Федеральное агенство по образованию

Санкт-Петербургский государственный
электротехнический университет "ЛЭТИ"

АНАЛИЗ ОДНОРОДНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

Санкт-Петербург
Издательство СПбГЭТУ "ЛЭТИ"
2005

Федеральное агенство по образованию

Санкт-Петербургский государственный
электротехнический университет "ЛЭТИ"

АНАЛИЗ ОДНОРОДНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ

Учебное пособие

Санкт-Петербург
Издательство СПбГЭТУ "ЛЭТИ"
2005

УДК 519.2
ББК В17
А 64

А 64 Анализ однородных статистических данных: Учеб. пособие / Сост: Егоров В. А., Ингстер Ю. И., Лившиц А. Н., Малова И. Ю., Малов С. В. : Изд-во СПбГЭТУ "ЛЭТИ 2005. 56 с.

ISBN 5-7629-0669-8

А 64 Содержит обзор классических методов обработки однородных статистических данных. Предназначено студентам ФКТИ.

УДК 519.2
ББК В17

Рецензенты: кафедра высшей математики Санкт-Петербургского государственного университета растительных полимеров; д-р. физ.-мат. наук, профессор Невзоров В. Б. (СПбГУ).

Утверждено
редакционно-издательским советом университета
в качестве учебного пособия

ISBN 5-7629-0669-8

Введение

В данном учебном пособии рассматриваются некоторые классические методы статистического анализа данных. Предполагается, что читатель знаком с основными понятиями теории вероятностей, математического анализа, линейной алгебры, теории меры и интеграла.

Задача статистики состоит в сборе данных, их анализе и интерпретации. Методы математической статистики позволяют реализовать второй этап статистического анализа – анализ данных. Иными словами, по полученным данным требуется сделать те или иные выводы о свойствах статистического эксперимента, результатом которого они являются. Поскольку детальный достоверный ответ на вопрос о свойствах эксперимента обычно дать не удается, в математической статистике используются вероятностные методы. В некотором смысле задачи математической статистики являются обратными к задачам теории вероятностей. Если задача теории вероятностей заключается в изучении результатов наблюдений в условиях рассматриваемого эксперимента, то задача математической статистики состоит в изучении свойств эксперимента на основании полученных данных. Известный тезис, что критерий истины есть практика, имеет непосредственное отношение к математической статистике. Отметим, что при изучении свойств эксперимента, помимо полученных данных, могут использоваться также некоторые известные свойства эксперимента, которые образуют его математическую вероятностную модель. В рамках этой модели мы приступаем к анализу данных. Таким образом, выбор совокупности вероятностных моделей, лежащих в основе статистического эксперимента, является весьма важным фактором. Выбор слишком широкой совокупности приводит к невозможности серьезного статистического анализа, тогда как выбор слишком узкого семейства может вызвать естественные сомнения в его правомерности и привести к необоснованным статистическим выводам. Ясно, что на основании одного наблюдения и без введения каких-либо жестких предположений о свойствах статистического эксперимента невозможно сделать какие-либо серьезные выводы о его вероятностных свойствах. В этом смысле естественным решением является многократное проведение статистического эксперимента в одних и тех же жестких условиях. Обычно результат каждого отдельного эксперимента может быть записан в виде чис-

ла или набора чисел. Таким образом, простейшая модель статистического эксперимента строится на основании понятия выборки.

Набор независимых и одинаково распределенных случайных величин $X_1, \ldots, X_n$ с общим распределением P_θ, принадлежащим классу распределений $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$, будем называть однородной выборкой или просто *выборкой* из распределения P_θ. Материал данного методического пособия не будет выходить за рамки данного класса моделей статистического эксперимента.

Можно выделить три основных класса задач математической статистики – точечное оценивание параметра θ (или функции от параметра $g(\theta)$) распределения P_θ, интервальное оценивание, а также класс задач проверки статистических гипотез. Более подробно эти задачи будут сформулированы в соответствующих разделах данного учебного пособия. По характеру параметрического множества Θ среди статистических моделей рассматриваемого класса можно отметить параметрические и непараметрические модели. Семейство распределений $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ называется параметрическим, если $\Theta \subseteq \mathbb{R}^d$ при некотором значении d. В этом случае будем говорить, что семейство $\mathcal{P}$ является d-параметрическим, а d – размерность параметра. В противном случае будем говорить, что семейство распределений непараметрическое, а параметр θ бесконечномерный. В качестве примера непараметрического семейства распределений рассмотрим совокупность всех распределений на прямой $\mathbb{R}$. В этом случае естественно в качестве параметра выбрать функцию распределения. Другими примерами непараметрических семейств являются: совокупность всех абсолютно непрерывных распределений на прямой с параметром "плотность распределения" или совокупность всех дискретных распределений, где в качестве параметра выступает дискретная плотность (или совокупность вероятностей $p(i) = p_i = P(X_1 = i)$, $i \in \mathbb{N}$).

Важным свойством семейства распределений вероятностей является свойство доминированности некоторой σ-конечной мерой (μ, $P_\theta << \mu$) при любом значении $\theta \in \Theta$. Тогда, согласно теореме Радона–Никодима, существуют плотности $p_\theta \equiv \dfrac{dP_\theta}{d\mu}$ распределения (относительно доминирующей меры μ), такие, что

$$P_\theta(I) = \int\limits_I p_\theta(x) d\mu(x).$$

В качестве доминирующей меры обычно выбирают меру Лебега (для семейств абсолютно непрерывных распределений) или считающую меру (для семейств дискретных распределений, сконцентрированных на некотором не более чем счетном множестве).[1] В первом случае плотностью относительно меры Лебега является обычная плотность распределения, во втором случае плотностью относительно считающей меры являются соответствующие вероятности, образующие дискретную плотность распределения. Некоторые параметрические семейства распределений будут рассмотрены далее.

Работа с данными начинается с изучения некоторых эвристических оценок для характеристик случайных величин. Будет введено понятие эмпирического распределения и соответствующей функции распределения, а также гистограммы и полигона частот. Потом будет рассмотрена задача точечного и доверительного оценивания параметров соответствующего семейства вероятностных распределений. Далее последует ряд задач проверки статистических гипотез; будет рассмотрена задача построения наиболее мощного критерия проверки простой гипотезы при простой альтернативе. Затем мы займемся построением наиболее мощных критериев для односторонних альтернатив и рассмотрением случая двусторонних альтернатив. В заключение будут рассмотрены задачи проверки статистических гипотез с использованием непараметрических критериев χ^2 и Колмогорова.

1. Точечное оценивание

Пусть $X_1, \ldots, X_n$ – выборка из неизвестного распределения $P_\theta \in \mathcal{P}$. Задача точечного оценивания параметра состоит в том, чтобы на основании результатов наблюдений выбрать из множества $\mathcal{P}$ распределение, оптимальным образом соответствующее полученным данным. Часто требуется восстановить не все распределение P_θ, а лишь некоторые его числовые характеристики $g(\theta)$. В этом случае говорят о задаче оценивания функции $g(\theta)$.

Статистикой называется любая измеримая функция T, сопоставляющая каждому фиксированному набору наблюдений $X_1, \ldots, X_n$ элемент некоторого множества E, которое обычно является подмножеством неко-

[1] Отметим также, что любое конечное и даже счетное семейство может быть доминировано конечной мерой. В качестве доминирующей меры можно взять линейную комбинацию всех распределений с положительными коэффициентами, сумма которых конечна.

торого евклидова или функционального пространства. Каждая статистика порождает σ-алгебру на множестве наблюдений. Будем говорить, что статистики эквивалентны, если соответствующие порожденные σ-алгебры совпадают. Ясно, что если распределение статистики не зависит от оцениваемого параметра, то она не несет никакой информации о параметре. Такие статистики носят название *подчиненные.* В противоположность этому, статистика называется *достаточной*, если условные вероятности при условии, что значение данной статистики известно, не зависят от оцениваемого параметра. Иными словами, при известном значении достаточной статистики любая статистика является подчиненной, а следовательно, вся информация об оцениваемом параметре содержится в достаточной статистике. При этом естественно стремление уменьшить количество хранимой информации. Достаточная статистика U называется минимальной, если для любой другой достаточной статистики T существует отображение f такое, что $U = f(T)$.[1]

1. Пусть $X_1, \ldots, X_n$ – набор наблюдений (случайных величин). Отображение $T : \mathbb{R}^n \to \mathbb{R}^n$, сопоставляющее исходной совокупности *вариационный ряд* $X_{(1)} \leq \ldots \leq X_{(n)}$ (т. е. тот же набор величин, но упорядоченных по возрастанию), является статистикой. Элементы вариационного ряда называются *порядковыми статистиками.* Так, вариационный ряд выборки (3, 1, 4, 2, 3) имеет вид (1, 2, 3, 3, 4). Несмотря на то что вариационный ряд состоит из тех же элементов, что и исходный набор наблюдений, при переходе к вариационному ряду теряется информация о том, в каком порядке следовали элементы исходной выборки. Чтобы восстановить эту информацию, используются ранги $(R_1, \ldots, R_n)$, где R_i – порядковый номер элемента X_i в вариационном ряду, $i = 1, \ldots, n$. Отметим, что вектор рангов представляет собой перестановку чисел $(1, \ldots, n)$. Ранги элементов рассмотренного набора наблюдений – (3, 1, 5, 2, 4) (с аналогичным успехом в качестве вектора рангов можно взять (4, 1, 5, 2, 3), так как если наблюдения повторяются, то выбор рангов неоднозначен). Отметим, что если исходный набор наблюдений является выборкой, то вектор рангов – подчиненная статистика, тогда как вариационный ряд – достаточная статистика.[2]

[1] Имеется в виду, конечно, что U является T-измеримой.

[2] В данном учебном пособии рассматриваются лишь однородные независимые выборки, поэтому ранги использоваться не будут.

Оценкой параметра θ называется статистика, отображающая выборку в множество $E = \Theta$, т. е. сопоставляющая каждому набору наблюдений некоторое значение параметра $\delta(\vec{X}) \in \Theta$. Аналогично, оценкой параметрической функции $g(\theta)$ является статистика вида $\tau = \tau(\vec{X}) \in \mathbb{R}$.

Для измерения степени отклонения значения оценки от теоретического значения параметра вводится функция потерь $W(\delta, \theta)$, обладающая свойствами $W(\delta, \theta) \geq 0$ и $W(\theta, \theta) = 0$. Однако поскольку значение θ неизвестно, отклонение также является неизвестной величиной. В качестве критерия качества оценки вводится функция риска

$$R(\theta, \delta) = \mathbf{E}_\theta W(\delta(\theta), \theta).$$

Аналогично, в задаче оценивания функции $g(\theta)$ качество оценки $\tau = \tau(\vec{X})$ характеризуется функцией риска $R(\theta, \tau) = E_\theta(W(\tau, g(\theta))$. Здесь обычно функция потерь имеет вид $W(\tau, g(\theta)) = l(|\tau - g(\theta)|)$, где $l(0) = 0,\ l(t) \geq 0$ и $l(t)$ не убывает по $t \geq 0$. Например, если $l(t) = 0$ при $t \leq \varepsilon$ и $l(t) = 1$ при $t > \varepsilon,\ \varepsilon > 0$, то функция риска $R_\varepsilon(\theta, \tau)$ есть *вероятность отклонений* оценки τ от оцениваемой функции $g(\theta)$ за уровень $\varepsilon > 0$:

$$R_\varepsilon(\theta, \tau) = P_\theta(|\tau - g(\theta)| > \varepsilon).$$

При $l(t) = t^2$ мы получаем *квадратичный риск* оценки

$$R^{(2)}(\theta, \tau) = \mathbf{E}_\theta(\tau - g(\theta))^2.$$

Используя неравенство Маркова, легко получить, что

$$R_\varepsilon(\theta, \tau) \leq R^{(2)}(\theta, \tau)/\varepsilon^2, \quad \forall\ \varepsilon > 0.$$

Это неравенство позволяет оценивать вероятности отклонений через квадратичный риск.

Часто для характеристики качества оценки τ рассматривают ее смещение $b(\tau, g(\theta)) = \mathbf{E}_\theta(\tau) - g(\theta)$ и дисперсию $\mathbf{D}_\theta(\tau, g(\theta)) = \mathbf{D}_\theta(\tau)$. Оценка τ называется *несмещенной*, если $b(\tau, g(\theta)) = 0,\ \forall\ \theta \in \Theta$. При этом для квадратичного риска справедливо представление

$$R^{(2)}(\theta, \tau) = b^2(\tau, g(\theta)) + \mathbf{D}_\theta(\tau).$$

Те же понятия используются для характеристики качества оценки $\delta = \delta(\vec{X})$ одномерного параметра $\theta \in \mathbb{R}$, поскольку здесь можно рассмотреть функцию $g(\theta) = \theta$.

Следует отметить, что попытка минимизировать риск при каждом θ, не накладывая никаких ограничений на класс рассматриваемых оценок,

заведомо терпит неудачу. Действительно, оценка $\delta(\vec{X}) = \theta_0$, минимизирующая риск при $\theta = \theta_0$, не обладает таким свойством при других значениях θ и является абсолютно бессмысленной со статистической точки зрения, поскольку сама статистика не несет никакой информации о параметре.

Один из путей преодоления данной сложности – сужение класса рассматриваемых оценок. Пусть пространство Θ достаточно "хорошее" и снабжено некоторой нормой $\|\cdot\|$ для измерения близости θ и $\delta(\vec{X})$. Естественное предположение о том, что с ростом числа наблюдений оценка должна становиться точнее и точность ее должна стремиться к абсолютной, находит отражение в следующем понятии. Оценка $\delta_n = \delta_n(\vec{X})$ параметра θ называется *состоятельной*, если при каждом значении $\theta \in \Theta$ для любого $\varepsilon > 0$

$$P_\theta(\|\delta_n - \theta\| > \varepsilon) \xrightarrow[n\to\infty]{} 0,$$

и *сильно состоятельной*, если $\delta_n \to \theta$ при $n \to \infty$ с вероятностью 1.

Аналогично, оценка $\tau_n = \tau_n(\vec{X})$ функции $g(\theta)$ называется *состоятельной*, если при $n \to \infty$, $R_\varepsilon(\theta, \tau_n) \to 0 \quad \forall\, \varepsilon > 0,\ \theta \in \Theta$, и $\sqrt{n}$-*состоятельной*, если это соотношение выполнено для любой последовательности $\varepsilon_n \to 0$, такой что $\sqrt{n}\varepsilon_n \to \infty$. Оценка $\tau_n = \tau_n(\vec{X})$ называется *состоятельной в среднеквадратичном*, если $R^{(2)}(\theta, \tau_n) \to 0$ при $n \to \infty$, и $\sqrt{n}$-*состоятельной в среднеквадратичном*, если последовательность $nR^{(2)}(\theta, \tau_n)$ ограничена при $n \to \infty$ для всех $\theta \in \Theta$. Из неравенства, связывающего вероятности уклонений и квадратичный риск, следует, что состоятельность в среднеквадратичном влечет состоятельность, а $\sqrt{n}$-состоятельность в среднеквадратичном влечет $\sqrt{n}$-состоятельность оценки τ_n.

Для более тонкого исследования асимптотических свойств оценок при $n \to \infty$ полезно понятие асимптотической нормальности оценок. Оценка $\tau_n = \tau_n(\vec{X})$ называется *асимптотически нормальной с нормирующим множителем* $\sigma(\theta) > 0$, если при всех $\theta \in \Theta$ имеет место сходимость по P_θ-распределению

$$\sqrt{n}(\tau_n - g(\theta))/\sigma(\theta) \to \xi,\ n \to \infty,$$

где $\xi \sim N(0,1)$ есть стандартная нормальная случайная величина. Если, кроме того, имеет место сходимость вторых моментов,

$$n\mathbf{E}_\theta(\tau_n - g(\theta))^2/\sigma^2(\theta) \to E\xi^2 = 1,\ n \to \infty,$$

то оценку будем называть сильно асимптотически нормальной.

Асимптотически нормальные оценки являются $\sqrt{n}$-состоятельными. Для асимптотически нормальных оценок можно написать предельное вы-

ражение для вероятностей превышения уровня $\varepsilon_{n,\theta} = t\sigma(\theta)/\sqrt{n},\ t > 0$:

$$P_\theta(|\tau_n - g(\theta)| > t\sigma(\theta)/\sqrt{n}) \to P(|\xi| > t) = 1 - 2\Phi(-t), \quad n \to \infty,$$

где $\Phi(t)$ есть функция распределения стандартного нормального закона. Это предельное соотношение используют для оценки вероятностей отклонений за малый уровень порядка $1/\sqrt{n}$ при больших n.

Для сильно асимптотически нормальных оценок справедливо предельное выражение для квадратичного риска

$$nR^{(2)}(\theta, \tau_n) \to \sigma^2(\theta), \quad n \to \infty,$$

то есть $R^{(2)}(\theta, \tau_n) \approx \sigma^2(\theta)/n$ при больших n.

Эти соотношения позволяют сравнивать качество асимптотически нормальных оценок при больших n по величине нормирующего множителя $\sigma(\theta)$: чем меньше $\sigma(\theta)$, тем лучше оценка.

2. Эмпирическое распределение

Пусть $X_1, \ldots, X_n$ – выборка из распределения $P_\theta,\ \theta \in \Theta$. Истинное значение P_θ будем называть *теоретическим распределением*. Мы, естественно, предполагаем, что истинное значение параметра принадлежит Θ.

По исходной выборке построим дискретное распределение, имеющее атомы $1/n$ в точках $X_1, \ldots, X_n$. Оно называется *эмпирическим распределением*, построенным по данной выборке, а соответствующая функция распределения F_n называется *эмпирической функцией распределения*. Эмпирическая функция распределения имеет вид

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i < x\}},\ \ x \in \mathbb{R},$$

где

$$\mathbb{1}_A = \begin{cases} 1, \text{ если событие } A \text{ имеет место;} \\ 0, \text{ если событие } A \text{ не имеет места.} \end{cases}$$

Иными словами, значение эмпирической функции распределения в точке x есть отношение числа наблюдений, меньших x, к общему числу наблюдений.

Эмпирическая функция распределения является статистикой, и ее часто рассматривают в качестве оценки для истинной (или теоретической) функции распределения элементов выборки. Отметим, что эмпирическая

функция распределения однозначно определяет вариационный ряд, поэтому является достаточной статистикой в рассматриваемой модели. В терминах порядковых статистик эмпирическая функция распределения может быть записана в виде

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_{(i)} < x\}}, \quad x \in \mathbb{R}.$$

Для любого фиксированного $x \in \mathbb{R}$ значение $F_n(x)$ эмпирической функции распределения является несмещенной оценкой теоретической функции распределения $F(x)$; квадратичный риск этой оценки есть $R^{(2)}(F, F_n(x)) = = F(x)(1 - F(x))/n$; оценка $F_n(x)$ является сильно асимптотически нормальной с нормирующим множителем $\sigma^2(F) = F(x)(1 - F(x))$.

Рассмотрим отклонение эмпирической функции распределения от теоретической $D_n(\vec{X}) = \sup\limits_{-\infty<x<\infty} |F_n(x) - F(x)|$. Согласно теореме Гливенко–Кантелли, для эмпирической функции распределения F_n, построенной по выборке из распределения F, справедливо асимптотическое соотношение

$$\lim_{n\to\infty} D_n(\vec{X}) = \lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = 0$$

с вероятностью 1, что позволяет рассматривать эмпирическую функцию распределения как состоятельную оценку теоретической функции распределения (в равномерной норме).

Для выборки из дискретного распределения, носитель которого не имеет точек сгущения (например, распределения со значениями на множестве целых чисел), эмпирическое распределение может использоваться для оценивания дискретной плотности распределения. Для этого следует ввести частоты $\upsilon(x)/n$, где $\upsilon(x)$ – число наблюдений, имеющих значение x. С ростом n, согласно закону больших чисел, частоты будут сходиться к теоретическим значениям $q_\theta(x) = P_\theta(X_1 = x)$. С распределением частот и некоторых функций от частот мы еще встретимся при построении критерия χ^2. Кроме того, приближая распределения дискретными распределениями и используя частоты в качестве приближения для вероятностей, мы получаем во многих практических ситуациях наглядный способ представления данных. Например, в случае целочисленных данных отрезок вещественной прямой между наименьшим и наибольшим наблюдениями разбивается на равные интервалы I_j единичной длины с центрами в j, $j \in \mathbb{N}$. Функция

$h(x) = \upsilon(j)/n$, $x \in I_j$, носит название *гистограмма частот*. В случае если выборка взята из абсолютно непрерывного распределения, гистограмма частот строится следующим образом. Вещественная прямая разбивается на интервалы одинаковой длины $h > 0$. Пронумеруем соответствующие интервалы $\{I_i\}_{i\in\mathbb{N}}$ (их счетное число) натуральными числами. Далее вычислим выборочные аналоги вероятностей попадания в соответствующие интервалы υ_i/n, где υ_i – число наблюдений попавших в i-й интервал. Далее определим функцию $H(x;h) = \upsilon_i/(nh)$, $x \in I_i$, $i \in \mathbb{N}$, которая называется *гистограмма*. Отметим, что площадь подграфика данной функции равна единице. Более того, $\upsilon_i/(nh) \to P_\theta(X_1 \in I_i)$ по вероятности (и с вероятностью 1) при $n \to \infty$ при каждом фиксированном θ. Таким образом, если плотность непрерывна, то $H(x;h)$, $x \in I_j$, является оценкой некоторого среднего значения плотности на интервале I_j. Уменьшая подходящим образом $h = h(n)$ (так что $nh(n) \xrightarrow[n\to\infty]{} \infty$, $h(n) \to 0$), получаем состоятельную оценку теоретической плотности распределения. Если функция плотности достаточно гладкая, то ломаными ее график можно приблизить лучше, чем ступенчатыми функциями. Отсюда следует, что для оценки гладких плотностей целесообразно использовать так называемый *полигон частот* вместо гистограммы. Полигон частот – кусочно-линейная непрерывная функция, совпадающая с гистограммой частот в середине каждого интервала и линейная между серединами двух соседних интервалов. Нетрудно видеть, что площадь подграфика полигона частот тоже равна единице.[1]

3. Выборочные числовые характеристики

Напомним, что числовой характеристикой называется отображение из подмножества множества всех распределений в вещественную прямую, сопоставляющее каждому распределению некоторое вещественное число по определенному правилу. Наиболее важными числовыми характеристиками являются:

1) математическое ожидание $\mathbf{E}_\theta X = \int\limits_{-\infty}^{\infty} x dP_\theta(x)$;

[1] Более современные методы оценивания теоретической плотности распределения основаны на построении "ядерных оценок" вида $f_n(x) = (nh_n)^{-1} \sum\limits_{i=1}^{n} K\Big((x - X_i)/h_n\Big)$, где K – некоторая функция ограниченной вариации ("ядро").

2) дисперсия $\mathbf{D}_\theta X = \mathbf{E}_\theta(X - \mathbf{E}_\theta X)^2 = \mathbf{E}_\theta X^2 - (\mathbf{E}_\theta X)^2$ и среднеквадратическое отклонение $\sqrt{\mathbf{D}_\theta X}$;

3) квантили $\zeta_\theta(p)$, определяемые соотношениями $P_\theta(X \le \zeta_\theta(p)) \ge p$ и $P_\theta(X \ge \zeta_\theta(p)) \ge 1 - p$. (Среди квантилей наиболее часто используются выборочная медиана $\zeta_\theta(1/2)$ и квартили $\zeta_\theta(1/4)$ и $\zeta_\theta(3/4)$.)

Рассмотрим также другие числовые характеристики:

1) моменты, центральные моменты и абсолютные моменты порядка k: $\alpha_k = \mathbf{E}_\theta X^k$, $\mu_k = \mathbf{E}_\theta(X - \mathbf{E}X)^k$ и $\mathbf{E}_\theta|X|^k$, $k \in \mathbb{N}$;

2) асимметрия: $\mathbf{Asi}(\theta) = \mathbf{E}_\theta(X - \mathbf{E}_\theta X)^3/(\mathbf{D}_\theta X)^{3/2} = \mu_3/(\mu_2)^{3/2}$;

3) эксцесс: $\mathbf{Ex}(\theta) = (\mathbf{E}_\theta(X - \mathbf{E}_\theta X)^4)/(\mathbf{D}_\theta X)^2 - 3 = \mu_4/\mu_2^2 - 3$.

Понятно, что в условиях статистического эксперимента числовые характеристики, вообще говоря, зависят от параметра, а потому неизвестны. Наряду с теоретическими числовыми характеристиками можно рассматривать их выборочные аналоги, т. е. соответствующие числовые характеристики эмпирического распределения. Приведем примеры выборочных числовых характеристик.

1. Выборочное математическое ожидание: $\overline{X} = \int\limits_{-\infty}^{\infty} x dF_n(x) = n^{-1} \sum\limits_{i=1}^{n} X_i$,

2. Выборочная дисперсия: $s^2 = \int\limits_{-\infty}^{\infty} x dF_n(x) - \overline{X}^2 = n^{-1} \sum\limits_{i=1}^{n} X_i^2 - \overline{X}^2$,

3. Выборочная медиана:

$$Z_n = \begin{cases} \alpha X_{(n/2+1)} + (1-\alpha) X_{(n/2)}, & \text{если } n - \text{четное}, \\ X_{((n+1)/2)}, & \text{если } n - \text{нечетное}, \end{cases}$$

где $\alpha \in [0, 1]$ может быть выбрано произвольным.

Аналогично определяются выборочные начальные и центральные моменты $\alpha_{k,n}$, $\mu_{k,n}$.

Рассмотрим статистические свойства выборочных моментов. Обозначим $\mathcal{F}_k$ множество функций распределения, для которых конечны абсолютные момента k-го порядка, $k > 0$. Выборочные начальные моменты $\alpha_{k,n}$ являются несмещенными состоятельными оценками начальных моментов $\alpha_k(F)$, $F \in \mathcal{F}_k$. Если $F \in \mathcal{F}_{2k}$, то эти оценки асимптотически нормальны с нормирующим множителем $\sigma^2(F) = \alpha_{2k} - \alpha_k^2$; квадратичный риск для этих оценок есть $\sigma^2(F)/n$. Например, выборочное среднее $\overline{X}$ есть несмещенная состоятельная оценка математического ожидания, если оно существует. Если существует второй центральный момент, то выборочное среднее

асимптотически нормально и нормирующим множителем является среднеквадратическое отклонение $\sigma(F) = \sqrt{\mathbf{D}_F(X)}$.

Выборочные центральные моменты уже не являются несмещенными оценками. Так, смещение выборочной дисперсии s^2 есть $b(s^2, F) = \mathbf{E}_F s^2 - \mathbf{D}_F(X) = -\mathbf{D}_F(X)/n$. Можно легко построить несмещенный вариант оценки дисперсии: $ns^2/(n-1)$, который обычно используется на практике. Асимптотические свойства этой оценки одинаковы: они состоятельны при $F \in \mathcal{F}_2$, если $F \in \mathcal{F}_4$, то они асимптотически нормальны с нормирующим множителем $\sigma^2(F) = \mu_4 - \mu_2^2$.

Выборочная медиана Z_n как оценка медианы $z = \zeta_F(1/2)$ асимптотически нормальна в классе функций распределения F, имеющих плотность распределения $f(x)$ непрерывную и положительную в точке $x = z$, с нормирующим множителем $\sigma(F) = 1/(2f(z))$.

Отметим, что асимметрия симметричного распределения равна 0, хотя не любое распределение с нулевой асимметрией является симметричным. Тем не менее асимметрия используется в качестве характеристики симметричности распределения. Эксцесс же характеризует, в некотором роде, степень отличия распределения от нормального, у которого данная числовая характеристика равна нулю. В то же время равенство нулю эксцесса распределения совсем не говорит о его нормальности.

Выборочные асимметрия $\mathbf{Asi}_n = \mu_{3,n}/\mu_{2,n}^{3/2}$ и эксцесс $\mathbf{Ex}_n = \mu_{4,n}/\mu_{2,n}^2 - 3$ также есть состоятельные и асимптотически нормальные оценки асимметрии и эксцесса. Эти характеристики обычно используются для поверки гипотезы о нормальности распределения выборки. Именно, если X имеет нормальное распределения, то случайные величины $\xi_n = \mathbf{Asi}_n\sqrt{n/6}$ и $\eta_n = \mathbf{Ex}_n\sqrt{n/24}$ имеют в пределе стандартное нормальное распределение. Поэтому большие по абсолютной величине значения ξ_n или η_n (скажем, $|\xi_n| > 3$ или $|\eta_n| > 3$) говорят о вероятном отклонении закона распределения от нормального (см. 6).

4. Параметрические классы распределений

В данном разделе рассматриваются некоторые семейства распределений, использующиеся в математической статистике, и устанавливаются связи между ними. Прежде чем перейти к рассмотрению наиболее важных параметрических семейств распределений, рассмотрим один естественный

метод построения семейств распределений путем сдвига и масштабирования. В основном, данный метод характерен для абсолютно непрерывных распределений.

Выберем некоторое "базовое" абсолютно непрерывное распределение с плотностью распределения p_0. Рассмотрим семейство распределений с плотностями вида $p_{a,b}$ такими, что

$$p_{a,b}(x) = b^{-1}p_0((x-a)/b),\ x \in \mathbb{R},$$

$a \in \mathbb{R}$, $b > 0$. В этом случае a носит название *параметр сдвига*, а b – *параметр масштаба*. Нетрудно убедиться, что если случайная величина X имеет базовое распределение с плотностью p_0, то величина $Y = bX + a$ имеет распределение с плотностью $p_{a,b}$. Можно получить формулы, связывающие некоторые числовые характеристики различных распределений данного семейства. В частности $\mu_k(a,b) = b^k\mu_{0k}$, где $\mu_k(a,b)$ и μ_{0k} – k-й центральный момент распределения с параметрами a и b и "базового" распределения соответственно, а следовательно, асимметрия и эксцесс не зависят от a и b. Что касается математического ожидания и квантилей, то они связаны следующими соотношениями: $\alpha_1(a,b) = a + b\alpha_{01}$, $\zeta_{a,b}(p) = a + b\zeta_0(p)$. Для характеристических функций будет выполнено равенство $f_{a,b}(t) = \exp(ita)f_0(bt)$.

Основные характеристики рассматриваемых распределений даны в прил. 2. Центральную роль в математической статистике играет семейство *нормальных* распределений $\mathcal{N}(a, \sigma^2)$ (см. прил. 2). Очевидно, что параметр $a \in \mathbb{R}$ нормального распределения является параметром сдвига, а параметр $\sigma > 0$ – параметром масштаба. Роль базового распределения играет стандартное нормальное распределение $\mathcal{N}(0,1)$.

Выборка из нормального распределения. Пусть $\xi, \xi_1, \ldots, \xi_n$ – выборка из стандартного нормального распределения $\mathcal{N}(0,1)$. Распределение случайной величины $\chi^2 = \sum\limits_{i=1}^{n} \xi_i^2$ носит название χ^2-*распределения* с n степенями свободы, а распределение случайной величины $\xi\Big(\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i^2\Big)^{-1/2}$ носит название распределения Стьюдента с n степенями свободы. Конечно, эти два типа распределений были введены искусственно для работы с нормально распределенными случайными величинами.

Пусть $X_1, \ldots, X_n$ – выборка из нормального распределения $\mathcal{N}(a, \sigma^2)$. По известной лемме Фишера

1) $\sqrt{n}\,\dfrac{\overline{X} - a}{\sigma}$ имеет стандартное нормальное $\mathcal{N}(0,1)$ распределение;

2) $\overline{X}$ и s^2 – независимые статистики;

3) $\dfrac{ns^2}{\sigma^2}$ имеет χ^2-распределение с $n-1$ степенью свободы;

4) $\sqrt{n-1}\,\dfrac{\overline{X} - a}{s}$ имеет распределение Стьюдента с $n-1$ степенью свободы.

Далее будет показано, что χ^2-распределения включаются в более широкое семейство γ-распределений, о котором разговор пойдет позднее. Что касается распределения Стьюдента, то оно симметрично относительно нуля и при большом числе степеней свободы данное распределение мало отличается от стандартного нормального.

Введем еще одно распределение, связанное с выборкой из стандартного нормального распределения. Распределение *Фишера–Снедекора* $F(n_1, n_2)$ определяется как распределение случайной величины $\dfrac{n_2}{n_1}\,\zeta$, где $\zeta = \eta_1/\eta_2$, а η_1, η_2 – независимые случайные величины, имеющие $\chi^2_{n_1}$- и $\chi^2_{n_2}$-распределения соответственно.

Замечание. Если $X_1, \ldots, X_{n_1}$ и $Y_1, \ldots, Y_{n_2}$ – две независимые нормальные выборки из $N(a_1, \sigma_1^2)$ и $N(a_2, \sigma_2^2)$, а s_1^2 и s_2^2 – выборочные дисперсии, полученные по первой и второй выборке соответственно, то отношение $\dfrac{n_1(n_2-1)\sigma_2^2 s_1^2}{n_2(n_1-1)\sigma_1^2 s_2^2}$ имеет распределение Фишера–Снедекора $F(n_1-1, n_2-1)$.

γ-распределения. Неполная гамма-функция при $x > 0$ и гамма-функция Эйлера определяются равенствами

$$\Gamma_x(p) = \int\limits_0^x t^{p-1}\exp(-t)\,dt \quad \text{и} \quad \Gamma(p) = \Gamma_\infty(p)$$

соответственно. Гамма-функция обладает следующими свойствами:

1. $\Gamma(1) = 1$, $\Gamma(1/2) = \sqrt{\pi}$;

2. $\Gamma(p+1) = p\,\Gamma(p)$ для всех $p > 0$.

Следовательно, $(n-1)! = \Gamma(n)$.

Если рассматривать значение x как переменную, то $\Gamma_x(p)/\Gamma(p)$, до-

определенная нулем при $x < 0$, является функцией распределения при каждом фиксированном $p > 0$. Семейство γ-*распределений* объединяет все распределения, имеющие такие функции распределения, и пополняется введением параметра масштаба b (прил. 2). Замечательное свойство γ-распределений состоит в том, что распределение суммы $X+Y$ двух независимых случайных величин $X \sim \Gamma(b, q_1)$ и $Y \sim \Gamma(b, q_2)$[1] с одинаковым параметром масштаба снова является γ-распределением $\Gamma(b, q_1 + q_2)$. Это следует непосредственно из соотношения $f_{b,q_1+q_2}(t) = f_{b,q_1}(t) f_{b,q_2}(t)$ для характеристических функций γ-распределений $\Gamma(b, q_1+q_2)$, $\Gamma(b, q_1)$ и $\Gamma(b, q_2)$ соответственно. Рассмотрим следующую задачу.

1. Пусть случайная величина ξ имеет стандартное нормальное распределение $\mathcal{N}(0, 1)$. Найдем распределение случайной величины $\eta = \xi^2$. Носитель распределения η будет совпадать с интервалом $[0, \infty)$. Таким образом, $p_\eta(x) = 0$ при $x \le 0$. Отметим, что преобразование $g : g(x) = x^2$, является монотонным на $(-\infty, 0]$ и на $[0, \infty)$. Тогда плотность распределения η выражается через плотность распределения ξ следующим образом:

$$\begin{aligned} p_\eta(t) &= |(g_-^{-1}(t))'|\, p_\xi(g_-^{-1}(t)) + |(g_+^{-1}(t))'|\, p_\xi(g_-^{-1}(t)) = \\ &= \frac{1}{2\sqrt{t}} \frac{1}{\sqrt{2\pi}} \exp(-t/2) + \frac{1}{2\sqrt{t}} \frac{1}{\sqrt{2\pi}} \exp(-t/2) = \frac{1}{\sqrt{2\pi t}} \exp(-t/2),\ t > 0. \end{aligned}$$

Итак, плотность распределения случайной величины η имеет вид

$$p_\eta(x) = \frac{1}{\sqrt{2\pi x}} \exp(-x/2) \mathbb{I}_{[0,\infty)}(x).$$

Следовательно, $\eta \sim \Gamma(2, 1/2)$.

Используя результат последней задачи и замечательное свойство γ-распределений, заключаем, что χ^2_n распределение является γ-распределением $\Gamma(2, n/2)$. Другим важным подсемейством данного семейства распределений является однопараметрическое семейство *показательных* распределений $\mathrm{E} = \{\mathrm{E}(b),\ b > 0\} = \{\Gamma(b, 1),\ b > 0\}$.[2]

t-распределение. Основные характеристики t-распределений даны в прил. 2. Отметим, что $\mathrm{T}(b, p)$ – симметричное распределение. Последовательность распределений $\mathrm{T}(b, p)$ сходится к $\mathcal{N}(0, b^2)$ при $p \to \infty$. Параметр

[1] $X \sim \Gamma(b, q)$ означает, что случайная величина X имеет γ-распределение с параметром формы, равным p и параметром масштаба, равным b.

[2] В некоторых ситуациях имеет смысл расширить семейство показательных распределений путем введения дополнительного параметра сдвига.

b является параметром масштаба. Чтобы пояснить значение параметра p рассмотрим следующую задачу.

2. Пусть ξ, η – независимые случайные величины; $\xi \sim \mathcal{N}(0, \alpha/2)$, $\eta \sim$ $\sim \Gamma(\alpha, p/2)$. Вычислим распределение случайной величины $\upsilon = \xi/\sqrt{\eta/p}$. Принимая во внимание монотонность квадратного корня на положительной полуоси, для вычисления плотности величины $\zeta = \sqrt{\eta}$ можно воспользоваться стандартной формулой $p_\zeta(x) = 2x\,\gamma(x^2; \alpha, p/2)$, где $\gamma(\,\cdot\,; \alpha, p/2)$ – плотность γ-распределения $\Gamma(\alpha, p/2)$. Величина $\sqrt{p}\,\xi$ имеет плотность $p_{\sqrt{p}\,\xi}(x) = \sqrt{2}(\alpha p)^{-1/2} \cdot \varphi(2x/p\alpha)$, где φ – плотность стандартного нормального распределения. Тогда плотность распределения частного равна

$$p_\upsilon(x) = \int\limits_{-\infty}^{\infty} \frac{y\sqrt{2}}{\sqrt{\alpha p}}\,\varphi\Big(\frac{2xy}{p\alpha}\Big) 2y\,\gamma(y^2; \alpha, p/2)dy = \frac{\int\limits_0^\infty 2y^p \exp\Big(-\big(1+\frac{2x^2}{\alpha p}\big)\,\frac{y^2}{\alpha}\Big)\,dy}{\sqrt{\pi p}\;\alpha^{p/2+1/2}\,\Gamma(p/2)}.$$

После замены переменных $u = (1 + 2x^2/(\alpha p))y^2/\alpha$ получаем, что

$$p_\upsilon(x) = \frac{\int\limits_0^\infty u^{p/2-1/2}\exp(-u)\,du}{\sqrt{\pi p}\,\Gamma\left(\frac{p}{2}\right)(1+2x^2/(\alpha p))^{p/2+1/2}} = \frac{\Gamma\Big((p+1)/2\Big)}{\sqrt{\pi p}\,\Gamma\left(\frac{p}{2}\right)(1+2x^2/(\alpha p))^{p/2+1/2}}.$$

Отметим, что распределение $\mathrm{T}(2, n)$ – распределение Стьюдента с n степенями свободы.

Двухстороннее показательное (Лапласа) распределение. Основные характеристики этого и других распределений приведены в прил. 2. Очевидно, что распределение Лапласа $\mathrm{DE}\,(a, b)$ обладает свойством симметричности относительно точки a. Согласно введенной терминологии параметр $a \in \mathbb{R}$ является параметром сдвига, а параметр $b > 0$ – параметром масштаба. Отметим, что если $X \sim \mathrm{E}\,(1)$ и $Y \sim \mathrm{E}\,(1)$ – независимые случайные величины, то $X - Y \sim \mathrm{DE}\,(0, 1)$.

Равномерное распределение $U(a, b)$. В отличие от распределений, рассмотренных ранее, равномерное распределение имеет ограниченный носитель $[a, b]$. Пусть случайная величина ξ имеет непрерывное распределение с функцией распределения F. Тогда по теореме Смирнова случайная величина $F(\xi)$ имеет равномерное $U(0, 1)$ распределение. В этом смысле равномерное распределение является "связующим" для класса всех непрерывных распределений.

Рассмотрим также ряд дискретных распределений. Дискретное распределение, сконцентрированное в точках 0 и 1, носит название *распределение Бернулли*. Параметр $p \in (0,1)$ распределения Бернулли равен вероятности попадания в 1.

Выборка из распределения Бернулли. Пусть $X_1, \ldots, X_k$ – выборка из распределения Бернулли с параметром p. Согласно формуле Бернулли, сумма $Y = \sum\limits_{i=1}^{k} X_i$, т. е. число единиц (успехов) в схеме Бернулли, имеет дискретное распределение $\mathrm{Bi}(p,k)$, сконцентрированное в точках $\{0, \ldots, k\}$, носящее название *биномиальное распределение*. Биномиальные распределения образуют семейство $\{\mathrm{Bi}(p,k),\ k \in \mathbb{N},\ p \in (0,1)\}$. Отметим, что на практике обычно рассматриваются однопараметрические семейства биномиальных распределений при фиксированном k, например семейство распределений Бернулли получается при $k = 1$.

Распределение Пуассона. Рассмотрим последовательность биномиальных распределений $\mathrm{Bi}(p,k)$, $k \in \mathbb{N}$ и $p = \lambda/k$. Согласно теореме Пуассона, имеет место сходимость данной последовательности распределений при $m \to \infty$ к распределению $\mathrm{P}(\lambda)$, сконцентрированному в целых неотрицательных точках, которое называется *распределение Пуассона*. Отметим, что если $X \sim \mathrm{P}(\lambda_1)$ и $Y \sim \mathrm{P}(\lambda_2)$ – независимые случайные величины, имеющие распределение Пуассона, то $X + Y \sim \mathrm{P}(\lambda_1 + \lambda_2)$.

Геометрическое распределение. Теперь рассмотрим случайные величины, описывающие длину серии из единиц до первого появления нуля в последовательности независимых случайных величин, имеющих одинаковое распределение Бернулли $\mathrm{Bi}(p,1)$. Носитель распределения данной случайной величины, как и у распределения Пуассона – целые неотрицательные числа, а вероятности образуют геометрическую прогрессию $p^i(1-p)$, $i = 0, 1, \ldots$. Распределения такого вида называются *геометрическими*. Данные распределения входят в класс отрицательных биномиальных (Паскаля) распределений $\{\mathrm{Nb}(p,m),\ p \in (0,1),\ m \in \mathbb{N}\}$.

Рассмотрим, также некоторые классы многомерных распределений.

Мультиномиальное распределение. Пусть $X_1, \ldots, X_n$ – выборка из распределения, сконцентрированного в точках $\{1, \ldots, s\}$, с атомами $p_1, \ldots, p_s$, $p_1 + \ldots + p_s = 1$, $p_i > 0$, $i = 1, \ldots, s$. Тогда распределение

величин $(\upsilon_1, \ldots, \upsilon_s)$, $\upsilon_k = \sum\limits_{i=1}^{n} \mathbb{1}_{\{X_i=k\}}$, $k = 1, \ldots, n$ является *мультиномиальным* (прил. 3).

Многомерное нормальное распределение является предельным для распределений нормированных сумм случайных векторов. В частности, распределение (мультиномиальное) вектора нормированных частот $(n_1^*, \ldots, n_{N-1}^*) = \vec{n}^*$, $n_i^* = (\upsilon_i - np_i)/\sqrt{n}$ стремится к вырожденному нормальному распределению $N(0, \Sigma = ||\sigma_{i,j}||_1^{N-1})$, где $\sigma_{i,i} = p_i(1-p_i)$, $\sigma_{i,j} = -p_i p_j$, $i \neq j$ (ковариации мультиномиального распределения). Известно также, что квадратичная форма $Y^{\mathrm{T}}\Sigma^{-1}Y$, стоящая в показателе экспоненты, выражающей плотность невырожденного нормального распределения $N(0, \Sigma)$, распределена по закону χ^2_{N-1}.[1] Данное свойство лежит в основе χ^2 критерия.

Экспоненциальные семейства. Семейства распределений (одномерных или многомерных), доминированных мерой μ (σ-конечной) и параметризованное m-мерным параметром $\theta = (\theta_1, \ldots, \theta_m) \in \Theta$ с плотностями (относительно меры μ), допускающими представление

$$p_\theta(\vec{x}) = h(\vec{x})\exp\Big(-\sum_{j=1}^{s} \eta_j(\theta)T_j(\vec{x}) + r(\theta)\Big),$$

называются m-параметрическими *экспоненциальными семействами ранга* s. Свойство принадлежности теоретического распределения к экспоненциальному семейству естественным образом сохраняется и для выборки.

3. Пусть $X_1, \ldots, X_n$ – выборка из экспоненциального семейства, плотности которого допускают представление

$$p_\theta(x) = h(x)\exp\Big(-\sum_{j=1}^{s} \eta_j(\theta)T_j(x) + r(\theta)\Big).$$

Тогда случайный вектор $(X_1, \ldots, X_n)$ имеет плотность распределения

$$p_\theta(\vec{x}) = \prod_{i=1}^{n} h(x_i)\exp\Big(-\sum_{i=1}^{n}\sum_{j=1}^{s} \eta_j(\theta)T_j(x_i) + nr(\theta)\Big).$$

Следовательно, распределение выборки принадлежит экспоненциальному

[1] В общем случае квадратичная форма $Y^{\mathrm{T}}AY$ имеет χ^2_r-распределение, если $A\Sigma$ – симметричная идемпотентная матрица ранга r.

семейству с $h(\vec{x}) = \prod\limits_{i=1}^{n} h(x_j)$ и $T_j(\vec{x}) = \sum\limits_{i=1}^{n} T_j(x_i)$. Отметим, что $T(\vec{x}) = (T_i(\vec{x}), \ldots, T_s(\vec{x}))$, $T_j(\vec{x}) = \sum\limits_{i=1}^{n} T_j(x_i)$ – достаточная статистика для выборки из экспоненциального семейства. Более того, если $m = s$ и существует набор $(\theta_0, \ldots, \theta_s)$ таких, что матрица $\|a_{i,j}\|_{i,j=1}^{d}$, где $a_{i,j} = \eta_i(\theta_j)$, невырождена, то эта достаточная статистика является минимальной.

Среди рассмотренных классов распределений большинство являются экспоненциальными семействами. Экспоненциальными семействами не являются лишь класс равномерных распределений, класс t-распределений и класс показательных распределений с параметром сдвига (прил. 2).

5. Параметрическое оценивание

Теория оценивания для этого случая является наиболее разработанной. Она содержит как частные (пригодные лишь для конкретных распределений), так и общие методы оценивания. Мы сосредоточим внимание на общих методах оценивания, исследуем их точность, рассмотрим асимптотические свойства оценок. Помимо точечного мы рассмотрим также доверительное оценивание (на примере доверительных интервалов).

5.1. Методы построения статистических оценок

Одним из наиболее простых методов построения статистических оценок является метод моментов. Идея состоит в приравнивании нескольких первых теоретических моментов и их выборочных аналогов. Количество уравнений в системе по методу моментов выбирается так, чтобы получить единственное решение.

1. Пусть $X_1, \ldots, X_n$ – выборка из нормального распределения $\mathcal{N}(a, \sigma^2)$. Поскольку $\mathbf{E}_{a,\sigma^2}(X_1) = a$, а $\mathbf{E}_{a,\sigma^2}(X_1)^2 = a^2 + \sigma^2$, то соответствующая система уравнений следующим образом:

$$\begin{cases} a = \overline{X}, \\ \sigma^2 + a^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{cases}$$

Решение этой системы $(\tilde{a}, \tilde{\sigma}^2) = (\overline{X}, s^2)$ будет оценкой параметров a, σ^2 по методу моментов.

Метод моментов отличается простотой в использовании, но не всегда он дает достаточно хорошие оценки. Лучшие результаты, как правило, обеспечивает метод максимального правдоподобия.

При каждом фиксированном значении $\theta \in \Theta$ естественно задавать достоверности исходов значением плотности (или производной Радона–Никодима) относительно некоторой доминирующей меры получения реализовавшегося исхода $\vec{X}$. Если величины имеют абсолютно непрерывное распределение, то это – плотность относительно меры Лебега (или просто плотность). Если же распределение исходного вектора дискретно, то в качестве доминирующей можно выбрать считающую меру на множестве возможных значений рассматриваемого вектора. Тогда степень достоверности (правдоподобие) исхода $\vec{X}$ определяется значением дискретной плотности в данной точке или вероятностью того, что исход эксперимента именно такой. Для сравнения степени достоверности того или иного исхода при различных значениях параметра $\theta \in \Theta$ вводится, если это возможно, мера, доминирующая все семейство распределений. Естественно, что отношение правдоподобия при различных θ не зависит от выбора доминирующей меры.

Пусть $X_1, \ldots, X_n$ – набор независимых наблюдений с плотностями (дискретными плотностями, плотностями относительно доминирующей меры μ) $f_{\theta,1}, \ldots, f_{\theta,n}$ соответственно. *Функцией правдоподобия*, построенной по исходным наблюдениям, будем называть

$$L(\vec{X};\, \theta) = \prod_{i=1}^{n} f_{\theta,\, i}(X_i), \quad \theta \in \Theta.$$

Значение $L(\vec{X};\, \theta)$ при фиксированном θ будем называть правдоподобием исходного набора наблюдений. Идея, лежащая в основе *метода максимального правдоподобия*, состоит в том, чтобы по результатам наблюдений отыскать значение $\hat{\theta}(\vec{X}) \in \Theta$, максимизирующее правдоподобие, т. е. $L(\vec{X}; \hat{\theta}(\vec{X})) \geq L(\vec{X}; \theta)$ для любого $\theta \in \Theta$. Далее отметим, что задача нахождения точки максимума для функции правдоподобия сводится к аналогичной задаче для ее логарифма. Это позволяет дифференцировать по всем $\theta \in \Theta$ не произведение, а сумму

$$\ln L(\vec{X};\, \theta) = \sum_{i=1}^{n} \ln(f_{\theta,\, i}(X_i)).$$

Если $\theta = (\theta_1, \ldots, \theta_n)$ – n-мерный параметр и $f_{\theta,i}$ дифференцируемы по θ, то максимум надо искать среди решений системы уравнений

$$U(\vec{X};\, \theta) = \frac{\partial}{\partial \theta_i} \ln L(\vec{X};\, \theta) = 0.$$

Отметим, что по теореме факторизации статистика T достаточна тогда и только тогда, когда функция правдоподобия допускает представление в виде

$$L(\vec{x};\, \theta) = g(T(\vec{x}),\, \theta)\, h(\vec{x}),$$

где $g(t,\, \theta)$ и h – некоторые неотрицательные функции. Следовательно, оценка максимального правдоподобия является функцией от минимальной достаточной статистики.

Рассмотрим некоторые примеры. Известно, что для всех этих примеров функция правдоподобия имеет единственный максимум, поэтому найденные ниже нами стационарные точки являются точками максимума.

2. Пусть $X_1, \ldots, X_n$ – выборка из двухпараметрического ($\theta = (a, \sigma^2)$) нормального распределения $N(a, \sigma^2)$. Логарифм функции правдоподобия имеет вид

$$\ln L(\vec{x};\, \theta) = -\frac{n}{2} \ln(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - a)^2}{2\sigma^2}.$$

Тогда максимум правдоподобия находится из системы уравнений

$$\begin{cases} 2 \sum\limits_{i=1}^{n} (X_i - a) = 0, \\ n/\sigma - \sum\limits_{i=1}^{n} \dfrac{(X_i - a)^2}{\sigma^3} = 0. \end{cases}$$

Получаем, что оценка максимального правдоподобия имеет вид $(\hat{a},\, \hat{\sigma}^2) = (\overline{X},\, s^2)$.

3. Пусть $X_1, X_2 \ldots, X_n$ – выборка из распределения Лапласа $\mathrm{DE}\,(a, b)$, $\theta = (a, b)$. Вычисляем логарифм функции правдоподобия:

$$\ln L(\vec{x};\, \theta) = -n \ln 2b - \sum_{i=1}^{n} |x_i - a|/b.$$

Данная функция при фиксированном значении x лишь кусочно дифференцируема по a. Тем не менее, для нахождения максимума по a при каждом фиксированном b можно использовать кусочную производную при $x \neq x_i$, $i = 1, \ldots, n$,

$$\frac{\partial}{\partial a} \ln L(\vec{x};\, \theta) = -\sum_{i=1}^{n} \operatorname{sign}(x_i - a), \quad \text{где} \quad \operatorname{sign}(t) = \begin{cases} 1, & t > 0, \\ 0, & t = 0, \\ -1, & t < 0. \end{cases}$$

Очевидно, что данная функция монотонно убывает по a, поэтому экстремум (максимум) достигается в точке, где производная меняет знак.[1] Итак, оценка максимального правдоподобия для a при каждом фиксированном b – выборочная медиана Z_n, и ее значние не зависит от b. Оценку максимального правдоподобия для b получаем из уравнения

$$\frac{\partial}{\partial a}\ln L(\vec{x};\,\theta) = -\frac{n}{b} + \frac{1}{b^2}\sum_{i=1}^{n}|X_i - a| = 0,$$

подставляя вместо a выборочную медиану Z_n. Таким образом, оценка максимального правдоподобия для θ имеет вид $(\hat{a},\hat{b})=(Z_n, n^{-1}\sum\limits_{i=1}^{n}|X_i - Z_n|)$.

4. Предположим, что $X_1,\ldots,X_n$ – выборка из двухпараметрического ($\theta = (a,b)$) равномерного распределения на интервале (a,b). Функция правдоподобия имеет вид

$$L(\vec{x};\,\theta) = \frac{1}{(b-a)}\prod_{i=1}^{n}\mathbb{1}_{[a,b]}(x_i) = \frac{1}{b-a}\,\mathbb{1}_{[a,b]}(\min_i(x_i))\,\mathbb{1}_{[a,b]}(\max_i(x_i)),$$

где

$$\mathbb{1}_A(x) = \begin{cases} 1, \text{ если } x \in A; \\ 0, \text{ если } x \notin A. \end{cases}$$

В данной задаче нет смысла логарифмировать функцию правдоподобия. Заметим, что решение исходной задачи оптимизации сводится к минимизации разности $(b-a)$ при условии $a \le \min\limits_{1\le i\le n}(X_i) \le \max\limits_{1\le i\le n}(X_i) \le b$. Ясно, что решение данной задачи $a = \min\limits_{1\le i\le n}(X_i)$, $b = \max\limits_{1\le i\le n}(X_i)$. Таким образом, оценка максимального правдоподобия имеет вид $(\hat{a},\hat{b}) = (\min\limits_{1\le i\le n}(X_i), \max\limits_{1\le i\le n}(X_i))$.

Рассмотрим более подробно задачу оценки одномерного параметра $\theta \in \Theta \subset \mathbb{R}$ по выборке объема n с плотностью распределения одного наблюдения f_θ. Пусть выполнены достаточно общие условия регулярности, которые мы не будем здесь точно формулировать. Определим *информацию Фишера*, содержащуюся в одном наблюдении, выражением

$$I(\theta) = \mathbf{E}_\theta\left(\frac{\partial \log f_\theta}{\partial\theta}\right)^2 = \int\limits_{-\infty}^{\infty}\frac{(\partial f_\theta/\partial\theta(x))^2}{f_\theta(x)}\mu(dx).$$

[1] Отметим, что если n – четное, то функция правдоподобия постоянна на интервале между двумя средними порядковыми статистиками.

Например, для нормального распределения с $\mathcal{N}(a,d)$, $d=\sigma^2$ информация Фишера о параметре $\theta=a$ при известном $d>0$ есть $I(a)=1/d$, а информация Фишера о параметре $\theta=d>0$ при известном a есть $I(d)=1/2d^2$; для распределения Пуассона $I(\lambda)=1/\lambda$, для распределения Бернулли $I(p)=1/p(1-p)$. Если в выражении для $I(\theta)$ вместо f_θ подставить совместную плотность распределения всей выборки, то мы получим информацию Фишера, содержащуюся во всей выборке. Известно, что информация по Фишеру во всей выборке равна сумме информаций, содержащихся в каждом ее наблюдении. Поэтому информация по Фишеру во всей выборке равна $nI(\theta)$.

Для любой несмещенной оценки $\delta_n=\delta(X_1,...,X_n)$ справедлива следующая нижняя граница ее дисперсии и, следовательно, квадратичного риска:

$$\mathbf{D}_\theta(\delta_n)\geq\frac{1}{nI(\theta)},\quad\forall\,\theta\in\Theta,$$

называемая *неравенством Рао–Крамера*. Несмещенная оценка δ_n называется *эффективной*, если $\mathbf{D}_\theta(\delta_n)\leq\mathbf{D}_\theta(\tilde{\delta}_n)$ для всех $\theta\in\Theta$ и любой несмещенной оценки $\tilde{\delta}_n$. Таким образом, равенство в неравенстве Рао–Крамера есть достаточное (но не необходимое) условие эффективности несмещенной оценки. Используя его, можно показать, что выборочное среднее есть эффективная оценка среднего нормального распределения, параметра λ распределения Пуассона, параметра $u=1/\lambda$ показательного распределения; относительная частота есть эффективная оценка вероятности p для распределения Бернулли.

К сожалению, эффективные оценки существуют лишь в специальных случаях. В этой связи вводится понятие асимптотической эффективности. Оценка δ_n называется *асимптотически эффективной*, если

$$nR^{(2)}(\theta,\delta_n)\to 1/I(\theta),\ n\to\infty,\quad\forall\,\theta\in\Theta.$$

При выполнении достаточно общих предположений регулярности *оценки максимальноого правдоподобия являются асимптотически эффективными и асимптотически нормальными с нормирующим множителем* $\sigma^2(\theta)=1/I(\theta)$.

Эти результаты распространяются на случай многомерного параметра $\theta \in \Theta \subset \mathbb{R}^d$ с использованием понятия информационной матрицы Фишера

$$I(\theta) = \left\| E_\theta \left(\frac{\partial \log f_\theta}{\partial \theta_k} \; \frac{\partial \log f_\theta}{\partial \theta_j} \right) \right\|_{j,k=1}^{d}.$$

Замечание. В случае, если $\mathcal{P}$ – семейство всевозможных распределений наборов независимых одинаково распределенных случайных величин, то правдоподобие в указанной ранее постановке ввести не удается, поскольку не существует меры, доминирующей $\mathcal{P}$. В то же время, если сравнивать вероятности полученного результата наблюдений при различных распределениях выборки, то максимум достигается при равномерном распределении на множестве полученных наблюдений. Таким образом, эмпирическое распределение, в каком-то смысле, является оценкой максимального правдоподобия теоретического распределения.

5.2. Доверительное оценивание

Пусть $X_1, \ldots, X_n$ – выборка из распределения $P_\theta \in \mathcal{P}$. Задача доверительного оценивания заключается в том, чтобы построить статистику, значение которой $I(\vec{X})$ – подмножество множества Θ, накрывающее теоретическое значение θ с достаточно большой наперед заданной вероятностью при любом значении θ, т. е.

$$P_\theta(\theta \in I(\vec{X})) \geq 1 - \alpha, \quad \text{при любом} \quad \theta \in \Theta.$$

Значение $1 - \alpha$ носит название доверительный уровень (или доверительная вероятность) данного интервала. Обычно доверительная вероятность достаточно велика, например 0.9, 0.95, 0.99.

Мы ограничимся доверительным оцениванием одномерного параметра. Как и при точечном оценивании, необходимо сузить класс возможных доверительных оценок, чтобы задача стала содержательной. В одномерном случае естественно ограничиться доверительными оценками, которые являются интервалами.

Интервал $[T_1, T_2]$, образованный парой статистик $T_1(X_1, \ldots, X_n)$ и $T_2(X_1, \ldots, X_n)$, называется *доверительным интервалом* надежности (или с доверительным уровнем) $1 - \alpha$, если при всех $\theta \in \Theta$ выполняется неравенство $P_\theta(T_1 \leqslant \theta \leqslant T_2) \geqslant 1 - \alpha$.

Очевидно, что в такой постановке задача имеет множество решений. Следует отметить, что если задаться целью, то можно выбирать довери-

тельный интервал, содержащий произвольное наперед заданное значение параметра (вовсе не обязательно истинное). При этом он с большой вероятностью должен содержать и истинное значение параметра. Поэтому его длина может быть достаточно велика вне зависимости от длины выборки. Таким образом, естественной мерой качества доверительного интервала является его длина. Однако возможны и другие подходы к выбору доверительного интервала. Например, если нужно оценить истинное значение параметра снизу (или сверху), то имеет смысл рассматривать односторонние интервалы, а если длина фиксирована, то мерой качества будет минимальный объем выборки, необходимый для того, чтобы данный интервал имел уровень $1 - \alpha$.

Рассмотрим некоторые методы построения доверительного интервала. Начнем с простейшего варианта, когда найдена случайная величина $G(\vec{X};\, \theta)$, монотонно зависящая от параметра, распределение которой не зависит от θ. Поскольку функция распределения F_G не зависит от θ, можно найти значения g_1 и g_2 из условия (так чтобы в интервале $[g_1, g_2]$ не содержался другой интервал, удовлетворяющий этому условию)

$$P_\theta(g_1 \le G(\vec{X};\, \theta) \le g_2) = F_G(g_2+) - F_G(g_1) \ge 1 - \alpha.$$

Далее, границы доверительного интервала $T_1(\vec{X})$ и $T_2(\vec{X})$ для параметра θ выбираются как решения относительно θ уравнений

$$G(\vec{X};\, \theta) = g_i, \quad i = 1, 2.$$

Тогда $P_\theta(T_1(\vec{X}) \le \theta \le T_2(\vec{X})) = P_\theta(g_1 \le G(\vec{X};\, \theta) \le g_2) \ge 1 - \alpha$. Если функции распределения исходного семейства $\mathcal{P}$ непрерывны и монотонно меняются по параметру (например, по параметру сдвига), то можно выбрать

$$G(\vec{X};\, \theta) = -\sum_{i=1}^{n} \ln F(X_i;\, \theta).$$

Нетрудно показать, что если случайная величина X имеет непрерывную функцию распределения F, то случайная величина $F(X)$ равномерно распределена на интервале $[0, 1]$ (данное свойство носит название *теорема Смирнова*, а соответствующее преобразование называется преобразованием Смирнова). Таким образом, $-\ln F(X_i;\, \theta)$ имеет γ-распределение $\Gamma(1, 1)$. Следовательно, используя свойства γ-распределения, заключаем, что $G(\vec{X};\, \theta)$ имеет известное распределение $\Gamma(1, n)$. Далее находим квантили γ_1 и γ_2, удовлетворяющие соотношению $P(\xi \in [\gamma_1, \gamma_2]) = 1 - \alpha$,

где $\xi \sim \Gamma(1, n)$, и решаем неравенства $G(\vec{X};\, \theta) \in [\gamma_1, \gamma_2]$ относительно θ. Очевидно, что основная сложность данного метода – нахождение решения данных неравенств.

5. Пусть $X_1, \ldots, X_n$ – выборка из абсолютно непрерывного распределения с плотностью

$$p_\theta(x) = \theta\, e^{\theta x} \mathbb{I}_{(-\infty,0]}(x).$$

В этом случае

$$F_\theta(x) = \begin{cases} e^{\theta x}, & x \le 0 \\ 1, & x > 0 \end{cases}$$

и $-\ln F(x) = \theta x \mathbb{I}_{(-\infty,0]}(x)$. Таким образом,

$$G(\vec{X}; \theta) = -\theta \sum_{i=1}^{n} X_i$$

имеет $\Gamma(1, n)$ распределение и соответствующий доверительный интервал для θ с доверительной вероятностью $1 - \alpha$ будет $[-\gamma_1/(n\overline{X}), \gamma_2/(n\overline{X})]$.

Рассмотрим другие примеры.

6. Пусть $X_1, \ldots, X_n$ – выборка из двухпараметрического нормального распределения $N(a, \sigma^2)$. Будем строить доверительные интервалы (уровня доверия $1-\alpha$) для параметров a и σ^2 с использованием леммы Фишера. Согласно п. 3 теоремы Фишера величина nS^2/σ^2 имеет распределение χ^2_{n-1}. Из уравнения

$$P\Big(g_1 \le \frac{ns^2}{\sigma^2} \le g_2\Big) = \int_{g_1}^{g_2} k_{n-1}(x)\, dx = 1 - \alpha$$

находим константы g_1 и g_2. Если требуется односторонний доверительный интервал, то выбор константы (g_1 или g_2) однозначен. При построении двухстороннего интервала обычно выбирают константы по принципу:

$$\int_{-\infty}^{g_1} k_{n-1}(x)\, dx = \int_{g_2}^{\infty} k_{n-1}(x)\, dx = \alpha/2.$$

Если задаться целью построить наикратчайший интервал, то необходимо минимизировать разность $(1/g_1) - (1/g_2)$ при условии $\int_{g_1}^{g_2} k_{n-1}(x)\, dx = 1 - \alpha$. Для этого можно воспользоваться методом Лагранжа.

Чтобы построить доверительный интервал для среднего, воспользуемся функцией $G(\vec{X};\, \theta) = \sqrt{n-1}(\overline{X} - a)/s^2$, которая, согласно

п. 4 теоремы Фишера, имеет распределение Стьюдента с $n-1$ степенью свободы. Находим константы $g_{\alpha,1}$ и $g_{\alpha,2}$ из уравнений $S_{n-1}(g_{\alpha,1}) = 1-$ $-S_{n-1}(g_{\alpha,2}) = \alpha/2$. Отметим, что в силу симметричности распределения Стьюдента $g_{\alpha,1} = -g_{\alpha,2} = t_{\alpha/2}$. Тогда доверительный интервал для a имеет вид $[\overline{X} - (st_{\alpha/2})/\sqrt{n-1},\ \overline{X} + (st_{\alpha/2})/\sqrt{n-1}]$.

Рассмотрим иную постановку задачи построения доверительного интервала. Теперь требуется достигнуть нужной точности доверительной оценки надежности $1-\alpha$ за счет накопления достаточного количества наблюдений. Такой подход носит название *адаптивный*.

7. Пусть имеется выборка $X_1, \ldots, X_n$ из нормального распределения $\mathcal{N}(a, \sigma)$, где σ известно. Известна длина симметричного доверительного интервала S. Найдем минимальный объем выборки, необходимый для того, чтобы данный интервал имел уровень α. В качестве $G(\vec{X};\, \theta)$ здесь можно взять $\dfrac{\overline{X} - a}{\sigma/\sqrt{n}}$, принадлежащую $\mathcal{N}(0,1)$ при любом a. Если $u_{-\alpha/2} = -u_{\alpha/2}$ – соответствующие квантили, то $P_a\Big(-u_{\alpha/2} < \dfrac{\overline{X} - a}{\sigma/\sqrt{n}} < -u_{\alpha/2}\Big) = 1-\alpha$ и наш доверительный интервал есть $\overline{X} - \dfrac{u_{\alpha/2}\sigma}{\sqrt{n}} < a < \overline{X} + \dfrac{u_{\alpha/2}\sigma}{\sqrt{n}}$. Теперь можно найти n из уравнения $S = 2\dfrac{u_{\alpha/2}\sigma}{\sqrt{n}}$. Для целых n и фиксированного α это уравнение редко имеет точное решение. В качестве допустимого n следует брать $n = \left[\left(2\dfrac{u_{\alpha/2}\sigma}{S}\right)^2\right] + 1$, где $[x]$ – это наибольшее целое число, меньшее x.

Более сложная ситуация возникает, если не удается найти монотонно зависящую от параметра функцию $G(\vec{X};\, \theta)$, распределение которой не зависит от параметра. В этом случае для построения доверительных интервалов используется определенный класс статистик. Будем говорить, что распределение статистики T монотонно зависит от параметра, если функция распределения $F_T(x;\, \theta) = P_\theta(T < x)$ монотонно возрастает (или убывает) по параметру θ при каждом фиксированном $x \in \mathbb{R}$. Обычно все разумные точечные оценки параметра обладают этим свойством. В дальнейшем будем считать, что исходная статистика удовлетворяет этому свойству и является точечной оценкой параметра. Более того, будем считать, что $F_T(x;\, \theta)$ – непрерывная функция θ. В силу непрерывности функции F_T

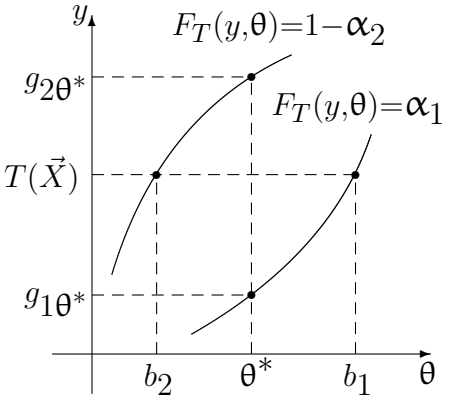
по θ уравнения $F_T(x;\,\theta)=\gamma$, $\gamma\in(0,1)$ разрешимы относительно параметра θ. Пусть $b(x,\gamma)$ – корень соответствующего уравнения. Тогда интервал с границами $b_2=b(T(\vec{X}),\,1-\alpha_2)$ и $b_1=b(T(\vec{X}),\,\alpha_1)$ является доверительным уровня $1-\alpha$. Данный метод применим и в случае, если $F_T(x;\,\theta)$ дискретны. При определении квантилей $F_T^{-1}(\gamma;\,\theta)$ следует добиться выполнения неравенства

$$F_T(F_T^{-1}(1-\alpha_2))-F_T(F_T^{-1}(\alpha_1)_-)\geq 1-\alpha.$$

Приведем графическую интерпретацию данного факта. Пусть при каждом значении параметра найдены константы $g_1(\theta)$ и $g_2(\theta)$ такие, что

$$P_\theta(g_1(\theta)\leq T(\vec{X})\leq g_2(\theta))\geq 1-\alpha$$

(если это возможно, то надо добиваться равенства). Для определенности выбираем их из соотношений $F_T(g_1(\theta))\leq\alpha/2$ и $1-F_T(g_2(\theta))\leq\alpha/2$. Рассмотрим множество $\Theta\times\Theta$ – конечный или бесконечный прямоугольник на плоскости. Рассмотрим множество точек плоскости (см. рисунок)



$$\mathrm{D}=\{(\theta',\,\theta):g_1(\theta)\leq\theta'\leq g_2(\theta)\}.$$

Тогда сечение этого множества на уровне статистики T – это соответствующий доверительный интервал.

8. Пусть $X_1,\ldots,X_n$ – выборка из распределения Бернулли $\mathrm{Bi}(\theta,\,1)$ с вероятностью успеха θ. В качестве статистики, непрерывно зависящей от параметра по распределению, выберем $\overline{X}$. Очевидно, что

$$F_T(k/n;\,\theta)=\sum_{j=0}^{k}C_n^j\theta^j(1-\theta)^{n-j}$$

(имеет место монотонное убывание по θ). Находим значения θ_1 и θ_2 такие, что

$$\sum_{j=n\overline{X}}^{n}C_n^j\theta_1^j(1-\theta_1)^{n-j}=\alpha/2\quad\text{и}\quad\sum_{j=0}^{n\overline{X}}C_n^j\theta_2^j(1-\theta_2)^{n-j}=\alpha/2.$$

Тогда, в силу приведенных аргументов, интервал $[\theta_1,\theta_2]$ будет доверительным интервалом уровня $1-\alpha$.

В общем случае для построения доверительного интервала (области) уровня значимости $1-\alpha$ можно использовать следующий метод. Рассмот-

рим набор подмножеств выборочного пространства $\{\mathfrak{D}(\theta)\}_{\theta\in\Theta}$ такой, что $P_\theta(\mathfrak{D}(\theta)) \geq 1-\alpha$, $\theta \in \Theta$. Тогда множество всех θ, при которых результат эксперимента $\vec{X}$ попадает в $\mathfrak{D}(\theta)$, будет доверительной областью (а если это интервал, то доверительным интервалом) уровня значимости $1-\alpha$. Однако данный метод слишком общий, и, вообще говоря, трудно ожидать от него хороших результатов.

При больших объемах выборок можно использовать асимптотический подход к построению доверительных интервалов. Формально, разговор об асимптотических доверительных интервалах конечной выборки не имеет смысла. С другой стороны, если выборка бесконечна, то, по всей вероятности, доверительный интервал будет состоять из одной точки – теоретического значения параметра. Однако можно рассматривать последовательности интервалов, которые аппроксимируют последовательности доверительных интервалов, построенных по конечным выборкам. При больших объемах выборок их, с определенными оговорками, можно использовать вместо доверительных интервалов.

Пусть $X_1, X_2, \ldots$ – последовательность независимых одинаково распределенных случайных величин, и для заданного $\alpha > 0$ существуют статистики (точнее, последовательности статистик) $T^n_{1,\alpha}(X_1,\ldots,X_n)$ и $T^n_{2,\alpha}(X_1,\ldots,X_n)$, такие что

$$\lim_{n\to\infty} P_\theta\Big(T^n_{1,\alpha}(X_1,\ldots,X_n) < \theta < T^n_{2,\alpha}(X_1,\ldots,X_n)\Big) \geqslant 1-\alpha, \quad \forall\, \theta\in\Theta.$$

Тогда мы говорим, что задан асимптотический доверительный интервал с уровнем доверия $1-\alpha$.

Рассмотрим метод построения асимптотических доверительных интервалов на базе асимптотически нормальной оценки δ параметра θ, т. е. оценки, для которой с ростом объема выборки имеет место сходимость по распределению

$$\sqrt{n}(\delta(\vec{X}) - \theta) \Longrightarrow \mathcal{N}(0, \sigma^2(\theta)),$$

где $\sigma^2(\theta)$ – известная функция параметра θ. Тогда

$$\frac{\sqrt{n}(\delta(\vec{X}) - \theta)}{\sigma^2(\theta)} \Longrightarrow \mathcal{N}(0,1).$$

Выберем x_α из уравнения $\Phi(x_\alpha) = 1-\alpha/2$, где $\Phi(x)$ функция распределения стандартного нормального закона. Тогда

$$P_\theta(-x_\alpha \leq \sqrt{n}(\delta(\vec{X}) - \theta)/\sigma^2(\theta) \leq x_\alpha) \xrightarrow[n\to\infty]{} 1-\alpha.$$

Доверительная область для θ получается как решение неравенств под знаком вероятности. В случае если решение не является интервалом или если найти его затруднительно, то можно использовать вместо $\sigma(\theta)$ состоятельную оценку $\sigma^*(\vec{X})$. Тогда

$$\frac{\sqrt{n}(\delta(\vec{X})-\theta)}{\sigma^*(\vec{X})} \Longrightarrow \mathcal{N}(0,1),$$

и получаем последовательность асимптотических доверительных интервалов

$$[\delta(\vec{X}) - x_{\alpha/2}\sigma^*(\vec{X})/\sqrt{n},\ \delta(\vec{X}) + x_{\alpha/2}\sigma^*(\vec{X})/\sqrt{n}].$$

Если исходная последовательность распределений непрерывно меняется по θ, то из состоятельности оценки δ ($\delta(\vec{X}) \xrightarrow[n\to\infty]{} \theta$ по вероятности) немедленно следует, что $\sigma(\delta(\vec{X})) \xrightarrow[n\to\infty]{} \sigma(\theta)$.

9. Построим доверительный интервал для оценки вероятности биномиального распределения по частоте. Имеет место асимптотическая нормальность. Пусть $\upsilon = m/n$ – наблюденная частота успеха. Имеем $P(|\upsilon - p| \le t\sqrt{p(1-p)/n}) \approx 1-\alpha$, где $\Phi(t) = 1-\alpha/2$. Чтобы построить доверительный интервал, решим неравенство $|\upsilon - p| < t\sqrt{p(1-p)/n}$ или $((t^2/n)+1)p^2 - 2(\upsilon + (t^2/n))p + \upsilon^2 < 0$ относительно p. Решением будет интервал $[p_1, p_2]$, где

$$p_1 = \frac{n}{t^2+n}\left(\upsilon + \frac{t^2}{2n} - t\sqrt{\frac{\upsilon(1-\upsilon)}{n} + \left(\frac{t}{2n}\right)^2}\right),$$

$$p_2 = \frac{n}{t^2+n}\left(\upsilon + \frac{t^2}{2n} + t\sqrt{\frac{\upsilon(1-\upsilon)}{n} + \left(\frac{t}{2n}\right)^2}\right).$$

Итак, $[p_1, p_2]$ – асимптотический доверительный интервал для p.

Можно избежать решения квадратного уравнения, подставив вместо теоретического значения параметра p состоятельную оценку υ. Тогда из неравенства $|\upsilon - p| \le x_\alpha\sqrt{\upsilon(1-\upsilon)/n}$ получаем последовательность асимптотических доверительных интервалов

$$[\upsilon - x_\alpha\sqrt{\upsilon(1-\upsilon)/n}, \upsilon + x_\alpha\sqrt{\upsilon(1-\upsilon)/n}\,].$$

Нетрудно увидеть, что данная последовательность асимптотических доверительных интервалов асимптотически эквивалентна $[p_1, p_2]$ и отличается от нее лишь членами порядка $1/n$, тогда как обе последовательности

этих доверительных интервалов стягиваются в точку со скоростью порядка $1/\sqrt{n}$.

Для полноты картины рассмотрим еще один способ построения доверительного интервала в рассматриваемой ситуации. Поскольку $p(1-p)$ – дисперсия распределения Бернулли, то s^2 является состоятельной оценкой для $p(1-p)$. Используя s^2 для оценивания величины $p(1-p)$, получаем еще одну последовательность асимптотических доверительных интервалов $[\upsilon - x_\alpha s/\sqrt{n}, \upsilon + x_\alpha s/\sqrt{n}]$.

Имеется универсальный метод построения асимптотических доверительных интервалов на основе оценок максимального правдоподобия. Как уже отмечалось, при достаточно общих предположениях регулярности оценка максимального правдоподобия θ_n^* параметра $\theta \in \Theta \subset \mathbb{R}$ асимптотически нормальна с нормирующим множителем $1/\sqrt{I(\theta)}$; кроме того информация Фишера $I(\theta)$ положительна и непрерывна по θ. Тогда $1/\sqrt{I(\theta_n^*)}$ есть состоятельная оценка нормирующего множителя, и из предыдущих рассуждений мы получаем асимптотически доверительные интервалы вида $[\theta_n^* - x_\alpha/\sqrt{nI(\theta_n^*)},\ \theta_n^* + x_\alpha/\sqrt{nI(\theta_n^*)}]$. Можно показать, что эти доверительные интервалы являются асимптотически наикратчайшими при любом заданном уровне значимости $1-\alpha \in (0,1)$.

6. Проверка статистических гипотез

В обыденной жизни под гипотезой понимается некоторое предположение. В математической статистике понятие гипотезы несколько сужается и интерес представляют лишь предположения об истинном значении параметра. В терминах параметрического семейства распределений $\mathcal{P} = \{P_\theta,\ \theta \in \Theta\}$ любая статистическая гипотеза может быть записана в виде

$$H : \theta \in \Theta_0,$$

где Θ_0 – некоторое подмножество параметрического множества Θ. Статистическая гипотеза вида $H : \theta = \theta_0$ (т. е. определяет значение параметра однозначно) при некотором наперед заданном значении θ_0 называется *простой*. В противном случае статистическая гипотеза *сложная*.

При постановке задачи проверки статистических гипотез различают основную гипотезу $H_0 : \theta \in \Theta_0$ и альтернативную – $H_1 : \theta \in \Theta \backslash \Theta_1$, где Θ_1 – непустое подмножество множества Θ. Задача состоит в том, чтобы по

экспериментальным данным принять или отвергнуть основную гипотезу. Обычно если $\Theta_1 = \Theta \backslash \Theta_0$, то говорят о задаче проверки согласия с гипотезой H_0.

Конечно, по выборке конечного объема, вообще говоря, нельзя построить процедуру, дающую достоверный ответ на вопрос, какая из гипотез верна. В таблице приведены возможные ситуации: по вертикали указана фактическая ситуация, а по горизонтали – возможные решения исследователя.

Фактическая ситуация	**Решение исследователя**	
	Принять H_0	Принять H_1
Верна H_0	Верное решение	Ошибка 1-го рода
Верна H_1	Ошибка 2-го рода	Верное решение

Итак, помимо верного решения возможна ошибка 1-го рода, заключающаяся в отвержении основной гипотезы при ее справедливости, или ошибка 2-го рода, заключающаяся в принятии основной гипотезы при справедливости альтернативной.

Статистика φ со значениями в интервале $[0, 1]$, сопоставляющая каждому набору наблюдений соответствующую вероятность отвергнуть гипотезу H_0, называется *критерием* (или тестом).

Если множество значений критерия $\{0, 1\}$, то критерий называется *нерандомизованным* (т. е. результаты наблюдений однозначно определяют решение). Если же возможны промежуточные значения, то критерий называется рандомизованным (например, если результаты эксперимента не прояснили ситуацию, то рекомендуется осуществить случайный выбор так, чтобы с вероятностью φ отвергнуть нулевую гипотезу). Таким образом выборочное пространство разбивается на три области – *допустимая* $\{\vec{x} \in \mathfrak{X} : \varphi(\vec{x}) = 0\}$, или область принятия гипотезы, *критическая* $\{\vec{x} \in \mathfrak{X} : \varphi(\vec{x}) = 1\}$, или область отвержения гипотезы, и *область сомнения* $\{\vec{x} \in \mathfrak{X} : \varphi(\vec{x}) \in (0, 1)\}$.

Идея подхода Пирсона состоит в том, чтобы ограничить вероятность ошибки 1-го рода некоторым наперед заданным числом α, называемым *уровнем значимости критерия*. При этом вероятность ошибки 2-го рода должна быть по возможности минимальной.

Нетрудно видеть, что если справедлива гипотеза H_0, т. е. $\theta \in \Theta_0$, то $P_\theta(\text{ош. 1-го рода}) = \mathbf{E}_\theta \varphi(\vec{X})$. В противном случае, если $\theta \in \Theta_1$,

то P_θ(ош. 2-го рода) $= \mathbf{E}_\theta(1-\varphi(\vec{X}))$. *Мощностью критерия* называется функция $\beta : \Theta_1 \to [0,1]$, задаваемая равенством $\beta(\theta) = \mathbf{E}_\theta\varphi(\vec{X}) = 1-$ $-P_\theta$(ош. 2-го рода). Возникает экстремальная задача

$$\sup_{\theta\in\Theta_0} \mathbf{E}_\theta\varphi(\vec{X}) \le \alpha,$$
$$\beta(\theta) = \mathbf{E}_\theta\varphi(\vec{X}) \to \max_{\varphi},\ \theta \in \Theta_1.$$

Если решение этой задачи существует, то оно называется *равномерно наиболее мощным* критерием. Равномерно наиболее мощные критерии существуют довольно редко. Однако в случае проверки простой гипотезы против простой альтернативы наиболее мощный критерий (для простых гипотез понятие равномерности вырождается) всегда существует и строится явно с использованием статистики отношения правдоподобия.

Необходимым атрибутом качественного критерия является его состоятельность, определяемая асимптотическими свойствами критерия, а потому рассматривается уже не один критерий при фиксированном значении n, а последовательность критериев $\{\varphi_n\}_{n\in\mathbb{N}}$ определенного вида. Критерий называется *состоятельным*, если мощность критерия $\beta_\varphi(\theta) \to 1$ при каждом фиксированном значении $\theta \in \Theta\backslash\Theta_0$. Можно говорить также о равномерной состоятельности при фиксированной альтернативе $\sup_{\theta\in\Theta_1} \beta_\varphi(\theta) \to 1$, однако это условие часто нарушается из-за невозможности различать близкие распределения при фиксированном n. Тогда имеет смысл рассматривать наряду с основной гипотезой последовательности альтернатив $\{H_{A,i}\}_{i\in\mathbb{N}}$ такие, что из $H_{A,i}$ следует H_A. Для того чтобы установить границы различимости гипотез состоятельным критерием при конечном n на уровне значимости α, находится область $\Theta_{A,n}$ в множестве Θ_A, в которой вероятность ошибки второго рода не превышает некоторого наперед заданного малого числа α_{II}.

Среди задач проверки статистических гипотез можно выделить класс параметрических задач, т. е. задач, в которых рассматриваемое параметрическое семейство $\mathcal{P}$ параметризовано (или допускает параметризацию) векторно-значным параметром $\theta = (\theta_1, \ldots, \theta_m)$. Обычно $\mathcal{P}$ – некоторый класс распределений с параметром сдвига, масштаба и т. д. (например, двухпараметрический класс нормальных распределений). В противном случае задача называется непараметрической. Критерий, позволяющий решать непараметрическую задачу, называется *непараметрическим*.

6.1. Методы построения статистических критериев

Поставим задачу проверки гипотезы $H_0 : \theta \in \Theta_0$ при альтернативе $H_1 : \theta \in \Theta_1$. Для построения критерия уровня значимости α не требуется знания функции мощности, так как он определяется только значением величины α. В основе общего метода построения критериев лежит понятие статистики критерия.

Статистикой критерия называется функция $G(\vec{X}) = G(\vec{X}, H_0)$, которая при справедливости H_0 является подчиненной статистикой (т. е. ее распределение не зависит от θ). Остается лишь вычислить или оценить квантили этого распределения, чтобы построить нерандомизованный критерий вида

$$\varphi(\vec{X}) = \begin{cases} 0, & G(\vec{X}) \in I_\alpha, \\ 1, & G(\vec{X}) \notin I_\alpha, \end{cases}$$

а I_α выбирается из условия $P_\theta(G(\vec{X}) \notin I_\alpha) = P_0(G(\vec{X}) \notin I_\alpha) < \alpha$, $\theta \in \Theta_0$.

Отметим, что условие подчиненности статистики критерия относительно подсемейства $\{P_\theta, \theta \in \Theta_0\}$ может быть ослаблено. В этом случае критерий строится аналогично, но множество I_α, определяющее допустимую и критическую области, должно быть выбрано из соотношения $\sup_{\theta \in \Theta_0} P_\theta(G(\vec{X}) \notin I_\alpha) < \alpha$. Возможны также рандомизованные версии подобного критерия.

Обычно (за исключением некоторых специальных ситуаций) вычисление распределения статистики критерия при каждом фиксированном n оказывается непростой задачей, поэтому при достаточно больших объемах выборки имеет смысл использовать асимптотический подход, заключающийся в том, что вместо точного распределения статистики критерия, используется асимптотическое распределение. При этом вместо подчиненности статистики критерия следует использовать асимптотическую подчиненность (асимптотическое распределение не зависит от параметра при справедливости H_0). Впрочем, как было отмечено ранее, условие подчиненности часто не является существенным. Асимптотические распределения наиболее часто используемых статистик критерия затабулированы.

Теперь займемся изучением качества критерия. Интуитивно ясно, что если статистика критерия остается подчиненной и при $\theta \notin \Theta_0$, то критерий

не может быть использован для проверки статистической гипотезы H_0 при альтернативе H_1.

6.2. Наиболее мощные критерии

Пусть $X_1, \ldots, X_n$ – выборка из распределения P_θ, $\theta \in \{\theta_0, \theta_1\}$. Предположим, что меры P_{θ_0} и P_{θ_1} имеют плотности $f((\cdot); \theta_0)$ и $f((\cdot); \theta_1)$ соответственно относительно доминирующей меры μ. Обычно в качестве μ выбирается либо мера Лебега (абсолютно непрерывный случай), либо считающая мера (дискретный случай). Поставим задачу проверки простой гипотезы $H_0 : P = P_{\theta_0}$ против простой альтернативы $H_1 : P = P_{\theta_1}$. Фундаментальная лемма Неймана–Пирсона гарантирует, что критерий вида

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

где $l(\vec{X}) = l(\vec{X}, \theta_1, \theta_0) = \dfrac{L(\vec{X}, \theta_1)}{L(\vec{X}, \theta_0)}$, а константа c и вероятность $p \in [0, 1]$ находятся из уравнения

$$\mathbf{E}_{\theta_0} \varphi(\vec{X}) = P_{\theta_0}(l(\vec{X}) > c) + p\, P_{\theta_0}(l(\vec{X}) = c) = \alpha,$$

является наиболее мощным критерием уровня α для проверки простой гипотезы H_0 при простой альтернативе H_1. Наиболее мощный критерий определен однозначно на множестве $l(\vec{X}) \neq c$.

Замечание 1. Критическая область с границей c определяется однозначно, с точностью до множеств нулевой вероятности, из приведенного уравнения. Если $P_{\theta_0}(l(\vec{X}) = c) > 0$, то константа p также находится однозначно. В противном случае выбор p не имеет значения, поскольку событие $\{l(\vec{X}) = c\}$ имеет нулевую вероятность.

Замечание 2. Нельзя гарантировать единственность наиболее мощного критерия (с точностью до множеств нулевой вероятности).

1. Пусть $X_1, \ldots, X_n$ – выборка из распределения Бернулли $\mathrm{Bi}(\theta, 1)$. Функция правдоподобия в этом случае имеет вид

$$L(\vec{X}; \theta) = \theta^{\sum_{i=1}^n X_i} (1 - \theta)^{(n - \sum_{i=1}^n X_i)},$$

и, таким образом, статистика отношения правдоподобия будет иметь вид

$$l(\vec{X}) = \frac{L(\vec{X}; \theta_1)}{L(\vec{X}; \theta_0)} = \left(\frac{\theta_1(1 - \theta_0)}{\theta_0(1 - \theta_1)} \right)^{\sum_{i=1}^n X_i} \left(\frac{1 - \theta_1}{1 - \theta_0} \right)^n.$$

В силу монотонности статистики $l(\vec{X})$ относительно значения минимальной достаточной статистики $\sum\limits_{i=1}^{n} X_i$ неравенства $l(\vec{X}) > c$, $l(\vec{X}) < c$ и равенство $l(\vec{X}) = c$ могут быть переписаны в более удобной форме в терминах $\overline{X}$. Так, в случае $\theta_0 > \theta_1$ получаем соответственно $\overline{X} < c^*$, $\overline{X} > c^*$ и $\overline{X} = c^*$ при

$$c^* = \frac{\log c - n(\log(1-\theta_1) - \log(1-\theta_0))}{\log(\theta_1(1-\theta_0)) - \log(\theta_0(1-\theta_1))}.$$

В противном случае получаем $\overline{X} > c^*$, $\overline{X} < c^*$ и $\overline{X} = c^*$ соответственно. Здесь c^* – некоторая, вообще говоря, отличная от c постоянная. Предположим для определенности $\theta_0 > \theta_1$. Тогда наиболее мощный критерий имеет вид

$$\varphi(\vec{x}) = \begin{cases} 1, & \text{при } \overline{X} < c^*; \\ p, & \text{при } \overline{X} = c^*; \\ 0, & \text{при } \overline{X} > c^*. \end{cases}$$

Для нахождения константы c^* воспользуемся формулой Бернулли

$$P_{\theta_0}(\overline{X} \leqslant c^*) = P(n\overline{X} \leqslant nc^*) = \sum_{i=0}^{[nc^*]} C_n^i \theta^i (1-\theta)^{n-i},$$

где $[nc]$ – наибольшее целое число, меньшее nc. Константа c^* находится из соотношения

$$\sum_{i=0}^{n^*-1} C_n^i \theta^i (1-\theta)^{n-i} \leqslant \alpha < \sum_{i=0}^{n^*} C_n^i \theta^i (1-\theta)^{n-i},$$

где $n^* = [nc^*]$ – целое число. В свою очередь, константа p находится из соотношения

$$p = (\alpha - \textstyle\sum_{i=0}^{n^*-1} C_n^i \theta^i (1-\theta)^{n-i}) / (C_n^{n^*} \theta^{n^*} (1-\theta)^{n-n^*}).$$

Наиболее мощный критерий определен однозначно на множестве $\{\overline{X} \neq c^*\}$, однако на множестве $\{\overline{X} = c^*\}$ он не всегда определен неоднозначно. Например, если $p_0 = 1/2$, $n = 2k$ – четное число, а c^* – нечетное, то можно построить нерандомизованный критерий φ_1, совпадающий с φ всюду, кроме случая $\overline{X} = c^*$ и

$$\varphi_1(X) = \begin{cases} 1, & \sum\limits_{i=1}^{k} X_i > \sum\limits_{i=k+1}^{n} X_i, \ \overline{X} = c^*; \\ 0, & \sum\limits_{i=1}^{k} X_i < \sum\limits_{i=k+1}^{n} X_i, \ \overline{X} = c^*. \end{cases}$$

Конечно, это очень специальный случай. Отметим, что при $n \to \infty$ вероятности $P_{\theta_0}(n\overline{X} = k)$ имеют тенденцию к уменьшению. Нерандомизованный асимптотический наиболее мощный критерий можно получить с использованием интегральной формулы Муавра–Лапласа

$$P_{\theta_0}(\overline{X} < c^*) = P\left(\sqrt{n}\frac{\overline{X} - \theta_0}{\sqrt{\theta_0(1-\theta_0)}} \leqslant c_n^*\right) \approx \Phi(c_n^*),$$

где $c_n^* = \sqrt{n}(c^* - \theta_0)/\sqrt{\theta_0(1-\theta_0)}$. Константу c^* находим из условия $\Phi(c_n^*) \approx \alpha$, а p принимаем равным 0 (или 1). Тогда критерий φ с соответствующими значениями c^* и p будет асимптотическим критерием уровня α.

Пусть θ – скалярный параметр, $\theta_* \in \Theta$. На практике довольно часто критерий отношения правдоподобия является наиболее мощным для проверки односторонней гипотезы $H_0 : \theta \leq \theta_*$ при альтернативе $H_1 : \theta > \theta_*$. Будем говорить, что семейство $\mathcal{P}$ имеет монотонное (относительно θ_* и T) отношение правдоподобия, если для любых $\theta_0, \theta_1 \in \Theta : \theta_0 \leqslant \leqslant \theta_* \leqslant \theta_1$, статистика отношения правдоподобия $l(\vec{X}; \theta_1, \theta_0)$ является монотонной функцией некоторой одномерной статистики $T(\vec{X})$ ($l(\vec{X}; \theta_1, \theta_0) = = l^*(T(\vec{X}); \theta_1, \theta_0)$). В этом случае решение уравнения $l(\vec{x}; \theta_1, \theta_*) < c$ может быть записано с использованием статистики T в виде $T < c^*$, если l^* возрастает, или в виде $T > c^*$, если l^* убывает. Пусть в поставленной задаче семейство $\mathcal{P}$ имеет монотонное отношение правдоподобия относительно θ_* и некоторой статистики T (для определенности считаем, что l^* возрастает). В этом случае критерий

$$\varphi(\vec{x}) = \begin{cases} 1, & \text{если } T(\vec{X}) > c; \\ p, & \text{если } T(\vec{X}) = c; \\ 0, & \text{если } T(\vec{X}) < c, \end{cases}$$

где константы c и $p \in [0,1]$ выбираются из уравнения

$$\sup_{\theta \leq \theta_*} \mathbf{E}_\theta \varphi(\vec{X}) = \mathbf{E}_{\theta_*} \varphi(\vec{X}) = P_{\theta_*}(T(\vec{X}) > c) + p\, P_{\theta_*}(T(\vec{X}) = c) = \alpha,$$

является равномерно наиболее мощным критерием уровня α для проверки гипотезы H_0 при альтернативе H_1.

Рассмотрим сначала случай простой основной гипотезы и односторонних альтернатив (Θ – полупрямая). Например, пусть по выборке $X_1, ..., X_n$ проверяется гипотеза $H_0 : a = 0$ о среднем нормального распределения $N(a, \sigma^2)$ при известном $\sigma > 0$ против альтернативы $H_1 : \mathrm{a} = a_1,\ a_1 > 0$.

Тогда оптимальный критерий уровня значимости $\alpha \in (0,1)$ не зависит от $a_1 > 0$ и имеет вид

$$\varphi_\alpha^+ = \begin{cases} 1, & \text{при } \lambda_n > c_\alpha; \\ 0, & \text{при } \lambda_n < c_\alpha, \end{cases}$$

где $\lambda = \sqrt{n}\bar{X}_n$, $\bar{X}_n$ – выборочное среднее, а c_α есть $1-\alpha$-квантиль стандартного нормального распределения, т. е. $\Phi(c_\alpha) = 1-\alpha$ (при $\lambda_n = c_\alpha$ можно принимать любое решение). Этот же критерий является равномерно наиболее мощным уровня значимости α для проверки сложной гипотезы $H_0 : a \le 0$ против сложной правосторонней альтернативы $H_1^+ : a > 0$.

Аналогично, для гипотезы $H_0 : a > 0$ и левосторонней альтернативы $H_1^- : a < 0$ равномерно наиболее мощный критерий уровня значимости $\alpha \in (0,1)$ имеет вид

$$\varphi_\alpha^- = \begin{cases} 1, & \text{при } \lambda_n < -c_\alpha; \\ 0, & \text{при } \lambda_n > -c_\alpha, \end{cases}$$

(при $\lambda_n = c_\alpha$ решение произвольно). Отметим, что для простой гипотезы $H_0 : a = 0$ и двухсторонней альтернативы $H_1^\pm : a \ne 0$ уже не существует равномерно наиболее мощных критериев.

Наложим дополнительное ограничение на класс рассматриваемых критериев, потребовав, чтобы $\beta(\theta_1) \ge \alpha(\theta_0)$ для всех $\theta_0 \in \Theta_0$, $\theta_1 \in \Theta_1$, где $\alpha(\theta_0)$ – вероятность ошибки 1-го рода. Это требование называют *условием несмещенности критерия*. Для гипотезы $H_0 : a = 0$ о среднем нормального распределения и двусторонней альтернативы $H_1^\pm : a \ne 0$ равномерно наиболее мощный несмещенный критерий уровня значимости α имеет вид

$$\varphi_\alpha^\pm = \begin{cases} 1, & \text{при } |\lambda_n| > x_\alpha; \\ 0, & \text{при } |\lambda_n| < x_\alpha, \end{cases}$$

где порог x_α выбирается из условия $\Phi(x_\alpha) = 1 - \alpha/2$.

Если $\mathcal{P}$ – однопараметрическое экспоненциальное семейство с плотностями вида

$$f(\vec{x};\theta) = h(\vec{x}) \exp(a(\theta)U(\vec{x}) + r(\theta)),$$

то оно имеет монотонное отношение правдоподобия относительно $T(\vec{X}) = U(\vec{X})$ при любом θ_*. Более того, существует равномерно наиболее мощный несмещенный критерий уровня значимости α для проверки

гипотезы $H_0 : \theta \in [\theta_1, \theta_2]$ при альтернативе $H_1 : \theta \notin [\theta_1, \theta_2]$, который строится по правилу

$$\varphi(\vec{x}) = \begin{cases} 1, & \text{если } c_1 < U(\vec{X}) < c_2, \\ p_i, & \text{если } U(\vec{X}) = c_i, \\ 0, & \text{если } U(\vec{X}) \notin [c_1, c_2], \end{cases}$$

где константы c_i, p_i, $i = 1, 2$, выбираются из уравнений

$$\mathbf{E}_{\theta_1} \varphi(\vec{X}) = \mathbf{E}_{\theta_2} \varphi(\vec{X}) = \alpha.$$

Рассмотрим пример наиболее мощного критерия проверки сложной односторонней гипотезы при сложной односторонней альтернативе.

2. Для целей некоторого химического производства желательно, чтобы вода содержала не более одной бактерии на единицу объема $v = 1$. Для проверки чистоты воды отбирается n проб объема v. Каждая из этих проб добавляется в пробирку с питательной средой. Если проба была загрязнена (т. е. содержала хоть одну бактерию), то раствор в соответствующей пробирке потемнеет. Считаем, что бактерии случайным образом распределены по исходному объему жидкости. Концентрацией υ будем называть среднее число бактерий на единицу объема. Положим, что $m = \upsilon V$ бактерий случайным образом распределены в объеме V. Тогда вероятность того, что в отобранной пробе объема $v = 1$ будет в точности k бактерий, вычисляется по формуле Бернулли

$$P(\mu_m = k) = C_m^k p^k (1-p)^{m-k},$$

где $p = v/V$. Далее отметим, что по теореме Пуассона приближенно (почти точно, если $v << V$)

$$P(\mu_m = k) \approx \frac{\lambda^k}{k!} e^{-\lambda},$$

где $\lambda = mp = \upsilon V(v/V) = \upsilon v = \upsilon$. В частности $P(\mu_m = 0) = e^{-\upsilon}$. Кроме того, при $v << V$ можно считать, что отбор проб происходит независимо.

Итак, исходный набор наблюдений представляет собой выборку из распределения Бернулли (успех – проба чистая) с параметром $p = e^{-\upsilon}$. Построим наиболее мощный критерий проверки гипотезы $H_0 : p \geq e^{-1} (\upsilon \leq 1)$ при альтернативе $H_1 : p < e^{-1}$ $(\upsilon > 1)$ для выборки из распределения Бернулли. Функция правдоподобия имеет вид

$$L(\vec{x};\, p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{(1-x_i)} = p^{\sum_{k=1}^n x_k} (1-p)^{(n - \sum_{k=1}^n x_k)}.$$

Тогда статистика отношения правдоподобия представляется в виде

$$l(\vec{X}; p, p_0) = \Big(p(1-p_0)/(p_0(1-p))\Big)^{\sum_{k=1}^n X_k}\Big((1-p)/(1-p_0)\Big)^n.$$

Очевидно, что данная статистика монотонно зависит (убывает) от статистики $T(\vec{X}) = \sum\limits_{k=1}^{n} X_k$ – число зараженных проб в выборке. Тогда наиболее мощный критерий имеет такой же вид, как и в предыдущем примере, только знаки неравенств будут противоположными.

К сожалению, даже для случая скалярного параметра $\theta \in \mathbb{R}$ равномерно наиболее мощные или равномерно наиболее мощные несмещенные критерии существуют лишь в специальных случаях, и даже для этих случаев обычно не просто определить пороговое значение статистики критерия. Вместе с тем, обычно существуют асимптотически равномерно наиболее мощные и равномерно наиболее мощные несмещенные критерии (будем их называть асимптотически оптимальными), которые можно построить на основе оценок максимального правдоподобия. Например, пусть проверяется простая гипотеза $H_0 : \theta = \theta_0$, где θ_0 – внутренняя точка конечного или бесконечного интервала $\Theta \subset \mathbb{R}$. Рассмотрим односторонние альтернативы $H_1^+ : \theta > \theta_0$, $H_1^- : \theta < \theta_0$ и двустороннюю альтернативу $H_1^\pm : \theta \neq \theta_0$. Тогда при достаточно общих предположениях регулярности асимптотически оптимальные критерии для этих альтернатив имеют тот же вид, что и критерии $\varphi^+, \varphi^-, \varphi^\pm$ для нормального распределения с тем отличием, что статистика критерия имеет вид

$$\lambda_n = \sqrt{nI(\theta_0)}(\theta_n^* - \theta_0),$$

где θ_n^* – оценка максимального правдоподобия, а $I(\theta_0)$ – значение информации Фишера при нулевой гипотезе.

6.3. Критерий отношения правдоподобия при проверке сложных гипотез

Рассмотрим еще одну статистику, базирующуюся на отношении правдоподобия. Поставим задачу проверки согласия результатов наблюдений с гипотезой $H_0 : \theta \in \Theta_0$, где $\Theta_0 \subset \Theta \subset \mathbb{R}^k$, $\dim(\Theta) = k$, $\dim(\Theta_0) = l$. Рассмотрим статистику отношения правдоподобия

$$\lambda_n = \frac{\sup_{\theta\in\Theta_0} L(\vec{X};\, \theta)}{\sup_{\theta\in\Theta} L(\vec{X};\, \theta)}.$$

При выполнении ряда условий регулярности в условиях H_0 имеет место асимптотическое соотношение

$$\lim_{n\to\infty} P(-2\ln\lambda_n > t) = P(\chi^2_{k-l} > t),\ \ t \geq 0.$$

Данное соотношение позволяет при больших n приближенно вычислять границу критической области с заданным уровнем значимости.

3. Пусть $X_1, \ldots, X_n$ – выборка из двухпараметрического нормального распределения $N(\theta, \sigma^2)$. Рассмотрим гипотезу $H_0 : \theta = \theta_0$ (при неизвестном параметре σ^2 – это сложная гипотеза). Функция правдоподобия имеет вид

$$L_n(\vec{X};\, \theta,\, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2\Big).$$

Отметим, что $\sup(L_n(\vec{X};\, \theta,\, \sigma^2),\ \theta = \theta_0,\ \sigma^2 > 0)$ достигается при $\sigma^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i-\theta_0)^2 = s_0^2$, а $\sup(L_n(\vec{X};\, \theta,\, \sigma^2),\ \theta \in \mathbb{R},\ \sigma^2 > 0)$ – при $\theta = \overline{X}$, $\sigma^2 = s^2$ (оценки максимального правдоподобия). Тогда

$$-2\ln\lambda_n = -2\ln\frac{(s\sqrt{2\pi})^n \exp(-n/2)}{(s_0\sqrt{2\pi})^n \exp(-n/2)} = n\ln(s_0^2/s^2).$$

Далее отметим, что $s^2 = s_0^2 - (\theta_0 - \overline{X})^2$. Следовательно,

$$-2\ln\lambda_n = n\ln\left(\frac{s^2 + (\overline{X}-\theta_0)^2}{s^2}\right) = n\ln\left(1 + \frac{(\overline{X}-\theta_0)^2}{s^2}\right).$$

Данная статистика является монотонной функцией от выражения $\dfrac{(\overline{X}-\theta_0)^2}{s^2}$. Таким образом, асимптотически данный критерий эквивалентен двухстороннему критерию Стьюдента.

Рассмотрим принципиально иную ситуацию. Предположим, что мы имеем дело с двумя различными семействами распределений $\mathcal{P}_0 = \{P_{0,\theta}, \theta \in \Theta\}$ и $\mathcal{P}_1 = \{P_{1,\theta}, \theta \in \Theta\}$, параметризованными, например, одним и тем же параметром и доминированные σ-конечной мерой μ, $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$. Поставим задачу проверки гипотезы $H_0 : P \in \mathcal{P}_1$ при альтернативе $H_1 : P \in \mathcal{P}_1$. Пусть T_0 и T_1 – достаточные статистики данных семейств распределений. Переход к условным распределениям фактически сводит данную задачу к задаче проверки простой гипотезы при простой альтернативе. Введем статистику условного отношения правдоподобия

$$l(\vec{X}, T_1, T_2; H_0, H_1) = \frac{L(\vec{X}|T_1, H_1)}{L(\vec{X}|T_0, H_0)},$$

где $L(\cdot|T_i, H_i)$ – условная плотность распределения при условии статистики T_i и $P \in \mathcal{P}_i$, $i = 0, 1$. Согласно определению достаточной статистики, $L(\cdot|T_i, H_i)$ не зависит от θ, но зависит от семейства распределений. Подобно критерию отношения правдоподобия для проверки простой гипотезы при простой альтернативе, строим статистический критерий значимости

$$\varphi(\vec{X}) = \begin{cases} 0, \ l(\vec{X}, T_1, T_2; H_0, H_1) < c; \\ p, \ l(\vec{X}, T_1, T_2; H_0, H_1) = c; \\ 1, \ l(\vec{X}, T_1, T_2; H_0, H_1) > c, \end{cases}$$

а константы $c > 0$ и $p \in [0, 1]$ выбираются из соотношения $\mathbf{E}_{H_0}\varphi(\vec{X}) = \alpha$. Остается отметить, что если область $\Omega(T_1, T_2)$ обладает свойством, что $P_{H_0}(l(\vec{X}, T_1, T_2) \in \Omega(u_1, u_2)|T_1 = u_1, T_2 = u_2) = \alpha$, то

$$P_{0,\theta}(l(\vec{X}, T_1, T_2) \in \Omega(T_1, T_2)) =$$

$$= \iint\limits_{\mathbb{R}^2} P_{H_0}(l(\vec{X}, T_1, T_2) \in \Omega(u_1, u_2)|T_1 = u_1, T_2 = u_2) p_{T_1,T_2}(u_1, u_2) du_1 du_2 = \alpha$$

для любого значения θ. Следовательно, полученный критерий имеет уровень значимости α. При выборе альтернативы, а также при выборе достаточных статистик T_1 и T_2, следует помнить о том, что условные распределения, стоящие в числителе и знаменателе статистики критерия, должны различаться (хотя бы на каком-либо множестве ненулевой вероятности значений (T_1, T_2)). Например, нельзя брать в качестве достаточных статистик выборку или вариационный ряд.

4. (геометрическое распределение, как альтернатива пуассоновскому). Пусть основная гипотеза H_0 состоит в том, что исходная выборка $X_1, \ldots, X_n$ является выборкой из семейства распределений Пуассона $\mathrm{P}(\theta)$, $\theta > 0$. В качестве альтернативы H_1 положим, что исходная выборка – выборка из геометрического распределения $\mathrm{Nb}(\theta/(1+\theta), 1)$. Минимальная достаточная статистика семейства распределений Пуассона, как и семейства геометрических распределений, $T = n\overline{X}$. Таким образом, $T_1(\vec{X}) = T_2(\vec{X}) = T(\vec{X})$. Условная дискретная плотность распределения $\mathrm{P}(\theta)$ при условии T имеет вид

$$p_{\vec{X}|T}(\vec{x}|H_0) = \frac{T!}{X_1! \cdots X_n!} n^{-T},$$

тогда как условная дискретная плотность распределения $\mathrm{Nb}(\theta/(1+\theta), 1)$

$$p_{\vec{X}|T}(\vec{x}|H_1) = 1/C_{n+T-1}^{T}.$$

Получаем статистику условного отношения правдоподобия

$$l(\vec{X}, T; H_0, H_1) = C^T_{n+T-1} n^{-T} T!/(X_1! \cdots X_n!).$$

Решение уравнения $l(\vec{X}, T; H_0, H_1) > c$, определяющего границу критической области, может быть записано в виде $\sum\limits_{i=1}^{n} \ln X_j! \geq c^*_\alpha$. Следовательно, в качестве статистики критерия удобно выбрать $\sum\limits_{i=1}^{n} \ln(X_j!)$. Доказано, что статистика

$$L_n = \sqrt{n}\Big(\sum_{i=1}^{n} \ln(X_j!) - \sqrt{n} \sum_{k=0}^{\infty} \ln(k!)(\overline{X}^k) e^{-\overline{X}}/k!\Big)$$

является асимптотически нормальной. Для вычисления состоятельной оценки асимптотической дисперсии σ следует подставить в качестве параметра распределения Пуассона θ значение $\overline{X}$ и вычислить соответствующее предельное значение дисперсии. Опуская достаточно сложные выкладки, получаем, что $\sigma_n = \mathbf{D}(\log Y!) - \mathbf{cov}(Y, \log Y)/\overline{X}$, где Y – случайная величина, имеющая распределение Пуассона с параметром $\theta = \overline{X}$. Далее строим асимптотический критерий на базе статистики L_n/σ_n, которая имеет стандартное нормальное асимптотическое распределение.

6.4. Непараметрические критерии

Напомним, что критерии, позволяющие решать задачи проверки статистических гипотез в непараметических семействах (т. е. при бесконечномерном параметре), носят название *непараметрические*. При этом, конечно, непараметрические критерии могут использоваться и в ситуациях, когда одна или даже обе гипотезы параметрические. Мы рассмотрим ряд наиболее популярных критериев проверки согласия с параметрической гипотезой H_0.

Критерий Колмогорова. Данный критерий предназначен для проверки согласия с простой гипотезой $H : \theta = \theta_0$ и позволяет работать с непараметрическими семействами *непрерывных* распределений. Учитывая характер статистики критерия, исходную гипотезу удобно записывать в терминах функций распределения

$$H_0 : F \equiv F_0,$$

где F_0 – некоторая фиксированная функция распределения. В качестве статистики критерия выбираем статистику Колмогорова

$$D_n = \sup_{x\in(-\infty,\infty)} |F_n(x) - F_0(x)|.$$

Теорема Колмогорова дает предельное распределение нормированной статистики Колмогорова. Если $F(x)$ непрерывна, то для любого положительного значения t

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

При вычислении значения статистики Колмогорова следует использовать тот факт, что выборочная функция распределения $F_n(x)$ является кусочно-постоянной функцией, поэтому

$$D_n = \max_{i\in\{1,\ldots,n\}} \max(|F_n(X_i) - F(X_i)|, |F_n(X_{i+}) - F(X_i)|).$$

Функция $K(t)$, которая носит название *функция распределения Колмогорова*, затабулирована, и таблицы ее значений присутствуют в во всех статистических справочниках и в некоторых задачниках по теории вероятностей и математической статистике. Отметим, что критерий Колмогорова состоятелен для любой фиксированной простой альтернативы. Более того, он будет состоятельным для любой последовательности альтернатив $H_n : F \in \Theta_{An}$, где $\Theta_{An} \subset \{G : \sqrt{n}\sup_x |G(x) - F_0(x)| > M_n\}$, $M_n \xrightarrow[n\to\infty]{} \infty$.

Критерий χ^2 для проверки простой и сложной гипотез согласия. Пусть $\vec{X} = (X_1, \ldots, X_n)$ – выборка из распределения P с функцией распределения F. Рассмотрим задачу проверки согласия с гипотезой $H_0 : F = F_0$ (подразумевается, что семейство $\mathcal{P}$ непараметрическое). Предположим, что множество возможных значений X_1 (определяемое видом семейства $\mathcal{P}$) разбито на N непересекающихся подмножеств, например само состоит из N точек (N исходов эксперимента, N состояний системы) или множество $-\infty < x < \infty$ разбито на интервалы $I_i = [t_{i-1}, t_i]$, $i = 1, \ldots, N$ с границами $-\infty = t_1 < t_2 < \ldots < t_{N-1} < t_N = \infty$. Пусть n_i – число элементов выборки $\vec{X} = (X_1, \ldots, X_n)$, попавших в интервал I_i. Обозначим вероятности $P_{H_0}(X_1 \in I_i)$ через p_i (теоретические частоты). Частота n_i/n является состоятельной оценкой p_i. Используем следующую меру отклонения выборочных значений от теоретических $S = \sum_{i=1}^{N} f_i(\frac{n_i}{n} - p_i)^2$, где

f_i – некоторые веса. Если в качестве таких весов взять n/p_i, то получится статистика так называемого критерия *хи-квадрат* (χ^2)

$$X^2 = \sum_{i=1}^{N} \frac{n}{p_i}\left(\frac{n_i}{n} - p_i\right)^2 = \sum_{i=1}^{N} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{N} \frac{n_i^2}{np_i} - n.$$

Отметим, что распределение вектора $n_1, \ldots, n_N$ является мультиномиальным. В терминах нормированных частот $(n_1^*, \ldots, n_N^*) = \vec{n}^*$, $n_i^* = (n_i - np_i)/\sqrt{n}$ статистика X^2 имеет вид

$$X^2 = \sum_{i=1}^{N} \frac{1}{p_i}(n_i^*)^2 + \frac{1}{p_N}(n_1^* + \ldots + n_{N-1}^*)^2 = \vec{n}^{*\mathrm{T}}\Omega\vec{n}^*,$$

где $\Omega = ||g_{i,j}||_1^N$, $g_{i,i} = \frac{1}{p_i} + \frac{1}{p_N}$, $g_{i,j} = \frac{1}{p_N}$, $i \neq j$ и, как легко проверить, $\Omega = \Sigma^{-1}$, где Σ – ковариационная матрица соответствующего предельного (многомерного нормального) распределения. Это объясняет тот факт, что X^2 сходится по распределению к χ^2_{N-1} (см. 4). Данный критерий не может различать распределения с одинаковыми теоретическими частотами. Имеются рекомендации (в методе группировки) разбивать вещественную прямую на интервалы, вероятности которых достаточно близки между собой. Если n и n_i достаточно велики, то критерий согласия выглядит так: если α – уровень значимости, то при $X^2(\vec{n}) \geqslant \chi^2_{1-\alpha,N-1}$, где $\chi^2_{1-\alpha,N-1}$ – квантиль порядка $1-\alpha$ распределения χ^2_{N-1}, гипотезу H_0 отвергают, в противном случае – нет. При справедливости альтернативной гипотезы предельным распределением для X^2 будет нецентральное χ^2-распределение с параметром нецентральности $\lambda^2 = \sqrt{n}\sum\limits_{i=1}^{N}(p_i - p_i^*)^2/p_i^*$, где p_i^* и p_i – истинный и гипотетический параметры мультиномиального распределения соответственно.[1]

Широко применяемый метод проверки сложных гипотез с помощью критерия χ^2 включает в себя оценки максимума правдоподобия для параметров.

Пусть $\vec{X} = \{X_1, \ldots, X_n\}$ – выборка из распределения с полностью неизвестной функцией распределения. Проверим гипотезу H_0 о принадлежности теоретического распределения исходной выборки некоторому параметрическому семейству $F \in \{P_\theta,\ \theta \in \Theta, \Theta \subset R^m\}$. Пусть осуществ-

[1] Нецентральное χ^2-распределение имеет достаточно сложную структуру и здесь не приводится.

лена группировка данных (т. е. задача сведена к мультиномиальному распределению) или изначально речь идет о мультиномиальном распределении, зависящем от многомерного параметра. Обозначим через $\vec{n}$ вектор $\{n_1, \ldots, n_N\}$, $\sum\limits_{j=1}^{N} n_j = n$. В указанной постановке значение X^2 при справедливости основной гипотезы зависит от θ:

$$X^2(\theta) = \sum_{i=1}^{N} \frac{(n_i - np_i(\theta))^2}{np_i(\theta)},$$

поэтому мы не можем использовать данное выражение в качестве статистики критерия. Выберем из всех значений $X^2(\theta)$ наиболее подходящее. В качестве статистики критерия будем использовать $\widetilde{X}^2 = X^2(\widetilde{\theta}) = \min_{\theta \in \Theta} X^2(\theta)$. При определенных условиях регулярности статистика $\widetilde{X}^2$ асимптотически эквивалентна статистике $\widehat{X}^2 = X^2(\widehat{\theta})$, если $\widehat{\theta}$ – мультиномиальная оценка максимального правдоподобия[1] параметра θ и при справедливости H_0 распределения $\widetilde{X}^2$ и $\widehat{X}^2$ сходятся к χ^2_{N-m-1}, при $n \to \infty$, где m – число оцениваемых параметров (которое должно совпадать с рангом матрицы $\left\| \dfrac{\partial p_i(\theta)}{\partial \theta_k} \right\|_{i=1\ k=1}^{N \quad m}$.

Итак, имея вектор $\vec{n} = \{n_1, \ldots, n_N\}$, для нахождения максимума $L_n(\widehat{\theta})$ функции правдоподобия соответствующего мультиномиального распределения по $\theta \in \theta$

$$L(\vec{n};\, \theta) = \frac{n!}{n_1! \ldots n_N!} \prod_{j=1}^{N} p_j^{n_i}(\theta)$$

необходимо решить уравнения

$$\sum_{j=1}^{N} \frac{n_j}{p_j(\theta)} \frac{\partial p_j(\theta)}{\partial \theta_k} = 0,\ k = 1, \ldots, m.$$

Отдельно отметим, что если использовать оценки максимального правдоподобия относительно исходного семейства распределений, то асимптотическое распределение может отличаться от χ^2.

Рассмотрим ряд примеров.

[1] Мультиномиальной оценкой максимального правдоподобия будем называть оценку, максимизирующую функцию правдоподобия соответствующего мультиномиального распределения по параметру $\theta \in \Theta$.

5. (Генетическая модель Фишера). При самоскрещивании кукурузы по двум характеристикам (крахмалистая, сахаристая, с белым основанием листа, с зеленым основанием листа) возникают 4 типа потомства. Если N_i – число растений типа i среди общего числа n экземпляров потомства и θ_i – вероятность такого типа, то (N_1, N_2, N_3, N_4) имеет мультиномиальное (прил. 3) распределение с вероятностями

$$P(N_1, N_2, N_3, N_4) = \frac{n!\theta_1^{N_1}\theta_2^{N_2}\theta_3^{N_3}\theta_4^{N_4}}{N_1!N_2!N_3!N_4!}.$$

В модели Фишера величины θ_i выбираются так, что $\theta_1 = (2+\theta)/4$; $\theta_2 = \theta_3 = (1-\theta)/4$; $\theta_4 = \theta/4$, где θ – некий параметр. Нам надо проверить гипотезу о том, что в нашей мультиномиальной модели θ_i являются указанными функциями θ, т. е. о том, что параметры теоретического мультиномиального распределения лежат на данной прямой в трехмерной гиперплоскости $(\theta_1, \theta_2, \theta_3, \theta_4)$ с $\sum\limits_{i=1}^{4} \theta_i = 1$.

В соответствии с описанной ранее процедурой мы должны, располагая выборкой N_1, N_2, N_3, N_4, оценить параметр θ методом максимума правдоподобия, т. е. выбрать θ, на котором реализуется $\max \prod\limits_{i=1}^{4} \theta_i(\theta)^{N_i}$, или, что то же самое, $\max \sum\limits_{i=1}^{4} N_i \ln(\theta_i(\theta))$. Находим производную $\sum\limits_{i=1}^{4} N_i\, \theta_i'(\theta)/\theta_i(\theta)$ и приравниваем ее к нулю:

$$\frac{N_1}{2+\theta} - \frac{N_2+N_3}{1-\theta} + \frac{N_4}{\theta} = 0.$$

Отсюда $n\theta^2+(N_4+2N_2+2N_3-N_1)\theta-2N_4 = 0$. Но левая часть меньше нуля при $\theta = 0$; больше нуля при $\theta = 1$; поэтому в интервале (0, 1) уравнение имеет единственный корень $\widehat{\theta}_n$. Таким образом, критерий согласия χ^2 при уровне значимости α отклоняет гипотезу H_0 лишь в том случае, если

$$\sum_{j=1}^{4} \frac{N_j^2}{(n\theta_j(\widehat{\theta}_n))} - n \geqslant \chi^2_{1-\alpha,\,2}$$

или, что эквивалентно,

$$\frac{N_1^2}{n(2+\widehat{\theta}_n)} + \frac{N_2^2+N_3^2}{n(1-\widehat{\theta}_n)} + \frac{N_4^2}{n\widehat{\theta}_n} \geqslant (\chi^2_{1-\alpha,\,2} + n)/4.$$

6. (Проверка гипотезы о показательном распределении). По выборке $\vec{X} = (X_1, \ldots, X_n)$ надо проверить гипотезу H_0 о том, что распределение имеет функцию распределения $F_\xi(x) = 1 - e^{-x/\theta}$, $x > 0$ (параметр $\theta > 0$ неизвестен). Применяя метод группировки данных с интервалами $E_j = [(j-1)a, ja]$, $j = 1, \ldots, N-1$, $E_N = [(N-1)a, \infty)$, где $a > 0$, N – заданные числа, построим критерий согласия χ^2 для гипотезы H_0. Обозначим через $p_j(\theta)$, $j = 1, \ldots, N$, вероятности $P(\xi \in E_j | H_0)$. Имеем $p_j(\theta) = e^{-(j-1)a/\theta}(1 - e^{-(a/\theta)})$, $j = 1, \ldots, N-1$, $p_N(\theta) = e^{-(N-1)a/\theta}$. Как и во всех подобных задачах, метод максимума правдоподобия нужно применять к схеме независимых испытаний с N исходами и вышеописанными вероятностями. Располагая выборкой объема n, мы знаем число n_j элементов выборки, попавших в интервал E_j, $j = 1, \ldots, N$; $\sum\limits_{j=1}^{N} n_j = n$. Функция правдоподобия имеет вид $\prod\limits_{i=1}^{N} p_i(\theta)^{n_i}$, ее логарифм – $\sum\limits_{i=1}^{N} n_i \ln p_i(\theta)$, и уравнение максимума правдоподобия есть снова $\sum\limits_{i=1}^{N} n_i \dfrac{p_i'(\theta)}{p_i(\theta)} = 0$. Вынося за скобки общий множитель $(-a/\theta)'$ получаем, что $\sum\limits_{j=1}^{N-1} n_j \dfrac{j - 1 - je^{-a/\theta}}{1 - e^{-a/\theta}} + (N-1)n_N = 0$. Обозначая $z = e^{-a/\theta}$, имеем $\hat{z}_N = (\sum\limits_{j=1}^{N} jn_j - n)/(\sum\limits_{j=1}^{N} jn_j - n_N)$ и оценки максимума правдоподобия для вероятностей $\hat{p}_j = p_j(\hat{\theta})$ имеют вид $\hat{p}_j = \hat{z}_N^{j-1}(1 - \hat{z}_N)$, $j = 1, \ldots, N-1$, $\hat{p}_N = \hat{z}_N^{N-1}$.

Соответствующий критерий согласия χ^2 отвергает гипотезу H_0 тогда и только тогда, когда $\widehat{X}^2 = \sum\limits_{j=1}^{N} \dfrac{(n_j - n\hat{p}_j)^2}{n\hat{p}_j} \geqslant \chi^2_{1-\alpha, N-2}$, где $\chi^2_{1-\alpha}$ – квантиль χ^2 распределения с $N-2$ степенями свободы порядка $1-\alpha$. В соответствии с общей теорией данный критерий является асимптотическим критерием уровня α, если $\sup_{i \in \{1,\ldots,N\}} np_i \to \infty$ при $n \to \infty$.

Список рекомендованной литературы

Боровков А. А. Математическая статистика. М.: Наука, 1984.

Ивченко Г. И., Медведев Ю. И. Математическая статистика. М.: Высш. шк., 1984.

Крамер Г. Математические методы статистики. 2-е изд. М.: Мир, 1975.

Кокс Д., Хинкли Д. Теоретическая статистика. М. Мир, 1978.

Леман Э. Теория точечного оценивания. М.: Наука, 1991.

Леман Э. Проверка статистических гипотез. М.: Наука, 1964.

ПРИЛОЖЕНИЯ

1. Таблицы распределений

Квантили $t\gamma$ распределения Стьюдента: $S_n(t\gamma) = \gamma$

n	γ						n	γ					
	0.8	0.9	0.95	0.975	0.99	0.995		0.8	0.9	0.95	0.975	0.99	0.995
1	1.376	3.078	6.314	12.71	31.82	63.66	12	0.873	1.356	1.782	2.179	2.681	3.055
2	1.061	1.886	2,920	4.303	6.965	9.925	15	0.866	1.341	1.753	2.131	2.602	2.947
3	0.978	1.638	2.353	3.182	4.541	5.841	18	0.862	1.330	1.734	2.101	2.552	2.878
4	0.941	1.533	2.132	2.776	3.747	4.604	20	0.860	1.325	1.725	2.086	2.528	2.845
5	0.920	1.476	2.015	2.571	3.365	4.032	25	0.856	1.316	1.708	2.059	2.485	2.787
6	0.906	1.440	1.943	2.447	3.143	3.707	30	0.854	1.310	1.697	2.042	2.457	2.750
7	0.896	1.415	1.895	2.365	2.998	3.499	40	0.851	1.303	1.684	2.021	2.423	2.704
8	0.889	1.397	1.860	2.306	2.896	3.355	50	0.849	1.299	1.676	2.009	2.403	2.678
9	0.883	1.383	1.833	2.262	2.821	3.250	100	0.845	1.290	1.660	1.984	2.364	2.626
10	0.879	1.372	1.812	2.228	2.764	3.169	∞^*	0.842	1.282	1.645	1.960	2.326	2.576

Квантили $x\gamma$ распределения χ^2_n: $K_n(x\gamma) = \gamma$

n	γ												
	0.005	0.01	0.02	0.05	0.1	0.2	0.5	0.8	0.9	0.95	0.98	0.99	0.995
1	0.000	0.000	0.001	0.004	0.016	0.064	0.455	1.642	2.706	3.841	5.412	6.635	7.879
2	0.010	0.020	0.040	0.103	0.211	0.446	1.386	3.219	4.605	5.991	7.824	9.210	10.60
3	0.072	0.115	0.185	0.352	0.584	1.005	2.366	4.642	6.251	7.815	9.837	11.34	12.84
4	0.207	0.297	0.429	0.711	1.064	1.649	3.357	5.989	7.779	9,487	11.67	13.28	14.86
5	0.412	0.554	0.752	1.145	1.610	2.343	4.351	7.289	9.236	11.07	13.39	15.09	16.75
6	0.676	0.872	1.134	1.635	2.204	3.070	5.348	8.558	10.64	12.59	15.03	16.81	18.55
7	0.989	1.239	1.564	2.167	2.833	3.822	6.346	9.803	12.02	14.07	16.62	18.48	20.28
8	1.344	1.646	2.032	2.733	3.490	4.594	7.344	11.03	13.36	15.51	18.17	20.09	21.95
9	1.735	2.088	2.532	3.325	4.168	5.380	8.343	12.24	14.68	16.92	19.68	21.67	23.59
10	2.156	2.558	3.059	3.940	4.865	6.179	9.342	13.44	15.99	18.31	21.16	23.21	25.19
11	2.603	3.053	3.609	4.575	5.578	6.989	10.34	14.63	17.28	19.68	22.62	24.72	26.76
12	3.074	3.571	4.178	5.226	6.304	7.807	11.34	15.81	18.55	21.03	24.05	26.22	28.30
13	3.565	4.107	4.765	5.892	7.042	8.634	12.34	16.98	19.81	22.36	25.47	27.69	29.82
14	4.075	4.660	5.368	6.571	7.790	9.467	13.34	18.15	21.06	23.68	26.87	29.14	31.32

Критические значения λ_α распределения статистики Колмогорова: $P(\sqrt{n}\,D_n > \lambda_\alpha) = \alpha$

n	α					n	α				
	0.2	0.1	0.05	0.02	0.01		0.2	0.1	0.05	0.02	0.01
5	0.9995	1.1392	1.2595	1.4024	1.4949	40	1.0465	1.1962	1.3289	1.4859	1.5941
10	1.0202	1.1658	1.2942	1.4440	1.5461	50	1.0493	1.1992	1.3323	1.4897	1.5983
20	1.0356	1.1839	1.3152	1.4698	1.5760	100	1.0563	1.2067	1.3403	1.4987	1.6081
30	1.0424	1.1916	1.3238	1.4801	1.5876	∞^{**}	1.0727	1.2238	1.3581	1.5174	1.6276

* Случай $n = \infty$ соответствует стандартному нормальному распределению.

** Случай $n = \infty$ соответствует предельному распределению Колмогорова.

2. Основные классы распределений и их характеристики (начало)

Абсолютно нерперывные распределения

Класс распределения	Обозначение	Плотность распределения	Характеристическая функция
Нормальное /Гауссовское/	$N(a,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$	$\exp\left(ita-\frac{\sigma^2t^2}{2}\right)$
Равномерное	$U(a,b)$	$\mathbb{1}_{[a,b]}(x)/(b-a)$	$\frac{\exp(itb)-\exp(ita)}{it(b-a)}$
Показательное	E(a,b)	$\frac{\exp(-(x-a)/b)}{b}\mathbb{1}_{[a,\infty)}(x)$	$\frac{\exp(ita)}{1-itb}$
Двухстороннее показательное /Лапласа/	DE (a,b)	$\frac{1}{2b}\exp\left(-\frac{\lvert x-a\rvert}{b}\right)$	$\frac{\exp(ita)}{1+t^2b^2}$
Гамма	$\Gamma(b,p)$	$\frac{x^{p-1}\exp(-x/b)}{b^p\Gamma(p)}\mathbb{1}_{[0,\infty)}(x)$	$\frac{1}{(1-ibt)^p}$
Коши	C(a,b)	$b/(\pi(b^2+(x-a)^2))$	$\exp(ita-\lvert bt\rvert)$
Бета	B(a,b)	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}\mathbb{1}_{[0,1]}(x)$	$\frac{\Gamma(a+b)}{\Gamma(a)}\sum_{k=0}^{\infty}\frac{(it)^k\Gamma(a+k)}{k!\,\Gamma(a+b+k)}$
Парето	P(a,b)	$ba^bx^{-(b+1)}\mathbb{1}_{[a,\infty)}(x)$	...
Логистическое	L(a,b)	$\frac{\exp(-(x-a)/b)}{b(1+\exp(-(x-a)/b))^2}$	$\exp(ita)\Gamma(1+itb)\Gamma(1-itb)$
Треугольное	Tr(a,b)	$b^{-1}(1-\lvert x-a\rvert/b)\mathbb{1}_{[-b,b]}(x)$	$2(bt)^{-2}\exp(ita)(1-\cos(bt))$
Хинчина	–	$(1-\cos(x))/(\pi x^2)$	$1-\lvert t\rvert$
t-распределение (Стьюдента)	T(b,p)	$\frac{\sqrt{2}\Gamma((p+1)/2)}{\Gamma(p/2)\sqrt{\pi p\,b}}\left(1+\frac{2x^2}{p\,b}\right)^{-(p+1)/2}$	...

Дискретные распределения

Класс распределения	Обозначение	Дискретная плотность распределения	Характеристическая функция
Биномиальное	Bi(p,m)	$C_m^kp^k(1-p)^{m-k}$, $k=0,\dots,m$	$(1+p(\exp(it)-1))^m$
Отрицательное биномиальное /Паскаля/	Nb(p,m)	$C_{m+k-1}^kp^m(1-p)^k$; $k=0,1,\dots$	$\left(\frac{p}{1-(1-p)\exp(it)}\right)^m$
Пуассона	P(a)	$\frac{a^k}{k!}\exp(-a)$, $k=0,1,\dots$	$\exp(a(\exp(it)-1))$

2. Основные классы распределений и их характеристики (продолжение)

Абсолютно нерперывные распределения

Распределение	Числовые характеристики $\mathbf{E}X$; $\mathbf{D}X$; $\mathbf{Asi}X$; $\mathbf{Ex}X$	Информация Фишера
$N(a,\sigma^2)$	$a;\ \sigma^2;\ 0;\ 0$	$I(a)=\sigma^{-2},\ I(\sigma^2)=1/2\sigma^4$
$U(a,b)$	$(b+a)/2;\ (b-a)^2/12;\ 0;\ -6/5$	не существует
E(a,b)	$a+b;\ b^2;\ 2;\ 6$	$I(a)-\nexists;\ I(b)=1/b^2$
DE (a,b)	$a;\ 2b^2;\ 0;\ 3$	$I(a)=I(b)=1/b^2$
$\Gamma(b,p)$	$bp;\ b^2p;\ 2p^{-1/2};\ 6p^{-1}$	$I(b)=p/b^2;\ I(p)=\psi'(p)^*$
C(a,b)	не существуют	$I(a)=I(b)=1/(2b^2)$
B(a,b)	$\frac{a}{a+b};\ \frac{ab}{(a+b)^2(a+b+1)};\ \frac{2(b-a)}{2+a+b}\left(\frac{1+a+b}{ab}\right)^{1/2};$ $\frac{(2(a^2-ab+b^2)+ab(a+b))(1+a+b)}{(2+a+b)(3+a+b)}-3$	$I(a)=\psi'(a)-\psi'(a+b)^*;$ $I(b)=\psi'(b)-\psi'(a+b)$
P(a,b)	$\frac{ab}{b-1}, b>1;\ \frac{a^2b}{(b-1)^2(b-2)}, b>2;$ $\frac{2(b-2)^{3/2}(1+b)}{\sqrt{b}(b^2-5b+6)}, b>3;\ \frac{6(b^3+b^2-6b-2)}{b(b^2-7b+12)}, b>4$	$I(a)-\nexists;\ I(b)=1/b^2$
L(a,b)	$a;\ b^2\pi^2/3;\ 0;\ 6/5$	$I(a)=\frac{1}{3b^2};\ I(b)=\frac{3+\pi^2}{9b^2}$
Tr(a,b)	$a;\ b^2/6;\ 0;\ -3/5$	не существует
Хинчина	не существуют	не существует
T(b,p)	$0, p>1;\ 2(p-2)/(bp\,(p-1)^2), p>2;\ 0, p>3;$ $6(7-6p+p^2)/((p-2)(p-3)^2), p>4$	$I\left(\sqrt{\frac{bp}{2}}\right)=\frac{2}{bp}-\frac{3\sqrt{\pi}\,\Gamma\left(\frac{p+1}{2}\right)}{2bp\,\Gamma(p/2)}$

Дискретные распределения

Распределение	Числовые характеристики $\mathbf{E}X$; $\mathbf{D}X$; $\mathbf{Asi}X$; $\mathbf{Ex}X$	Информация Фишера
Bi(p,m)	$mp;\ mp(1-p);\ (1-2p)/(mp(1-p))^{1/2};$ $3(m^2-1)+m(1-6p+6p^2)/(p(1-p))$	$I(p)=\frac{m}{p(1-p)}$
Nb(p,m)	$m(1-p)/p;\ m(1-p)/p^2;\ (p-2)/(m(1-p))^{1/2}$ $3(m^2-1)+m(6-6p+p^2)/(1-p)$	$I(p)=\frac{m}{p^2(1-p)}$
P(a)	$a;\ a;\ 1/\sqrt{a};\ 1/a$	$I(a)=1/a$

*$\psi(p)=\Gamma'(p)/\Gamma(p)$.

3. Некоторые многомерные распределения

Абсолютно непрерывные многомерные распределения

Класс	Обозначение	Плотность распределения	Характеристическая функция	Числовые характеристики $\mathbf{E}\vec{X}=(a_1,\ldots,a_s)$ $\mathbf{D}\vec{X}=\|\sigma_{i,j}\|_{i,j=1}^{s}$.
Нормальное /Гауссовское/	$\mathcal{N}(\vec{a},R)$	$p(\vec{x};\vec{a},R)=(2\pi)^{-n/2}\det(R)^{-1/2}\times$ $\exp(-(\vec{x}-\vec{a})R^{-1}(\vec{x}-\vec{a})^T)/2$	$f(\vec{t};\vec{a},R)=\exp(i<\vec{t},\vec{a}>-<\vec{t}R,\vec{t}>)$	$\vec{a};\ R.$
Дирихле	–	$p(\vec{x};\vec{a},R)=\dfrac{\Gamma(p_1+\ldots+p_s)}{\Gamma(p_1)\cdots\Gamma(p_s)}x_1^{p_1-1}\cdots x_s^{p_s-1};$ $x_i\geqslant 0,\quad \sum x_i=1.$	$\ldots$	$(p_1/p,\ldots,p_s/p);$ $\sigma_{i,i}=\dfrac{p_i(p-p_i)}{p^2(p+1)},$ $\sigma_{i,j}=\dfrac{p_ip_j}{p(p+1)};$ $p=p_1+\ldots+p_s$

Дискретные многомерные распределения

Класс	Обозначение	Дискретная плотность распределения	Характеристическая функция	Числовые характеристики $\mathbf{E}\vec{X}=(a_1,\ldots,a_s)$ $\mathbf{D}\vec{X}=\|\sigma_{i,j}\|_{i,j=1}^{s}$.
Полиномиальное /мультиномиальное/	$\mathrm{M}(\vec{p},n)$	$q(\vec{k};\vec{p},n)=\dfrac{n!}{k_1!\ldots k_s!}p_1^{k_1}\ldots p_s^{k_s},$ $k_1+\ldots+k_s=n$	$f(\vec{t};\vec{p},n)=\left(1+\sum\limits_{r=1}^{s}p_r(e^{it_r}-1)\right)^n$	$(np_1,\ldots,np_s);$ $\sigma_{i,i}=np_i(1-p_i);$ $\sigma_{i,j}=-np_ip_j,\ i\neq j$

4. Статистические критерии

Название	Семейство	Гипотеза	Статистика критерия G(θ,X)	Примечания	Предельное распределение при H_0
Стьюдента	$\mathcal{N}(a, \sigma^2)$ $a \in \mathbb{R}$, $\sigma > 0$.	параметрическая сложная согласия с $H_0 : a = a_0$.	$S_n = \dfrac{\overline{X} - a}{s}$	σ неизвестно	Распределение Стьюдента
Хи-квадрат (согласия с простой гипотезой)	Непарам.	согласия частот (простая)	$X^2 = \sum\limits_{i=1}^{n} \dfrac{(n_i - np_i)^2}{np_i}$	r – число зон n_i – число эл-тов в i-й зоне p_i – теоретическая вероятность попадания в i-ю зону при H_0	χ^2_{r-1} – хи-квадрат
Хи-квадрат (согласия для сложной гипотезы)	Непарам.	согласия частот (сложная, параметрическая)	$X^2 = \sum\limits_{i=1}^{n} \dfrac{(n_i - np_i(\theta_n))^2}{np_i(\theta_n)}$	r – число зон n_i – число эл-тов в i-й зоне p_i – теоретическая вероятность попадания в i-ю зону при $\theta = \theta_n$ θ_n – мультиномиальная ОМП или оценка $\min X^2$	χ^2_{r-l-1} – хи-квадрат l-размерность параметра
Колмогорова	Непарам.	простая согласия с $H_0 : F \equiv F_0$	$\sqrt{n} \max\limits_{t \in (-\infty, \infty)} \lvert F_n(t) - F_0(t) \rvert^*$	F_n– эмпирическая функция распределения F_0– непрерывная гипотетическая функция распределения	Распределение Колмогорова
Крамера – фон-Мизеса – Смирнова	Непарам.	простая согласия с $H_0 : F \equiv F_0$	$\dfrac{1}{12n^2} + \dfrac{1}{n} \sum\limits_{i=1}^{n} \left(F_0(X_{(i)}) - \dfrac{2j-1}{2n} \right)^2$	F_n– эмпирическая функция распределения F_0– непрерывная гипотетическая функция распределения	Распределение ω^2 (омега-квадрат)

* Критерий Колмогорова, как и критерий Крамера–фон Мизеса–Смирнова, входит в группу, так называемых критериев расстояний. В эту группу также входят весьма популярные асимптотические критерии, основанные на U-статистиках.

Оглавление

Введение ... 3
1. Точечное оценивание ... 5
2. Эмпирическое распределение ... 9
3. Выборочные числовые характеристики ... 11
4. Параметрические классы распределений ... 13
5. Параметрическое оценивание ... 20
5.1. Методы построения статистических оценок ... 20
5.2. Доверительное оценивание ... 25
6. Проверка статистических гипотез ... 32
6.1. Методы построения статистических критериев ... 35
6.2. Наиболее мощные критерии ... 36
6.3. Критерий отношения правдоподобия при проверке сложных гипотез ... 41
6.4. Непараметрические критерии ... 44
Список рекомендованной литературы ... 50
Приложения
1. Таблицы распределений ... 51
2. Основные классы распределений и их характеристики ... 52
3. Некоторые многомерные распределения ... 54
4. Статистические критерии ... 55

Владимир Алексеевич Егоров, Юрий Измайлович Ингстер,
Александр Нахимович Лившиц, Ирина Юрьевна Малова,
Сергей Васильевич Малов

Анализ однородных статистических данных

Учебное пособие

Редактор Н. В. Лукина

Подписано в печать . Формат 60 × 84 1/16.
Бумага офсетная. Печать офсетная. Печ. л. 3.5
Гарнитура "Times New Roman". Тираж 250 экз. Заказ .

Издательство СПбГЭТУ "ЛЭТИ"
197376, С.-Петербург, ул. Проф. Попова, 5