция распределения, которая для каждого значения t является непрерывной и монотонно[14] убывающей функцией параметра θ. Пусть $t = T(x)$ – реализация статистики $T(X)$ при $X = x$. Тогда если уравнения

$$F_T(t+0;\theta_2) = \frac{1-\gamma}{2},$$
$$F_T(t;\theta_1) = \frac{1+\gamma}{2}$$

имеют решения $\theta_1 < \theta_2$, то будем утверждать, что $\theta \in (\theta_1, \theta_2)$. Тогда $(\theta_1,\, \theta_2)$ является γ-доверительным интервалом параметра θ. Для абсолютно непрерывных статистик $T(X)$ выполнено $F_T(t+0;\theta) = F_T(t;\theta)$.

[11] Боровков А.А. Математическая статистика : учебник. 4-е изд., стер., 2010. С. 353–363.

[12] Леман Э. Проверка статистических гипотез. Москва : Наука. Главная редакция физико-математической литературы, 1979. С. 94–99.

[13] Боровков А.А. Математическая статистика : учебник. 4-е изд., стер., 2010. С. 285–296.

[14] Необязательно строго монотонно.

Замечание. Если функция распределения произвольной случайной величины ξ определяется как $F_\xi(x) = \mathbb{P}(\xi \le x)$, то есть с нестрогим неравенством, тогда вышеуказанные два уравнения на θ следует заменить уравнениями

$$F_T(t;\theta_2) = \frac{1-\gamma}{2},$$
$$F_T(t-0;\theta_1) = \frac{1+\gamma}{2}.$$

Например, в учебнике Г.И. Ивченко и Ю.И. Медведева всюду подразумевается именно эта версия определения функции распределения (с нестрогим неравенством). В курсах С.А. Гуза и О.Г. Горбачева неравенство строгое.

Задача. Найти γ-доверительный интервал для параметра θ модели $\mathrm{Be}(\theta)$.

Решение. Для оценивания воспользуемся оценкой $T(X) = \bar{X}$ параметра θ. Статистика $T(X)$ принимает значения k/n при $k = 0, \dots, n$, а ее функция распределения

$$F_T(0;\theta) = 0,\ F_T\left(k/n;\theta\right) = \sum_{j=0}^{k-1} C_n^j \cdot \theta^j (1-\theta)^{n-j},\quad k = 1, \dots, n.$$

При $k = 0$ эта функция является константой, а потому непрерывной и монотонно убывающей (нестрого) функцией θ. Для $k > 0$ эта функция является полиномом относительно θ и потому является непрерывной. Докажем ее монотонность относительно θ. Наиболее просто доказать ее монотонность можно, если расписать каждое слагаемое в отдельности и взять производную. Итак,

$$F_T\left(k/n;\theta\right) = C_n^0(1-\theta)^n + C_n^1\theta(1-\theta)^{n-1} + \dots + C_n^{k-1}\theta^{k-1}(1-\theta)^{n-k+1}.$$

Теперь по отдельности вычислим производную каждого слагаемого:

$$\frac{d}{d\theta} C_n^0\theta^0(1-\theta)^{n-0} = -n(1-\theta)^{n-1},$$
$$\frac{d}{d\theta} C_n^1\theta^1(1-\theta)^{n-1} = n(1-\theta)^{n-1} - n(n-1)\cdot\theta(1-\theta)^{n-2},$$
$$\frac{d}{d\theta} C_n^2\theta^2(1-\theta)^{n-2} = n(n-1)\cdot\theta(1-\theta)^{n-2} - \frac{n(n-1)(n-2)}{2}\theta^2(1-\theta)^{n-3},$$
$$\dots$$

Очевидно, что при сложении производных почти все слагаемые «по цепочке» сократятся, и в итоге останется

$$\frac{d}{d\theta}F_T\left(k/n;\theta\right) = -C_n^{k-1}(n-k+1)\cdot\theta^{k-1}(1-\theta)^{n-k} < 0$$

для всех $k = 1,\ldots,n$. Итак, монотонность по θ для всех $k = 1,\ldots,n$ доказана. Более того, легко видеть, что $F_T(k/n;\theta)$ принимает все значения из отрезка $[0,1]$. Если $t = k/n$ для $k = 1,\ldots,n-1$, то остается решить уравнения

$$F_T((k+1)/n;\theta_2) = \frac{1-\gamma}{2},$$
$$F_T(k/n;\theta_1) = \frac{1+\gamma}{2}$$

и найти $\theta_1 < \theta_2$; решение этих уравнений всегда существует, так как при вариации θ_1 и θ_2 на интервале $(0,1)$ левые части этих уравнений принимают непрерывное множество значений на интервале $(0,1)$.

Что касается крайних случаев ($k = 0,n$), то при $k = 0$ второе уравнение не имеет решений; тогда в качестве левой границы доверительного интервала можно выбрать минимально возможное значение: $\theta_1 = 0$. Правая же граница существует и находится из первого уравнения. Точно так же при $k = n$ первое уравнение не имеет решений, поэтому выберем максимально возможное значение[15]: $\theta_2 = 1$. Причина этих «заплаток» находится в том, что статистика T дискретная, принимает конечное число значений, и при крайних значениях статистики левая и правая границы для θ находятся «в бесконечности», но так как параметрическое множество $\Theta = [0,1]$ ограничено слева и справа конечными значениями, то они и выбираются в качестве границ в соответствующих случаях.

Теперь несколько слов о том, как на практике вычисляются искомые θ_1 и θ_2. С одной стороны, так как функция распределения статистики T строго монотонно зависит от θ, получить θ_1 и θ_2 можно и методом Ньютона, он сойдется из любого начального приближения. Однако можно поступить хитрее, если заметить, что функция распределения статистики связана с квантилями бета-распределения:

$$F_T\left(k/n;\theta\right) = \sum_{j=0}^{k-1} C_n^j\cdot\theta^j(1-\theta)^{n-j} = 1-\mathrm{Beta}_\theta(k,\,n-k+1),\ k = 1,\ldots,n.$$

[15] В справедливости этих допущений для $k = 0$ и $k = n$ предлагается убедиться самостоятельно.

Поэтому θ_1 и θ_2 можно рассчитать просто по формулам

$$\theta_1 = F_{\mathrm{Beta}(k,n-k+1)}\left(\frac{1-\gamma}{2}\right),\ \theta_2 = F_{\mathrm{Beta}(k+1,n-k)}\left(\frac{1+\gamma}{2}\right).$$

Отметим наконец что искомые значения θ_1 и θ_2 являются функциями выборки X. Для них будет справедливо неравенство

$$\forall\theta \in [0,1]\ \ \mathbb{P}_\theta(\theta_1(X) < \theta < \theta_2(X)) \geq \gamma.$$

8.5. Приложение

Решим оптимизационную задачу:

$$g_2 - g_1 \to \min,\ \int\limits_{g_1}^{g_2} f_G(g)dg = \gamma,$$

где $f_G(g)$ – функция плотности распределения Стьюдента $\mathrm{St}(n-1)$. Введем функцию Лагранжа:

$$\mathcal{L}(g_1, g_2, \lambda) = g_2 - g_1 + \lambda\left(\int\limits_{g_1}^{g_2} f_G(g)dg - \gamma\right).$$

Необходимые условия оптимальности тогда запишутся так:

$$\frac{\partial\mathcal{L}}{\partial g_1} = -1 - \lambda\cdot f_G(g_1) = 0,$$

$$\frac{\partial\mathcal{L}}{\partial g_2} = 1 + \lambda\cdot f_G(g_2) = 0,$$

$$\frac{\partial\mathcal{L}}{\partial\lambda} = \int\limits_{g_1}^{g_2} f_G(g)dg - \gamma = 0.$$

В силу симметричности функции плотности распределения Стьюдента $f_G(g)$ относительно нуля из первых двух уравнений следует, что либо $g_1 = g_2$, но это противоречит третьему уравнению, либо $g_2 = -g_1$. Но тогда из третьего уравнения следует, что $g_1 = \mathrm{St}_{\frac{1-\gamma}{2}}(n-1)$ и $g_2 = \mathrm{St}_{\frac{1+\gamma}{2}}(n-1)$.

Теперь проверим достаточные условия оптимальности. Для этого вычислим матрицу вторых производных функции Лагранжа. Так как

$$\lambda = -1/f_G(g_1) = -1/f_G(g_2),$$

то матрица вторых производных

$$\frac{\partial^2 L}{\partial w^2} = \begin{bmatrix} \dfrac{f_G^{'}(g_1)}{f_G(g_1)} & 0 \\ \\ 0 & -\dfrac{f_G^{'}(g_2)}{f_G(g_2)} \end{bmatrix},$$

где вектор переменных $w = [g_1, g_2]$. Если эта матрица является положительно определенной на решении, то решение является оптимальным. Функция плотности распределения Стьюдента $\mathrm{St}(n-1)$ выражается по формуле

$$f_G(g) = C \cdot \left(1 + \frac{g^2}{n-1}\right)^{-n/2},$$

где C не зависит от g. Тогда

$$\frac{f_G^{'}(g)}{f_G(g)} = \frac{d}{dg} \ln f_G(g) = -\frac{n}{2} \frac{2g/(n-1)}{1 + g^2/(n-1)}.$$

Далее замечаем, что $g_1 < 0$ и $g_2 > 0$, значит

$$\frac{f_G^{'}(g_1)}{f_G(g_1)} > 0, \quad -\frac{f_G^{'}(g_2)}{f_G(g_2)} > 0,$$

значит все верхние левые угловые минора положительные, что и доказывает оптимальность найденных решений.

Замечание. Для доказательства положительной определенности матрицы вторых производных достаточно было знать лишь то, что плотность распределения $f_G(g)$ монотонно возрастает при $g < 0$ и монотонно убывает при $g > 0$. Таким образом, решение оптимизационной задачи является общим для всех симметричных относительно $g = 0$ функций плотности $f_G(g)$ с таким свойством монотонности, например, для распределения $\mathrm{N}(0, 1)$.

9. Метод наименьших квадратов

Есть две скалярные величины, x и y, которые связаны неизвестным соотношением $y = y(x)$. Эти величины могут вместе или по

отдельности быть случайными или неслучайными. Задача состоит в том, чтобы определить связь $y = y(x)$. Для этого сначала ограничим поиск на следующем классе функций

$$y = f(x, \beta) = \sum_{j=1}^{m} \beta_j g_j(x). \tag{4}$$

Здесь $\beta_1, \ldots, \beta_m$ – неизвестные параметры, а $g_j(x)$ – известные функции, которые в общем случае могут быть произвольными. Так как выражение (4) линейно по параметрам β_j, то модель $f(x, \beta)$ называют *линейной моделью*, молчаливо подразумевая, что линейность понимается относительно параметров β_j. Функции $g_j(x)$ могут нелинейно зависеть от x; например, это могут быть полиномы: $g_1(x) = 1$, $g_2(x) = x$, $g_3(x) = x^2$ и т.д. В случаях, когда $g_j(x) = x^{j-1}$, такую модель называют *полиномиальной (параболической) моделью* (здесь слово «полиномиальная» относится уже к функциям $g_j(x)$).

Теперь, когда модель $f(x, \beta)$ введена, задача поиска зависимости $y = y(x)$ свелась к поиску конечного числа неизвестных параметров $\beta_1, \ldots, \beta_m$. Чтобы их определить, нужны *данные*, то есть пары (x_i, y_i), $i = 1, \ldots, n$, значений исследуемых переменных. Так как модель $f(x, \beta)$ не обязана точно определять зависимость $y = y(x)$ для каких-то значений β_j, мы не можем довольствоваться m парами (x_i, y_i), чтобы потом решить m уравнений с m неизвестными. Поэтому количество данных n обычно берут много большим, чем число неизвестных m. Так как систему уравнений решить мы не можем, параметры β_j подбираются так, чтобы они минимизировали некоторый функционал. Рассмотрим квадратичный функционал

$$J = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \beta_j g_j(x_i) \right)^2$$

и будем его минимизировать относительно β_j. Решение задачи безусловной оптимизации $J \to \min$ выражается аналитически по формуле

$$\beta^* = (X^T X)^{-1} X^T Y,$$

где $Y = [y_1, y_2, \ldots, y_n]$, $X_{ij} = g_j(x_i)$ и $\beta^* = [\beta_1^*, \beta_2^*, \ldots, \beta_m^*]$ – вектор искомых параметров. Конечно, здесь предполагается, что матрица $X^T X$ имеет обратную, хотя даже если обратная не существует, то все равно есть простой способ найти решение[16]. Далее будет предполагаться, что

[16] Рао С.Р. Линейные статистические методы и их применение. Москва : Наука. Главная редакция физико-математической литературы, 1968.

X^TX не вырождена. Кстати говоря, если было бы $n < m$, то матрица X^TX всегда была бы вырожденной.

Метод оценивания неизвестной зависимости $y = y(x)$, изложенный выше, называется *методом наименьших квадратов*. Найденная оценка β^* называется *оценкой по методу наименьших квадратов*, или сокращенно *МНК-оценкой*. Есть очень простое правило запоминания формулы для β^*. Сначала в матрично-векторном виде надо записать уравнение модели:

$$Y = X\beta,$$

затем домножить слева обе части этого уравнения на X^T:

$$X^TY = X^TX\beta,$$

откуда уже следует формула для β^*. Конечно, эти выкладки не доказывают то, что полученная формула для β минимизирует функционал J, но они помогают вспомнить формулу.

Матрица X называется иногда *матрицей плана*. В нашем изложении она определяется через функции $g_j(x)$, что естественно, ибо это параметризация функции $y = y(x)$ через комбинацию известных зависимостей $g_j = g_j(x)$. Но бывает, что в задаче вместе с данными (x_i, y_i) дают сразу матрицу X_{ij}; такой подход в изложении представлен в учебнике Ивченко Г.И. и Медведева Ю.И., там она обозначается буквой Z и связана с нашим X соотношением $Z = X^T$.

Функционал J представляет собой меру разности между эмпирическими данными y_i и гипотетическими $f(x_i, \beta)$. Невязку для каждого i-го элемента данных обозначим

$$\varepsilon_i(\beta) = y_i - f(x_i, \beta).$$

Природа ε может быть как неслучайной (несоответствие данных модели при неслучайных x и y), так и случайной (когда x или y случайные). В математической статистике обычно предполагается, что есть случайная составляющая ошибки ε_i, причем

$$\mathbb{E}\varepsilon_i = 0,\ \mathbb{D}\varepsilon_i = \sigma^2,\ i = 1, \ldots, n,$$
$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j,$$

то есть средние ошибок равны нулю, они имеют одинаковую дисперсию и некоррелированы. Если дополнительно предполагается, что ε_i образуют нормальный случайный вектор, то тогда модель (с ошибками) называется *нормальной*.

Теперь перейдем к свойствам МНК-оценок. Введем класс несмещенных линейных оценок вектора β:

$$K = \{\tilde{\beta} : \tilde{\beta} = AY,\ \forall \beta \in \mathbb{R}^m\ \mathbb{E}_\beta \tilde{\beta} = \beta\}$$

и более широкий класс несмещенных оценок (необязательно линейных):

$$\overline{K} = \{\tilde{\beta} : \forall \beta \in \mathbb{R}^m\ \mathbb{E}_\beta \tilde{\beta} = \beta\}.$$

Теорема Гаусса–Маркова.

$$\forall \beta \in \mathbb{R}^m\ \forall \tilde{\beta} \in K\ \forall j \in \{1, \ldots, m\}\ \mathbb{D}_\beta \beta_j^* \le \mathbb{D}_\beta \tilde{\beta}_j.$$

Если модель нормальная, то верно более сильное утверждение:

$$\forall \beta \in \mathbb{R}^m\ \forall \tilde{\beta} \in \overline{K}\ \forall j \in \{1, \ldots, m\}\ \mathbb{D}_\beta \beta_j^* \le \mathbb{D}_\beta \tilde{\beta}_j.$$

Замечание. Словами это значит, что в классе K несмещенных линейных оценок вектора β, метод наименьших квадратов дает оценку β^*, все компоненты которой имеют дисперсии не больше, чем дисперсии соответствующих компонент любой другой оценки из класса K. Здесь есть важное словосочетание: *линейных оценок* (линейных по вектору Y). В произвольной модели может оказаться, что есть несмещенная, но нелинейная относительно Y оценка параметров β, с меньшими дисперсиями. Но оказывается, что если модель является нормальной, то МНК-оценка является эффективной оценкой вектора β среди всех несмещенных (не только линейных по Y) оценок. Кроме того, отмечу, что оптимальная оценка (в классе несмещенных) является функцией достаточной статистики, но оптимальная оценка в классе несмещенных линейных оценок этим свойством может не обладать.

Перечислим еще ряд фактов. Введем обозначение $\varepsilon^* = \varepsilon(\beta^*)$.
1) $\tilde{\sigma}^2 = \frac{\varepsilon^{*T}\varepsilon^*}{n-m}$ является несмещенной оценкой дисперсии σ^2,
а если дополнительно предположить нормальность модели, то
2) $\beta^* = \hat{\beta}$, то есть МНК-оценка является оценкой максимального правдоподобия,
3) $\hat{\sigma}^2 = \frac{\varepsilon^{*T}\varepsilon^*}{n}$ является оценкой максимального правдоподобия дисперсии σ^2,
4) $\frac{\varepsilon^{*T}\varepsilon^*}{\sigma^2} \in \chi^2(n-m)$ – центральная статистика для σ^2,
5) $\frac{\beta_j^* - \beta_j}{\sqrt{a_{jj}\tilde{\sigma}^2}} \in \mathrm{St}(n-m)$ – центральная статистика для β_j,
где a_{jj} – это j-й диагональный элемент матрицы $X^T X$. Величина $\varepsilon^{*T}\varepsilon^*$ носит название *остаточной суммой квадратов*, это просто значение функционала J на МНК-оценке.

Задача 1. Результаты измерений значений нормально распределенной случайной величины Y при десяти значениях 1, 2, 3, ..., 10 неслучайной величины x есть 2.2, 3.1, 4.1, 5.0, 5.8, 6.9, 7.8, 9.0, 10.2, 11.1 соответственно. Построить уравнение линейной регрессии. Проверить гипотезу о целесообразности уточнения полученного уравнения в случае известной дисперсии случайной величины Y: $\mathbb{D}Y = 0.02$. Уточняет ли член $0.01x^2$ полученное уравнение регрессии?

Решение. Рассмотрим модель

$$y = \beta_1 + \beta_2 x,$$

где β_1 и β_2 – неизвестные параметры, а x и y – скалярные переменные, среди которых x неслучаен, а y – случайная величина, причем имеющая нормальное распределение. Будем считать, что y_i – независимые величины. Здесь также $g_1(x) = 1$, $g_2(x) = x$. Чтобы оценить β по МНК, сначала выпишем матрицу X:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & 10 \end{bmatrix}.$$

Далее, вычислим матрицу $X^T X$ и ее обратную:

$$X^T X = \begin{bmatrix} 10 & 55 \\ 55 & 385 \end{bmatrix}, \quad (X^T X)^{-1} = \begin{bmatrix} 7/15 & -1/15 \\ -1/15 & 2/165 \end{bmatrix}.$$

Выпишем вектор измерений Y:

$$Y = [2.2,\ 3.1,\ 4.1,\ 5.0,\ 5.8,\ 6.9,\ 7.8,\ 9.0,\ 10.2,\ 11.1]^T.$$

В итоге получаем МНК-оценку параметров:

$$\beta^* = (X^T X)^{-1} X^T Y = [1.0600,\ 0.9927]^T.$$

Значит, искомое уравнение линейной регрессии

$$y = 1.0600 + 0.9927 \cdot x.$$

Остаточная сумма квадратов равна ${\varepsilon^*}^T \varepsilon^* = 0.1916$.

Теперь поставим вопрос о целесообразности уточнения модели, если известна точная дисперсия $\sigma^2 = \mathbb{D}Y = 0.02$. Пусть предположения о том, что модель нормальная и линейная, истинны и сомнению

не подвергаются. Теперь если число неизвестных параметров $m = 2$ (только это мы проверяем), то мы ожидаем, что

$$T_{\chi^2} = \frac{\varepsilon^{*T}\varepsilon^*}{\sigma^2} \in \chi^2(n-2).$$

На уровне значимости α критическая область

$$\Omega_2 = \{t : t < \chi^2_{\alpha/2}(8)\} \cup \{t : t > \chi^2_{1-\alpha/2}(8)\}.$$

Пусть $\alpha = 0.05$, тогда $\chi^2_{\alpha/2}(8) = 2.18$, $\chi^2_{1-\alpha/2}(8) = 17.53$, а значение статистики T_{χ^2} равно $t = 9.58$ и не попадает в критическую область. Значит данные не противоречат гипотезе о том, что все факторы учтены. Может быть, если увеличить выборку, то гипотезу придется отклонить, следовательно тогда будет целесообразно уточнить модель, увеличив число факторов. Еще тут нужно обратить внимание на то, что в критическую область мы отнесли оба хвоста распределения хи-квадрат (а не только правый хвост распределения, как в классическом методе хи-квадрат Пирсона). Критическая область в произвольной задаче не определяется хвостами распределения статистики критерия, она строится по значениям статистики критерия как меры отклонения эмпирических данных от гипотетических. В нашем случае в критическую область добавлены как малые, так и большие значения статистики критерия, потому что эти значения говорят либо о слишком малом отклонении от данных (вследствие переобучения), либо о слишком большом отклонении от данных (из-за недообучения) соответственно.

Теперь добавим слагаемое $0.01x^2$ в полученное уравнение регрессии:

$$y = 1.0600 + 0.9927 \cdot x + 0.01 \cdot x^2.$$

В этом случае функционал J станет равным $2.5609 > 0.1916$, значит это слагаемое не уточняет уравнение регрессии.

Это не удивительно, так как набор параметров

$$\beta = [1.0600,\ 0.9927,\ 0.01]$$

не является МНК-оценкой в модели

$$y = \beta_1 + \beta_2 x + \beta_3 x^2.$$

Чтобы получить МНК-оценку в этой новой модели, необходимо соста-

вить новую матрицу X:

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ \vdots & \vdots & \vdots \\ 1 & 10 & 100 \end{bmatrix},$$

рассчитать матрицы $X^T X$ и $(X^T X)^{-1}$:

$$X^T X = \begin{bmatrix} 10 & 55 & 385 \\ 55 & 385 & 3025 \\ 385 & 3025 & 25333 \end{bmatrix},$$

$$(X^T X)^{-1} = \begin{bmatrix} 83/60 & -21/40 & 1/24 \\ -21/40 & 637/2640 & -1/48 \\ 1/24 & -1/48 & 1/528 \end{bmatrix}$$

и получить новую оценку

$$\beta^* = (X^T X)^{-1} X^T Y = [1.4017,\, 0.8219,\, 0.0155]^T.$$

Остаточная сумма квадратов равна ${\varepsilon^*}^T \varepsilon^* = 0.0643$, что приблизительно в три раза меньше, чем в модели, линейной по x.

Важное замечание. Может показаться, что чем больше параметров берешь в выражении (4), тем лучше, потому что функционал J уменьшается и данные лучше приближаются более сложной моделью. Но неограниченно увеличивать число параметров неразумно, потому что можно столкнуться с *переобучением модели* – ситуацией, когда модель отлично приближает данные, которые были ей поданы для оптимизации параметров, но плохо приближает новые данные. Поэтому число параметров не должно быть слишком большим, как и слишком малым.

Численная оптимизация. Мы видели, что МНК-оценка для линейных моделей определяется по аналитической формуле. Иногда, впрочем, эту оценку ищут через численное решение оптимизационной задачи. Так поступают, например, тогда, когда число параметров в задаче большое или матрица $X^T X$ плохо обусловлена и аккуратное вычисление обратной матрицы затруднено. Численное решение также применяется и в случаях, когда модель нелинейная относительно неизвестных параметров. У процедуры численной оптимизации есть и свои особенности: следует грамотно выставить критерии ее остановки. Неверно было бы думать, что чем меньше оптимизируемый функционал, то

тем лучше, опять же из-за проблем с возможным переобучением модели. Чтобы этого не произошло, рекомендуется в процессе оптимизации следить за значениями функционала как на обучающей выборке, так и на валидирующей (вспомогательной, не участвующей в оптимизации) выборке, и останавливать процедуру, когда между этими значениями достигается значительный разлад. Особенности выстраивания оптимизационных процедур в статистическом обучении возникают из-за того, что здесь мы имеем дело не просто с задачами оптимизации, а с задачами построения хорошего аппроксиматора данных. Аппроксиматор должен не исходные данные хорошо аппроксимировать, а вообще все данные. Отсюда и особенности оптимизации.

10. Вопросы на понимание

Вопросы по первому заданию:

1. Что такое распределение хи-квадрат?
2. Что такое квантиль распределения?
3. Опишите детально процедуру проверки монетки и кубика на симметричность.
4. Какое распределение имеют порядковые статистики из равномерного на отрезке $[0, 1]$ распределения? Как это распределение получить?
5. Зачем при использовании критерия хи-квадрат Пирсона проверять то, что объем выборки и частоты исходов достаточно большие?
6. Объяснить, почему у критериев Неймана–Пирсона с ростом вероятности ошибки первого рода уменьшается вероятность ошибки второго рода.
7. Пусть в задаче с двумя простыми гипотезами H_1 и H_2 при первой гипотезе $\mathbb{P}_{H_1}(L_2(X)/L_1(X) = c) = 0$. Верно ли, что эта вероятность равна нулю и при второй гипотезе?
8. Привести пример, когда даны две простые гипотезы с непрерывными распределениями, а отношение правдоподобия имеет дискретное распределение.
9. Как показать, что критерий Неймана–Пирсона является несмещенным?
10. Как выглядит критерий Неймана–Пирсона в случае, когда носители распределений не совпадают?
11. Что делать в случае, если реализация выборки находится за пределами носителей и первой, и второй гипотезы?

12. Привести пример двух простых гипотез, когда наиболее мощный критерий не единственен.
13. Привести пример двух гипотез с непрерывными распределениями, когда для наиболее мощного критерия необходима рандомизация.
14. В каких случаях не существует равномерно наиболее мощный критерий среди всех критериев фиксированного уровня значимости?
15. Как можно использовать критерий Неймана–Пирсона, чтобы найти равномерно наиболее мощный критерий?
16. Какую оптимизационную задачу решает последовательный критерий Вальда?
17. Можно ли говорить, что критерий Вальда всегда делает меньше шагов, чем требуется при применении критерия Неймана–Пирсона для получения тех же вероятностей ошибок первого и второго рода?
18. Как доказать, что распределение статистики Колмогорова–Смирнова не зависит от истинной функции распределения?
19. Как эмпирическая функция распределения связана с броуновским мостом?

Вопросы по второму заданию:

1. Докажите устно, что выборочная дисперсия – состоятельная оценка истинной дисперсии.
2. В каких случаях эффективные оценки существуют? В каких случаях они не существуют?
3. Привести пример, когда эффективная оценка не существует.
4. Привести пример, когда оптимальная оценка не существует.
5. Привести пример, когда оценка максимального правдоподобия существует не в каждой точке.
6. Как связаны наиболее мощный критерий, байесовский критерий и минимаксный критерий?
7. В каком случае оценка метода наименьших квадратов совпадает с оценкой максимального правдоподобия?
8. Что значат слова «линейная», «полиномиальная» и «нормальная» для регрессии? К каким математическим объектам эти слова относятся?
9. Как интервально оценить дисперсию ошибок в нормальной модели в методе наименьших квадратов?
10. Как ищется оптимальная оценка по полной достаточной статистике?
11. Привести пример, когда неравенство Рао–Крамера не выполнено.

12. Как связано интервальное оценивание параметров распределения с задачей проверки гипотез?
13. Предложите и обоснуйте оптимальный способ оценки вероятности выпадения монеты на решку.
14. Какие у нас есть инструменты для исследования на состоятельность?
15. Как связаны байесовские и минимаксные решающие правила?
16. Бывают ли рандомизированные байесовские решающие правила? А минимаксные?
17. Что такое апостериорные вероятности гипотез?
18. Как связаны байесовские решающие правила с принципом максимума апостериорной вероятности?
19. Чем отличаются друг от друга тесты Стьюдента и Фишера (Снедекора)?

11. Вопросы для диктантов

Вопросы по первому заданию:

1. Простая выборка, статистика, порядковые статистики и их плотности. Эмпирическая функция распределения: определение, мат. ожидание, дисперсия, поточечная сходимость к истинной функции распределения, нормальное приближение. Статистика Колмогорова–Смирнова. Теорема Гливенко. Теорема Колмогорова.

2. Выборочное пространство, простые и сложные статистические гипотезы, статистический критерий, критическая область, вероятности ошибок 1-го и 2-го рода, уровень значимости, квантиль.

3. Критерий согласия Колмогорова. Критерий хи-квадрат для простой гипотезы. Критерий хи-квадрат для сложной параметрической гипотезы. Критерий независимости хи-квадрат. Критерий однородности хи-квадрат. Критерий инверсий. p-value.

4. Функция правдоподобия, функция отношения правдоподобия. Рандомизированный критерий, критическая функция, мощность критерия. Лемма Неймана–Пирсона (для наиболее мощного критерия), критерий Неймана–Пирсона. Равномерно наиболее мощный критерий, несмещенность критерия, состоятельность критерия. Теорема о равномерно наиболее мощном несмещенном критерии.

Вопросы по второму заданию:

1. Функция риска. Байесовский риск. Байесовское решающее правило. Теорема о виде байесовского решающего правила в случае несколь-

ких простых гипотез. Минимаксное решающее правило. Связь байесовского и минимаксного решающих правил. Лемма Неймана–Пирсона (для наиболее мощного критерия, критерия Байеса и минимаксного критерия).

2. Несмещенность, состоятельность, оптимальность точечных оценок. Достаточные статистики. Критерий факторизации. Полные достаточные статистики. Теорема Рао–Блекуэлла–Колмогорова. Теорема о полных достаточных статистиках. Теорема и неравенство Рао–Крамера. Эффективные оценки. Оценка максимального правдоподобия. Байесовская оценка и алгоритм ее построения.

3. Доверительный интервал. Построение доверительного интервала с помощью центральной статистики. Построение доверительного интервала с помощью точечной оценки. Интервальное оценивание параметров нормальной модели.

4. Метод наименьших квадратов. Модели линейной, параболической и нормальной регрессии. Теорема Гаусса–Маркова.

12. Задачи повышенной сложности

12.1. Эмпирическая функция распределения

Задача 1*. Пусть $U_1, \ldots, U_n$ – простая выборка из равномерного распределения $\mathrm{U}(0,1)$, $G_n(u)$ – эмпирическая функция распределения. Доказать, что

$$n \int\limits_0^1 \left(G_n(s) - s\right)^2 ds = \frac{1}{12n} + \sum_{k=1}^{n} \left(U_{(k)} - \frac{2k-1}{2n}\right)^2,$$

$$\int\limits_0^1 \left(G_n(s) - s\right)^2 ds \leq \frac{1}{3}.$$

Задача 2. Доказать, что для выборки с непрерывной функцией распределения $F(y)$ и эмпирической функцией распределения $F_n(y)$ при любом $t \in [0,1]$ справедливо равенство

$$\mathbb{P}\left(\sup_{y\in\mathbb{R}} |F_n(y) - F(y)| > t\right) = \mathbb{P}\left(\sup_{0\leq y\leq 1} |G_n(y) - y| > t\right),$$

где $G_n(y)$ – эмпирическая функция распределения, построенная по выборке из равномерного распределения $\mathrm{U}(0,1)$.

Задача 3. Пусть $X = (X_1, \ldots, X_n)$ – выборка из непрерывного распределения F над $\mathbb{R}$, F_n – эмпирическая функция распределения и

$$C_n(F) = \int (F_n(x) - F(x))^2 dF(x).$$

Докажите, что распределение $C_n(F)$ не зависит от F.

12.2. Порядковые статистики

Задача 4*. Пусть случайная величина X имеет распределение $\mathrm{U}(a, b)$. Найти совместное распределение порядковых статистик $X_{(i)}$ и $X_{(j)}$. Вычислить их математические ожидания, дисперсии и коэффициент корреляции.

Задача 5*. Пусть дана выборка $X_1, X_2, \ldots, X_n$ из показательного распределения с параметром α. Доказать, что случайные величины $X_{(1)}$, $X_{(2)} - X_{(1)}$, $\ldots$, $X_{(n)} - X_{(n-1)}$ независимы. Выяснить распределение статистики $X_{(1)}$ и распределение разности соседних порядковых статистик $X_{(k+1)}$ и $X_{(k)}$. Доказать справедливость равенства $\mathbb{E}X_{(k)} = \alpha^{-1}\left((n-k+1)^{-1} + \cdots + n^{-1}\right)$.

Задача 6*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из распределения с плотностью $f(x) = (1/2)\cos x$, $|x| \leq \pi/2$. Вывести формулу

$$\mathbb{E}X_{(n)} = \frac{\pi}{2}(1 - 2^{-2n+1}C_{2n}^n).$$

Доказать, что при $n \to \infty$:

$$\mathbb{E}X_{(n)} = \frac{\pi}{2} - \sqrt{\frac{\pi}{n}}\left(1 + O\left(\frac{1}{n}\right)\right),$$

$$\mathbb{P}\left(\frac{\sqrt{n}}{2}\left(X_{(n)} - \frac{\pi}{2}\right) \leq x\right) \to e^{-x^2},\ x < 0.$$

Задача 7. Пусть дана выборка из распределения F такого, что

$$\lim_{y\to\infty} y(1 - F(y) + F(-y)) = 0.$$

Доказать, что $X_{(1)}/n \xrightarrow{\mathbb{P}} 0$ и $X_{(n)}/n \xrightarrow{\mathbb{P}} 0$ при $n \to \infty$.

Задача 8. Предполагается выполнить $n+1$ независимых измерений случайной величины X, имеющей непрерывную функцию распределения $F(x)$. Найти а) априорную вероятность того, что значение X_{n+1}, полученное в $(n+1)$-м измерении, окажется больше, чем k-е по величине значение X, полученное в предшествующих n измерениях;

б) априорную вероятность того, что значение X_{n+1} окажется в k-м блоке выборки, то есть вероятность $\mathbb{P}(X_{(k)} < X_{n+1} < X_{(k+1)})$. Зависит ли она от номера блока?

Задача 9. Пусть $X_1, \ldots, X_{2n-1}$ – простая выборка случайной величины, имеющей непрерывную функцию распределения $F(x)$, а $W_1, \ldots, W_{2n}$ – ее доли, $W_i = F(X_{(i)}) - F(X_{(i-1)})$. Найти распределение вероятностей суммы $S = \sum\limits_{k=1}^{n} W_{2k}$, то есть суммы четных долей.

Задача 10*. Пусть дана выборка $X = (X_1, \ldots, X_n)$ из распределения с плотностью $\theta^{-1} e^{(a-x)/\theta} I_{(a,\infty)}(x)$, где параметры $a \in \mathbb{R}$ и $\theta > 0$. Пусть $X_{(1)} \leq \cdots \leq X_{(n)}$ – порядковые статистики, $X_{(0)} = 0$, и $Z_i = X_{(i)} - X_{(i-1)}$, $i = 1, \ldots, n$. Покажите, что $Z_1, \ldots, Z_n$ независимые, $2(n-i+1)Z_i/\theta$ имеет распределение $\chi^2(2)$ и

$$2\left[\sum_{i=1}^{r} X_{(i)} + (n-r)X_{(r)} - na\right]/\theta$$

имеет распределение $\chi^2(2r)$, $r = 1, \ldots, n$.

Задача 11*. Пусть $X = (X_1, \ldots, X_n)$ выборка из непрерывного распределения F над $\mathbb{R}$, F_n – эмпирическая функция распределения, $D_n^+ = \sup\limits_{x \in \mathbb{R}} [F_n(x) - F(x)]$, $D_n^- = \sup\limits_{x \in \mathbb{R}} [F(x) - F_n(x)]$. Докажите, что D_n^+ и D_n^- имеют одинаковое распределение, и что для любого $t \in (0,1)$:

$$\mathbb{P}(D_n^+ \leq t) = n! \prod_{i=1}^{n} \int\limits_{\max\{0, \frac{n-i+1}{n} - t\}}^{u_{n-i+2}} du_1 \ldots du_n.$$

12.3. Статистические критерии

Задача 12. Для проверки гипотезы $H_1 : \mathrm{U}(0,1)$ используется статистика

$$T = \int\limits_0^1 \left(F_n(y) - y\right)^2 dy,$$

где $F_n(y)$ – эмпирическая функция распределения. Построить критерий проверки гипотезы на уровне значимости α.

Задача 13. По выборке из геометрического распределения с параметром p построить наиболее мощный критерий асимптотического размера ε, различающий гипотезу $p = p_1$ и альтернативу $p = p_2$, если $p_1 < p_2$. Вычислить предел мощности построенного критерия при $n \to \infty$.

Задача 14*. Пять независимых одинаково распределенных случайных величин приняли значения: -0.46, $+0.11$, -0.32, $+0.19$, $+0.17$. Проверить гипотезу о равномерном распределении $H_1 : \mathrm{U}(-0.5, 0.5)$ при альтернативе $H_2 : \mathrm{N}(0, 0.03)$. Величину вероятности ошибки первого рода α положить равной 0.20. Найти вероятность ошибки второго рода β.

Задача 15*. Для классификации состояния объекта, который может находиться в одном из двух состояний H_1 или H_2, может использоваться один из двух скалярных признаков Y_1 и Y_2, представляющих собой нормально распределенные случайные величины, связанные с состояниями объекта известными распределениями:

$$H_1 : Y_1 \in \mathrm{N}(0, 1),\ Y_2 \in \mathrm{N}(0, 1),$$
$$H_2 : Y_1 \in \mathrm{N}(0, 16),\ Y_2 \in \mathrm{N}(2, 1).$$

Для классификации состояния объекта используется критерий Неймана–Пирсона. Следует установить, какой из этих признаков предпочтительней. Найти зависимость мощности при альтернативе от вероятности ошибки первого рода. Сравнить эффективность признаков при допустимых значениях вероятности ошибки первого рода $\alpha_1 = 0.01$ и $\alpha_2 = 0.15$.

Задача 16*. Пусть произведены две серии из n_1 и n_2 независимых испытаний, в каждом из которых наблюдается либо исход A, либо исход $\bar{A}$. Результаты сведены в таблицу, в столбцах которой указано число реализации соответствующих исходов для каждой серии. Убедиться, что статистика X_n^2 для проверки гипотезы H_1 об однородности испытаний представима в виде $X_n^2 = Z_n^2$, где статистика

$$Z_n = \left(\frac{v_{11}}{n_1} - \frac{v_{12}}{n_2} \right) \sqrt{\frac{n n_1 n_2}{v_{1\bullet} v_{2\bullet}}}.$$

Доказать, что $\mathcal{L}(Z_n | H_1) \to \mathrm{N}(0, 1)$ при $n_1, n_2 \to \infty$, и, основываясь на этом, построить критерий проверки гипотезы $H_1 : p_1 = p_2$ против односторонней альтернативы $H_2 : p_1 > p_2$ (здесь p_i – вероятность реализации A в испытаниях i-й серии, $i = 1, 2$).

	(1)	(2)	Σ
A	v_{11}	v_{12}	$v_{1\bullet}$
A	v_{21}	v_{22}	$v_{2\bullet}$
Σ	$v_{\bullet 1} = n_1$	$v_{\bullet 2} = n_2$	$n = n_1 + n_2$

Задача 17*. Как выглядит критерий проверки гипотезы H_1 : $\mathrm{U}(-a, a)$ против альтернативы $H_2 : \mathrm{N}(0, \sigma^2)$ (параметры a и σ заданы), если известно, что наблюдаемая непрерывная случайная величина ξ имеет распределение, симметричное относительно нуля? Рассмотреть случай большой выборки. Провести с этих позиций анализ следующих данных: -0.460 -0.114 -0.325 $+0.196$ -0.174 при $a = 1/2$, $\sigma^2 = 0.09$.

Задача 18*. Пусть $\xi = (\xi_1, \ldots, \xi_r)$ – нормальный случайный вектор, имеющий при гипотезе H_i распределение $\mathrm{N}(\mu^{(i)}, A)$, $i = 1, 2$ (общая ковариационная матрица A предполагается невырожденной). Постройте критерий Неймана–Пирсона для проверки гипотезы H_1 при альтернативе H_2 по одному наблюдению над ξ, а также критерий, минимизирующий сумму вероятностей ошибок.

Задача 19. Пусть $X = (X_1, \ldots, X_n)$ – выборка из нормального распределения $\mathrm{N}(\mu, \theta^2)$. Постройте р.н.м. несмещенный критерий проверки простой гипотезы $H_1 : \theta = \theta_0$ при двусторонней альтернативе $H_2 : \theta \neq \theta_0$.

Задача 20. Пусть $X = (X_1, \ldots, X_n)$ – выборка из распределения $\Gamma(\theta, 1)$. Постройте р.н.м. несмещенный критерий проверки простой гипотезы $H_1 : \theta = \theta_0$ при двусторонней альтернативе $H_2 : \theta \neq \theta_0$.

Задача 21*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из равномерного распределения $\mathrm{U}(\theta, \theta + 1)$, $\theta \in \mathbb{R}$, $n \geq 2$. Докажите, что р.н.м. критерий размером α для проверки $H_1 : \theta \leq 0$ против $H_2 : \theta > 0$ имеет вид

$$\pi(x) = \begin{cases} 0, & x_{(1)} < 1 - \alpha^{1/n},\ x_{(n)} < 1, \\ 1, & \text{иначе}, \end{cases}$$

где $x_{(j)}$ – реализация j-й порядковой статистики.

Задача 22*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из дискретного равномерного распределения на точках $1, \ldots, \theta$, где $\theta = 1, 2, \ldots$. Докажите, что р.н.м. критерий размером α для проверки $H_1 : \theta \leq \theta_0$ против $H_2 : \theta > \theta_0$ при известном $\theta_0 > 0$ имеет вид

$$\pi(x) = \begin{cases} 1, & x_{(n)} > \theta_0, \\ \alpha, & x_{(n)} \leq \theta_0, \end{cases}$$

где $x_{(n)}$ – реализация n-й порядковой статистики.

Задача 23*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из дискретного равномерного распределения на точках $1, \ldots, \theta$, где $\theta = 1, 2, \ldots$. Докажите, что р.н.м. критерий размером α для проверки $H_1 : \theta = \theta_0$ против $H_2 : \theta \neq \theta_0$ при известном $\theta_0 > 0$ имеет вид

$$\pi(x) = \begin{cases} 1, & x_{(n)} > \theta_0 \text{ или } x_{(n)} \leq \theta_0 \alpha^{1/n}, \\ 0, & \text{иначе}, \end{cases}$$

где $x_{(n)}$ – реализация n-й порядковой статистики.

Задача 24*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из дискретного равномерного распределения на точках $1, \ldots, \theta$, где целое число $\theta \geq 2$. Найти критерий отношения правдоподобия уровня α для задачи проверки $H_1 : \theta \leq \theta_0$ против $H_2 : \theta > \theta_0$, где θ_0 известное целое ≥ 2.

Задача 25*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из дискретного равномерного распределения на точках $1, \ldots, \theta$, где целое число $\theta \geq 2$. Найти критерий отношения правдоподобия уровня α для задачи проверки $H_1 : \theta = \theta_0$ против $H_2 : \theta \neq \theta_0$, где θ_0 известное целое ≥ 2.

Задача 26*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из показательного распределения на интервале (a, ∞) с параметром θ. Предположим, что θ известно. Найти критерий отношения правдоподобия уровня α для проверки $H_1 : a \leq a_0$ против $H_2 : a > a_0$, где a_0 известная константа.

Задача 27*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из показательного распределения на интервале (a, ∞) с параметром θ. Предположим, что θ известно. Найти критерий отношения правдоподобия уровня α для проверки $H_1 : a = a_0$ против $H_2 : a \neq a_0$, где a_0 известная константа.

Задача 28*. Пусть $X = (X_1, \ldots, X_{n1})$ и $Y = (Y_1, \ldots, Y_{n2})$ две независимые выборки из равномерных распределений на $(0, \theta_1)$ и $(0, \theta_2)$ соответственно, причем $\theta_1 > 0$ и $\theta_2 > 0$ неизвестны. Найти критерий отношения правдоподобия размером α для проверки $H_1 : \theta_1 = \theta_2$ против $H_2 : \theta_1 \neq \theta_2$.

12.4. Разные задачи

Задача 29*. Пусть имеется конечная совокупность U, число элементов которой N неизвестно. Из этой совокупности n раз извлекается простая бесповторная выборка объемом m (каждый раз любая из C_N^m возможных комбинаций элементов U может быть извлечена с равной вероятностью). Обозначим через $\mu_r = \mu_r(n, m, N)$ число наблюдавшихся элементов, каждый из которых повторился ровно r раз $(r = 1, 2, \ldots, n)$. Оценить параметрические функции $\tau(N)$ по выборочным данным $(\mu_1, \ldots, \mu_n)$.

Задача 30. Дана выборка $X = (X_1, X_2, X_3)$ из распределения $\mathrm{N}(0, \theta^2)$, и пусть $T(X) = \sqrt{X_1^2 + X_2^2 + X_3^2}$. Рассмотрим статистику $p_T(x) = \frac{1}{2T} I(|x| \leq T)$, где $I(\cdot)$ – индикатор, который как функция переменного x представляет собой плотность равномерного распределения на отрезке $[-T, T]$. Убедитесь в том, что $p_T(x)$ при любом x является несмещенной оценкой для плотности исходного распределения $\mathrm{N}(0, \theta^2)$.

Задача 31. Покажите, что в логистической модели $f(x;\theta) = e^{-x+\theta}(1+e^{-x+\theta})^{-2}$, $-\infty < x < \infty$, $\theta \in (-\infty, \infty)$, выборочное среднее $\overline{X}$ является несмещенной и состоятельной оценкой θ.

Задача 32. Пусть $X_1,\ldots,X_n$ – выборка из распределения Пуассона с параметром λ. Какое распределение имеет выборка $Y_1,\ldots,Y_n$, где $Y_i = F(X_i)$, где $F(y)$ – функция распределения Пуассона? Тот же вопрос для распределения Бернулли.

Задача 33*. Пусть $X = (X_1,\ldots,X_n)$ – выборка из нормального $\mathrm{N}(\mu,\sigma^2)$ распределения. Найти распределение случайной величины

$$\eta = \frac{X_1 - \bar{X}}{\sqrt{n-1}S},$$

где $S^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \bar{X})^2$.

Задача 34*. Основываясь на следующем представлении производящей функции статистики T_n – числа инверсий в случайной выборке объемом n:

$$\Phi(z) = \frac{1}{n!}\prod_{r=1}^{n-1}(1 + z + \cdots + z^r)$$

доказать, что

$$\mathcal{L}(T_n) \approx \mathrm{N}\left(\frac{n(n-1)}{4}, \frac{n^3}{36}\right),$$

при $n \to \infty$.

Задача 35*. Доказать теорему Колмогорова. См. Биллингсли П. Сходимость вероятностных мер. Москва : Наука, 1977. 352 с.

12.5. Оптимальные оценки

Задача 36*. Продолжительность горения электрических ламп имеет распределение $\Gamma(\theta, 1)$. Чтобы оценить θ, берут выборку из n ламп и наблюдают «времена жизни» первых r перегоревших ламп $X_{(1)} < X_{(2)} < \cdots < X_{(r)}$. Постройте оптимальную несмещенную оценку вида $T(X) = \sum\limits_{k=1}^{r}\lambda_k X_{(k)}$.

Задача 37. По выборке $X = (X_1,\ldots,X_n)$ из распределения $\mathrm{U}(\theta, 2\theta)$ требуется оценить параметр θ. Рассмотреть класс несмещенных оценок параметра θ вида $T(X) = \alpha X_{(n)} + \beta X_{(1)}$, $\alpha, \beta \geq 0$ и найти в этом классе оптимальную оценку.

Задача 38*. Рассматривается задача оценивания дисперсии θ_2^2 в нормальной модели $\mathrm{N}(\theta_1, \theta_2^2)$ по выборке $X = (X_1, \ldots, X_n)$. Пусть

$$S_0^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_i - \overline{X})^2.$$

Найти оптимальную оценку вида $T_\lambda = \lambda S_0^2$, минимизирующую меру $\mathbb{E}_\theta (T_\lambda - \theta_2^2)^4$.

Задача 39*. Рассматривается задача оценивания дисперсии θ_2^2 в нормальной модели $\mathrm{N}(\theta_1, \theta_2^2)$ по выборке $X = (X_1, \ldots, X_n)$. Пусть

$$S_0^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_i - \overline{X})^2.$$

Найти оптимальную оценку вида $T_\lambda = \lambda S_0^2$, минимизирующую меру $\mathbb{E}_\theta |T_\lambda - \theta_2^2|$.

Задача 40. Докажите следующие свойства оптимальных оценок: если $T^* = T^*(X)$ – оптимальная несмещенная оценка некоторой функции $\tau = \tau(\theta)$, то

1) для любой статистики $\psi = \psi(X)$ с $\mathbb{E}_\theta \psi = 0 \ \forall \theta \in \Theta$ выполняется равенство $\mathrm{cov}_\theta(T^*, \psi) = 0 \ \forall \theta \in \Theta$;
2) для любой другой несмещенной оценки $T = T(X)$ имеем

$$\mathrm{cov}(T^*, T) = \mathbb{D}_\theta T^*.$$

Задача 41. Проверьте непосредственно, что выборочное среднее $\overline{X}$ в логистической модели $f(x;\theta) = e^{-x+\theta}(1 + e^{-x+\theta})^{-2}$, $x \in \mathbb{R}$, $\theta \in \mathbb{R}$, не является эффективной оценкой θ.

Задача 42*. Рассматривается задача оценивания функции $\tau(\theta) = \theta^2$ в модели $\Gamma(\theta, \lambda)$ по выборке $X = (X_1, \ldots, X_n)$. Докажите, что

$$T^*(X) = \frac{n}{\lambda(\lambda n + 1)} \overline{X}^2$$

– оптимальная несмещенная оценка $\tau(\theta)$; вычислив $\mathbb{D}T^*$, убедиться в том, что эта оценка не является эффективной.

Задача 43. Найти оптимальную оценку параметра $\theta > 0$ в модели $\mathrm{U}(0, \theta)$.

Задача 44. Найти оптимальную оценку функции $\tau(\theta) = 1 + \theta + \theta^2$ в пуассоновской модели $\mathrm{Po}(\theta)$, $\theta > 0$.

Задача 45. Найти оптимальную оценку функции $\tau(\theta) = e^\theta$ в пуассоновской модели $\mathrm{Po}(\theta)$, $\theta > 0$.

Задача 46. Найти оптимальную оценку функции $\tau(\theta) = (1+\theta)^2$ в биномиальной модели $\mathrm{Bi}(k, \theta)$, $\theta > 0$.

Задача 47*. Пусть $(X_1, \ldots, X_n)$ – выборка из распределения $\mathrm{Be}(\theta)$, $\theta \in (0, 1)$.

1. Найти оптимальную оценку функции θ^m, где m – положительное целое число и $m \leq n$.
2. Найти оптимальную оценку вероятности $\mathbb{P}(X_1 + \cdots + X_m = k)$, где m и k – положительные целые и $k \leq m \leq n$.

Задача 48*. Пусть $(X_1, \ldots, X_n)$ – выборка из распределения $\mathrm{N}(\theta, \sigma^2)$ с неизвестным $\theta \in \mathbb{R}$ и известным σ^2.

1. Найти оптимальные оценки функций θ^3 и θ^4.
2. Найти оптимальные оценки вероятностей $\mathbb{P}(X_1 \leq t)$ и $\frac{d}{dt}\mathbb{P}(X_1 \leq t)$ для фиксированного $t \in \mathbb{R}$.

Задача 49*. Пусть $(X_1, \ldots, X_n)$, $n > 2$, – выборка из равномерного распределения на интервале $(\theta_1 - \theta_2, \theta_1 + \theta_2)$, где $\theta_1 \in \mathbb{R}$ и $\theta_2 > 0$. Найти оптимальные оценки параметров θ_1 и θ_2 и оптимальную оценку θ_1/θ_2.

Задача 50. Докажите, что если для модели $F(x; \theta)$ существует полная достаточная статистика T и если статистика T_1 имеет распределение, не зависящее от параметра θ, то T_1 и T независимы.

Задача 51*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из распределения $\Gamma(\theta, \lambda)$ и $T = X_1 + \cdots + X_n$. Доказать, что статистика

$$T^* = \frac{\Gamma(\lambda n)}{\Gamma(\lambda n - a)} T^{-a}$$

– оптимальная несмещенная оценка функции $\tau(\theta) = \theta^{-a}$ при любом $a < \lambda n$. Доказать, что при целом $a - \lambda n \geq 0$ несмещенных оценок для $\tau(\theta) = \theta^{-a}$ не существует.

Задача 52*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из распределения $\Gamma(\theta, \lambda)$, $T = X_1 + \cdots + X_n$ и $\varphi(x)$ – заданная функция, для которой $\tau(\theta) = \mathbb{E}_\theta \varphi(\xi)$ существует. Докажите, что оптимальная оценка функции $\tau(\theta)$ имеет вид

$$T^* = \frac{\Gamma(\lambda n)}{\Gamma(\lambda)\Gamma(\lambda(n-1))} \int_0^1 \varphi(xT) x^{\lambda-1} (1-x)^{(n-1)\lambda-1} dx.$$

Задача 53*. Докажите, что для распределения Вейбулла $W(0, \lambda, \theta)$ с неизвестным параметром масштаба θ полной достаточной статистикой является $T(X) = \sum\limits_{i=1}^{n} X_i^\lambda$, а оптимальная оценка для $\tau(\theta) = \mathbb{E}_\theta \varphi(\xi)$,

где $\varphi(x)$ – заданная функция, имеет вид

$$T^* = (n-1)\int\limits_0^1 \varphi\left((tT)^{1/\lambda}\right)(1-t)^{n-2}dt.$$

Распределение Вейбулла $W(0,\lambda,\theta)$ определяется функцией распределения

$$F(x) = 1 - e^{-(x/\theta)^\lambda},\ \theta > 0,\ \lambda > 0,\ x \geq 0.$$

Задача 54*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из распределения $\mathrm{U}(0,\theta)$. Докажите, что $X_{(n)}$ – полная достаточная статистика для θ. Найдите оптимальную несмещенную оценку T^* дифференцируемой функции $\tau(\theta)$. Рассмотреть класс статистик $T_\lambda = \lambda T^*$ и для случая $\tau(\theta) = \theta$ убедиться в том, что в нем имеются оценки с равномерно по θ меньшей среднеквадратической ошибкой, чем у T^*.

12.6. Оценки максимального правдоподобия

Задача 55. Докажите, что для общей нормальной модели $\mathrm{N}(\theta_1, \theta_2^2)$ оценка максимального правдоподобия $\hat{\theta} = (\overline{X}, S)$, где

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

Задача 56. Докажите, что для распределения Лапласа, задаваемого плотностью $f(x;\theta) = \frac{1}{2}e^{-|x-\theta|}$, $x \in \mathbb{R}$, оценка максимального правдоподобия $\hat{\theta}$ совпадает с выборочной медианой. Можно ли здесь воспользоваться теоремой об асимптотической нормальности оценки максимального правдоподобия?

Задача 57*. Пусть $X = (X_1, \ldots, X_n)$ – выборка из распределения $\mathrm{N}(\theta, 1)$. В качестве оценки θ рассмотреть статистику

$$T_n = \begin{cases} \overline{X}, & |\overline{X}| \geq a_n, \\ b\overline{X}, & |\overline{X}| < a_n, \end{cases}$$

где константа $a_n \to 0$, но $\sqrt{n}a_n \to \infty$ при $n \to \infty$, и вычислите ее асимптотическую эффективность.

Задача 58*. Для моделей $\mathrm{Bi}(k,\theta)$, $\mathrm{Po}(\theta)$, $\mathrm{N}(\mu,\theta^2)$ и $\Gamma(\theta,\lambda)$ найдите такие параметрические функции $\tau(\theta)$, чтобы асимптотические дисперсии соответствующих оценок максимального правдоподобия не зависели от параметра θ.

12.7. Доверительное оценивание

Задача 59. Пусть $X = (X_1, \ldots, X_n)$ и $Y = (Y_1, \ldots, Y_m)$ – две независимые выборки, причем первая из распределения $\mathrm{N}(\theta_1, \sigma_1^2)$, а вторая из распределения $\mathrm{N}(\theta_2, \sigma_2^2)$. Постройте γ-доверительный интервал для разности $\tau = \theta_1 - \theta_2$. Решить задачу также для распределений $\mathrm{N}(\theta_1, \theta^2)$ и $\mathrm{N}(\theta_2, \theta^2)$ с неизвестными, но равными дисперсиями.

Задача 60. Пусть $X = (X_1, \ldots, X_n)$ и $Y = (Y_1, \ldots, Y_m)$ – две независимые выборки, причем первая из распределения $\Gamma(\theta_1, 1)$, а вторая из распределения $\Gamma(\theta_2, 1)$. Постройте γ-доверительный интервал для отношения $\tau = \theta_2/\theta_1$.

Задача 61. Убедиться в том, что $\left(X_{(1)} + \dfrac{\ln(1-\gamma)}{n}, X_{(1)}\right)$ есть γ-доверительный интервал для параметра θ экспоненциального распределения с плотностью

$$f(x;\theta) = e^{-(x-\theta)},\ x \geq \theta.$$

Убедиться в том, что $(X_{(n)}, X_{(n)}/\sqrt[n]{1-\gamma})$ есть γ-доверительный интервал для параметра θ модели $\mathrm{U}(0,\theta)$ по выборке объемом n.

Задача 62. Пусть $X = (X_1, \ldots, X_n)$ – выборка из распределения $\mathrm{Bi}(1,\theta)$. Основываясь на точечной статистике $T = \overline{X}$ параметра θ, покажите, что центральный γ-доверительный интервал для него (T_1, T_2) определяется условиями

$$\sum_{r=nT}^{n} C_n^r T_1^r (1-T_1)^{n-r} = \sum_{r=0}^{nT} C_n^r T_2^r (1-T_2)^{n-r} = \frac{1-\gamma}{2};$$

при этом

$$T_1 = \mathrm{Beta}_{\frac{1-\gamma}{2}}(nT, n-nT+1),\ T_2 = \mathrm{Beta}_{\frac{1+\gamma}{2}}(nT+1, n-nT)$$

есть квантили бета-распределений.

Задача 63*. Пусть X – выборка объемом 1 из отрицательного биномиального распределения $\overline{\mathrm{Bi}}(r,\theta)$, $\theta \in (0,1)$. Используя точечную оценку $T = X - r$, показать, что γ-доверительный интервал для θ есть

$$\left[\frac{1}{1 + \frac{T+1}{r}\mathrm{F}_{1-\alpha_2}(2(T+1), 2r)}, \frac{\frac{r}{T}\mathrm{F}_{1-\alpha_1}(2r, 2T)}{1 + \frac{r}{T}\mathrm{F}_{1-\alpha_1}(2r, 2T)}\right],$$

где $\alpha_1 + \alpha_2 = 1 - \gamma$ и $\mathrm{F}_\alpha(a,b)$ есть α-квантиль распределения Фишера–Снедекора $\mathrm{F}(a,b)$ и $\mathrm{F}_\alpha(a,0) = \infty$.

Задача 64. Пусть $X = (X_1, \ldots, X_n)$ – выборка из распределения $\mathrm{Po}(\theta)$. С помощью статистики $T = \overline{X}$ найти центральный γ-доверительный интервал для θ. Доказать, что интервал

$$\overline{X} \pm c_\gamma \sqrt{\overline{X}/n}$$

является асимптотическим γ-доверительным интервалом.

Задача 65. Построить асимптотические γ-доверительные интервалы для параметра θ пуассоновской модели $\mathrm{Po}(\theta)$, доказав и воспользовавшись нормальными аппроксимациями

$$2\sqrt{n}\left(\sqrt{\overline{X}} - \sqrt{\theta}\right) \xrightarrow{d} N(0,1),$$

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

при объеме выборки $n \to \infty$. Сравнить скорости сходимости полученных интервалов к истинному значению параметра θ с асимптотическим интервалом из предыдущей задачи.

Задача 66. Постройте асимптотический γ-доверительный интервал для параметра θ модели $\mathrm{N}(\mu, \theta^2)$, доказав и воспользовавшись аппроксимацией

$$\sqrt{2n}(\ln \hat{\theta}_n - \ln \theta) \xrightarrow{d} N(0,1).$$

Задача 67. Постройте асимптотический γ-доверительный интервал функции $\tau(\theta) = \Phi\left(\dfrac{x_0 - \theta_1}{\theta_2}\right)$ в модели $\mathrm{N}(\theta_1, \theta_2^2)$. Для этого воспользоваться оценкой максимального правдоподобия $\hat{\tau}_n = \Phi\left(\frac{x_0 - \overline{X}}{S}\right)$, $S^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{X})^2$.

12.8. Решающие правила

Задача 68*. Дана модель $\mathrm{Be}(\theta)$, $\Theta = \{\theta_1, \theta_2\}$, и матрица штрафов $(a, b > 0)$:

$$C = \begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix}.$$

Рассмотреть два случая: $\theta_1 = 2/3$, $\theta_2 = 1/2$ и $\theta_1 = 3/4$, $\theta_2 = 1/2$. Убедиться в том, что в обоих случаях множества допустимых решающих правил совпадают. Вычислить и сравнить полные риски байесовских

решающих правил для обоих случаев и для всех априорных распределений.

Задача 69. Дана модель $\mathrm{Bi}(3,\theta)$, $\Theta = \{1/3, 1/2\}$, и матрица штрафов

$$C = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}.$$

Рассмотреть решающие правила $\delta_i = (\delta_i(0), \delta_i(1), \delta_i(2), \delta_i(3))$: $\delta_2 = (d_1, d_2, d_2, d_2)$, $\delta_3 = (d_1, d_1, d_2, d_2)$,$\delta_1 = (d_1, d_1, d_1, d_2)$. Убедиться в том, что эти правила между собой несравнимы, и найдите среди них минимаксное.

Задача 70*. Дана модель $\overline{\mathrm{Bi}}(1,\theta)$, $\Theta = \{\theta_1, \theta_2\}$, и матрица штрафов $(a, b > 0)$

$$C = \begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix}$$

Определите минимаксную решающую функцию среди функций $(i = 1, 2, \dots)$

$$\delta_i(x) = \begin{cases} d_1 & x = 0, 1, \dots, i-1; \\ d_2 & x = i, i+1, \dots \end{cases}$$

Задача 71*. Рассматривается задача оценивания неизвестной вероятности успеха θ по наблюдениям числа успехов X в n испытаниях Бернулли (таким образом, в этом случае $\Theta = D = (0,1)$). Пусть функция потерь имеет вид

$$L(\theta, d) = \frac{(d-\theta)^2}{\theta(1-\theta)},$$

а априорное распределение θ является равномерным на интервале $(0,1)$. Найдите байесовское решение и докажите, что оно же является и минимаксным, при этом полный риск равен $1/n$.

Задача 72*. Рассмотрим задачу оценивания параметра $\theta > 0$ экспоненциального распределения с плотностью $f(x;\theta) = \theta e^{-\theta x}$, $x > 0$, по выборке $X = (X_1, \dots, X_n)$. Пусть функция потерь $L(\theta, d) = (1/\theta - d)^2$ и априорное распределение $\Gamma(a, \lambda)$, $\lambda > 2$. Докажите, что байесовская оценка имеет вид

$$\delta^*(X) = \frac{1}{a(\lambda + n - 1)} \left(a \sum_{i=1}^{n} X_i + 1 \right)$$

и ее риск

$$r(\delta^*) = \frac{1}{a^2(\lambda + n - 1)(\lambda - 1)(\lambda - 2)}.$$

Задача 73*. Рассматривается задача оценивания неизвестной вероятности «успеха» θ по наблюдению числа «успехов» r в n испытаниях Бернулли. Пусть функция потерь имеет вид

$$L(\delta,\theta)=\frac{(\delta-\theta)^2}{\theta(1-\theta)},$$

а априорное распределение параметра θ есть $\mathrm{U}(0,1)$. Найти байесовскую оценку и доказать, что она также является минимаксной.

Задача 74*. Пусть испытания Бернулли продолжаются до получения r-го «неуспеха» и X – число «успехов» в этих испытаниях, то есть $X\in\overline{\mathrm{Bi}}(r,\theta)$. Найти байесовскую оценку параметра θ, если функция потерь равна $L(\delta,\theta)=(\delta-\theta)^2$, а априорное распределение $\theta\in\mathrm{Beta}(a,b)$.

Задача 75*. По выборке $X_1,\ldots,X_n$ из распределения Пуассона $\mathrm{Po}(\theta)$ оценивается параметр θ при функции потерь $L(\delta,\theta)=(\delta-\theta)^2$ и априорном распределении $\theta\in\Gamma(a,\lambda)$. Найти байесовскую оценку и ее риск.

Задача 76*. Пусть $X=(X_1,\ldots,X_n)$ – выборка из равномерного распределения $\mathrm{U}(0,\theta)$, где априорное распределение параметра θ есть распределение Парето с параметрами x_0 и $\alpha>2$. Найти байесовскую оценку и соответствующий риск при функции потерь $L(\delta,\theta)=(\delta-\theta)^2$.

Задача 77. Найти байесовскую оценку и соответствующий риск параметра θ нормального распределения $\mathrm{N}(\theta,b^2)$ при функции потерь $L=(\delta-\theta)^2$ и априорном распределении $\theta\in\mathrm{N}(\mu,\sigma^2)$.

12.9. Задачи на программирование

Задача 78*. Проверить гипотезы об однородности, независимости и случайности встроенного в язык программирования генератора псевдослучайных чисел.

Задача 79. Пусть $X_{N1},\ldots,X_{Nn}$ – приближенно нормальные $\mathrm{N}(\mu,\sigma^2)$ числа, каждое из которых получено суммированием N равномерно распределенных на $(0,1)$ слагаемых. Получить три реализации (при $N=2,4,12$) выборок с $n=100$, $\mu=0$, $\sigma^2=1$. Для каждой выборки построить эмпирические функции распределения и гистограммы; получить оценки μ и σ^2; вычислить 3-й и 4-й выборочные центральные моменты и сравнить их с истинными значениями теоретических моментов.

Задача 80. Смоделировать выборку объемом $n=100$ из показательного распределения $\mathrm{Exp}(1)$. Построить эмпирическую функцию распределения и гистограмму, вычислить выборочные моменты первого и второго порядков, сравнить их с теоретическими значениями.

Задача 81*. Моделируя выборки одинакового объема $X = (X_1, \ldots, X_n)$ из нормального распределения $\mathrm{N}(\mu, \sigma^2)$, убедиться в результатах теоремы Фишера, то есть в том, что $\mathcal{L}(\bar{X}) = \mathrm{N}(\mu, \sigma^2/n)$ и $\mathcal{L}(nS^2/\sigma^2) = \chi^2(n-1)$, где $S^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \bar{X})^2$. Для этого построить эмпирические функции распределения и гистограммы случайных величин $\bar{X}$ и nS^2/σ^2 и сравнить их с теоретическими зависимостями. Кроме того, воспользоваться критерием хи-квадрат для проверки возникающих гипотез.

Задача 82*. Моделируя выборки одинакового объема $U = (U_1, \ldots, U_n)$ из равномерного на отрезке $\mathrm{U}(0,1)$ распределения, убедиться в том, что порядковые статистики $U_{(k)}$ имеют бета-распределение. Для этого построить эмпирические функции распределения и гистограммы случайных величин $U_{(k)}$ и сравнить их с теоретическими зависимостями. Кроме того, воспользоваться критерием хи-квадрат для проверки возникающих гипотез; достаточно рассмотреть какую-то одну порядковую статистику.

Задача 83*. Моделируя выборки одинакового объема $X = (X_1, \ldots, X_n)$ из нормального распределения $\mathrm{N}(0,1)$ и выбирая идемпотентные матрицы A размером $n \times n$ рангов r от 1 до n, убедиться в том, что случайная величина $X^T A X$ имеет распределение $\chi^2(r)$. Для этого построить эмпирические функции распределения и гистограммы случайных величин $X^T A X$ и сравнить их с теоретическими зависимостями. Рассмотреть также случаи неидемпотентных матриц.

Задача 84. Моделируя выборки одинакового объема $U_1, \ldots, U_n$ из равномерного на отрезке $\mathrm{U}(0,1)$ распределения, убедиться в том, что статистика $\sqrt{n}D_n$, где D_n – статистика Колмогорова–Смирнова, имеет приближенно колмогоровское распределение. Для этого выбрать $n = 100$ и построить гистограмму случайной величины $\sqrt{n}D_n$, и сравнить ее с теоретической зависимостью.

Задача 85. Смоделировать выборку $X = (X_1, \ldots, X_n)$ из распределения случайной величины, принимающей значения $1, \ldots, N$ с вероятностями $0 < p_j < 1$, $j = 1, \ldots, N$, убедиться в том, что распределение статистики хи-квадрат $T_{\chi^2} = \sum\limits_{j=1}^{N} \frac{(\nu_j - np_j)^2}{np_j}$ приближенно совпадает с распределением $\chi^2(N-1)$. Взять для этого $n = 100$, построить гистограмму случайной величины T_{χ^2} и сравнить ее с теоретической зависимостью.

Задача 86. Используя таблицу значений какой-либо функции ($\cos x, e^x, \ln x$ и т.д.) выписать 100 цифр, выбирая из каждого значения функции второй знак справа, и проверить для такой выборки по

критерию χ^2 гипотезу о равновероятности цифр 0, 1,. . . ,9 при уровне значимости 0.01.

Задача 87. Выписать 100 цифр после запятой числа π и по критерию χ^2 проверить гипотезу о дискретном равномерном распределении цифр на множестве $0, \ldots, 9$ при уровне значимости 0.01. Повторить то же самое для каких-нибудь значений функций $\cos x$, e^x, $\ln x$.

Задача 88. Смоделировать выборку объемом $n = 1000$ из распределения Бернулли $\mathrm{Be}(3/5)$ и проверить по критерию χ^2 соответствие данных теоретической модели.

Задача 89*. Смоделировав выборку объемом $n = 100$ из показательного распределения $\mathrm{Exp}(1)$ и сгруппировав полученные данные по $N = 4$ равновероятным интервалам (при гипотезе $H_1 : \mathrm{Exp}(1)$), проверить по критерию χ^2 гипотезу H_1 при уровне значимости $\alpha = 0.1$.

Задача 90. Смоделировать выборку объемом $n = 200$ равномерно распределенных на $(0, 1)$ чисел. С помощью критерия χ^2 проверить, что две подвыборки $(X_{2i}, i = 1, \ldots, 100)$ и $(X_{2i+1}, i = 0, \ldots, 99)$ являются выборками из одного и того же распределения.

Задача 91*. Построить критерий Неймана–Пирсона для проверки двух простых гипотез: $H_1 : \mathrm{N}(-1, 1)$, $H_2 : \mathrm{N}(2, 4)$. Объем выборки $n = 1$, уровень значимости $\alpha = 0.1$ Определить величину β вероятности ошибки второго рода. Построить зависимость $\beta(\alpha)$.

Задача 92*. Разработать алгоритм построения псевдослучайных последовательностей бросков одной и двух игральных костей, чтобы на любом шаге эмпирическое распределение было максимально близко к «правильному» (правильному – в смысле равномерного распределения выпадения граней одной кости).

Задача 93*. Реализовать последовательный критерий Вальда для проверки простых гипотез

$$H_1 : \mathrm{Be}(p_1)$$
$$H_2 : \mathrm{Be}(p_2)$$

Вывести график зависимости функции отношения правдоподобия от номера шага. Получить выборку значений числа шагов $(\nu_1, \ldots, \nu_n)$ когда процедура завершается, сравнить выборочное среднее $\bar{\nu}$ с приближенными теоретическими зависимостями. Построить по выборке ν гистограмму.

Задача 94*. Реализовать последовательный критерий Вальда для проверки простых гипотез о матрице переходов марковской цепи

$$H_1 : P = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

$$H_2 : P = \begin{bmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{bmatrix}$$

Начальным состоянием цепи считать $\pi_0 = [1, 0]$. Вывести график зависимости функции отношения правдоподобия от номера шага. Получить выборку значений числа шагов $(\nu_1, \ldots, \nu_n)$ когда процедура завершается, сравнить выборочное среднее $\bar{\nu}$ с приближенными теоретическими зависимостями. Построить по выборке ν гистограмму.

Задача 95*. Реализовать критерий Байеса для проверки 10 гипотез вида $\mathrm{Be}(p)$ для $p = p_1$, $p = p_2$,..., $p = p_{10}$ и априорными вероятностями $q_1, \ldots, q_{10}$. Для конкретных значений $p_1, \ldots, p_{10}$, $q_1, \ldots, q_{10}$, элементов матрицы штрафов и объема выборки вычислить байесовский (полный) риск.

Задача 96*. Посчитать коэффициент детерминации и корреляции между рублем и нефтью на двух интервалах времени: от 10.11.2014 до 18.11.2015 и от 10.11.2014 до 31.10.2015. Сравнить полученные результаты с расчитанными независимыми экспертами (0.77 детерминация, 0.88 корреляция): http://www.rbc.ru/finances/18/11/2015/564c56e89a79470df4076f59

Задача 97*. Сформируйте двумерную выборку (X_1, Y_1), ..., (X_n, Y_n) из двумерного нормального распределения $\mathrm{N}\left((0,0), \begin{Vmatrix} 1 & \theta \\ \theta & 1 \end{Vmatrix}\right)$, $\theta \in (-1, 1)$. Составьте и численно решите уравнение правдоподобия, при этом возьмите объем выборки достаточно большим, чтобы решение уравнения было единственным. Постройте гистограмму случайной величины $\sqrt{n}(\hat{\theta} - \theta)$ и сравните ее распределение с $\mathrm{N}\left(0, \dfrac{(1-\theta^2)^2}{1+\theta^2}\right)$.

Задача 98*. Запишите уравнение правдоподобия для оценки параметра θ распределения Коши $\mathrm{K}(\theta)$ и решите его методом накопления.

Задача 99*. Смоделируйте выборки, объемы которых $n = 10, 100, 1000$, и получите оценки максимального правдоподобия параметров следующих распределений:

1. $\mathrm{N}(\theta_1, \theta_2^2)$, при моделировании положите $\theta_1 = 1, \theta_2^2 = 4$.
2. $\mathrm{Be}(\theta)$, при моделировании положите $\theta = 0.7$.
3. $\mathrm{U}(0, \theta)$, при моделировании положите $\theta = 1$.

Задача 100*. Убедиться в асимптотической нормальности оценки максимального правдоподобия. Для этого выбрать распределение $\Gamma(\theta, 1)$ и для $k = 100$ выборок объемом 200 получить выборку $\hat{\theta}_1, \ldots, \hat{\theta}_k$, построить гистограмму и сравнить ее с теоретической зависимостью.

Задача 101*. Известно, что γ-доверительным интервалом для среднеквадратического отклонения θ модели $\mathrm{N}(\mu, \theta^2)$ является любой интервал $(T/a_2, T/a_1)$, где $T^2 = \sum_{i=1}^{n}(X_i - \mu)^2$, где числа $a_1 < a_2$ выбираются из условия $\int_{a_1}^{a_2} x k_n(x^2)dx = \gamma/2$, $k_n(t)$ – плотность распределения $\chi^2(n)$. Требуется определить оптимальный (наикратчайший) в этом классе интервал. Оптимальность понимать в смысле минимизации a_2/a_1.

Задача 102*. С помощью метода наименьших квадратов оцените параметр a в линейной модели $y = ax$. Измерения производятся в каждой из r точек x_i по n_i раз, $i = 1, 2, \ldots, r$. Ошибки измерения e_{ij} следует считать некоррелированными случайными величинами с распределением $\mathrm{N}(0, \sigma^2)$, так что результаты измерения дают $y_{ij} = ax_i + e_{ij}$. Отобразите на плоскости (x, y) прямую $y = ax$, точки (x_i, y_{ij}) для всех i, j, а также прямую после оценки параметра $y = a^*x$.

Задача 103*. Имеются две правильных игральных кости с 6 гранями. Результат бросания этих костей представляет собой некоторую случайную величину, принимающую значения от 2 до 12. Вычислить вероятность того, что число $i \in \{2, \ldots, 12\}$ за n бросков выпадет чаще чем $j \in \{2, \ldots, 12\}$, $i \neq j$. Найти асимптотику этой вероятности с ростом числа бросков n.

Задача 104*. В задаче построения доверительного интервалла для оценки параметра $\theta \in [0, 1]$ в модели $\mathrm{Be}(\theta)$ построить множество D_γ для произвольного $\gamma \geq 0.95$ (см. Г.И. Ивченко, Ю.И. Медведев. Введение в математическую статистику : учебник. Москва : Издательство ЛКИ, 2010. – С. 282).

Задача 105*. Пусть имеются k предметов, веса которых $\beta_1, \ldots, \beta_k$ неизвестны. Для определения этих весов взвешиваются комбинации предметов: каждая операция (взвешивание) состоит в том, что несколько предметов кладут на одну чашу весов, несколько – на другую и добавляют разновес для приведения весов в равновесие. В результате получают соотношения

$$\beta_1 z_1^{(i)} + \cdots + \beta_k z_k^{(i)} = x_i$$

(для i-го взвешивания, $i = 1, \ldots, n$), где $z_j^{(i)} = 1, -1, 0$ в зависимости от того, лежит j-й предмет на левой чаше весов, на правой или вообще не участвует в данном (i-м) взвешивании, а x_i – добавляемый вес. Считая погрешности измерений независимыми и нормальными $\mathrm{N}(0, \sigma^2)$,

оценить веса четырех ($k = 4$) предметов по данным следующей таблицы за восемь ($n = 8$) взвешиваний:

$z_1^{(i)}$	1	1	1	1	1	1	1	1
$z_2^{(i)}$	1	−1	1	−1	1	−1	1	−1
$z_3^{(i)}$	1	1	−1	−1	1	1	−1	−1
$z_4^{(i)}$	1	−1	−1	1	1	−1	−1	1
x_i	20.2	8.1	9.7	1.9	19.9	8.3	10.2	1.8

Вычислить вторые моменты оценок, а также найти оценку для σ^2. Построить систему совместных доверительных интервалов для $\beta_1, \ldots, \beta_4$ с доверительным уровнем ≥ 0.95.

Заключение

В учебном пособии приведены задачи с решениями по курсу «Математическая статистика». Эти задачи рекомендуется предлагать студентам в семестровом курсе. Разобраны стандартные темы, входящие в курс: проверка гипотез о согласии распределений, независимости случайных величин, однородности выборок, двух простых гипотез и сложных гипотез, а также свойства и методы изучения точечных и интервальных оценок. Кроме задач с решениями здесь представлены задачи, собранные автором в разное время, и которые можно предлагать студентам во время приема заданий, на контрольных работах и на экзаменах. Литература в конце пособия рекомендуется для углубления и расширения знаний в вышеуказанных разделах математической статистики.

Литература

1. Боровков А.А. Математическая статистика: учебник. 4-е изд., стер. Санкт-Петербург : Издательство «Лань», 2010.
2. Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику. Изд. 2-е, испр. и доп. Москва : ЛЕНАНД, 2017.
3. Крамер Г. Математические методы статистики. Москва : Мир, 1975.
4. Лагутин М.Б. Наглядная математическая статистика : учебное пособие. Москва : БИНОМ. Лаборатория знаний, 2019.
5. Леман Э. Проверка статистических гипотез. Москва : Наука, 1979.
6. Леман Э. Теория точечного оценивания. Москва : Наука, 1991.
7. Натан А.А., Горбачев О.Г., Гуз С.А. Математическая статистика. Москва : МЗ Пресс, 2005.
8. Чернова Н.И. Математическая статистика : учебное пособие. 2-е изд., испр. и доп. Новосибирск : РИЦ НГУ, 2014.
9. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2014.

Выходные данные

Для заметок