Министерство Российской Федерации
по связи и информатизации

Сибирский государственный университет
телекоммуникаций и информатики

Н. И. Чернова

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие

Новосибирск
2009

УДК 519.2

Доцент, канд. физ.-мат. наук Н. И. Чернова. Математическая статистика: Учебное пособие / СибГУТИ.— Новосибирск, 2009.— 90 с.

Учебное пособие содержит полугодовой курс лекций по математической статистике для студентов экономических специальностей. Учебное пособие соответствует требованиям Государственного образовательного стандарта к профессиональным образовательным программам по специальности 080116 — «Математические методы в экономике».

Кафедра ММБП
Табл. 7, рисунков — 9, список лит. — 8 наим.
Рецензенты: А. П. Ковалевский, канд. физ.-мат. наук, доцент кафедры высшей математики НГТУ
В. И. Лотов, д-р физ.-мат. наук, профессор кафедры теории вероятностей и математической статистики НГУ

Для специальности 080116 — «Математические методы в экономике»

Утверждено редакционно-издательским советом СибГУТИ в качестве учебного пособия

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Учебное пособие содержит полный курс лекций по математической статистике для студентов, обучающихся по специальности «Математические методы в экономике» Сибирского государственного университета телекоммуникаций и информатики. Содержание курса полностью соответствует образовательным стандартам подготовки бакалавров по указанной специальности.

Курс математической статистики опирается на семестровый курс теории вероятностей и является основой для годового курса эконометрики. В результате изучения предмета студенты должны овладеть математическими методами исследования различных моделей математической статистики.

Курс состоит из восьми глав. Первая глава является главной для понимания предмета. Она знакомит читателя с основными понятиями математической статистики. Вторая глава посвящена методам точечного оценивания неизвестных параметров распределения: моментов и максимального правдоподобия.

Третья глава рассматривает сравнение оценок в среднеквадратичном смысле. Здесь же изучается неравенство Рао — Крамера как средство проверки эффективности оценок.

В четвёртой главе рассматривается интервальное оценивание параметров, которое завершается в следующей главе построением интервалов для параметров нормального распределения. Для этого вводятся специальные статистические распределения, которые затем используются в критериях согласия в восьмой главе. Глава шестая даёт необходимые основные понятия теории проверки гипотез, поэтому изучить её читателю следует весьма тщательно.

Наконец, главы седьмая и восьмая дают перечень наиболее часто используемых на практике критериев согласия. В девятой главе рассмотрены простые модели и методы регрессионного анализа и доказаны основные свойства полученных оценок.

Практически каждая глава завершается списком упражнений по тексту главы. Приложение содержит таблицы с перечнем основных характеристик дискретных и абсолютно непрерывных распределений, таблицы основных статистических распределений.

В конце книги приведен подробный предметный указатель. В списке литературы перечислены учебники, которые можно использовать в дополнение к курсу, и сборники задач для практических занятий.

Нумерация параграфов в каждой главе отдельная. Формулы, примеры, утверждения и т. п. имеют сквозную нумерацию. При ссылке на объект из другой главы для удобства читателя указан номер страницы, на которой содержится объект. При ссылке на объект из той же главы приводится только номер формулы, примера, утверждения. Окончание доказательств отмечено значком $\square$.

ГЛАВА I
ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Математическая статистика опирается на методы теории вероятностей, но решает иные задачи. В теории вероятностей рассматриваются случайные величины с *заданным* распределением или случайные эксперименты, свойства которых целиком известны. Но откуда берутся знания о распределениях в практических экспериментах? С какой вероятностью, например, выпадает герб на данной монете? Для определения этой вероятности мы можем подбрасывать монету много раз. Но в любом случае выводы придётся делать по результатам конечного числа наблюдений. Так, наблюдая 5 035 гербов после 10 000 бросаний монеты, нельзя сделать точный вывод о вероятности выпадения герба: даже если эта вероятность отличается от 0,5, герб может выпасть 5035 раз. Точные выводы о распределении можно делать лишь тогда, когда проведено бесконечное число испытаний, что неосуществимо. Математическая статистика позволяет по результатам конечного числа экспериментов делать более-менее точные выводы о распределениях случайных величин, наблюдаемых в этих экспериментах.

§ 1. Задачи математической статистики

Предположим, что мы повторяем один и тот же случайный эксперимент в одинаковых условиях. В результате каждого повторения эксперимента наблюдается некоторый набор данных (числовых или каких-то иных).

При этом возникают следующие вопросы.

1. Если наблюдается одна случайная величина — как по набору её значений в нескольких экспериментах сделать возможно более точный вывод о её распределении?

2. Если наблюдается проявление двух или более признаков, — что можно сказать о виде и силе зависимости наблюдаемых случайных величин?

Часто можно высказать некие предположения о наблюдаемом распределении или о его свойствах. В этом случае по опытным данным требуется подтвердить или опровергнуть эти предположения («гипотезы»). При этом надо помнить, что ответ «да» или «нет» может быть дан лишь с определенной степенью достоверности, и чем дольше мы можем продолжать эксперимент, тем точнее могут быть выводы. Иногда можно заранее утверждать о наличии

некоторых свойств наблюдаемого эксперимента — например, о функциональной зависимости между наблюдаемыми величинами, о нормальности распределения, о его симметричности, о наличии у распределения плотности или о его дискретном характере и т. д.

Итак, математическая статистика работает там, где есть случайный эксперимент, свойства которого частично или полностью неизвестны, и где мы умеем воспроизводить этот эксперимент в одних и тех же условиях некоторое (а лучше — какое угодно) число раз.

Результаты экспериментов могут носить количественный или качественный характер. Количественные результаты можно, например, складывать. Так, одной из их осмысленных характеристик является среднее арифметическое наблюдений. Качественные результаты складывать бессмысленно, хотя они и могут быть облечены в числовую форму. Скажем, месяц рождения опрошенного — качественное, а не количественное наблюдение: его хоть и можно задать числом, но среднее арифметическое этих чисел несёт столько же разумной информации, как сообщение о том, что в среднем человек родился между июнем и июлем.

В первых главах мы будем изучать работу с количественными результатами наблюдений.

§ 2. Выборка

Пусть $\xi : \Omega \to \mathbb{R}$ — случайная величина, наблюдаемая в случайном эксперименте. Проводя n раз этот эксперимент в одинаковых условиях, мы получим числа $X_1,\ X_2,\ \ldots,\ X_n$ — значения наблюдаемой случайной величины в первом, втором и т. д. экспериментах. Случайная величина ξ имеет некоторое распределение $\mathcal{F}$, которое нам *частично или полностью неизвестно*. Рассмотрим подробнее набор $\vec{X} = (X_1, \ldots, X_n)$, называемый *выборкой*.

В серии *уже произведённых* экспериментов выборка — это набор чисел. Но до того, как эксперимент проведён, имеет смысл считать выборку набором случайных величин (независимых и распределённых так же, как ξ). Действительно, до проведения опытов мы не можем сказать, какие значения примут элементы выборки: это будут какие-то из значений случайной величины ξ. Поэтому имеет смысл считать, что до опыта X_i — случайная величина, одинаково распределенная с ξ, а после опыта — число, которое мы наблюдаем в i-м по счёту эксперименте, т. е. одно из возможных значений *случайной величины* X_i.

О п р е д е л е н и е 1. *Выборкой* $\vec{X} = (X_1, \ldots, X_n)$ объёма n из распределения $\mathcal{F}$ называется набор из n независимых и одинаково распределенных случайных величин, имеющих распределение $\mathcal{F}$.

Элементы выборки часто преобразуют для удобства работы с большим набором данных — упорядочивают или группируют.

Если элементы выборки $X_1, \ldots, X_n$ упорядочить по возрастанию, получится набор новых случайных величин, называемый *вариационным рядом*:

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

Здесь $X_{(1)} = \min\{X_1, \ldots, X_n\}$, $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Элемент $X_{(k)}$ называется k-м членом вариационного ряда или k-й *порядковой статистикой*.

При группировке данных выделяют несколько групп значений элементов выборки, подсчитывают количество элементов в каждой группе и далее имеют дело только с этим новым набором данных. Как группировка, так и упорядочение данных отбрасывают часть содержащейся в выборке информации.

Задачей математической статистики является получение по выборке выводов о неизвестном распределении $\mathcal{F}$, из которого она извлечена. Распределение характеризуется функцией распределения, плотностью или таблицей, набором числовых характеристик: $\mathsf{E}\,\xi = \mathsf{E}\,X_1$, $\mathsf{D}\,\xi = \mathsf{D}\,X_1$, $\mathsf{E}\,\xi^k = \mathsf{E}\,X_1^k$. По выборке нужно уметь строить приближения для всех этих характеристик. Такие приближения называют *оценками*. Термин «оценка» не имеет никакого отношения к неравенствам. Оценкой для некоторой неизвестной характеристики распределения называют построенную по выборке случайную величину, которая в каком-то смысле является приближением этой неизвестной характеристики распределения.

Пример 1. Шестигранный кубик подброшен 100 раз. Первая грань выпала 25 раз, вторая и пятая — по 14 раз, третья — 21 раз, четвёртая — 15 раз, шестая — 11 раз. Мы имеем дело с *числовой* выборкой, которая для удобства сгруппирована по количеству выпавших очков.

По данным результатам эксперимента нельзя определить вероятности $p_1, \ldots, p_6$ выпадения граней. Можно лишь сказать, что получены числовые оценки для этих вероятностей: 0,25 для p_1, 0,14 для p_2 и для p_5 и т. д.

Даже не проводя такой эксперимент, мы могли бы заранее сказать, что оценкой для неизвестной вероятности p_1 будет случайная величина

$$p_1^* = \frac{\text{число выпавших на кости единиц}}{100},$$

а оценкой для вероятности p_2 будет случайная величина

$$p_2^* = \frac{\text{число выпавших на кости двоек}}{100}.$$

В данной серии экспериментов эти случайные величины приняли значения 0,25 и 0,14 соответственно. В другой серии их значения изменятся.

§ 3. Выборочные характеристики

Из теории вероятностей нам известно универсальное средство для приближённого вычисления всевозможных математических ожиданий: закон больших чисел. Этот закон гарантирует, что средние арифметические независимых и одинаково распределённых слагаемых в некотором смысле сближаются с математическим ожиданием типичного слагаемого (если, конечно, это математическое ожидание существует).

Поэтому в качестве приближения (оценки) для неизвестного математического ожидания $\mathsf{E}\,X_1$ можно использовать среднее арифметическое всех элементов выборки: *выборочное среднее*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \ldots + X_n}{n}. \tag{1}$$

В качестве оценки для $\mathsf{E}\,X_1^k$ годится *выборочный k-й момент*

$$\overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k = \frac{X_1^k + \ldots + X_n^k}{n}, \tag{2}$$

а в качестве оценки для дисперсии $\mathsf{D}\,X_1 = \mathsf{E}\,(X_1 - \mathsf{E}\,X_1)^2 = \mathsf{E}\,X_1^2 - (\mathsf{E}\,X_1)^2$ используется *выборочная дисперсия*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \overline{X^2} - \left(\overline{X}\right)^2. \tag{3}$$

В общем случае величину

$$\overline{g(X)} = \frac{1}{n}\sum_{i=1}^{n} g(X_i) = \frac{g(X_1) + \ldots + g(X_n)}{n}$$

можно использовать для оценивания величины $\mathsf{E}\,g(X_1)$.

Точно так же закон больших чисел Бернулли позволяет нам оценивать различные вероятности. Например, вероятность события $\{X_1 < 3\}$ можно заменить на долю успешных испытаний в схеме Бернулли: если для каждого элемента выборки успехом считать событие $\{X_i < 3\}$, то доля успехов

$$p^* = \frac{\text{количество } X_i < 3}{n}$$

будет сходиться (по вероятности) к вероятности успеха $\mathsf{P}(X_1 < 3)$.

Оценивать неизвестную функцию распределения $F(y) = \mathsf{P}(X_1 < y)$ можно с помощью *эмпирической функции распределения*

$$F_n^*(y) = \frac{\text{количество } X_i < y}{n}. \tag{4}$$

Познакомимся подробно с каждой из введённых выше выборочных характеристик и изучим её свойства. К ожидаемым свойствам оценок относят следующие два: *несмещённость* и *состоятельность*.

Свойство *состоятельности* оценки гарантирует, что оценка приближается (по вероятности) к оцениваемой величине с ростом объёма выборки.

Оценку называют *несмещённой*, если её математическое ожидание совпадает с оцениваемой величиной. Это свойство означает отсутствие систематического смещения в большую или меньшую сторону при многократном использовании данной оценки.

§ 4. Свойства эмпирической функции распределения

П р и м е р 2. Пусть дана числовая выборка

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

Построим по ней вариационный ряд

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

и эмпирическую функцию распределения (рис. 1).

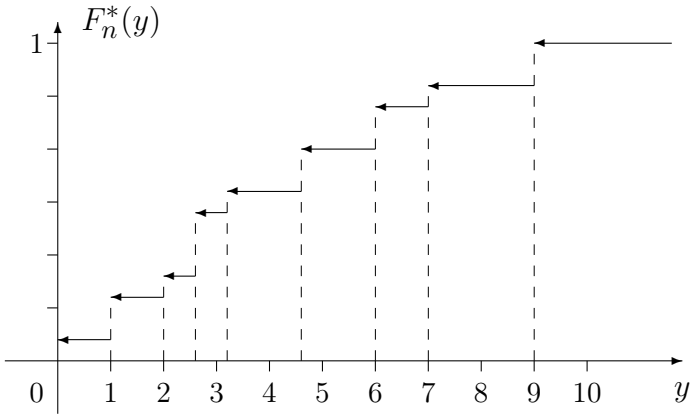

Рис. 1. Эмпирическая функция распределения

Эмпирическая функция распределения имеет скачки в точках выборки (вариационного ряда), величина скачка в точке X_i равна $\dfrac{m}{n}$, где m — количество элементов выборки, совпадающих с X_i. Эмпирическая функция распределения по вариационному ряду строится так:

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

Т е о р е м а 1. *Пусть дана выборка из распределения с функцией распределения F и пусть F_n^* — эмпирическая функция распределения, построенная по этой выборке. Тогда для любой фиксированной точки y выполнены свойства:*

1) $F_n^*(y) \xrightarrow{\text{p}} F(y)$ *при* $n \to \infty$, *т. е.* $F_n^*(y)$ *является состоятельной оценкой для* $F(y)$;

2) $\mathsf{E}\, F_n^*(y) = F(y)$, *т. е.* $F_n^*(y)$ *является несмещённой оценкой для* $F(y)$.

Д о к а з а т е л ь с т в о. По определению (4),

$$F_n^*(y) = \frac{\text{количество } X_i < y}{n}.$$

Свяжем с выборкой схему Бернулли: в i-м испытании произошёл успех, если $X_i < y$. В таком случае величина ν_n, равная количеству X_i меньших y, есть число успехов в n независимых испытаниях Бернулли с вероятностью успеха $p = \mathsf{P}(X_1 < y) = F(y)$. По закону больших чисел Бернулли

$$F_n^*(y) = \frac{\nu_n}{n} \xrightarrow{\text{p}} p = F(y).$$

Величина ν_n имеет биномиальное распределение с параметрами n и p. Поэтому

$$\mathsf{E}\, F_n^*(y) = \frac{\mathsf{E}\,\nu_n}{n} = \frac{np}{n} = p = F(y). \qquad \square$$

На самом деле сходимость эмпирической функции распределения к теоретической имеет даже «равномерный» характер: *наибольшее* из расхождений между этими функциями распределения стремится к нулю.

Т е о р е м а 2 (Г л и в е н к о — К а н т е л л и). *В условиях теоремы 1*

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

§ 5. Свойства выборочных моментов

Выборочное среднее $\overline{X}$, определённое формулой (1), является несмещённой и состоятельной оценкой для теоретического среднего (математического ожидания), которое для удобства мы будем обозначать m_1.

Т е о р е м а 3. *Пусть имеется выборка из распределения с конечным первым моментом* $\mathsf{E}\, X_1 = m_1$. *Тогда*

1) $\mathsf{E}\overline{X} = m_1$, *т. е. выборочное среднее* $\overline{X}$ *является несмещённой оценкой для истинного математического ожидания* m_1;

2) $\overline{X} \xrightarrow{\text{p}} m_1$ *при* $n \to \infty$, *т. е. выборочное среднее* $\overline{X}$ *является состоятельной оценкой для* m_1.

Доказательство. Первое утверждение следует из свойств математического ожидания:

$$\mathsf{E}\overline{X} = \frac{1}{n}(\mathsf{E}X_1 + \ldots + \mathsf{E}X_n) = \frac{1}{n} \cdot n\, m_1 = m_1.$$

Из ЗБЧ в форме Хинчина получаем второе утверждение:

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n} \xrightarrow{\text{p}} \mathsf{E}X_1 = m_1. \qquad \square$$

Выборочный k-й момент $\overline{X^k}$, определённый формулой (2), является несмещённой и состоятельной оценкой для теоретического k-го момента. Обозначим теоретический k-й момент буквой m_k.

Теорема 4. *Пусть имеется выборка из распределения с конечным k-м моментом $\mathsf{E}\,X_1^k = m_k$. Тогда*

1) $\mathsf{E}\overline{X^k} = m_k$, *т. е.* $\overline{X^k}$ *является несмещённой оценкой для* m_k;

2) $\overline{X^k} \xrightarrow{\text{p}} m_k$ *при* $n \to \infty$, *т. е.* $\overline{X^k}$ *является состоятельной оценкой для* m_k.

Упражнение. Доказать теорему 4.

Выше мы определили формулой (3) выборочную дисперсию S^2. Оказывается однако, что оценка S^2, будучи состоятельной, обладает систематическим смещением в меньшую сторону по сравнению с истинной дисперсией распределения $\mathsf{D}\,X_1 = \sigma^2$. Введём поэтому ещё одну, теперь уже несмещённую, оценку для дисперсии. Величину

$$S_0^2 = \frac{n}{n-1}\,S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{5}$$

называют *несмещённой выборочной дисперсией*. Убедимся в адекватности её названия.

Теорема 5. *Пусть дана выборка из распределения с конечной дисперсией* $\mathsf{D}\,X_1 = \sigma^2$. *Тогда*

1) *обе выборочные дисперсии* S^2 *и* S_0^2 *являются состоятельными оценками для истинной дисперсии:*

$$S^2 \xrightarrow{\text{p}} \sigma^2, \quad S_0^2 \xrightarrow{\text{p}} \sigma^2 \quad \textit{при} \quad n \to \infty;$$

2) *величина* S^2 *— смещённая оценка дисперсии, а* S_0^2 *— несмещённая:*

$$\mathsf{E}S^2 = \frac{n-1}{n}\,\sigma^2 < \sigma^2, \qquad \mathsf{E}S_0^2 = \sigma^2.$$

Доказательство. Докажем первое утверждение теоремы. Воспользуемся вторым равенством из формулы (3): $S^2 = \overline{X^2} - (\overline{X})^2$. Используя состоятельность первого и второго выборочных моментов и свойства сходимости

по вероятности, получаем

$$S^2 = \overline{X^2} - (\overline{X})^2 \xrightarrow{\text{P}} m_2 - (m_1)^2 = \mathsf{E}X_1^2 - (\mathsf{E}X_1)^2 = \sigma^2.$$

Далее, $\dfrac{n}{n-1} \to 1$, поэтому $S_0^2 = \dfrac{n}{n-1}\, S^2 \xrightarrow{\text{P}} \sigma^2$ при $n \to \infty$.

Для доказательства второго утверждения теоремы воспользуемся несмещённостью первого и второго выборочных моментов:

$$\begin{aligned}
\mathsf{E}S^2 &= \mathsf{E}\left(\overline{X^2} - (\overline{X})^2\right) = \mathsf{E}\overline{X^2} - \mathsf{E}(\overline{X})^2 = m_2 - \mathsf{E}(\overline{X})^2 = \\
&= m_2 - \left(\left(\mathsf{E}\overline{X}\right)^2 + \mathsf{D}\overline{X}\right) = m_2 - (m_1)^2 - \mathsf{D}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \\
&= \sigma^2 - \frac{1}{n^2}\, n\mathsf{D}X_1 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\,\sigma^2,
\end{aligned}$$

откуда сразу следует $\mathsf{E}S_0^2 = \dfrac{n}{n-1}\,\mathsf{E}S^2 = \sigma^2$. □

§ 6. Гистограмма как оценка плотности

Важной характеристикой для абсолютно непрерывного распределения является плотность распределения.

Эмпирическим аналогом плотности распределения является так называемая *гистограмма*.

Гистограмма строится по *группированным* данным. Область на прямой, занимаемую элементами выборки, делят на k интервалов. Пусть $A_1, \ldots, A_k$ — интервалы на прямой, называемые *интервалами группировки*. Обозначим для $j = 1, \ldots, k$ через ν_j число элементов выборки, попавших в интервал A_j. Случайная величина ν_j равна числу успехов в n испытаниях схемы Бернулли, если в i-м испытании успехом считать событие $\{X_i \in A_j\}$.

На каждом из интервалов A_j строят прямоугольник, площадь которого пропорциональна ν_j. Общая площадь всех прямоугольников должна равняться единице. Поэтому высота f_j прямоугольника над интервалом A_j равна

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

где через l_j обозначена длина интервала A_j.

Полученная фигура, состоящая из объединения прямоугольников, называется гистограммой.

Пример 3. Имеется вариационный ряд из примера 2:

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

Разобьём отрезок $[0,\ 10]$ на четыре равных отрезка. Отрезку $[0,\ 2{,}5)$ принадлежат четыре элемента выборки, отрезку $[2{,}5,\ 5)$ — шесть, отрезку

[5, 7,5) — три, и отрезку [7,5, 10] — два элемента выборки. Строим гистограмму (рис. 2). На рис. 3 — гистограмма для той же выборки, но при разбиении области на пять равных отрезков.

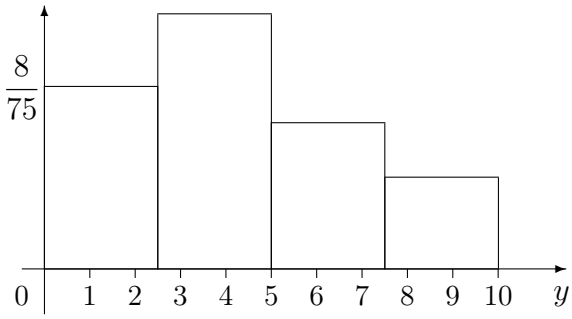

Рис. 2. Гистограмма при $k = 4$

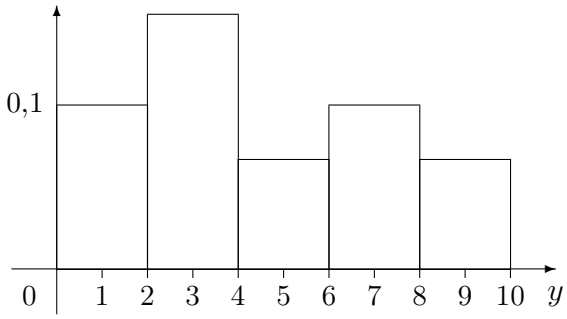

Рис. 3. Гистограмма при $k = 5$

Чем больше интервалов группировки, тем лучше: фигура, состоящая из более узких прямоугольников, точнее приближает истинную плотность распределения. С другой стороны, бессмысленно брать число интервалов $k = k(n)$ порядка n: тогда в каждый интервал попадёт в среднем по одной точке и гистограмма не будет приближаться к плотности с ростом n. Справедливо следующее утверждение.

Пусть плотность распределения элементов выборки является непрерывной функцией. Если количество интервалов группировки стремится к бесконечности таким образом, что $k(n)/n \to 0$, *то имеет место сходимость по вероятности гистограммы к плотности в каждой точке* y.

Обычно берут число интервалов порядка $c \cdot \sqrt[3]{n}$ (или длину интервала порядка $c/\sqrt[3]{n}$).

§ 7. Вопросы и упражнения

1. Дана числовая выборка (0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0) из распределения Бернулли. Вычислить по ней значения выборочного среднего, выборочных k-х моментов, выборочной дисперсии и несмещённой выборочной дисперсии. Построить график эмпирической функции распределения.

2. Дана числовая выборка (1, 3, 2, 5, 0, 0, 1, 2, 1, 1, 3, 2) из распределения Пуассона. Вычислить по ней значения выборочного среднего, выборочной дисперсии и несмещённой выборочной дисперсии. Построить график эмпирической функции распределения.

3. Выборка объёма $n = 100$ задана таблицей:

X_i	-1	0	1	3	5
N_i	20	15	10	20	35

.

Числа N_i соответствуют количеству значений, равных X_i, в выборке. Вычислить по этой выборке значения выборочного среднего, выборочной дисперсии и несмещённой выборочной дисперсии. Построить график эмпирической функции распределения.

4. Выборка объёма $n = 100$ утеряна, осталась лишь информация по интервалам группировки:

A_j	$[-10, -7)$	$[-7, 0)$	$[0, 5)$	$[5, 11)$	$[11, 15]$
ν_j	10	20	30	25	15

.

Построить гистограмму.

5. Пусть $(-0{,}8;\ 2{,}9;\ 4{,}3;\ -5{,}7;\ 1{,}1;\ -3{,}2)$ — наблюдавшиеся значения выборки. Построить график эмпирической функции распределения и проверить, что $F_6^*(-5) = 1/6$, $F_6^*(0) = 1/2$ и $F_6^*(4) = 5/6$.

6. Пусть $(3,\ 0,\ 4,\ 3,\ 6,\ 0,\ 3,\ 1)$ — наблюдавшиеся значения выборки. Построить график эмпирической функции распределения и проверить, что $F_8^*(1) = 1/4$, $F_8^*(3) = 3/8$ и $F_8^*(5) = 7/8$.

7. Указать какую-нибудь выборку объёма $n = 12$, которая имеет ту же эмпирическую функцию распределения, что и выборка из упражнения 5.

8. Указать какую-нибудь (отличную от выборки из упражнения 6) выборку объёма $n = 8$, которая имеет ту же эмпирическую функцию распределения, что и выборка из упражнения 6.

ГЛАВА II
ТОЧЕЧНОЕ ОЦЕНИВАНИЕ

Ситуация, когда о распределении наблюдений не известно совсем ничего, встречается довольно редко. Проводя эксперимент, мы можем предполагать или утверждать что-либо о распределении его результатов. Например, может оказаться, что это распределение нам известно с точностью до значений одного или нескольких числовых параметров. Так, в широких предположениях рост юношей одного возраста имеет нормальное распределение с неизвестными средним и дисперсией, а число покупателей в магазине в течение часа — распределение Пуассона с неизвестной «интенсивностью» λ. Рассмотрим задачу оценивания по выборке неизвестных параметров распределения. Оказывается, различными способами бывает возможно построить даже не одну, а множество оценок для одного и того же неизвестного параметра.

§ 1. Точечные оценки и их свойства

Параметрические семейства распределений. Пусть имеется выборка $X_1, \ldots, X_n$ объёма n, извлечённая из распределения $\mathcal{F}_\theta$, которое известным образом зависит от неизвестного параметра θ.

Здесь $\mathcal{F}_\theta$ — некий класс распределений, целиком определяющихся значением скалярного или векторного параметра θ.

Примерами параметрических семейств распределений могут служить все известные нам распределения: распределение Пуассона Π_λ, где $\lambda > 0$; распределение Бернулли B_p, где $p \in (0, 1)$; равномерное распределение $\mathrm{U}_{a,\,b}$, где $a < b$; равномерное распределение $\mathrm{U}_{0,\,\theta}$, где $\theta > 0$; нормальное распределение $\mathrm{N}_{a,\,\sigma^2}$, где $a \in \mathbb{R}$, $\sigma > 0$ и т. д.

Точечные оценки. Пусть дана выборка объёма n из параметрического семейства распределений $\mathcal{F}_\theta$.

О п р е д е л е н и е 2. *Статистикой* (оценкой) называется произвольная функция $\theta^* = \theta^*(X_1, \ldots, X_n)$ от элементов выборки.

З а м е ч а н и е 1. Статистика есть функция от эмпирических данных, но никак не от параметра θ. Статистика, как правило, предназначена именно для оценивания неизвестного параметра θ (поэтому её иначе называют *оценкой*), и уже поэтому от него зависеть не может.

Свойства оценок. Дадим три определения хороших свойств оценок. Про два из них мы уже говорили ранее.

О п р е д е л е н и е 3. Статистика $\theta^* = \theta^*(X_1, \ldots, X_n)$ называется *несмещённой* оценкой параметра θ, если $\mathsf{E}\theta^* = \theta$.

О п р е д е л е н и е 4. Статистика $\theta^* = \theta^*(X_1, \ldots, X_n)$ называется *асимптотически несмещённой* оценкой параметра θ, если $\mathsf{E}\theta^* \to \theta$ при $n \to \infty$.

О п р е д е л е н и е 5. Статистика $\theta^* = \theta^*(X_1, \ldots, X_n)$ называется *состоятельной* оценкой параметра θ, если $\theta^* \xrightarrow{\text{p}} \theta$ при $n \to \infty$.

Несмещённость — свойство оценок при фиксированном n. Означает это свойство отсутствие ошибки «в среднем», т. е. при систематическом использовании данной оценки. Несмещённость является желательным, но не обязательным свойством оценок. Достаточно, чтобы смещение оценки (разница между её средним значением и истинным параметром) уменьшалось с ростом объёма выборки. Поэтому асимптотическая несмещённость является весьма желательным свойством оценок. Свойство состоятельности означает, что последовательность оценок приближается к неизвестному параметру при увеличении количества наблюдений. В отсутствие этого свойства оценка совершенно «несостоятельна» как оценка.

П р и м е р 4. Пусть $X_1, \ldots, X_n$ — выборка объёма n из нормального распределения N_{a,σ^2}, где $a \in \mathbb{R}$, $\sigma > 0$. Как найти оценки для параметров a и σ^2, если оба эти параметра (можно их считать одним двумерным параметром) неизвестны?

Мы уже знаем хорошие оценки для математического ожидания и дисперсии любого распределения. Оценкой для истинного среднего $a = \mathsf{E}X_1$ может служить выборочное среднее $a^* = \overline{X}$. Теорема 3 (с. 12) утверждает, что эта оценка несмещённая и состоятельная.

Для дисперсии $\sigma^2 = \mathsf{D}X_1$ у нас есть сразу две оценки:

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

Как показано в теореме 5 (с. 13), обе эти оценки состоятельны, и одна из них — несмещённая (*которая*?), а другая — асимптотически несмещённая.

§ 2. Метод моментов

Рассмотрим некоторые стандартные методы получения точечных оценок. Метод моментов предлагает для нахождения оценки неизвестного параметра использовать выборочные моменты вместо истинных. Этот метод заключается в следующем: любой момент случайной величины X_1 (например, k-й)

является функцией от параметра θ. Но тогда и параметр θ может оказаться функцией от теоретического k-го момента. Подставив в эту функцию вместо неизвестного теоретического k-го момента его выборочный аналог, получим вместо параметра θ его оценку θ^*.

Пусть $X_1, \ldots, X_n$ — выборка объёма n из семейства распределений $\mathcal{F}_\theta$, где θ — неизвестный числовой параметр.

Вычислим какой-нибудь из существующих моментов распределения. Пусть $\mathsf{E}\, X_1^k = m_k = h(\theta)$, причём функция $h(x)$ непрерывна и обратима (взаимно-однозначна). Тогда параметр θ можно выразить через k-й момент: $\theta = h^{-1}(m_k)$. В качестве *оценки метода моментов* для параметра θ берут величину $\theta^* = h^{-1}(\overline{X^k})$.

Замечание 2. Если параметр θ — вектор, а не число, т. е. если неизвестных параметров несколько, то в методе моментов берут не один момент m_k, а столько, сколько требуется для того, чтобы выразить через моменты все неизвестные параметры.

Пример 5. Пусть $X_1, \ldots, X_n$ — выборка объёма n из равномерного на отрезке $[0, \theta]$ распределения $\mathrm{U}_{0,\theta}$, где $\theta > 0$.

Найдём оценку метода моментов θ_1^* по первому моменту:

$$\mathsf{E} X_1 = m_1 = \frac{\theta}{2}, \qquad \theta = 2m_1, \qquad \theta_1^* = 2\overline{X}.$$

Найдём оценку метода моментов θ_k^* по k-му моменту:

$$\mathsf{E} X_1^k = m_k = \int\limits_0^\theta x^k \, \frac{1}{\theta}\, dx = \frac{\theta^k}{k+1}, \qquad \theta = \sqrt[k]{(k+1)m_k},$$

тогда

$$\theta_k^* = \sqrt[k]{(k+1)\overline{X^k}}. \tag{6}$$

Пример 6. Пусть $X_1, \ldots, X_n$ — выборка объёма n из нормального распределения N_{a,σ^2}.

Найдём оценки метода моментов для неизвестных параметров a и σ^2. Мы можем сразу записать выражения параметров через первые два момента:

$$\begin{cases} a = m_1, \\ \sigma^2 = m_2 - (m_1)^2, \end{cases} \quad \text{поэтому} \quad \begin{cases} a^* = \overline{X}, \\ (\sigma^2)^* = \overline{X^2} - \left(\overline{X}\right)^2 = S^2. \end{cases}$$

Пример 7. Пусть $X_1, \ldots, X_n$ — выборка объёма n из показательного распределения E_α.

Найдём оценку метода моментов по первому моменту $\mathsf{E} X_1 = m_1 = \frac{1}{\alpha}$. Выразим $\alpha = \frac{1}{m_1}$. Поэтому $\alpha^* = \frac{1}{\overline{X}}$.

§ 3. Свойства оценок метода моментов

Т е о р е м а 6. *Оценки, полученные методом моментов, являются состоятельными оценками.*

Д о к а з а т е л ь с т в о. Пусть $\theta^* = h^{-1}(\overline{X^k})$ — оценка для параметра θ, полученная методом моментов из равенства $m_k = h(\theta)$, где функция $y = h(x)$ непрерывна и обратима. По теореме 4 имеем

$$\overline{X^k} \xrightarrow{\text{p}} m_k.$$

Поскольку функция $y = h(x)$ непрерывна и обратима, то и обратная к ней функция $x = h^{-1}(y)$ также непрерывна. Поэтому

$$\theta^* = h^{-1}(\overline{X^k}) \xrightarrow{\text{p}} h^{-1}(m_k) = \theta. \qquad \square$$

Если оценки метода моментов обязаны быть состоятельными, то свойство несмещённости для них является скорее исключением, нежели правилом.

П р и м е р 8. Рассмотрим последовательность оценок для неизвестного параметра θ равномерного на отрезке $[0, \theta]$ распределения, полученную в примере 5 и исследуем напрямую их свойства.

Их состоятельность вытекает из теоремы 6. Проверим несмещённость полученных оценок. По теореме 3, $\mathsf{E}\overline{X} = m_1$, поэтому

$$\mathsf{E}\theta_1^* = \mathsf{E}2\overline{X} = 2\mathsf{E}\overline{X} = 2m_1 = 2 \cdot \theta/2 = \theta,$$

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

Рассмотрим оценку $\theta_2^* = \sqrt{3\overline{X^2}}$. Функция $y = \sqrt{x}$ является вогнутой в области $x > 0$, поэтому мы можем воспользоваться неравенством Йенсена:

$$\mathsf{E}\theta_2^* = \mathsf{E}\sqrt{3\overline{X^2}} \leqslant \sqrt{3\mathsf{E}\overline{X^2}} = \sqrt{3m_2} = \theta.$$

Полезно заметить, что знак равенства в неравенстве Йенсена возможен только для линейных функций либо для вырожденных случайных величин. В данном случае $y = \sqrt{x}$ нелинейна, а случайная величина $3\overline{X^2}$ имеет невырожденное распределение. Поэтому $\mathsf{E}\theta_2^* < \theta$, т. е. оценка θ_2^* является смещённой. Такими же смещёнными будут и оценки θ_k^* при всех $k > 2$.

П р и м е р 9. Оценка $\alpha^* = \dfrac{1}{\overline{X}}$ в примере 7 является смещённой оценкой. Действительно, применяя неравенство Йенсена к выпуклой на $(0, +\infty)$ функции $y = 1/x$, получим:

$$\mathsf{E}\,\alpha^* = \mathsf{E}\,\frac{1}{\overline{X}} > \frac{1}{\mathsf{E}\,\overline{X}} = \alpha.$$

§ 4. Метод максимального правдоподобия

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

Выясним сначала, что такое «вероятность получить данную выборку», т. е. что́ именно нужно максимизировать. Вспомним, что для абсолютно непрерывных распределений $\mathcal{F}_\theta$ их плотность $f_\theta(y)$ — «почти» (с точностью до dy) вероятность попадания в точку y:

$$\mathsf{P}(X_1 \in (y,\, y + dy)) = f_\theta(y)\, dy.$$

А для дискретных распределений $\mathcal{F}_\theta$ вероятность попасть в точку y равна $\mathsf{P}_\theta(X_1 = y)$. В зависимости от типа распределения $\mathcal{F}_\theta$ обозначим через $f_\theta(y)$ одну из следующих двух функций:

$$f_\theta(y) = \begin{cases} \text{плотность } f_\theta(y), & \text{если } \mathcal{F}_\theta \text{ абсолютно непрерывно,} \\ \mathsf{P}_\theta(X_1 = y), & \text{если } \mathcal{F}_\theta \text{ дискретно.} \end{cases} \tag{7}$$

В дальнейшем функцию $f_\theta(y)$, определённую формулой (7), мы будем называть *плотностью* распределения $\mathcal{F}_\theta$ независимо от того, является ли это распределение дискретным или абсолютно непрерывным.

О п р е д е л е н и е 6. Функция

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

называется *функцией правдоподобия*. При фиксированном θ эта функция является случайной величиной. Функция (тоже случайная)

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

называется *логарифмической* функцией правдоподобия.

В дискретном случае при фиксированных $x_1, \ldots, x_n$ значение функции правдоподобия $f(x_1, \ldots, x_n, \theta)$ равно вероятности, с которой выборка $X_1, \ldots, X_n$ в данной серии экспериментов принимает значения $x_1, \ldots, x_n$. Эта вероятность меняется в зависимости от θ:

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

В абсолютно непрерывном случае эта функция пропорциональна вероятности попасть «почти» в точку $x_1, \ldots, x_n$, а именно, в «кубик» со сторонами $dx_1, \ldots, dx_n$ вокруг точки $x_1, \ldots, x_n$.

О п р е д е л е н и е 7. *Оценкой максимального правдоподобия* (ОМП) θ^* для неизвестного параметра θ называют такое значение θ, при котором достигается максимум функции $f(\vec{X};\, \theta)$.

З а м е ч а н и е 3. Поскольку функция $\ln y$ монотонна, то точки максимума функций $f(\vec{X};\, \theta)$ и $L(\vec{X};\, \theta)$ совпадают (*обосновать*). Поэтому оценкой максимального правдоподобия можно называть точку максимума (по переменной θ) функции $L(\vec{X};\, \theta)$.

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

П р и м е р 10. Пусть $X_1, \ldots, X_n$ — выборка объёма n из распределения Пуассона Π_λ, где $\lambda > 0$. Найдём ОМП λ^* для неизвестного параметра λ. Здесь

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

поэтому функция правдоподобия равна

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

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

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

Тогда

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

Точку экстремума $\lambda^* = \overline{X}$ находим как решение уравнения $\dfrac{n\overline{X}}{\lambda} - n = 0$.

Проверим, что в точке $\lambda^* = \overline{X}$ достигается максимум функции L. Для этого достаточно выяснить, будет ли отрицательной вторая производная функции L в этой точке. Но вторая производная функции L равна

$$\frac{\partial^2}{\partial\lambda^2} L(\vec{X}, \lambda) = \frac{\partial}{\partial\lambda}\left(\frac{n\overline{X}}{\lambda} - n\right) = -\frac{n\overline{X}}{\lambda^2} < 0$$

и отрицательна при всех значениях λ, в том числе и в точке $\lambda^* = \overline{X}$.

Пример 11. Пусть $X_1, \ldots, X_n$ — выборка объёма n из нормального распределения N_{a,σ^2}, где $a \in \mathbb{R}$, $\sigma > 0$ — два неизвестных параметра.

Это распределение имеет плотность

$$f_{(a,\sigma^2)}(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(y-a)^2/2\sigma^2}.$$

Перемножив плотности в точках $X_1, \ldots, X_n$, получим функцию правдоподобия

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

а затем логарифмическую функцию правдоподобия

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

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

$$\begin{cases} \dfrac{\partial}{\partial a}\, L(\vec{X};\, a,\, \sigma^2) &= \dfrac{2\sum(X_i - a)}{2\sigma^2} = \dfrac{n\overline{X} - na}{\sigma^2}, \\[2ex] \dfrac{\partial}{\partial \sigma^2}\, L(\vec{X};\, a,\, \sigma^2) &= -\dfrac{n}{2\sigma^2} + \dfrac{\sum(X_i - a)^2}{2\sigma^4}. \end{cases}$$

Оценка максимального правдоподобия для (a, σ^2) является решением системы уравнений

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

Решая, получаем хорошо знакомые оценки

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

Упражнение. Проверить, что $(\overline{X},\, S^2)$ — точка максимума, а не минимума. Для этого вычислить матрицу вторых производных функции L в данной точке и проверить её отрицательную определённость, т. е. чередование знаков главных миноров (первый отрицательный, второй положительный).

Пример 12. Пусть $X_1, \ldots, X_n$ — выборка объёма n из равномерного распределения $\mathrm{U}_{0,\theta}$, где $\theta > 0$.

Плотность этого распределения равна

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

Запишем функцию правдоподобия

$$f(\vec{X};\,\theta)=\begin{cases}\dfrac{1}{\theta^n}, & \text{если } 0\leqslant X_i\leqslant\theta\ \ \forall i,\\ 0 & \text{иначе}\end{cases}\quad=\quad\begin{cases}\dfrac{1}{\theta^n}, & 0\leqslant X_{(1)}\leqslant X_{(n)}\leqslant\theta,\\ 0 & \text{иначе.}\end{cases}$$

Представим функцию $f(\vec{X};\,\theta)$ как функцию переменной θ:

$$f(\vec{X};\,\theta)=\begin{cases}\dfrac{1}{\theta^n}, & \text{если } \theta\geqslant X_{(n)},\ X_{(1)}\geqslant 0,\\ 0, & \text{если } \theta<X_{(n)} \text{ или } X_{(1)}<0.\end{cases}$$

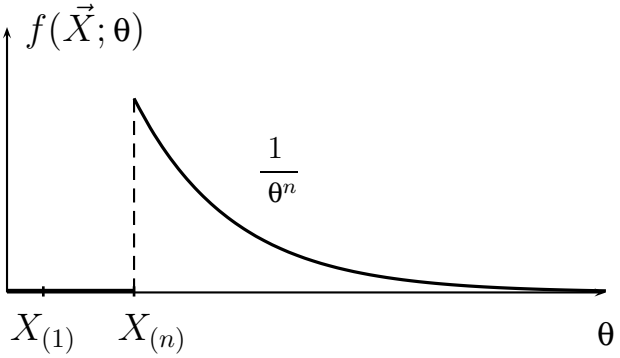

Рис. 4. График функции правдоподобия распределения $\mathrm{U}_{0,\,\theta}$

Видим на рис. 4, что максимум функции правдоподобия достигается в точке $X_{(n)}$. Она и будет ОМП: $\theta^*=X_{(n)}=\max\{X_1,\ldots,X_n\}$.

§ 5. Асимптотическая нормальность оценок

Ещё одно важное свойство оценок связано с их предельным поведением. Предположим, что разность оценки и параметра, подходящим образом нормированная, имеет распределение, которое с ростом n всё более похоже на стандартное нормальное распределение. В таком случае оценку (последовательность оценок) называют *асимптотически нормальной*. Асимптотическая нормальность оценок является важным свойством последовательностей оценок. В дальнейшем мы увидим, что это свойство используется при построении доверительных интервалов для неизвестных параметров, в задачах проверки гипотез о значениях этих параметров, а также позволяет сравнивать качества оценок.

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

О п р е д е л е н и е 8. Оценка θ^* называется *асимптотически нормальной оценкой* (АНО) параметра θ с коэффициентом $\sigma^2(\theta)$, если при $n\to\infty$ распределение случайной величины $\dfrac{\sqrt{n}(\theta^*-\theta)}{\sigma(\theta)}$ сходится к стандартному нор-

мальному распределению, т. е. для любого x

$$\mathsf{P}\left(\frac{\sqrt{n}(\theta^* - \theta)}{\sigma(\theta)} < x\right) \to \Phi(x) \quad \text{при } n \to \infty.$$

Пример 13. Пусть дана выборка из распределения с конечной дисперсией $\mathsf{D}\,X_1 = \sigma^2$.

Убедимся, что выборочное среднее $\overline{X}$ является асимптотически нормальной оценкой для истинного математического ожидания $m_1 = \mathsf{E}\,X_1$. При этом коэффициент асимптотической нормальности равен как раз $\sigma^2 = \mathsf{D}\,X_1$.

Действительно, по центральной предельной теореме распределение членов последовательности

$$\frac{\sqrt{n}(\overline{X} - m_1)}{\sigma} = \frac{X_1 + \ldots + X_n - nm_1}{\sigma\sqrt{n}} = \frac{X_1 + \ldots + X_n - n\mathsf{E}\,X_1}{\sqrt{n\mathsf{D}\,X_1}}$$

сближается со стандартным нормальным распределением.

Пример 14. Пусть $X_1, \ldots, X_n$ — выборка объёма n из равномерного распределения $\mathrm{U}_{0,\,\theta}$ с параметром $\theta > 0$. Проверим, являются ли оценки $\theta^* = 2\overline{X}$ и $\theta^{**} = X_{(n)}$ асимптотически нормальными.

Используем предыдущий пример. Величина $\theta^* = 2\overline{X} = \dfrac{2X_1 + \ldots + 2X_n}{n}$ есть среднее арифметическое случайных величин с математическим ожиданием $\mathsf{E}\,(2X_1) = 2 \cdot \theta/2 = \theta$ и дисперсией $\mathsf{D}\,(2X_1) = 4\mathsf{D}\,X_1 = \theta^2/3$.

Поэтому оценка $\theta^* = 2\overline{X}$ является АНО для параметра θ с коэффициентом асимптотической нормальности $\sigma^2(\theta) = \theta^2/3$.

Для проверки асимптотической нормальности оценки $\theta^{**} = X_{(n)}$ заметим, что величина $\sqrt{n}\,(X_{(n)} - \theta)$ при любом n принимает только отрицательные значения, поэтому её распределение не может приближаться ни к какому нормальному закону. Оценка θ^{**} не является асиптотически нормальной.

§ 6. Вопросы и упражнения

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

2. Дана выборка $X_1, \ldots, X_n$ из распределения Пуассона Π_λ с параметром $\lambda > 0$. Проверить, что X_1 является несмещённой оценкой для λ. Является ли эта оценка состоятельной?

3. Дана выборка $X_1, \ldots, X_n$ из равномерного распределения $\mathrm{U}_{0,\,\theta}$ с параметром $\theta > 0$. Проверить состоятельность и несмещённость оценок $\theta^* = X_{(n)},\ \theta^{**} = X_{(n)} + X_{(1)}$ для параметра θ.

4. Построить оценки неизвестных параметров по методу моментов для неизвестных параметров следующих семейств распределений: B_p — по первому моменту, Π_λ — по первому и второму моменту, $\mathrm{U}_{a,\,b}$ — по первому и второму моменту, E_α — по всем моментам, $\mathrm{E}_{1/\alpha}$ — по первому моменту, $\mathrm{U}_{-\theta,\,\theta}$ — как получится, $\Gamma_{\alpha,\,\lambda}$ — по первому и второму моменту, $\mathrm{N}_{a,\,\sigma^2}$ (для σ^2 при a известном и при a неизвестном).

5. Построить оценки неизвестных параметров по методу максимального правдоподобия для следующих семейств распределений: $\mathrm{B}_{m,\,p}$ при известном значении $m \in \mathbb{N}$, $\Pi_{\lambda+1}$, $\mathrm{U}_{0,\,2\theta}$, $\mathrm{E}_{2\alpha+3}$, $\mathrm{U}_{-\theta,\,\theta}$, $\mathrm{N}_{a,\,\sigma^2}$ при известном a.

6. Какие из оценок в упражнениях 4 и 5 несмещённые? Какие из них состоятельны?

7. Эмпирическая функция распределения $F_n^*(y)$ строится по выборке из равномерного распределения на отрезке $[0,\,a]$, где $a > 1$. Для какого параметра $\theta = \theta(a)$ статистика $F_n^*(1)$ является несмещённой оценкой? Является ли она состоятельной оценкой того же параметра?

8. Пусть элементы выборки $X_1, \ldots, X_n$ имеют распределение с плотностью

$$f_\theta(y) = \begin{cases} 3\theta y^2 e^{-\theta y^3}, & \text{если } y > 0, \\ 0, & \text{если } y \leqslant 0, \end{cases}$$

где $\theta > 0$ — неизвестный параметр. Найти ОМП для параметра θ.

9. Дана числовая выборка 0, 1, 6, 0, 1, 3, 2, 2, 1, 0, 3, 4, 4, 2 из распределения Пуассона с параметром λ. Вычислить значение оценок метода моментов для параметра λ, полученных по первому и второму моментам.

10. Дана выборка $X_1, \ldots, X_n$ из равномерного распределения $\mathrm{U}_{a,\,b}$ с параметрами $a < b$. Доказать, что оценками максимального правдоподобия для параметров a и b будут $X_{(1)}$ и $X_{(n)}$ соответственно.

11. Дана выборка $X_1, \ldots, X_n$ из распределения Пуассона Π_λ с параметром $\lambda > 0$. Проверить, что $\overline{X}$ является асимптотически нормальной оценкой для λ. Найти коэффициент асимптотической нормальности.

ГЛАВА III
СРАВНЕНИЕ ОЦЕНОК

Используя метод моментов и метод максимального правдоподобия, мы получили для каждого параметра достаточно много различных оценок. Каким же образом их сравнивать? Что должно быть показателем «хорошести» оценки? Понятно, что чем дальше оценка отклоняется от параметра, тем она хуже. Но величина $|\theta^* - \theta|$ для сравнения непригодна: во-первых, параметр θ неизвестен, во-вторых, θ^* — случайная величина, поэтому при разных значениях выборки эти расстояния будут, вообще говоря, различны. Для сравнения оценок используют обычно усреднённые характеристики рассеяния. Например, это может быть $\mathsf{E}(\theta^* - \theta)^2$, $\mathsf{E}|\theta^* - \theta|$, либо какие-то иные средние.

§ 1. Среднеквадратичный подход к сравнению оценок

Среднеквадратичный подход использует в качестве «расстояния» от оценки до параметра величину $\mathsf{E}(\theta^* - \theta)^2$.

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

Определение 9. Говорят, что оценка θ_1^* *не хуже* оценки θ_2^* *в среднеквадратичном смысле*, если для любого θ

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

Среди всех мыслимых оценок наилучшей в среднеквадратичном смысле не существует. Но если разбить оценки на отдельные классы, то наилучшая в каждом классе может и найтись. Обычно рассматривают классы оценок, имеющих одинаковое *смещение* $b(\theta) = \mathsf{E}\theta^* - \theta$.

Обозначим через $K_b = K_{b(\theta)}$ класс всех оценок со смещением, равным заданной функции $b(\theta)$:

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

Здесь K_0 — класс несмещённых оценок.

Определение 10. Оценка $\theta^* \in K_b$ называется *эффективной* оценкой в классе K_b, если она лучше (не хуже) всех других оценок класса K_b в среднеквадратичном смысле.

З а м е ч а н и е 4. Для оценки $\theta^* \in K_0$ по определению дисперсии

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

т. е. сравнение в среднеквадратичном несмещённых оценок есть просто сравнение их дисперсий. Для смещённых оценок $\theta^* \in K_b$

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

т. е. сравнение в среднеквадратичном оценок с одинаковым смещением также приводит к сравнению их дисперсий.

З а м е ч а н и е 5. Заметим без доказательства, что в классе оценок с одинаковым смещением не может существовать двух различных эффективных оценок: если эффективная оценка существует, она единственна.

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

П р и м е р 15. Пусть дана выборка объёма n из равномерного распределения $\mathrm{U}_{0,\theta}$, где $\theta > 0$. В примерах 5 и 12 мы нашли ОММ по первому моменту $\theta^* = 2\overline{X}$ и ОМП $\theta^{**} = X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Сравним их в среднеквадратичном смысле. Оценка $\theta^* = 2\overline{X}$ несмещённая, поэтому

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

Для θ^{**} имеем $\mathsf{E}(\theta^{**} - \theta)^2 = \mathsf{E}(\theta^{**})^2 - 2\theta\,\mathsf{E}\theta^{**} + \theta^2$. Найдём функцию и плотность распределения случайной величины θ^{**}:

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

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

Посчитаем первый и второй моменты случайной величины $\theta^{**} = X_{(n)}$:

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

Поэтому

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

При $n = 1,\ 2$ среднеквадратичные отклонения оценок θ^* и θ^{**} равны: ни одна из этих оценок не лучше другой в среднеквадратичном смысле, а при $n > 2$ оценка $X_{(n)}$ оказывается лучше, чем $2\overline{X}$:

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

Оценку $X_{(n)}$ можно превратить в несмещённую оценку $\dfrac{n+1}{n}\,X_{(n)}$, т. е. оценку из того же класса, что и $2\overline{X}$. Но и тогда исправленная оценка оказывается лучше в среднеквадратичном смысле, чем $2\overline{X}$ (см. упражнение 2 § 3).

§ 2. Неравенство Рао — Крамера

В классе одинаково смещённых оценок эффективной мы назвали оценку с наименьшим среднеквадратичным отклонением. Но попарное сравнение оценок — далеко не лучший способ отыскания эффективной оценки. Существует утверждение, позволяющее во многих случаях *доказать* эффективность оценки (если, конечно, она на самом деле эффективна). Это утверждение называется неравенством Рао—Крамéра и говорит о том, что в любом классе $K_{b(\theta)}$ существует нижняя граница для среднеквадратичного отклонения любой оценки. Таким образом, если найдётся оценка, отклонение которой в точности равно этой нижней границе (самое маленькое), то данная оценка — эффективна, поскольку у всех остальных оценок отклонение меньшим быть не может. К сожалению, данное неравенство верно не для всех семейств распределений. Например, оно не имеет места для равномерных распределений.

Более точно, исключим из рассмотрения любые семейства распределений, для которых область значений элементов выборки зависит от параметра θ.

Потребуем также, чтобы так называемая *информация Фишера*

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

была конечна, положительна и непрерывна по θ.

Если данные условия выполнены, справедливо следующее утверждение.

Теорема 7 (неравенство Рао — Крамера). *Для любой несмещённой оценки $\theta^* \in K_0$ справедливо неравенство*

$$\mathsf{D}\theta^* = \mathsf{E}\,(\theta^* - \theta)^2 \geqslant \frac{1}{nI(\theta)}\,.$$

Неравенство сформулировано для класса несмещённых оценок. Похожим образом выглядит неравенство Рао — Крамера для смещённых оценок.

Сформулируем очевидное следствие из неравенства Рао — Крамера.

С л е д с т в и е 1. *Если для оценки* $\theta^* \in K_0$ *достигается равенство в неравенстве Рао — Крамера*

$$\mathsf{E}\,(\theta^* - \theta)^2 = \frac{1}{nI(\theta)}\,,$$

то оценка θ^* *эффективна в классе* K_0.

П р и м е р 16. Пусть дана выборка объёма n из нормального распределения $\mathrm{N}_{a,\,\sigma^2}$. Проверим, является ли оценка $a^* = \overline{X} \in K_0$ эффективной.

Найдём информацию Фишера относительно параметра a. Плотность распределения равна

$$f_a(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(y-a)^2/(2\sigma^2)}, \quad \ln f_a(y) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y-a)^2}{2\sigma^2}\,.$$

Соответственно, $\dfrac{\partial}{\partial a}\ln f_a(y) = \dfrac{y-a}{\sigma^2}$. Найдя второй момент этого выражения при $y = X_1$, получим информацию Фишера

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

Найдём дисперсию оценки $\overline{X}$: $\mathsf{D}\overline{X} = \dfrac{1}{n}\,\mathsf{D}X_1 = \dfrac{\sigma^2}{n}$.

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

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

Итак, оценка $a^* = \overline{X}$ эффективна (т. е. обладает наименьшей дисперсией среди несмещённых оценок).

П р и м е р 17. Пусть дана выборка объёма n из нормального распределения $\mathrm{N}_{0,\,\sigma^2}$. Проверим, является ли эффективной оценка

$$\sigma^{2*} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \overline{X^2} \in K_0.$$

У п р а ж н е н и е. Получить эту оценку методом моментов и методом максимального правдоподобия.

Найдём информацию Фишера относительно параметра σ^2. Плотность распределения равна

$$f_{\sigma^2}(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-y^2/(2\sigma^2)}, \quad \ln f_{\sigma^2}(y) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{y^2}{2\sigma^2}\,.$$

Продифференцируем это выражение по параметру σ^2:

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

Вычислим информацию Фишера

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

Осталось найти $\mathsf{D}X_1^2 = \mathsf{E}X_1^4 - (\mathsf{E}X_1^2)^2 = \mathsf{E}X_1^4 - \sigma^4$. Используем тот факт, что величина $\xi = X_1/\sigma$ имеет стандартное нормальное распределение, и её четвёртый момент равен трём (мы вычисляли его в курсе теории вероятностей): $\mathsf{E}\xi^4 = 3$, $X_1 = \xi\cdot\sigma$, поэтому

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

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

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

Найдём дисперсию оценки $\sigma^{2*} = \overline{X^2}$ и сравним её с правой частью неравенства Рао — Крамера:

$$\mathsf{D}\overline{X^2} = \frac{1}{n^2}\mathsf{D}\sum_1^n X_i^2 = \frac{1}{n}\mathsf{D}X_1^2 = \frac{2\sigma^4}{n} = \frac{1}{nI(\sigma^2)},$$

Поэтому оценка $\sigma^{2*} = \overline{X^2}$ эффективна.

§ 3. Вопросы и упражнения

1. Дана выборка объёма n из распределения Пуассона с параметром λ. Сравнить оценки X_1, $\dfrac{X_1+X_2}{2}$ и $\overline{X}$ в среднеквадратичном смысле.

2. Используя вычисления из примера 15, сравнить в среднеквадратичном смысле оценки $\theta^* = 2\overline{X}$ и $\theta^{***} = \dfrac{n+1}{n}X_{(n)}$. Проверить, является ли оценка θ^{***} несмещённой.

3. Является ли эффективной несмещённая оценка $\theta^* = 2\overline{X}$, полученная по выборке из равномерного распределения на отрезке $[0,\,\theta]$?

4. Дана выборка из распределения Пуассона с параметром λ. Проверить эффективность оценки $\lambda^* = \overline{X}$ с помощью неравенства Рао — Крамера.

5. Дана выборка из биномиального распределения $\mathrm{B}_{m,\,p}$, где $m = 10$. Проверить по неравенству Рао — Крамера эффективность оценки $p^* = \overline{X}/10$.

ГЛАВА IV

ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

Пусть есть выборка из распределения $\mathcal{F}_\theta$ с неизвестным параметром θ. До сих пор мы занимались «точечным оцениванием» неизвестного параметра — находили оценку (для числовой выборки это число), способную в некотором смысле заменить параметр. Существует другой подход к оцениванию, при котором мы указываем случайный интервал, накрывающий параметр с заранее заданной вероятностью. Границы этого интервала зависят от выборки. Такой подход называется интервальным оцениванием. Сразу заметим: чем больше уверенность в том, что параметр лежит в интервале, тем шире интервал. Поэтому бессмысленно искать диапазон, внутри которого θ содержится гарантированно — таким интервалом будет вся область возможных значений параметра.

§ 1. Доверительные интервалы

Пусть $X_1, \ldots, X_n$ — выборка объёма n из распределения $\mathcal{F}_\theta$ с параметром $\theta \in \mathbb{R}$. Пусть задано число $0 < \varepsilon < 1$.

Определение 11. Интервал (θ^-, θ^+), границы которого зависят от заданного ε и от выборки $X_1, \ldots, X_n$, называется *доверительным интервалом* для параметра θ *уровня доверия* $1 - \varepsilon$, если при любом возможном значении θ

$$\mathsf{P}\left(\theta^- \leqslant \theta \leqslant \theta^+\right) = 1 - \varepsilon.$$

Замечание 6. Интервал из определения 11 называют также *точным* доверительным интервалом.

Определение 12. Интервал (θ^-, θ^+) называется *асимптотическим* доверительным интервалом для параметра θ (асимптотического) уровня доверия $1 - \varepsilon$, если при любом возможном значении θ

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

На самом деле в определении 12 речь идёт, конечно, не об одном интервале, но о *последовательности* интервалов, зависящих от n.

З а м е ч а н и е 7. Случайны здесь границы интервала (θ^-, θ^+), поэтому читают событие $\{\theta^- < \theta < \theta^+\}$ как «интервал (θ^-, θ^+) накрывает параметр θ», а не как «θ лежит в интервале...».

Прежде чем рассматривать какие-то регулярные способы построения точных и асимптотических доверительных интервалов, разберем два примера, а затем попробуем извлечь из них некоторую общую философию построения доверительных интервалов.

П р и м е р 18. Пусть $X_1, \ldots, X_n$ — выборка объёма n из нормального распределения $\mathrm{N}_{a,\,\sigma^2}$, где $a \in \mathbb{R}$ — неизвестный параметр, а значение $\sigma > 0$ известно. Требуется при произвольном n построить точный доверительный интервал для параметра a уровня доверия $1 - \varepsilon$.

Знаем, что нормальное распределение устойчиво по суммированию. Поэтому распределение суммы элементов выборки при любом её объёме n нормально: $n\overline{X} = X_1 + \ldots + X_n$ имеет нормальное распределение $\mathrm{N}_{na,\,n\sigma^2}$, а центрированная и нормированная величина

$$\eta = \frac{n\overline{X} - na}{\sigma\sqrt{n}} = \sqrt{n}\,\frac{\overline{X} - a}{\sigma}$$

имеет стандартное нормальное распределение.

По заданному $\varepsilon \in (0,\,1)$ найдём число $c > 0$ такое, что

$$\mathsf{P}(-c < \eta < c) = 1 - \varepsilon.$$

Число c является квантилью уровня $1 - \dfrac{\varepsilon}{2}$ стандартного нормального распределения (рис. 5):

$$\mathsf{P}(-c < \eta < c) = \Phi(c) - \Phi(-c) = 1 - \varepsilon, \quad \Phi(c) = 1 - \frac{\varepsilon}{2}.$$

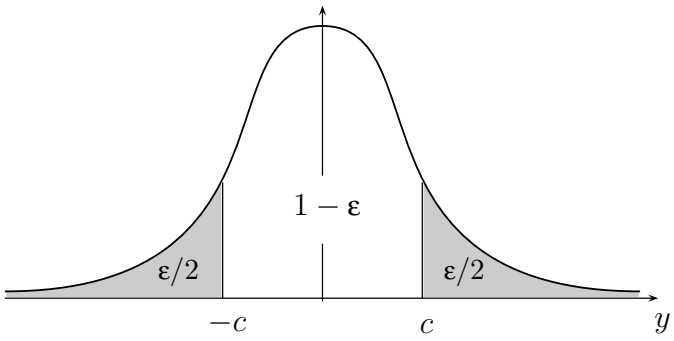

Рис. 5. Квантили стандартного нормального распределения

По заданному ε в таблице значений функции $\Phi(x)$ найдём квантили $c = \tau_{1-\varepsilon/2}$ или $-c = \tau_{\varepsilon/2}$. Разрешив затем неравенство $-c < \eta < c$ отно-

сительно a, получим точный доверительный интервал:

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

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

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

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

$$\left(\overline{X}-\frac{\sigma\,\tau_{1-\varepsilon/2}}{\sqrt{n}},\;\overline{X}+\frac{\sigma\,\tau_{1-\varepsilon/2}}{\sqrt{n}}\right). \tag{8}$$

У п р а ж н е н и е. Имеет смысл задать себе несколько вопросов.

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

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

3. Какова середина полученного в примере 18 доверительного интервала? Какова его длина? Что происходит с границами доверительного интервала при $n\to\infty$? Как быстро это с ними происходит?

П р и м е р 19. Пусть $X_1,\,\ldots,\,X_n$ — выборка объёма n из показательного распределения E_α, где $\alpha>0$. Требуется построить асимптотический доверительный интервал для параметра α уровня доверия $1-\varepsilon$.

Вспомним ЦПТ: распределение случайной величины

$$\frac{\sum X_i-n\,\mathsf{E}X_1}{\sqrt{n\,\mathsf{D}X_1}}=\sqrt{n}\,\frac{\overline{X}-1/\alpha}{1/\alpha}=\sqrt{n}\,\left(\alpha\overline{X}-1\right)$$

с ростом n становится всё более похоже на стандартное нормальное распределение. Возьмём $c=\tau_{1-\varepsilon/2}$ — квантиль стандартного нормального распределения. По ЦПТ при $n\to\infty$

$$\mathsf{P}\left(-c<\sqrt{n}\,\left(\alpha\overline{X}-1\right)<c\right)\to\Phi(c)-\Phi(-c)=1-\varepsilon.$$

Разрешив относительно α неравенство $-\tau_{1-\varepsilon/2}<\sqrt{n}\,\left(\alpha\overline{X}-1\right)<\tau_{1-\varepsilon/2}$, получим асимптотический доверительный интервал:

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

§ 2. Принципы построения доверительных интервалов

Чтобы построить точный доверительный интервал, необходимо реализовать следующие шаги.

1. Найти функцию $G(\vec{X}, \theta)$, распределение которой $\mathcal{G}$ не зависит от параметра θ.

2. Найти числа g_1 и g_2 — квантили распределения $\mathcal{G}$, для которых

$$1 - \varepsilon = \mathsf{P}(g_1 < G(\vec{X}, \theta) < g_2).$$

3. Разрешить неравенство $g_1 < G(\vec{X}, \theta) < g_2$ относительно θ.

Замечание 8. Часто в качестве g_1 и g_2 берут квантили распределения $\mathcal{G}$ уровней $\varepsilon/2$ и $1-\varepsilon/2$. Но, вообще говоря, квантили следует выбирать так, чтобы получить самый короткий доверительный интервал.

Совершенно аналогично выглядит общий принцип построения асимптотических доверительных интервалов. Отличие от построения точных доварительных интервалов лишь в том, что достаточно знать *предельное* распределение функции $G(\vec{X}, \theta)$, а не точное.

Следующий пример (как и пример 19) показывает, что ЦПТ дает универсальный вид функции G для построения *асимптотических* доверительных интервалов.

Пример 20. Пусть $X_1, \ldots, X_n$ — выборка объёма n из распределения Пуассона Π_λ, где $\lambda > 0$. Требуется построить асимптотический доверительный интервал для параметра λ уровня доверия $1 - \varepsilon$.

Согласно ЦПТ, распределение случайной величины

$$G(\vec{X}, \lambda) = \frac{X_1 + \ldots + X_n - n\mathsf{E}X_1}{\sqrt{n\mathsf{D}X_1}} = \sqrt{n}\,\frac{\overline{X} - \lambda}{\sqrt{\lambda}}$$

сближается с нормальным стандартным распределением. Пусть $c = \tau_{1-\varepsilon/2}$ — квантиль стандартного нормального распределения. При $n \to \infty$

$$\mathsf{P}\left(-c < \sqrt{n}\,\frac{\overline{X} - \lambda}{\sqrt{\lambda}} < c\right) \to \Phi(c) - \Phi(-c) = 1 - \varepsilon.$$

Но разрешить неравенство под знаком вероятности относительно λ не просто: мешает корень в знаменателе. Заменим λ под корнем на какую-нибудь состоятельную оценку для λ — например, на $\overline{X}$. Разрешив теперь неравенство под знаком вероятности относительно λ, получим

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

Итак, искомый асимптотический доверительный интервал имеет вид

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

Для построения асимптотических доверительных интервалов можно использовать асимптотически нормальные оценки (это тоже ЦПТ).

Т е о р е м а 8. *Пусть* θ^* *— АНО для параметра* θ *с коэффициентом* $\sigma^2(\theta)$, *и функция* $\sigma(\theta)$ *непрерывна по* θ. *Тогда интервал*

$$\left(\theta^* - \frac{\tau_{1-\varepsilon/2}\,\sigma(\theta^*)}{\sqrt{n}},\ \ \theta^* + \frac{\tau_{1-\varepsilon/2}\,\sigma(\theta^*)}{\sqrt{n}}\right)$$

является асимптотическим доверительным интервалом для параметра θ *уровня доверия* $1-\varepsilon$.

Д о к а з а т е л ь с т в о. По определению АНО, при $n \to \infty$

$$\mathsf{P}\left(-c < \sqrt{n}\,\frac{\theta^* - \theta}{\sigma(\theta)} < c\right) \to \Phi(c) - \Phi(-c),$$

где $c = \tau_{1-\varepsilon/2}$ — квантиль стандартного нормального распределения.

Заменим в знаменателе мешающее $\sigma(\theta)$ на $\sigma(\theta^*)$. Разрешив неравенство

$$-c < \sqrt{n}\,\frac{\theta^* - \theta}{\sigma(\theta^*)} < c$$

относительно θ, получим асимптотический доверительный интервал

$$\left(\theta^* - \frac{c\,\sigma(\theta^*)}{\sqrt{n}},\ \ \theta^* + \frac{c\,\sigma(\theta^*)}{\sqrt{n}}\right). \qquad \square$$

В следующей главе мы продолжим знакомство с точными доверительными интервалами. В частности, мы найдём такие интервалы для параметров нормального распределения.

§ 3. Вопросы и упражнения

1. Что больше: квантиль стандартного нормального распределения уровня 0,05 или уровня 0,1? Почему? Нарисовать их на графике плотности этого распределения.

2. По одному и тому же правилу построены два доверительных интервала уровней доверия 0,05 и 0,1. Какой из них шире?

3. По числовой выборке объёма $n = 10\,000$ из нормального распределения с параметрами a и 1 вычислили выборочное среднее $\overline{X} = 0,32$. Указать границы точного доверительного интервала для параметра a с уровнем доверия 0,95.

ГЛАВА V

РАСПРЕДЕЛЕНИЯ, СВЯЗАННЫЕ С НОРМАЛЬНЫМ

В предыдущей главе мы построили в числе других точный доверительный интервал для параметра a нормального распределения при известном σ^2. Остался нерешённым вопрос: как построить точные доверительные интервалы для σ при известном и при неизвестном a, а также для a при неизвестном σ? Мы уже видели, что для решения этих задач требуется отыскать такие функции от выборки и неизвестных параметров, распределения которых не зависят от этих параметров. При этом сами искомые функции не должны зависеть от мешающих параметров. Особый интерес к нормальному распределению связан, разумеется, с центральной предельной теоремой: почти всё в этом мире нормально (или близко к тому). В этой главе мы изучим новые распределения, связанные с нормальным, их свойства и свойства выборок из нормального распределения.

§ 1. Основные статистические распределения

Гамма-распределение. С гамма-распределением мы познакомились в курсе теории вероятностей (*вспомнить*!). Нам понадобится свойство устойчивости по суммированию этого распределения.

Лемма 1. *Пусть* $X_1, \ldots, X_n$ *независимы, и* ξ_i *имеет гамма-распределение* $\Gamma_{\alpha,\lambda_i}$, $i = 1, \ldots, n$. *Тогда их сумма* $S_n = \xi_1 + \ldots + \xi_n$ *имеет гамма-распределение с параметрами* α *и* $\lambda_1 + \ldots + \lambda_n$.

В курсе теории вероятностей мы доказали следующий факт: квадрат случайной величины со стандартным нормальным распределением имеет гамма-распределение.

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

Распределение χ^2 Пирсона. Из лемм 1 и 2 следует утверждение.

Лемма 3. *Если* $\xi_1, \ldots, \xi_k$ *независимы и имеют стандартное нормальное распределение, то случайная величина*

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

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

В статистике это распределение играет совершенно особую роль и имеет собственное название.

О п р е д е л е н и е 13. Распределение суммы k квадратов независимых случайных величин со стандартным нормальным распределением называется распределением χ^2 (*хи-квадрат*) или распределением Пирсона с k степенями свободы и обозначается H_k.

Согласно лемме 3, распределение H_k совпадает с $\Gamma_{1/2,\,k/2}$. Поэтому плотность распределения H_k равна

$$f(y) = \begin{cases} \dfrac{1}{2^{k/2}\,\Gamma(k/2)}\, y^{\frac{k}{2}-1} e^{-y/2}, & \text{если } y > 0; \\ 0, & \text{если } y \leqslant 0. \end{cases}$$

Плотности распределений H_k при $k = 1,\ 2,\ 4,\ 8$ показаны на рис. 6.

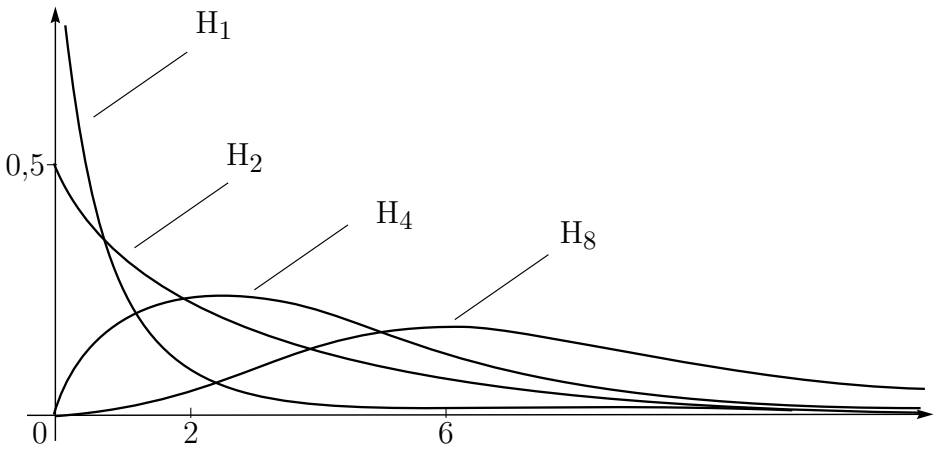

Рис. 6. Плотности χ^2-распределений с различным числом степеней свободы

Рассмотрим свойства χ^2-распределения. Устойчивость его относительно суммирования следует из устойчивости гамма-распределения.

С в о й с т в о 1. *Если случайные величины* $\chi^2 \sim \mathrm{H}_k$ *и* $\psi^2 \sim \mathrm{H}_m$ *независимы, то их сумма* $\chi^2 + \psi^2$ *имеет распределение* H_{k+m}.

С в о й с т в о 2. *Если величина* χ^2 *имеет распределение* H_k, *то*

$$\mathsf{E}\chi^2 = k \quad \textit{и} \quad \mathsf{D}\chi^2 = 2k.$$

Д о к а з а т е л ь с т в о. Пусть $\xi_1, \xi_2, \ldots$ независимы и имеют стандартное нормальное распределение. Тогда

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

Поэтому

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

Распределение H_n при небольших n табулировано. Однако при большом числе степеней свободы для вычисления функции этого распределения или, наборот, его квантилей пользуются различными аппроксимациями с помощью стандартного нормального распределения. Одно из приближений предлагается в следующем свойстве.

Свойство 3 (аппроксимация Фишера). *Пусть* $\chi_n^2 \sim \mathrm{H}_n$. *Тогда при* $n \to \infty$ *распределение случайной величины*

$$\sqrt{2\chi_n^2} - \sqrt{2n-1}$$

сближается со стандартным нормальным распределением. Поэтому при больших n *можно пользоваться аппроксимацией для функции распределения* $\mathrm{H}_n(x) = \mathsf{P}\left(\chi_n^2 < x\right)$*:*

$$\mathrm{H}_n(x) \approx \Phi\left(\sqrt{2x} - \sqrt{2n-1}\right). \tag{9}$$

Свойство 4. *Если случайные величины* $\xi_1, \ldots, \xi_k$ *независимы и имеют нормальное распределение* N_{a,σ^2}, *то*

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

Упражнение. Доказать свойство 4, вспомнив, как нормальное распределение превратить в стандартное нормальное.

Распределение Стьюдента. Английский статистик Госсет, публиковавший научные труды под псевдонимом Стьюдент, ввёл следующее распределение.

Определение 14. Пусть ξ_0, ξ_1, $\ldots$, ξ_k независимы и имеют стандартное нормальное распределение. Распределение случайной величины

$$t_k = \frac{\xi_0}{\sqrt{\dfrac{\xi_1^2 + \ldots + \xi_k^2}{k}}}$$

называется *распределением Стью́дента* с k степенями свободы и обозначается T_k.

Распределение Стьюдента совпадает с распределением случайной величины $t_k = \dfrac{\xi}{\sqrt{\chi_k^2/k}}$, где $\xi \sim \mathrm{N}_{0,1}$ и $\chi_k^2 \sim \mathrm{H}_k$ независимы.

Плотность распределения Стьюдента с k степенями свободы равна

$$f_k(y) = \frac{\Gamma\big((k+1)/2\big)}{\sqrt{\pi k}\,\Gamma(k/2)} \left(1 + \frac{y^2}{k}\right)^{-(k+1)/2}. \tag{10}$$

Свойство 5. *Распределение Стьюдента симметрично: если случайная величина t_k имеет распределение Стьюдента T_k с k степенями свободы, то и $-t_k$ имеет такое же распределение.*

Упражнение. Доказать, исходя из симметричности стандартного нормального распределения.

Свойство 6. *Распределение Стьюдента T_k сближается со стандартным нормальным распределением при $k \to \infty$.*

Доказательство. Для доказательства достаточно заметить, что знаменатель у случайной величины с распределением Стьюдента стремится к единице по ЗБЧ: $\dfrac{\xi_1^2 + \ldots + \xi_k^2}{k} \xrightarrow{\mathrm{p}} \mathsf{E}\,\xi_1^2 = 1$ при $k \to \infty$. □

Графики плотностей стандартного нормального распределения и распределения Стьюдента приведены для сравнения на рис. 7.

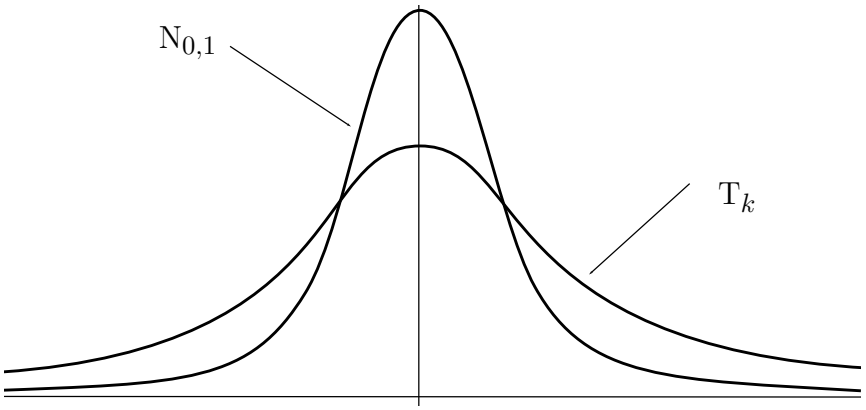

Рис. 7. Плотности распределений T_k и $\mathrm{N}_{0,\,1}$

Отметим, что распределение Стьюдента табулировано: если в каких-то доверительных интервалах появятся квантили этого распределения, то мы найдём их по соответствующей таблице, либо, при больших n, используем нормальную аппроксимацию для распределения Стьюдента.

Распределение Стьюдента с одной степенью свободы есть стандартное распределение Коши. Действительно, если подставить $k = 1$ в плотность (10) и учесть $\Gamma(1/2) = \sqrt{\pi}$ и $\Gamma(1) = 1$, то получится плотность распределения Коши:

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

Упражнение. Как получить случайную величину с распределением Коши, имея две независимые случайные величины со стандартным нормальным распределением?

Свойство 7. *У распределения Стьюдента существуют только моменты порядка $m < k$ и не существуют моменты порядка $m \geqslant k$. При этом все существующие моменты нечётного порядка равны нулю.*

У п р а ж н е н и е. Посмотреть на плотность (10) и убедиться в сходимости или расходимости на бесконечности при соответствующих m интегралов

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

Распределение Фишера. Следущее распределение тоже тесно связано с нормальным распределением, но понадобится нам не при построении доверительных интервалов, а чуть позже — в задачах проверки гипотез. Там же мы поймём, почему его называют *распределением дисперсионного отношения*.

О п р е д е л е н и е 15. Пусть χ^2_k имеет распределение H_k, а ψ^2_n — распределение H_n, причём эти случайные величины независимы. Распределение случайной величины

$$f_{k,\,n} = \frac{\chi^2_k \,/\, k}{\psi^2_n \,/\, n} = \frac{n}{k} \cdot \frac{\chi^2_k}{\psi^2_n}$$

называется *распределением Фишера* с k и n степенями свободы и обозначается $\mathrm{F}_{k,\,n}$.

Свойства распределения Фишера (или *Фишера — Снедекора*):

С в о й с т в о 8. *Если случайная величина $f_{k,\,n}$ имеет распределение Фишера $\mathrm{F}_{k,\,n}$, то $1/f_{k,\,n}$ имеет распределение Фишера $\mathrm{F}_{n,\,k}$.*

Заметим, что распределения $\mathrm{F}_{k,\,n}$ и $\mathrm{F}_{n,\,k}$ различаются, но связаны соотношением: для любого $x > 0$

$$\mathrm{F}_{k,\,n}(x) = \mathsf{P}(f_{k,\,n} < x) = \mathsf{P}\left(\frac{1}{f_{k,\,n}} > \frac{1}{x}\right) = 1 - \mathrm{F}_{n,\,k}\left(\frac{1}{x}\right).$$

Распределение Фишера также табулировано при многих k, n, причём свойство 8 позволяет приводить таблицы распределений только в половине случаев: например, при $k \geqslant n$.

С в о й с т в о 9. *Пусть $t_k \sim \mathrm{T}_k$ — случайная величина, имеющая распределение Стьюдента. Тогда $t_k^2 \sim \mathrm{F}_{1,\,k}$.*

§ 2. Преобразования нормальных выборок

Пусть $\vec{X} = (X_1, \ldots, X_n)$ — выборка из $\mathrm{N}_{0,\,1}$, т. е. набор независимых случайных величин со стандартным нормальным распределением. Там, где нам понадобятся операции матричного умножения, будем считать $\vec{X}$ вектором-столбцом. Пусть C — ортогональная матрица $(n \times n)$, т. е.

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

и $\vec{Y} = C\vec{X}$ — вектор с координатами $Y_i = C_{i1}X_1 + \ldots + C_{in}X_n$.

Координаты вектора $\vec{Y}$ имеют нормальные распределения как линейные комбинации независимых нормальных величин. Какие именно нормальные и с каким совместным распределением?

Оказывается, что *поворот*, т. е. умножение на *ортогональную* матрицу, не меняет совместного распределения нормального вектора! Это самое удивительное из всех свойств нормального распределения.

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

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

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

Следующее утверждение носит вспомогательный характер и по традиции называется леммой. Эта лемма является, пожалуй, самым главным вспомогательным утверждением во всех разделах теоретической статистики и эконометрики, связанных с нормальными наблюдениями.

Лемма 4 (лемма Фишера). *Пусть вектор $\vec{X}$ состоит из независимых случайных величин со стандартным нормальным распределением, C — ортогональная матрица, $\vec{Y} = C\vec{X}$.*

Тогда при любом $k = 1, \ldots, n-1$ случайная величина

$$T(\vec{X}) = \sum_{i=1}^{n} X_i^2 - Y_1^2 - \ldots - Y_k^2$$

не зависит от $Y_1, \ldots, Y_k$ и имеет распределение H_{n-k}.

Доказательство. Для ортогональной матрицы C нормы векторов $\vec{X}$ и $\vec{Y} = C\vec{X}$ совпадают: $X_1^2 + \ldots + X_n^2 = \|\vec{X}\|^2 = \|C\vec{X}\|^2 = Y_1^2 + \ldots + Y_n^2$. Поэтому

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

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

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

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

Действительно, обе эти величины являются функциями от одних и тех же наблюдений. Более того, в определение S_0^2 явным образом входит $\overline{X}$.

Теорема 10 (основное следствие леммы Фишера). *Пусть* $X_1, \ldots, X_n$ *независимы и имеют нормальное распределение с параметрами* a *и* σ^2. *Тогда:*

1) $\sqrt{n}\,\dfrac{\overline{X} - a}{\sigma} \sim \mathrm{N}_{0,\,1}$,

2) $\dfrac{(n-1)\,S_0^2}{\sigma^2} = \displaystyle\sum_{i=1}^{n} \frac{\big(X_i - \overline{X}\big)^2}{\sigma^2} \sim \mathrm{H}_{n-1}$,

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

Доказательство. Первое утверждение теоремы очевидно (*доказать, что очевидно*!). Докажем второе и третье. Убедимся сначала, что можно рассматривать выборку из стандартного нормального распределения вместо $\mathrm{N}_{a,\,\sigma^2}$:

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

где $\overline{z} = \dfrac{\overline{X} - a}{\sigma}$ — среднее арифметическое величин $z_i = \dfrac{X_i - a}{\sigma} \sim \mathrm{N}_{0,\,1}$. Итак, можно с самого начала считать, что X_i имеют стандартное нормальное распределение, $a = 0$, $\sigma^2 = 1$.

Применим лемму Фишера. Представим величину $(n-1)S_0^2$ в виде

$$T(\vec{X}\,) = (n-1)S_0^2 = \sum_{i=1}^{n}\big(X_i - \overline{X}\big)^2 = \sum_{i=1}^{n} X_i^2 - n(\overline{X})^2 = \sum_{i=1}^{n} X_i^2 - Y_1^2.$$

Здесь через Y_1 мы обозначили

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

Чтобы применить лемму Фишера, нужно найти ортогональную матрицу C такую, что Y_1 будет первой координатой вектора $\vec{Y} = C\vec{X}$.

Возьмём матрицу C с первой строкой $(1/\sqrt{n},\, \ldots,\, 1/\sqrt{n}\,)$. Так как длина (норма) этого вектора равна единице, его можно дополнить до ортонормального базиса в $\mathbb{R}^n$. Иначе говоря, этот столбец можно дополнить до ортогональной матрицы. Тогда величина $Y_1 = \sqrt{n}\;\overline{X}$ и будет первой координатой

вектора $\vec{Y} = C\vec{X}$. Осталось применить лемму Фишера и получить второе утверждение теоремы.

Из леммы Фишера следует также, что $(n-1)S_0^2 = \sum_{i=1}^n X_i^2 - Y_1^2$ не зависит от $Y_1 = \sqrt{n}\,\overline{X}$, т. е. $\overline{X}$ и S_0^2 независимы. □

Отметим без доказательства, что независимость величин $\overline{X}$ и S_0^2 — свойство, характерное *только* для нормального распределения. Так же, как и способность сохранять независимость координат после умножения на ортогональную матрицу.

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

Теорема 11 (полезное следствие леммы Фишера). *Пусть* $X_1, \ldots, X_n$ *независимы и имеют нормальное распределение с параметрами* a *и* σ^2. *Тогда*

1) $\sqrt{n}\,\dfrac{\overline{X}-a}{\sigma} \sim \mathrm{N}_{0,1}$ (*для* a *при* σ *известном*),

2) $\displaystyle\sum_{i=1}^{n}\Big(\frac{X_i-a}{\sigma}\Big)^2 \sim \mathrm{H}_n$ (*для* σ^2 *при* a *известном*),

3) $\dfrac{(n-1)\,S_0^2}{\sigma^2} \sim \mathrm{H}_{n-1}$ (*для* σ^2 *при* a *неизвестном*),

4) $\sqrt{n}\,\dfrac{\overline{X}-a}{S_0} \sim \mathrm{T}_{n-1}$ (*для* a *при* σ *неизвестном*).

Доказательство. Утверждения (1) и (3) следуют из леммы Фишера, (2) — из теоремы 4. Осталось воспользоваться леммой Фишера и определением распределения Стьюдента, чтобы доказать (4). Запишем

$$\sqrt{n}\,\frac{\overline{X}-a}{S_0} = \sqrt{n}\,\frac{\overline{X}-a}{\sigma}\cdot\frac{1}{\sqrt{\dfrac{(n-1)S_0^2}{\sigma^2}\cdot\dfrac{1}{n-1}}} = \frac{\xi_0}{\sqrt{\dfrac{\chi^2_{n-1}}{n-1}}}, \tag{11}$$

где величины

$$\xi_0 = \sqrt{n}\,\frac{\overline{X}-a}{\sigma} \sim \mathrm{N}_{0,1} \quad\text{и}\quad \chi^2_{n-1} = (n-1)S_0^2/\sigma^2 \sim \mathrm{H}_{n-1}$$

независимы по теореме 4. По определению 14, величина (11) имеет распределение Стьюдента T_{n-1}. □

§ 3. Доверительные интервалы для параметров нормального распределения

Пусть $X_1, \ldots, X_n$ — выборка объёма n из распределения $\mathrm{N}_{a,\,\sigma^2}$. Построим точные доверительные интервалы (ДИ) с уровнем доверия $1-\varepsilon$ для параметров нормального распределения, используя соответствующие утверждения теоремы 11.

Пример 21 (ДИ для a при известном σ^2). Этот интервал мы построили в примере 18 (с. 33):

$$\mathsf{P}\left(\overline{X} - \frac{\tau\sigma}{\sqrt{n}} < a < \overline{X} + \frac{\tau\sigma}{\sqrt{n}}\right) = 1-\varepsilon, \quad \text{где} \quad \Phi(\tau) = 1-\varepsilon/2.$$

Пример 22 (ДИ для σ^2 при известном a). По теореме 11

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

Пусть g_1 и g_2 — квантили распределения H_n уровней $\varepsilon/2$ и $1-\varepsilon/2$ соответственно. Тогда

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

Пример 23 (ДИ для σ^2 при неизвестном a). По теореме 11

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

Пусть g_1 и g_2 — квантили распределения H_{n-1} уровней $\varepsilon/2$ и $1-\varepsilon/2$ соответственно. Тогда

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

Упражнение. Найти 17 отличий примера 22 от примера 23.

Пример 24 (ДИ для a при неизвестном σ). По теореме 11

$$\sqrt{n}\,\frac{\overline{X}-a}{S_0} \sim \mathrm{T}_{n-1}.$$

Пусть c — квантиль распределения T_{n-1} уровня $1-\varepsilon/2$. Распределение Стьюдента симметрично. Поэтому

$$1-\varepsilon = \mathsf{P}\left(-c < \sqrt{n}\,\frac{\overline{X}-a}{S_0} < c\right) = \mathsf{P}\left(\overline{X} - \frac{c\,S_0}{\sqrt{n}} < a < \overline{X} + \frac{c\,S_0}{\sqrt{n}}\right).$$

Упражнение. Сравнить примеры 21 и 24.

§ 4. Вопросы и упражнения

1. Величины ξ_1 и ξ_2 независимы и имеют нормальное распределение с параметрами $a = 0$, $\sigma^2 = 16$. Найти k, при котором величины $\xi_1 - 3\xi_2$ и $k\xi_1 + \xi_2$ независимы. Можно использовать теорему 9 (с. 42).

2. Изобразить квантили уровней $\varepsilon/2$ и $1 - \varepsilon/2$ на графиках плотностей распределений H_n и T_n.

3. Имеется выборка объёма $n = 5$ из нормального распределения с неизвестными математическим ожиданием и дисперсией. Несмещённая выборочная дисперсия равна 1, а выборочное среднее равно 0. Построить точный доверительный интервал уровня 0,9: а) для неизвестной дисперсии, б) для неизвестного математического ожидания.

4. Решить упражнение 3, если а) $n = 25$, б) $n = 1\,000$.

5. Имеется выборка объёма $n = 5$ из нормального распределения с известным математическим ожиданием $a = 1$. Второй выборочный момент равен 2, а выборочное среднее равно 0,9. Построить точный доверительный интервал уровня 0,95 для неизвестной дисперсии.

6. Решить упражнение 5, если а) $n = 25$, б) $n = 1\,000$.

7. Как получить случайную величину с распределением Коши, имея две независимые случайные величины со стандартным нормальным распределением?

8. Пусть случайная величина χ^2_n имеет распределение хи-квадрат с n степенями свободы. Используя ЦПТ, доказать, что распределение случайной величины

$$\frac{\chi^2_n - n}{\sqrt{2n}}$$

при $n \to \infty$ сближается со стандартным нормальным распределением.

ГЛАВА VI

ПРОВЕРКА ГИПОТЕЗ

Имея выборку, мы можем выдвинуть несколько взаимоисключающих гипотез о теоретическом распределении, одну из которых следует предпочесть остальным. Задача выбора одной из нескольких гипотез решается построением статистического критерия. Как правило, по выборке конечного объёма безошибочных выводов о распределении сделано быть не может, поэтому всегда есть опасность выбрать неверную гипотезу. Так, бросая монету, можно выдвигать предположения об истинной вероятности выпадения герба. Допустим, есть две гипотезы: вероятность либо находится в пределах 0,45—0,55, либо нет. Получив после ста бросков ровно 51 герб, мы наверняка выберем первую гипотезу. Однако есть ненулевые шансы на то, что и при $p = 0{,}3$ выпадет 51 герб: выбирая первую гипотезу, мы можем ошибиться. Напротив, получив 33 герба, мы скорее всего предпочтём вторую гипотезу. И опять не исключена возможность, что столь далёкое от половины число гербов есть просто результат случайности, а монета на самом деле симметрична.

§ 1. Гипотезы и критерии

Пусть дана выборка $X_1, \ldots, X_n$ из распределения $\mathcal{F}$. Мы будем считать выборку набором независимых случайных величин с одним и тем же распределением, хотя в ряде задач и эти предположения нуждаются в проверке.

О п р е д е л е н и е 16. *Гипотезой* (H) называется любое предположение о распределении наблюдений:

$$H = \{\mathcal{F} = \mathcal{F}_1\} \qquad \text{или} \qquad H = \{\,\mathcal{F} \in \mathbb{F}\,\},$$

где $\mathbb{F}$ — некоторое подмножество в множестве всех распределений. Гипотеза называется *простой*, если она указывает на единственное распределение: $F = \mathcal{F}_1$. Иначе гипотеза называется *сложной*.

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

Пусть дана выборка $X_1, \ldots, X_n$, относительно распределения которой выдвинуты две гипотезы H_1 и H_2.

О п р е д е л е н и е 17. *Критерием* для проверки этих гипотез называется функция

$$\delta(\vec{X}) = \begin{cases} H_1, & \text{если } \vec{X} \notin S, \\ H_2, & \text{если } \vec{X} \in S. \end{cases}$$

которая при каждом возможном значении выборки определяет, какую из гипотез следует принимать. Область S в множестве всех возможных значений выборки, где принимается вторая (альтернативная) гипотеза, называется *критической областью*.

О п р е д е л е н и е 18. Говорят, что произошла *ошибка i-го рода* критерия δ, если критерий отверг *верную* гипотезу H_i. *Вероятностью ошибки i-го рода* критерия δ называется величина

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

З а м е ч а н и е 9. Говоря «H_i верна» и вычисляя $\mathsf{P}_{H_i}(\cdot)$, мы предполагаем, что распределение выборки именно такое, как указано в гипотезе H_i, и вычисляем вероятность в соответствии с этим распределением.

К сожалению, нам не известно, какая из гипотез верна в действительности, поэтому нам следует считаться с гипотетическими вероятностями ошибок критерия. Смысл этих ошибок состоит в следующем: если много раз применять критерий к выборкам из распределения, для которого гипотеза H_i верна, то в среднем доля α_i таких выборок будет признана противоречащей гипотезе H_i.

П р и м е р 25. Пусть любое изделие некоторого производства оказывается браком с вероятностью p. Контроль продукции допускает ошибки: годное изделие бракует с вероятностью γ, а бракованное пропускает (признаёт годным) с вероятностью ε.

Если ввести для проверяемого изделия гипотезы $H_1 = \{\text{изделие годное}\}$ и $H_2 = \{\text{изделие бракованное}\}$, а критерием выбора одной из них считать контроль продукции, то γ — вероятность ошибки первого рода этого критерия, а ε — второго рода:

$$\begin{aligned} \gamma &= \mathsf{P}_{H_1}(\delta = H_2) = \mathsf{P}(\text{контроль забраковал годное изделие}); \\ \varepsilon &= \mathsf{P}_{H_2}(\delta = H_1) = \mathsf{P}(\text{контроль пропустил бракованное изделие}); \end{aligned}$$

У п р а ж н е н и е. Вычислить вероятности ошибок первого и второго рода того же критерия, если гипотезы занумеровать иначе:

$$H_1 = \{\text{изделие бракованное}\}, \quad H_2 = \{\text{изделие годное}\}.$$

О п р е д е л е н и е 19. Вероятность ошибки первого рода $\alpha_1 = \alpha_1(\delta)$ иначе называют *размером* или *критическим уровнем* критерия δ:

$$\alpha_1 = \alpha_1(\delta) = \mathsf{P}_{H_1}(\delta(\vec{X}) \neq H_1) = \mathsf{P}_{H_1}(\delta(\vec{X}) = H_2) = \mathsf{P}_{H_1}(\vec{X} \in S).$$

Мощностью критерия δ называют величину $1 - \alpha_2$, где $\alpha_2 = \alpha_2(\delta)$ — вероятность ошибки второго рода критерия δ. Мощность критерия равна

$$1 - \alpha_2(\delta) = 1 - \mathsf{P}_{H_2}(\delta(\vec{X}) \neq H_2) = \mathsf{P}_{H_2}(\delta(\vec{X}) = H_2) = \mathsf{P}_{H_2}(\vec{X} \in S).$$

Итак, статистический критерий не отвечает на вопрос, верна или нет проверяемая гипотеза (гипотезы). Он лишь решает, противоречат или не противоречат выдвинутой гипотезе выборочные данные, можно ли принять или следует отвергнуть данную гипотезу. При этом вывод о том, что данные противоречат гипотезе, всегда весомее и категоричнее, нежели вывод «данные не противоречат гипотезе».

П р и м е р 26. Пусть критерий $\delta(\vec{X}) \equiv H_1$ всегда выбирает первую гипотезу. Тогда $\alpha_1 = \mathsf{P}_{H_1}(\delta = H_2) = 0,\ \alpha_2 = \mathsf{P}_{H_2}(\delta = H_1) = 1.$

Наоборот: пусть критерий $\delta(\vec{X}) \equiv H_2$ всегда выбирает вторую гипотезу. Тогда $\alpha_1 = \mathsf{P}_{H_1}(\delta = H_2) = 1,\ \alpha_2 = \mathsf{P}_{H_2}(\delta = H_1) = 0.$

П р и м е р 27. Имеется выборка объёма $n = 1$ из нормального распределения $\mathrm{N}_{a,\,1}$ и две простые гипотезы $H_1 = \{a = 0\}$ и $H_2 = \{a = 1\}$. Рассмотрим при некотором $b \in \mathbb{R}$ следующий критерий:

$$\delta(X_1) = \begin{cases} H_1, & \text{если } \ X_1 \leqslant b, \\ H_2, & \text{если } \ X_1 > b. \end{cases}$$

Изобразим на графике (рис. 8) соответствующие гипотезам плотности распределений и вероятности ошибок первого и второго рода критерия δ

$$\alpha_1 = \mathsf{P}_{H_1}(X_1 > b), \qquad \alpha_2 = \mathsf{P}_{H_2}(X_1 \leqslant b).$$

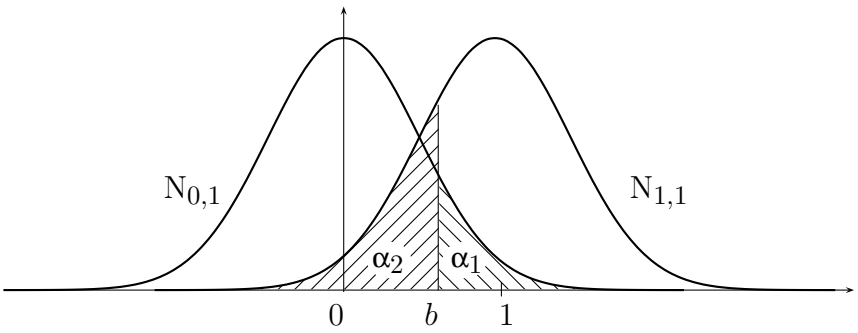

Рис. 8. Две простые гипотезы

О п р е д е л е н и е 20. Критерий δ называется критерием *асимптотического размера* ε, если $\alpha_1(\delta) \to \varepsilon$ при $n \to \infty$. Критерий δ называется *состоятельным*, если $\alpha_2(\delta) \to 0$ при $n \to \infty$.

§ 2. Вопросы и упражнения

1. Есть две гипотезы: основная состоит в том, что элементы выборки имеют нормальное распределение, а альтернатива — в том, что элементы выборки имеют распределение Пуассона. Построить критерий, обладающий нулевыми вероятностями ошибок первого и второго рода.

2. Пусть $X_1, \ldots, X_n$ — выборка из нормального распределения со средним a и единичной дисперсией. Для проверки основной гипотезы $a = 0$ против альтернативы $a = 1$ используется следующий критерий: основная гипотеза принимается, если $X_{(n)} < 2$, и отвергается в противном случае. Найти вероятности ошибок первого и второго рода.

3. Основная гипотеза состоит в том, что данный человек лишён телепатических способностей и угадывает мысли на расстоянии в каждом единичном эксперименте с вероятностью $1/2$. Гипотеза же о наличии телепатических способностей у данного человека принимается, если в 100 независимых однотипных экспериментах по угадыванию мыслей на расстоянии не менее 70 заканчиваются успехом. Чему равна вероятность признать телепатом человека без телепатических способностей?

4. Дана выборка $X_1, \ldots, X_n$ и две простые гипотезы: $H_1 = \{X_i$ имеют распределение с плотностью $f_1\}$, $H_2 = \{X_i$ имеют распределение с плотностью $f_2\}$, где

$$f_1(y) = \begin{cases} 3y^2, & \text{если } y \in [0,1], \\ 0 & \text{иначе;} \end{cases} \qquad f_2(y) = \begin{cases} 1, & \text{если } y \in [0,1], \\ 0 & \text{иначе.} \end{cases}$$

Критерий $\delta(X_1, \ldots, X_n)$ предписывает принимать гипотезу H_1, если все элементы выборки окажутся больше, чем $1/2$, и альтернативу H_2 — в противном случае. Найти вероятности ошибок первого и второго рода этого критерия.

5. Пусть $X_1, \ldots, X_{10}$ — выборка объема 10 из нормального распределения $N_{a,4}$. Используя соответствующий доверительный интервал, построить критерий согласия размера 0,1 для проверки гипотезы $a = 2$ против альтернативы $a \neq 2$. Принимается ли основная гипотеза при $\overline{X} = 2{,}5$?

ГЛАВА VII
КРИТЕРИИ СОГЛАСИЯ

Критериями согласия обычно называют критерии, предназначенные для проверки простой гипотезы $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ при сложной альтернативе, состоящей в том, что H_1 неверна. Мы рассмотрим более широкий класс основных гипотез, включающий в том числе и сложные гипотезы, а критериями согласия будем называть любые критерии, устроенные по одному и тому же принципу. А именно, пусть задана некоторая случайная величина, измеряющая отклонение эмпирического распределения от теоретического, распределение которой существенно разнится в зависимости от того, верна или нет основная гипотеза. Критерии согласия принимают или отвергают основную гипотезу исходя из величины этой функции отклонения.

§ 1. Общий вид критериев согласия

Мы опишем конструкцию критерия для случая простой основной гипотезы, а в дальнейшем будем её корректировать по мере изменения задачи.

Пусть $\vec{X} = (X_1, \ldots, X_n)$ — выборка из распределения $\mathcal{F}$. Проверяется основная гипотеза $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ при альтернативе $H_2 = \{\mathcal{F} \neq \mathcal{F}_1\}$.

Определение 21. Предположим, что нашлась функция $\rho(\vec{X})$ со следующими свойствами:

(К1) если гипотеза H_1 верна, т. е. если $X_i \sim \mathcal{F}_1$, то распределение величины $\rho(\vec{X})$ либо целиком известно, либо при больших n сближается с известным распределением $\mathcal{G}$;

(К2) если гипотеза H_1 неверна, т. е. если X_i имеют какое-то распределение $\mathcal{F}_2 \neq \mathcal{F}_1$, то $|\rho(\vec{X})| \xrightarrow{\mathrm{P}} \infty$ при $n \to \infty$ для любого такого $\mathcal{F}_2$.

Для случайной величины $\eta \sim \mathcal{G}$ определим постоянную C из равенства $\varepsilon = \mathsf{P}(|\eta| \geqslant C)$. Построим критерий

$$\delta(\vec{X}) = \begin{cases} H_1, & \text{если } |\rho(\vec{X})| < C, \\ H_2, & \text{если } |\rho(\vec{X})| \geqslant C. \end{cases} \tag{12}$$

Этот критерий называется *критерием согласия*.

Критерий согласия «работает» по принципу: если для данной выборки функция отклонения велика по абсолютному значению, то это свидетель-

ствует в пользу альтернативы, и наоборот. При этом степень «великости» определяется исходя из того, как функция отклонения должна себя вести, если бы основная гипотеза была верна. Действительно, если H_1 верна, статистика $\rho(\vec{X})$ имеет почти распределение $\mathcal{G}$. Следовательно, она должна себя вести подобно типичной случайной величине η из этого распределения. Но для случайной величины η попадание в область $\{|\eta| \geqslant C\}$ маловероятно: вероятность этого события равна малому числу ε. Поэтому попадание величины $\rho(\vec{X})$ в эту область заставляет подозревать, что гипотеза H_1 неверна. Тем более, что больших значений величины $|\rho(\vec{X})|$ следует ожидать именно при альтернативе H_2.

Убедимся в том, что этот критерий имеет (асимптотический) размер ε и является *состоятельным*.

Условие (K1) отвечает за размер критерия:

$$\alpha_1(\delta) = \mathsf{P}_{H_1}\big(|\rho(\vec{X})| \geqslant C\big) \to \mathsf{P}\,(|\eta| \geqslant C\,) = \varepsilon.$$

Расшифруем условие (K2), отвечающее за состоятельность критерия. По определению, запись $\xi_n \xrightarrow{\mathrm{p}} \infty$ означает, что для любого $C > 0$

$$\mathsf{P}(\xi_n < C) \to 0 \quad \text{при} \quad n \to \infty.$$

Согласно этому определению, для любого распределения $\mathcal{F}_2$ из числа альтернатив вероятность ошибки второго рода стремится к нулю:

$$\alpha_2(\delta,\, \mathcal{F}_2) = \mathsf{P}_{\mathcal{F}_2}\big(|\rho(\vec{X})| < C\big) \to 0,$$

что доказывает состоятельность критерия.

З а м е ч а н и е 10. Если в (K1) известно точное, а не приближённое распределение $\rho(\vec{X})$, то критерий (12) будет иметь точный размер ε.

Проверяя гипотезу, мы задали ε, затем по точному или предельному распределению $\rho(\vec{X})$ вычислили «барьер» C, с которым сравнили значение $|\rho(\vec{X})|$. На практике часто поступают иначе. Пусть по данной числовой выборке $\vec{x}$ вычислено число $\rho^* = \rho\big(\vec{x}\big)$. Число

$$\varepsilon^* = \mathsf{P}(|\eta\,| > |\rho^*|)$$

называют *реально достигнутым уровнем значимости* критерия. По величине ε^* можно судить о том, следует принять или отвергнуть основную гипотезу. *Именно это число является результатом проверки гипотезы в любом статистическом пакете программ.* Критерий (12) можно с её помощью записать так: H_1 отвергается при $\varepsilon^* \leqslant \varepsilon$.

Каков же смысл величины ε^*? Вероятность

$$\mathsf{P}_{H_1}(|\rho(\vec{X})| > |\rho^*|) \tag{13}$$

равна или приближённо равна ε^*. Вероятность (13) имеет следующий смысл: это вероятность, взяв выборку из распределения $\mathcal{F}_1$, получить по ней *бо́льшее* отклонение $|\rho(\vec{X})|$ эмпирического от истинного рапспределения, чем получено по проверяемой выборке. Больши́е значения вероятности (13) или ε^* свидетельствуют в пользу основной гипотезы. Напротив, малые значения вероятности (13) или ε^* свидетельствуют в пользу альтернативы.

Если, например, вероятность (13) равна 0,2, следует ожидать, что в среднем 20 % «контрольных» выборок, удовлетворяющих основной гипотезе (каждая пятая), будут обладать бо́льшим отклонением $|\rho(\vec{X})|$ по сравнению с тестируемой выборкой, в принадлежности которой распределению $\mathcal{F}_1$ мы не уверены. Можно отсюда сделать вывод, что тестируемая выборка ведёт себя не хуже, чем 20 % «правильных» выборок.

Но попадание в область вероятности 0,2 не является редким или «почти невозможным» событием. В статистике редкими обычно считают события с вероятностями $\varepsilon = 0{,}01$ или $\varepsilon = 0{,}05$ (это зависит от последствий ошибочного решения). Поэтому при $\varepsilon^* = 0{,}2 > 0{,}05$ основную гипотезу можно принять.

§ 2. Проверка простых гипотез о параметрах

Проверка гипотезы о среднем нормального распределения с известной дисперсией. Имеется выборка из нормального распределения $\mathrm{N}_{a,\,\sigma^2}$ с известной дисперсией σ^2. Проверяется гипотеза $H_1 = \{a = a_0\}$ против альтернативы $H_2 = \{a \neq a_0\}$. Построим критерий *точного* размера ε с помощью функции

$$\rho(\vec{X}) = \sqrt{n}\,\frac{\overline{X} - a_0}{\sigma}\,.$$

Очевидно свойство (K1): если H_1 верна, то $\rho(\vec{X}) \sim \mathrm{N}_{0,\,1}$.

По ε выберем $C = \tau_{1-\varepsilon/2}$ — квантиль стандартного нормального распределения. Критерий выглядит как все критерии согласия:

$$\delta(\vec{X}) = \begin{cases} H_1, & \text{если } |\rho(\vec{X})| < C, \\ H_2, & \text{если } |\rho(\vec{X})| \geqslant C. \end{cases} \tag{14}$$

У п р а ж н е н и е. Доказать, что критерий 14 имеет точный размер ε и является состоятельным.

Можно переписать этот критерий по-другому:

$$\delta(\vec{X}) = \begin{cases} H_1, & \text{если } \overline{X} \in \left(a_0 - \dfrac{C\sigma}{\sqrt{n}};\ a_0 + \dfrac{C\sigma}{\sqrt{n}}\right), \\ H_2 & \text{если } \overline{X} \leqslant a_0 - \dfrac{C\sigma}{\sqrt{n}} \quad \text{либо} \quad \overline{X} \geqslant a_0 + \dfrac{C\sigma}{\sqrt{n}}. \end{cases}$$

Вид критической области в критерии согласия зависит от вида альтернативной гипотезы. Так, для «двусторонней» альтернативы $H_2 = \{a \neq a_0\}$ критическая область имеет вид $|\rho| \geqslant C$, т. е. эта область есть объединение двух интервалов:

$$\overline{X} \leqslant a_0 - \frac{C\sigma}{\sqrt{n}} \quad \text{либо} \quad \overline{X} \geqslant a_0 + \frac{C\sigma}{\sqrt{n}}$$

Если же альтернатива будет односторонней, например, $H_2 = \{a < a_0\}$, то и критическую область следует брать одностороннюю: $\overline{X} \leqslant a_0 - \frac{C\sigma}{\sqrt{n}}$. При этом постоянную C следует выбирать так, чтобы при верной основной гипотезе вероятность попасть в критическую область равнялась ε. В данном случае $C = \tau_{1-\varepsilon}$, а не $\tau_{1-\varepsilon/2}$. Действительно: вероятность ε должна теперь соответствовать не двум «хвостам» нормального стандартного распределения, а одному.

Проверка гипотезы о среднем нормального распределения с неизвестной дисперсией. Проверяется та же гипотеза, что и в предыдущем разделе, но в случае, когда дисперсия σ^2 неизвестна. Критерий, который мы построим, называют одновыборочным критерием Стьюдента.

Введём функцию отклонения

$$\rho(\vec{X}) = \sqrt{n}\,\frac{\overline{X} - a_0}{\sqrt{S_0^2}}, \qquad \text{где } S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

По п. 4 *полезного следствия леммы Фишера* (с. 44) выполнено свойство (K1): если $a = a_0$, то ρ имеет распределение Стьюдента T_{n-1}.

Критерий строится в точности как в формуле (14), но в качестве C следует брать квантиль распределения Стьюдента, а не стандартного нормального распределения (*почему*?).

Критерии, основанные на доверительных интервалах. Имеется выборка из семейства распределений $\mathcal{F}_\theta$, где θ — неизвестный числовой параметр. Проверяется гипотеза $H_1 = \{\theta = \theta_0\}$ против альтернативы $H_2 = \{\theta \neq \theta_0\}$.

Пусть имеется точный доверительный интервал (θ^-, θ^+) для параметра θ уровня доверия $1 - \varepsilon$. Его можно использовать для проверки гипотезы H_1.

Критерий

$$\delta(\vec{X}) = \begin{cases} H_1, & \text{если } \ \theta_0 \in (\theta^-, \theta^+), \\ H_2, & \text{если } \ \theta_0 \notin (\theta^-, \theta^+) \end{cases}$$

имеет точный размер ε :

$$\alpha_1(\delta) = \mathsf{P}_{H_1}(\delta = H_2) = \mathsf{P}_{H_1}(\theta_0 \notin (\theta^-, \theta^+)) = 1 - \mathsf{P}_{H_1}(\theta^- < \theta_0 < \theta^+) = \varepsilon.$$

Критерий для проверки гипотезы о вероятности успеха или доле признака. Пусть дана выборка из распределения Бернулли с неизвестной вероятностью успеха p. Проверяется гипотеза $H_1 = \{p = p_0\}$ против альтернативы $H_2 = \{p \neq p_0\}$.

Используем ЦПТ для построения критерия. Пусть основная гипотеза верна. Тогда $X_i \sim \mathrm{B}_{p_0}$ и по теореме Муавра — Лапласа распределение случайной величины

$$\frac{n\overline{X} - np_0}{\sqrt{np_0(1-p_0)}}$$

с ростом n приближается к стандартному нормальному. Поэтому эту функцию можно взять в качестве функции $\rho(\vec{X})$.

Построим критерий:

$$\delta(\vec{X}) = \begin{cases} H_1, & \text{если } \left|\dfrac{n\overline{X} - np_0}{\sqrt{np_0(1-p_0)}}\right| < C, \\ H_2, & \text{иначе,} \end{cases}$$

где C выбирается как квантиль уровня $1 - \varepsilon/2$ стандартного нормального распределения, т. е. $\Phi(C) - \Phi(-C) = \varepsilon$.

Этот критерий имеет асимптотический размер ε и является состоятельным.

Пример 28. Монету подбросили 1 000 раз для проверки симметричности. При этом герб выпал 483 раза. Можно ли считать монету симметричной?

Сформулируем математическую задачу: по выборке из распределения Бернулли, в которой 483 единицы и 1 000-483 нуля, проверяется основная гипотеза $H_1 = \{p = 0{,}5\}$ против альтернативы $H_2 = \{p \neq 0{,}5\}$.

Мы можем либо воспользоваться критерием с каким-нибудь стандартным размером ε (например, 0,05), либо вычислить статистику критерия и реально достигнутый уровень значимости, по величине которого решить, принимать или отвергать основную гипотезу.

Пусть $\varepsilon = 0{,}05$. По таблице функции распределения стандартного нормального закона вычислим $C = 1{,}96$ — квантиль уровня $0{,}975 = 1 - \varepsilon/2$. Найдём статистику критерия. Величина $n\overline{X} = X_1 + \ldots + X_n$ равна числу единиц в выборке (числу гербов), поэтому

$$\rho(\vec{X}) = \frac{483 - 1\,000 \cdot 0{,}5}{\sqrt{1\,000 \cdot 0{,}5 \cdot 0{,}5}} = -1{,}075.$$

Поскольку $|\rho(\vec{X})| = 1{,}075$ меньше, чем $C = 1{,}96$, гипотезу H_1 можно принять. Таким образом, критерий с (асимптотическим) размером 0,05 сделал вывод о том, что выборочные данные не противоречат проверяемой гипотезе.

Вычислим реально достигнутый уровень значимости критерия:

$$\varepsilon^* = \mathsf{P}(|\xi| > 1{,}075) = 2\Phi(-1{,}075) = 2 \cdot 0{,}141 = 0{,}282,$$

где ξ имеет стандартное нормальное распределение.

Значение $\varepsilon^* = 0{,}282$ говорит о следующем: есть почти 30% шансов за то, что число гербов при бросании симметричной монеты будет больше отличаться от 500, чем полученное нами число гербов 483. Мы получили хорошее согласие с проверяемой гипотезой: есть целых тридцать процентов шансов получить худшее согласие даже для симметричной монеты. Любой критерий с вероятностью ошибки первого рода $\varepsilon < 0{,}282$ будет принимать основную гипотезу.

§ 3. Критерии для проверки гипотезы о распределении

Следующие два критерия используют для проверки гипотез о принадлежности выборки конкретному, целиком известному, распределению. Чаще всего таким распределением оказывается равномерное распределение. Например, оба этих критерии позволяют проверить равномерность результата генерации случайного числа на отрезке $[0,\,1]$. Критерий хи-квадрат Пирсона годится также для проверки гипотез о дискретном равномерном распределении. Например, о равномерности распределения дней рождения по дням недели, о симметричности монетки и т. п.

Критерий Колмогорова. Имеется выборка $\vec{X} = (X_1, \ldots, X_n)$ из распределения $\mathcal{F}$. Проверяется простая гипотеза $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ против сложной альтернативы $H_2 = \{\mathcal{F} \neq \mathcal{F}_1\}$. В том случае, когда распределение $\mathcal{F}_1$ имеет *непрерывную* функцию распределения F_1, можно пользоваться критерием Колмогорова. Критерий основан на следующей теореме.

Теорема 12 (Колмогорова). *Пусть дана выборка объёма n из распределения с непрерывной функцией распределения F_1, а F_n^* — эмпирическая функция распределения. Тогда при любом $y > 0$*

$$\mathsf{P}\left(\sqrt{n} \cdot \sup_{y \in \mathbb{R}} \left|F_n^*(y) - F_1(y)\right| < y\right) \to K(y) \quad \textit{при} \quad n \to \infty,$$

где $K(y)$ есть функция распределения Колмогорова

$$K(y) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2y^2}.$$

Значения этой функции (рис. 9) находят из соответствующих таблиц.

Положим $\rho(\vec{X}) = \sqrt{n} \sup\limits_y |F_n^*(y) - F_1(y)|$.

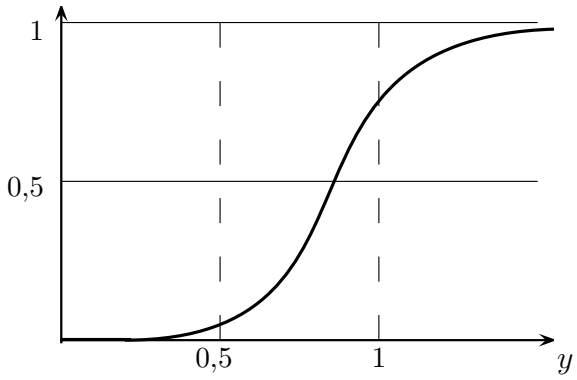

Рис. 9. График функции $K(y)$

По заданному ε с помощью таблицы значений функции $K(y)$ можно найти C такое, что $\varepsilon = K(y)$. Тогда критерий Колмогорова выглядит так:

$$\delta(\vec{X}) = \begin{cases} H_1, & \text{если } \rho(\vec{X}) < C, \\ H_2, & \text{если } \rho(\vec{X}) \geqslant C. \end{cases}$$

Этот критерий имеет асимптотический размер ε и является состоятельным.

Критерий χ^2 Пирсона. Критерий χ^2 основывается на группированных данных. Область значений предполагаемого распределения $\mathcal{F}_1$ делят на некоторое число интервалов. После чего строят функцию отклонения ρ по разностям *теоретических вероятностей* попадания в интервалы группировки и *эмпирических частот*.

Дана выборка объёма n из распределения $\mathcal{F}$. Проверяется простая гипотеза $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ при альтернативе $H_2 = \{\mathcal{F} \neq \mathcal{F}_1\}$.

Пусть $A_1, \ldots, A_k$ — интервалы группировки в области значений случайной величины с предполагаемым распределением $\mathcal{F}_1$. Пусть для каждого $j = 1, \ldots, k$ величина ν_j равна числу элементов выборки, попавших в интервал A_j:

$$\nu_j = \{\text{число } X_i \in A_j\} = \sum_{i=1}^{n} \mathrm{I}(X_i \in A_j),$$

Пусть число $p_j > 0$ равно теоретической вероятности попадания в интервал A_j случайной величины с распределением $\mathcal{F}_1$. Здесь $p_1 + \ldots + p_k = 1$. Как правило, длины интервалов выбирают так, чтобы $p_1 = \ldots = p_k = 1/k$. Пусть

$$\rho(\vec{X}) = \sum_{j=1}^{k} \frac{(\nu_j - np_j)^2}{np_j}. \tag{15}$$

Замечание 11. Поскольку мы строим критерий, опираясь только на частоту попадания элементов выборки в интервалы группировки, мы долж-

ны заранее понимать, что критерий не сможет отличить два распределения, у которых одинаковы вероятности попасть во все интервалы группировки.

Верна теорема.

Теорема 13 (Пирсона). *Если верна гипотеза* H_1, *то при фиксированном* k *и при* $n \to \infty$ *распределение величины* $\rho(\vec{X})$ *приближается к распределению* H_{k-1}, *где* H_{k-1} *есть* χ^2*-распределение с* $k-1$ *степенью свободы.*

Осталось построить критерий согласия по определению 21. Пусть случайная величина η имеет распределение H_{k-1}. По таблице распределения H_{k-1} найдём C, равное квантили уровня $1-\varepsilon$ этого распределения: $\varepsilon = \mathsf{P}(\eta \geqslant C)$. Критерий χ^2 устроен обычным образом:

$$\delta(\vec{X}) = \begin{cases} H_1, & \text{если } \rho(\vec{X}) < C, \\ H_2, & \text{если } \rho(\vec{X}) \geqslant C. \end{cases}$$

Число интервалов k выбирают так, чтобы значения $np_1 = \ldots = np_k$ были не менее 5—6. Если выборка уже сгруппирована, то группы, в которые попало менее пяти элементов выборки, объединяют с соседними, уменьшая тем самым число интервалов группировки.

Пример 29. Для проверки равномерности распределения дней рождения по месяцам года взят список дней рождений 683 студентов ИВТ СибГУТИ по данным на сентябрь 2007 г. Получено следующее распределение дней рождения: январь — 60, февраль — 62, март — 60, апрель — 63, май — 69, июнь — 59, июль — 62, август — 54, сентябрь — 41, октябрь — 45, ноябрь — 48 и декабрь — 60.

Итак, есть 12 интервалов группировки. Проверяемая гипотеза состоит в том, что вероятность элементу выборки попасть в каждый из них одна и та же и равна $1/12$ (можно было взять разные вероятности, пропорциональные числу дней каждого месяца).

Вычислим статистику критерия:

$$\rho(\vec{X}) = \frac{(60-683/12)^2}{683/12} + \frac{(62-683/12)^2}{683/12} + \ldots + \frac{(60-683/12)^2}{683/12} = 13{,}193.$$

Возьмём $\varepsilon = 0{,}05$ и найдём по таблице критических точек распределения χ^2_{11} величину C такую, что $\mathsf{P}(\chi^2_{11} > C) = 0{,}05$. Получим $C = 19{,}68$. Величина ρ оказалась меньше C, поэтому критерий принимает основную гипотезу. Реально достигнутый уровень значимости $\varepsilon^* = \mathsf{P}(\chi^2_{11} > 13{,}193)$ равен 0,281 (для его вычисления следует воспользоваться более подробными таблицами или любым подходящим пакетом программ). Он показывает, что достигнуто достаточно хорошее согласие с проверяемой гипотезой.

§ 4. Критерии для проверки параметрических гипотез

Очень часто требуется проверить, например, нормальность распределения выборки безо всякого знания о параметрах распределения. Предыдущие два критерия не годятся, поскольку проверяемая гипотеза является сложной. Следующий критерий является вариантом критерия Пирсона.

Критерий χ^2 для проверки параметрической гипотезы. Критерий χ^2 часто применяют для проверки гипотезы о принадлежности распределения выборки некоторому параметрическому семейству.

Пусть дана выборка из неизвестного распределения $\mathcal{F}$. Проверяется гипотеза о том, что это распределение принадлежит некоторому семейству распределений $\mathcal{F}_\theta$, где θ — неизвестный векторный параметр (размерности d).

Разобьём всю числовую ось на $k > d+1$ интервалов группировки $A_1, \ldots A_k$ и вычислим ν_j — число элементов выборки, попавших в интервал A_j. Но теперь вероятность $p_j = \mathsf{P}_{H_1}(X_1 \in A_j) = p_j(\theta)$ зависит от неизвестного параметра θ. Функция отклонения (15) также зависит от неизвестного параметра θ, и использовать её в критерии Пирсона нельзя:

$$\rho(\vec{X};\, \theta) = \sum_{j=1}^{k} \frac{(\nu_j - np_j(\theta))^2}{np_j(\theta)}. \tag{16}$$

Пусть θ^* — такое значение параметра θ, при котором функция $\rho(\vec{X};\, \theta)$ принимает наименьшее значение. Подставив вместо истинных вероятностей p_j их оценки $p_j(\theta^*)$, получим функцию отклонения

$$\rho(\vec{X};\, \theta^*) = \sum_{j=1}^{k} \frac{(\nu_j - np_j(\theta^*))^2}{np_j(\theta^*)}. \tag{17}$$

Теорема 14 (Р. Фишер). *Пусть верна гипотеза H_1. Если число d есть размерность вектора параметров θ и выполнены некоторые условия гладкости функций $p_j(\theta)$, то при фиксированном k и при $n \to \infty$ распределение величины $\rho(\vec{X};\, \theta^*)$ сближается с распределением H_{k-1-d}, где H_{k-1-d} есть χ^2-распределение с $k-1-d$ степенями свободы.*

Построим критерий χ^2. Пусть случайная величина η имеет распределение H_{k-1-d}. По заданному ε найдём C такое, что $\varepsilon = \mathsf{P}(\eta \geqslant C)$.

Критерий согласия χ^2 устроен обычным образом:

$$\delta(\vec{X}) = \begin{cases} H_1, & \text{если } \rho(\vec{X};\, \theta^*) < C, \\ H_2, & \text{если } \rho(\vec{X};\, \theta^*) \geqslant C. \end{cases}$$

Замечание 12. Вычисление точки минимума функции $\rho(\vec{X};\,\theta)$ в общем случае возможно лишь численно. Поэтому часто вместо оценки θ^* используют оценку максимального правдоподобия, построенную по выборке $X_1,\,\ldots,\,X_n$. Однако при такой замене предельное распределение величины $\rho(\vec{X};\,\theta)$ уже не равно H_{k-1-d} и зависит от θ.

Данный вариант критерия Пирсона годится для проверки любой параметрической гипотезы. Но для проверки нормальности распределения выборки можно использовать специальные критерии.

Критерий Андерсона — Дарлинга. Пусть $\vec{X} = (X_1,\,\ldots,\,X_n)$ — выборка из неизвестного распределения и $X_{(1)},\,\ldots,\,X_{(n)}$ — соответствующий вариационный ряд (выборка, упорядоченная по возрастанию).

Проверяется гипотеза H_1 о том, что распределение выборки принадлежит классу нормальных распределений (с неизвестными параметрами).

Вычислим выборочное среднее $\overline{X}$, выборочную дисперсию S^2 и преобразуем элементы *вариационного ряда*:

$$Y_i = \frac{X_{(i)} - \overline{X}}{S}\,.$$

Построим статистику критерия Андерсона — Дарлинга так:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\Big[(2i-1)\ln\Phi(Y_i) + (2n-2i+1)\ln\left(1-\Phi(Y_i)\right)\Big].$$

Обычно вводят поправочный коэффициент, необходимый для небольших объёмов выборки:

$$A^{*2} = A^2\left(1 + \frac{0{,}75}{n} + \frac{2{,}25}{n^2}\right).$$

Предельное распределение статистики Андерсона — Дарлинга при верной основной гипотезе имеет весьма сложный вид. Приведём значения квантилей этого распределения для нескольких часто используемых уровней:

$$h_{0,9} = 0{,}631; \quad h_{0,95} = 0{,}752; \quad h_{0,975} = 0{,}873; \quad h_{0,99} = 1{,}035.$$

Критерий Андерсона — Дарлинга принимает гипотезу о нормальности распределения выборки, если $A^{*2} < h_{1-\varepsilon}$, и отвергает в противном случае. Вероятность ошибки первого рода этого критерия с ростом n стремится к ε.

Заметим, что критерий Андерсона — Дарлинга годится не только для проверки нормальности: используя в статистике критерия вместо Φ другие непрерывные функции распределения, можно проверять принадлежность выборки соответствующему распределению. Однако предельное распределение статистики критерия зависит от теоретического распределения, поэтому для проверки других гипотез следует использовать другие квантили.

Критерий Жарка — Бера (Jarque — Bera). Пусть $\vec{X} = (X_1, \ldots, X_n)$ — выборка из неизвестного распределения. Проверяется гипотеза H_1 о том, что распределение выборки принадлежит классу нормальных распределений (с неизвестными параметрами). Критерий основан на величине выборочных коэффициентов асимметрии и эксцесса. Напомним, что для нормального распределения коэффициенты асимметрии и эксцесса равны нулю.

Вычислим выборочное среднее $\overline{X}$, выборочную дисперсию S^2 и введём выборочные асимметрию и эксцесс так:

$$\beta_1^* = \frac{\frac{1}{n}\sum (X_i - \overline{X})^3}{S^3}, \qquad \beta_2^* = \frac{\frac{1}{n}\sum (X_i - \overline{X})^4}{S^4} - 3.$$

Статистику критерия построим так:

$$\rho(\vec{X}) = n\left(\frac{\beta_1^{*2}}{6} + \frac{\beta_2^{*2}}{24}\right).$$

Если гипотеза о нормальности верна, распределение статистики критерия с ростом n приближается к распределению хи-квадрат с двумя степенями свободы, т. е. к показательному распределению с параметром $0{,}5$. Это означает, что для любого y

$$\mathsf{P}_{H_1}(\rho(\vec{X}) > y) \to e^{-y/2} \quad \text{при} \quad n \to \infty.$$

Поэтому критерий Жарка — Бера с асимптотическим размером ε предписывает отвергать основную гипотезу, как только $\rho > C$, где C есть решение уравнения $e^{-C/2} = \varepsilon$, т. е. $C = -2\ln\varepsilon$.

§ 5. Критерии для проверки однородности

Двувыборочный критерий Колмогорова—Смирнова. Даны две независимые выборки $\vec{X} = (X_1, \ldots, X_n)$ и $\vec{Y} = (Y_1, \ldots, Y_m)$ из неизвестных распределений $\mathcal{F}$ и $\mathcal{G}$ соответственно. Проверяется сложная гипотеза $H_1 = \{\mathcal{F} = \mathcal{G}\}$ при альтернативе $H_2 = \{H_1 \text{ неверна}\}$.

Критерий Колмогорова — Смирнова используют, если $\mathcal{F}$ и $\mathcal{G}$ имеют *непрерывные функции распределения*.

Пусть $F_n^*(y)$ и $G_m^*(y)$ — эмпирические функции распределения, построенные по выборкам $\vec{X}$ и $\vec{Y}$,

$$\rho(\vec{X}, \vec{Y}) = \sqrt{\frac{mn}{m+n}} \sup_y \left|F_n^*(y) - G_m^*(y)\right|.$$

Т е о р е м а 15. *Если гипотеза H_1 верна, то для любого $y > 0$*

$$\mathsf{P}\left(\rho(\vec{X}, \vec{Y}) < y\right) \to K(y) \quad \textit{при} \quad n,\ m \to \infty.$$

В таблице распределения Колмогорова по заданному ε найдём C такое, что $\varepsilon = K(C)$, и построим критерий Колмогорова — Смирнова

$$\delta(\vec{X}) = \begin{cases} H_1, & \text{если } \rho(\vec{X}) < C, \\ H_2, & \text{если } \rho(\vec{X}) \geqslant C. \end{cases}$$

Ранговый критерий Вилкоксона, Манна и Уитни. Даны две независимые выборки $\vec{X} = (X_1, \ldots, X_n)$ и $\vec{Y} = (Y_1, \ldots, Y_m)$ из неизвестных распределений $\mathcal{F}$ и $\mathcal{G}$. Проверяется сложная гипотеза $H_1 = \{\mathcal{F} = \mathcal{G}\}$ при альтернативе $H_2 = \{H_1 \text{ неверна}\}$.

Критерий Вилкоксона, Манна и Уитни (Wilcoxon, Mann, Whitney) используют, если $\mathcal{F}$ и $\mathcal{G}$ имеют *непрерывные функции распределения*. Составим из выборок $\vec{X}$ и $\vec{Y}$ общий вариационный ряд и подсчитаем статистику Вилкоксона W, равную сумме рангов $r_1, \ldots, r_m$ (номеров мест) элементов выборки $\vec{Y}$ в общем вариационном ряду. Зададим функцию U так (статистика Манна — Уитни):

$$U = W - \frac{1}{2}\, m(m+1).$$

Статистику критерия возьмём, центрировав и нормировав статистику U:

$$\rho(\vec{X}, \vec{Y}) = \frac{U - nm/2}{\sqrt{nm(n+m+1)/12}}.$$

Мы не будем доказывать следующее утверждение.

Т е о р е м а 16. *Если непрерывные распределения $\mathcal{F}$ и $\mathcal{G}$ таковы, что $\mathsf{P}(X_1 < Y_1) = 0{,}5$ то распределение величины $\rho(\vec{X}, \vec{Y})$ приближается к стандартному нормальному распределению при $n, m \to \infty$.*

Построим критерий асимптотического размера ε:

$$\delta(\vec{X}, \vec{Y}) = \begin{cases} H_1, & \text{если } |\rho(\vec{X})| < C, \\ H_2, & \text{если } |\rho(\vec{X})| \geqslant C, \end{cases}$$

где C — квантиль уровня $1 - \varepsilon/2$ распределения $\mathrm{N}_{0,\,1}$. Пользоваться этим критерием рекомендуют при $\min(n, m) > 25$.

Этот критерий может отличить от H_1 далеко не любую гипотезу. Например, если $\mathcal{F}$ и $\mathcal{G}$ — два нормальных распределения с одним и тем же средним, но разными дисперсиями, то разность $X_i - Y_j$ имеет нормальное распределение с нулевым средним, и примерно в половине случаев X_i будут меньше или больше Y_i.

Итак, на самом деле построенный выше критерий проверяет гипотезу

$$H_1' = \Big\{\text{распределения выборок таковы, что } \mathsf{P}(X_1 < Y_1) = \frac{1}{2}\Big\}.$$

Используя его для проверки первоначальной гипотезы однородности, следует помнить, какие альтернативы он не отличает от основной гипотезы.

Критерий Фишера. Критерий Фишера используют в качестве первого шага в задаче проверки однородности двух независимых нормальных выборок. Особенно часто возникает необходимость проверить равенство *средних* двух нормальных совокупностей: например, в медицине или биологии для выяснения наличия или отсутствия действия препарата. Эта задача решается с помощью критерия Стьюдента (с ним мы познакомимся на следующей странице), но только в случае, когда неизвестные дисперсии *равны*. Для проверки же равенства дисперсий пользуются сначала критерием Фишера. Самое печальное, если гипотеза равенства дисперсий отвергается критерием Фишера. Задачу о построении критерия точного размера ε (что особенно важно при маленьких выборках) для проверки равенства средних в этих условиях называют *проблемой Беренса — Фишера*. Её решение возможно лишь в частных случаях.

Пусть даны две независимые выборки из нормальных распределений: $\vec{X} = (X_1, \ldots, X_n)$ из $\mathrm{N}_{a_1,\,\sigma_1^2}$ и $\vec{Y} = (Y_1, \ldots, Y_m)$ из $\mathrm{N}_{a_2,\,\sigma_2^2}$, средние которых, вообще говоря, неизвестны. Критерий Фишера предназначен для проверки гипотезы $H_1 = \{\sigma_1 = \sigma_2\}$.

Обозначим через $S_0^2(\vec{X})$ и $S_0^2(\vec{Y})$ несмещённые выборочные дисперсии

$$S_0^2(\vec{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad S_0^2(\vec{Y}) = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2$$

и зададим функцию $\rho(\vec{X},\,\vec{Y})$ как их отношение $\rho(\vec{X},\,\vec{Y}) = S_0^2(\vec{X})/S_0^2(\vec{Y})$.

Удобно, если $\rho > 1$. С этой целью выборкой $\vec{X}$ называют ту из двух выборок, несмещённая дисперсия которой больше. Поэтому предположим, что $S_0^2(\vec{X}) > S_0^2(\vec{Y})$.

Т е о р е м а 17. *При верной гипотезе H_1 величина $\rho(\vec{X},\,\vec{Y})$ имеет распределение Фишера $\mathrm{F}_{n-1,\,m-1}$ с $n-1$ и $m-1$ степенями свободы.*

Д о к а з а т е л ь с т в о. По лемме Фишера, независимые случайные величины

$$\chi^2_{n-1} = \frac{(n-1)\,S_0^2(\vec{X})}{\sigma_1^2} \quad \text{и} \quad \psi^2_{m-1} = \frac{(m-1)\,S_0^2(\vec{Y})}{\sigma_2^2}$$

имеют распределения H_{m-1} и H_{n-1} соответственно. При $\sigma_1 = \sigma_2$ по определению распределения Фишера

$$\rho(\vec{X},\,\vec{Y}) = \frac{S_0^2(\vec{X})}{\sigma_1^2}\cdot\frac{\sigma_2^2}{S_0^2(\vec{Y})} = \frac{\chi^2_{n-1}/(n-1)}{\psi^2_{m-1}/(m-1)} \sim \mathrm{F}_{n-1,\,m-1}. \qquad \square$$

Возьмём квантиль $f_{1-\varepsilon}$ распределения Фишера $\mathrm{F}_{n-1,\,m-1}$. Критерием Фишера называют критерий

$$\delta(\vec{X},\,\vec{Y}) = \begin{cases} H_1, & \text{если } \ \rho(\vec{X},\,\vec{Y}) < f_{1-\varepsilon}, \\ H_2 & \text{если } \ \rho(\vec{X},\,\vec{Y}) \geqslant f_{1-\varepsilon}. \end{cases}$$

Критерий Стьюдента. Пусть имеются две независимые выборки: выборка $\vec{X} = (X_1,\,\ldots,\,X_n)$ из $\mathrm{N}_{a_1,\,\sigma^2}$ и выборка $\vec{Y} = (Y_1,\,\ldots,\,Y_m)$ из $\mathrm{N}_{a_2,\,\sigma^2}$ с неизвестными средними и *одной и той же* неизвестной дисперсией σ^2. Проверяется сложная гипотеза $H_1 = \{a_1 = a_2\}$.

Построим критерий Стьюдента *точного* размера ε.

Т е о р е м а 18. *Случайная величина* t_{n+m-2}, *равная*

$$t_{n+m-2} = \sqrt{\frac{nm}{n+m}} \cdot \frac{(\overline{X} - a_1) - (\overline{Y} - a_2)}{\sqrt{\dfrac{(n-1)S_0^2(\vec{X}) + (m-1)S_0^2(\vec{Y})}{n+m-2}}}$$

имеет распределение Стьюдента T_{n+m-2}.

Д о к а з а т е л ь с т в о. Легко видеть (*убедиться, что легко*!), что случайная величина $\overline{X} - a_1$ имеет распределение $\mathrm{N}_{0,\,\sigma^2/n}$, а случайная величина $\overline{Y} - a_2$ имеет распределение $\mathrm{N}_{0,\,\sigma^2/m}$. Тогда их разность распределена тоже нормально с нулевым средним и дисперсией

$$\mathsf{D}\big((\overline{X} - a_1) - (\overline{Y} - a_2)\big) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2 \cdot \frac{n+m}{nm}.$$

Нормируем эту разность:

$$\xi_0 = \frac{1}{\sigma}\sqrt{\frac{nm}{n+m}}\,\big((\overline{X} - a_1) - (\overline{Y} - a_2)\big) \sim \mathrm{N}_{0,\,1}$$

Из леммы Фишера следует, что независимые случайные величины $(n-1)\,S_0^2(\vec{X})/\sigma^2$ и $(m-1)\,S_0^2(\vec{Y})/\sigma^2$ имеют распределения H_{n-1} и H_{m-1} соответственно, а их сумма

$$S^2 = \frac{1}{\sigma^2}\,\big((n-1)S_0^2(\vec{X}) + (m-1)S_0^2(\vec{Y})\big)$$

имеет χ^2-распределение H_{n+m-2} с $n+m-2$ степенями свободы (*почему*?) и не зависит от $\overline{X}$ и от $\overline{Y}$ (*почему*?).

По определению 14 (с. 39), отношение $\dfrac{\xi_0}{\sqrt{S^2/(n+m-2)}}$ имеет распределение Стьюдента T_{n+m-2}. Осталось подставить в эту дробь ξ_0 и S^2 и убедиться, что σ сократится и получится t_{n+m-2} из теоремы 18. □

Введём функцию

$$\rho(\vec{X},\,\vec{Y}) \;=\; \sqrt{\frac{nm}{n+m}}\cdot\frac{\overline{X}-\overline{Y}}{\sqrt{\dfrac{(n-1)S_0^2(\vec{X})+(m-1)S_0^2(\vec{Y})}{n+m-2}}}.$$

Из теоремы 18 следует свойство (K1): если H_1 верна, т. е. если $a_1 = a_2$, то величина $\rho = t_{n+m-2}$ имеет распределение Стьюдента T_{n+m-2}.

Критерий Стьюдента выглядит как все критерии согласия: при двусторонней альтернативе $H_2 = \{a_1 \neq a_2\}$

$$\delta(\vec{X},\,\vec{Y}) = \begin{cases} H_1, & \text{если } \ |\rho(\vec{X},\,\vec{Y})| < C, \\ H_2, & \text{если } \ |\rho(\vec{X},\,\vec{Y})| \geqslant C, \end{cases}$$

где число $C = \tau_{1-\varepsilon/2}$ — квантиль распределения T_{n+m-2}.

При односторонней альтернативе $H_2 = \{a_1 < a_2\}$ или $H_2 = \{a_1 > a_2\}$ критерий имеет вид

$$\delta(\vec{X},\,\vec{Y}) = \begin{cases} H_1, & \rho(\vec{X},\,\vec{Y}) > -C, \\ H_2, & \rho(\vec{X},\,\vec{Y}) \leqslant -C \end{cases} \quad \text{или} \quad \delta(\vec{X},\,\vec{Y}) = \begin{cases} H_1, & \rho(\vec{X},\,\vec{Y}) < C, \\ H_2, & \rho(\vec{X},\,\vec{Y}) \geqslant C, \end{cases}$$

где число $C = \tau_{1-\varepsilon}$ — квантиль распределения T_{n+m-2}.

Пример 30. По двум независимым выборкам из нормальных распределений найдены выборочные средние $\overline{X} = 136{,}53$ и $\overline{Y} = 142{,}21$, а также несмещённые выборочные дисперсии $S_0^2(\vec{X}) = 2{,}7$ и $S_0^2(\vec{Y}) = 3{,}3$. Объёмы выборок равны 13 и 10 соответственно. При уровне значимости 0,05 проверить гипотезу равенства средних $H_1 = \{a_1 = a_2\}$ при односторонней альтернативе $H_1 = \{a_1 < a_2\}$.

Проверим сначала критерием Фишера гипотезу равенства дисперсий. Дисперсионное отношение равно (делим большую дисперсию на меньшую)

$$\rho(\vec{Y},\,\vec{X}) = \frac{S_0^2(\vec{Y})}{S_0^2(\vec{X})} = \frac{3{,}3}{2{,}7} = 1{,}222.$$

Найдём по таблице 5 приложения число C такое, что $\mathsf{P}(f_{9,\,12} > C) = 0{,}05$. Получаем $C = 2{,}8$. Поскольку значение статистики критерия 1,222 меньше, чем 2,8, нет оснований отвергнуть гипотезу равенства дисперсий.

Воспользуемся критерием Стьюдента для проверки равенства средних. Вычислим статистику критерия Стьюдента:

$$\rho(\vec{X},\,\vec{Y}) \;=\; \sqrt{\frac{13\cdot 10}{13+10}}\cdot\frac{136{,}53-142{,}21}{\sqrt{\dfrac{(13-1)2{,}7+(10-1)3{,}3}{13+10-2}}} = -7{,}853.$$

Поскольку альтернатива односторонняя, критическая область будет иметь вид $\rho \leqslant -C$, где C таково, что $\mathsf{P}(t_{21} > C) = 0{,}05$. В таблице 4 приложения приведены границы лишь для двусторонних критических областей. Поэтому воспользуемся этой таблицей с удвоенным $\alpha = 2\varepsilon = 0{,}1$. Получим $C = 1{,}72$. Видим, что значение статистики критерия попадает в критическую область: $-7{,}853 < -1{,}72$, поэтому основную гипотезу о равенстве средних следует отвергнуть в пользу альтернативы: истинное математическое ожидание у первой выборки меньше, чем у второй.

Заметим, что при такой большой разнице средних даже весьма «нетребовательный» критерий с размером $\varepsilon = 0{,}0005$, почти никогда не отвергающий основную гипотезу, всё равно вынужден будет отвергнуть наше предположение, поскольку $-7{,}853 < -3{,}82$.

Однофакторный дисперсионный анализ. Предположим, что влияние некоторого «фактора» на наблюдаемые нормально распределённые величины может сказываться только на значениях их математических ожиданий. Мы наблюдаем несколько выборок при различных «уровнях» фактора. Требуется определить, влияет или нет изменение уровня фактора на математическое ожидание.

Говоря формальным языком, однофакторный дисперсионный анализ решает задачу проверки равенства средних нескольких независимых нормально распределённых выборок с одинаковыми дисперсиями. Для двух выборок эту задачу мы решили с помощью критерия Стьюдента.

Пусть даны k независимых выборок

$$X^{(1)} = (x_1^{(1)}, \ldots, x_{n_1}^{(1)}), \quad \ldots, \quad X^{(k)} = (x_1^{(k)}, \ldots, x_{n_k}^{(k)})$$

из нормальных распределений $x_i^{(j)} \sim \mathrm{N}_{a_j,\,\sigma^2}$ с одной и той же дисперсией. Верхний индекс у наблюдений отвечает номеру выборки. Проверяется основная гипотеза $H_1 = \{a_1 = \ldots = a_k\}$.

Для каждой выборки вычислим выборочные среднее и дисперсию

$$\overline{X}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i^{(j)}, \qquad S^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} \bigl(x_i^{(j)} - \overline{X}^{(j)}\bigr)^2.$$

Положим $n = n_1 + \ldots + n_k$. Определим также общее выборочное среднее и общую выборочную дисперсию

$$\overline{X} = \frac{1}{n} \sum_{i,\,j} x_i^{(j)} = \frac{1}{n} \sum_{j=1}^{k} n_j \overline{X}^{(j)}, \qquad S^2 = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} \bigl(x_i^{(j)} - \overline{X}\bigr)^2.$$

Критерий для проверки гипотезы H_1 основан на сравнении внутригрупповой и межгрупповой дисперсий. Вычислим так называемую *межгрупповую дисперсию*, или *дисперсию выборочных средних*

$$S^2_{\text{м}} = \frac{1}{n}\sum_{j=1}^{k} n_j\big(\overline{X}^{(j)} - \overline{X}\big)^2.$$

Она показывает, насколько отличны друг от друга выборочные средние при разных уровнях фактора. Именно эта дисперсия отражает влияние фактора. При этом каждое выборочное среднее вносит в дисперсию вклад, пропорциональный объёму соответствующей выборки: выбросы средних могут быть вызваны малым числом наблюдений.

Вычислим так называемую *внутригрупповую дисперсию*

$$S^2_{\text{в}} = \frac{1}{n}\sum_{j=1}^{k} n_j\, S^{(j)} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\big(x_i^{(j)} - \overline{X}^{(j)}\big)^2.$$

Она показывает, насколько велики разбросы внутри выборок относительно выборочных средних. Эти разбросы определяются случайностью внутри выборок. Вывод о том, что средние существенно различны, т. е. присутствует влияние фактора на среднее, может быть сделан, если межгрупповая дисперсия оказывается существенно больше внутригрупповой. Чтобы понять, насколько больше, следует рассмотреть распределения этих случайных величин при верной основной гипотезе.

По основному следствию из леммы Фишера при любом $j = 1, \ldots, k$ величина $n_j\, S^{(j)}/\sigma^2$ имеет распределение H_{n_j-1} и не зависит от $\overline{X}^{(j)}$. Из независимости выборок и устойчивости χ^2-распределения относительно суммирования получаем

$$\frac{nS^2_{\text{в}}}{\sigma^2} = \sum_{j=1}^{k}\frac{n_j\, S^{(j)}}{\sigma^2} \sim \text{H}_{n-k}, \quad \text{где } \; n-k = n_1 - 1 + \ldots + n_k - 1.$$

Кроме того, величина $S^2_{\text{в}}$ не зависит от $\overline{X}^{(1)}, \ldots, \overline{X}^{(k)}$. Поэтому она не зависит и от их взвешенного среднего $\overline{X}$, а также (что уже совсем невероятно) от межгрупповой дисперсии $S^2_{\text{м}}$, поскольку последняя является функцией *только* от перечисленных средних. Эти свойства никак не связаны с проверяемой гипотезой и верны независимо от равенства или неравенства истинных средних.

Пусть гипотеза H_1 верна. Тогда выборки можно считать одной выборкой объёма n. По основному следствию леммы Фишера $nS^2/\sigma^2 \sim \text{H}_{n-1}$.

Величины S^2, $S^2_{\text{м}}$ и $S^2_{\text{в}}$ удовлетворяют легко проверяемому *основному дисперсионному соотношению*

$$\frac{nS^2}{\sigma^2} = \frac{nS^2_{\text{м}}}{\sigma^2} + \frac{nS^2_{\text{в}}}{\sigma^2}.$$

Величина в левой части имеет распределение H_{n-1}, справа — сумма двух независимых слагаемых, второе из которых имеет распределение H_{n-k}. Оказывается, что тогда первое распределено по закону H_{k-1}.

Итак, при верной гипотезе H_1 мы получили два χ^2-распределения независимых случайных величин

$$\chi^2 = \frac{nS^2_{\text{м}}}{\sigma^2} \sim \text{H}_{k-1} \quad \text{и} \quad \psi^2 = \frac{nS^2_{\text{в}}}{\sigma^2} \sim \text{H}_{n-k}.$$

Построим по ним статистику из распределения Фишера $\text{F}_{k-1,\,n-k}$

$$\rho = \frac{\chi^2}{k-1} \cdot \frac{n-k}{\psi^2} = \frac{n-k}{k-1} \cdot \frac{S^2_{\text{м}}}{S^2_{\text{в}}} \sim \text{F}_{k-1,\,n-k}.$$

По заданному ε найдём квантиль C уровня $1-\varepsilon$ распределения Фишера $\text{F}_{k-1,\,n-k}$ и устроим следующий критерий точного размера ε:

$$\delta = \begin{cases} H_1, & \text{если } \rho < C, \\ H_2, & \text{если } \rho \geqslant C. \end{cases}$$

Замечание 13. Предположение о равенстве дисперсий проверяют, например, с помощью критерия Бартлетта (см. [6]).

Сравнение долей признака в двух выборках. Пусть есть две независимые выборки $\vec{X} = (X_1, \ldots, X_{n_1})$ и $\vec{Y} = (Y_1, \ldots, Y_{n_2})$ из распределений Бернулли. Как проверить гипотезу о совпадении вероятностей успеха этих распределений? Обычно даны даже не выборки, а общее число успехов в каждой серии испытаний $m_1 = n_1\overline{X}$ и $m_2 = n_2\overline{Y}$, либо доли успехов $w_1 = m_1/n_1$, $w_2 = m_2/n_2$.

Пусть p_1 и p_2 — гипотетические вероятности успеха. Проверяется основная гипотеза $H_1 = \{p_1 = p_2\}$ при возможных альтернативах $p_1 \neq p_2$, $p_1 < p_2$ или $p_1 > p_2$.

Если основная гипотеза верна, выборки можно соединить в одну большую выборку из распределения Бернулли, и вероятность успеха оценить общей долей $p^* = (m_1 + m_2)/(n_1 + n_2)$.

Построим статистику аналогично тому, как это делалось в критерии Стьюдента. При верной основной гипотезе статистика

$$\rho(\vec{X}, \vec{Y}) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, \frac{w_1 - w_2}{\sqrt{p^*(1-p^*)}}$$

имеет при больших n_1 и n_2 распределение, близкое к стандартному нормальному. Это сразу следует из теоремы Муавра — Лапласа.

Построим критерий для случая двусторонней альтернативы:

$$\delta(\vec{X},\,\vec{Y}) = \begin{cases} H_1, & \text{если } \ |\rho(\vec{X},\,\vec{Y})| < C, \\ H_2, & \text{если } \ |\rho(\vec{X},\,\vec{Y})| \geqslant C, \end{cases}$$

где число $C = \tau_{1-\varepsilon/2}$ — квантиль стандартного нормального распределения. При односторонней альтернативе критерий использует одностороннюю критическую область, и квантиль следует брать уровня $1-\varepsilon$.

Сравнение долей признака в нескольких выборках. Если имеется k независимых выборок из распределений Бернулли, в каждой из которых наблюдается своя доля успехов $w_j = m_j/n_j$ (число успехов, делённое на объём выборки), то для проверки гипотезы о совпадении истинных вероятностей успеха $p_1 = \ldots = p_k$ пользуются одним из вариантов критерия хи-квадрат. Пусть p^* — оценка для истинной вероятности успеха в предположении, что основная гипотеза верна:

$$p^* = \frac{m_1 + \ldots + m_k}{n_1 + \ldots + n_k}\,.$$

Статистика критерия выглядит так:

$$\rho = \frac{1}{p^*(1-p^*)} \sum_{j=1}^{k} n_j (w_j - p^*)^2.$$

При верной основной гипотезе распределение статистики ρ с ростом объёмов всех выборок приближается к распределению χ^2_{k-1}. Поэтому критерий с асимптотическим размером ε будет отвергать основную гипотезу, если $\rho > C$, где $\mathsf{P}(\chi^2_{k-1} > C) = \varepsilon$.

Пользоваться данным критерием рекомендуется, если количества успехов и неудач в каждой выборке оказываются не менее десятка, а лучше — нескольких десятков.

Пример 31. Исследован процент юношей в группах ИВТ СибГУТИ 2006 и 2007 года поступления. Проверяется гипотеза о независимости гендерного состава от направления специализации. Получены следующие данные (число юношей/число всех студентов) по потокам специализации: П — 184/233, ВМ — 102/109, ММ — 44/102. Заметим сразу, что при таком числе девушек на потоке ВМ, вообще говоря, критерий использовать нельзя.

Проверяем гипотезу о равенстве трёх вероятностей $p_1 = p_2 = p_3$. Вычислим оценку для истинной доли:

$$p^* = \frac{184 + 102 + 44}{233 + 109 + 102} = \frac{330}{444}\,.$$

Вычислим значение статистики критерия:

$$\rho = \frac{1}{p^*(1-p^*)}\left(233\Big(\frac{184}{233} - p^*\Big)^2 + 109\Big(\frac{102}{109} - p^*\Big)^2 + 102\Big(\frac{44}{102} - p^*\Big)^2\right) = 75{,}8.$$

Квантиль любого разумного уровня для распределения χ^2_2 (то же, что показательное распределение с параметром $0{,}5$) гораздо меньше, чем полученное значение статистики критерия. Действительно, реально достигнутый уровень значимости есть примерно

$$\varepsilon^* = \mathsf{P}(\chi^2_2 > 75{,}8) = e^{-75{,}8/2}.$$

Это настолько мало, что гипотеза об однородности (о совпадении долей юношей для трёх потоков) отвергается категорически.

§ 6. Критерий χ^2 для проверки независимости

Есть выборка (X_1, Y_1), …, (X_n, Y_n) значений двух наблюдаемых совместно случайных величин X и Y в n независимых экспериментах. Проверяется гипотеза $H_1 = \{X \text{ и } Y \text{ независимы}\}$.

Введём k интервалов группировки Δ_1, …, Δ_k для значений X и m интервалов группировки ∇_1, …, ∇_m для значений Y:

$\vec{X}$ \ $\vec{Y}$	∇_1	∇_2	…	∇_m	$\sum_{j=1}^{m}$
Δ_1	ν_{11}	ν_{12}	…	ν_{1m}	$\nu_{1\cdot}$
⋮			…		
Δ_k	ν_{k1}	ν_{k2}	…	ν_{km}	$\nu_{k\cdot}$
$\sum_{i=1}^{k}$	$\nu_{\cdot 1}$	$\nu_{\cdot 2}$	…	$\nu_{\cdot m}$	n

Посчитаем эмпирические частоты:

$$\nu_{ij} = \text{число пар } (X_l,\, Y_l), \text{ попавших в } \Delta_i \times \nabla_j,$$

$$\nu_{\cdot j} = \text{число } Y_l, \text{ попавших в } \nabla_j, \quad \nu_{i\cdot} = \text{число } X_l, \text{ попавших в } \Delta_i.$$

Если гипотеза H_1 верна, то теоретические вероятности попадания пары $(X,\, Y)$ в любую из областей $\Delta_i \times \nabla_j$ равны произведению вероятностей: для всех i и j

$$p_{ij} = \mathsf{P}\big((X,\, Y) \in \Delta_i \times \nabla_j\big) = \mathsf{P}\big(X \in \Delta_i\big) \cdot \mathsf{P}\big(Y \in \nabla_j\big) = p^x_i \cdot p^y_j$$

По ЗБЧ при $n \to \infty$

$$\frac{\nu_{i\cdot}}{n} \xrightarrow{\ \mathrm{p}\ } p^x_i, \qquad \frac{\nu_{\cdot j}}{n} \xrightarrow{\ \mathrm{p}\ } p^y_i, \qquad \frac{\nu_{ij}}{n} \xrightarrow{\ \mathrm{p}\ } p_{ij}.$$

Поэтому большая разница между $\frac{\nu_{ij}}{n}$ и $\frac{\nu_{i\cdot}}{n} \times \frac{\nu_{\cdot j}}{n}$ (или между ν_{ij} и $\frac{\nu_{i\cdot}\,\nu_{\cdot j}}{n}$) служит основанием для отклонения гипотезы независимости. Пусть

$$\rho(\vec{X},\,\vec{Y}) = n \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\bigl(\nu_{ij} - (\nu_{i\cdot}\nu_{\cdot j})/n\bigr)^2}{\nu_{i\cdot}\nu_{\cdot j}}\,. \tag{18}$$

Т е о р е м а 19. *Если гипотеза H_1 верна, то при $n \to \infty$ распределение величины $\rho(\vec{X},\,\vec{Y})$ приближается к распределению* $\mathrm{H}_{(k-1)(m-1)}$.

Критерий согласия асимптотического размера ε строится как обычно: по заданному ε найдём C, равное квантили уровня $1-\varepsilon$ распределения $\mathrm{H}_{(k-1)(m-1)}$. Тогда критерий имеет вид

$$\delta(\vec{X},\,\vec{Y}) = \begin{cases} H_1, & \text{если } \ \rho(\vec{X},\,\vec{Y}) < C, \\ H_2, & \text{если } \ \rho(\vec{X},\,\vec{Y}) \geqslant C. \end{cases}$$

Количество интервалов группировки следует выбирать таким, чтобы в каждую ячейку попадало минимум 5–6 элементов выборки.

Мы рассмотрели некоторые типичные задачи проверки гипотез. Разумеется, полностью охватить все возможные виды задач нельзя. Например, мы не рассматривали критерии, проверяющие качества самой выборки: независимость и/или одинаковую распределённость элементов выборки друг от друга, мы ничего не сказали о том, как можно определять силу зависимости двух выборок друг от друга и т. п. Критерии для решения этих и многих других проблем читатель сможет найти самостоятельно.

§ 7. Вопросы и упражнения

1. Построить критерий для проверки равенства дисперсий двух независимых нормальных выборок с *известными* средними, статистика которого имеет при верной основной гипотезе распределение Фишера с n и m степенями свободы.

2. Построить критерий для проверки гипотезы о равенстве средних двух независимых нормальных выборок с произвольными *известными* дисперсиями, статистика которого имеет при верной основной гипотезе стандартное нормальное распределение.

3. Построить критерий точного размера ε для различения трёх гипотез о среднем нормального распределения с неизвестной дисперсией: $H_1 = \{a = a_0\},\ H_2 = \{a < a_0\}$ и $H_3 = \{a > a_0\}$.

4. Какие из приведённых в главе VII критериев можно сформулировать, используя доверительные интервалы? Сделать это.

5. Проверяется простая гипотеза о параметре $H_1 = \{\theta = \theta_0\}$ против альтернативы $H_2 = \{\theta \neq \theta_0\}$. Какими свойствами должен обладать доверительный интервал, чтобы критерий, построенный с его помощью, был состоятелен?

6. Имеется выборка из распределения Бернулли. Построить критерий для проверки гипотезы $p = 1/2$ при альтернативе $p \neq 1/2$.

7. Подбросить игральную кость 300 раз и проверить её правильность с помощью подходящего критерия.

8. Подбросить симметричную монету 200 раз и проверить своё умение правильно её подбрасывать с помощью критерия χ^2.

9. Построить критерий асимптотического размера ε для проверки гипотезы однородности двух независимых выборок с разными объёмами из распределения Бернулли.

10. Показать, что при $k = 2$ критерий для решения задачи однофакторного дисперсионного анализа совпадает с критерием Стьюдента.

11. Доказать основное дисперсионное соотношение.

ГЛАВА VIII
ИССЛЕДОВАНИЕ СТАТИСТИЧЕСКОЙ ЗАВИСИМОСТИ

Часто требуется определить, как зависит наблюдаемая случайная величина от одной или нескольких других величин. Самый общий случай такой зависимости — зависимость статистическая: например, $X = \xi + \eta$ и $Z = \xi + \varphi$ зависимы, но эта зависимость не функциональная. Для зависимых случайных величин имеет смысл рассмотреть математическое ожидание одной из них при фиксированном значении другой и выяснить, как влияет на среднее значение первой величины изменение значений второй. Так, стоимость квартиры зависит от площади, этажа, района и других параметров, но не является функцией от них. Зато можно считать её среднее функцией от этих величин. Разумеется, наблюдать это среднее значение мы не можем — в нашей власти лишь наблюдать значения результирующей случайной величины при разных значениях остальных. Эту зависимость можно воображать как вход и выход некоторой машины — «ящика с шуршавчиком». Входные данные (*факторы*) известны. На выходе мы наблюдаем результат преобразования входных данных в ящике по каким-либо правилам.

§ 1. Математическая модель регрессии

Пусть наблюдаемая случайная величина X зависит от случайной величины или случайного вектора Z. Значения Z мы либо задаём, либо наблюдаем. Обозначим через $f(t)$ функцию, отражающую зависимость среднего значения X от значений Z:

$$\mathsf{E}(X \mid Z = t) = f(t). \tag{19}$$

Функция $f(t)$ называется *линией регрессии X на Z*, а уравнение $x = f(t)$ — уравнением регрессии. После n экспериментов, в которых Z последовательно принимает значения $Z = t_1, \ldots, Z = t_n$, получим значения наблюдаемой величины X, равные $X_1, \ldots, X_n$. Обозначим через ε_i разницу $X_i - \mathsf{E}(X \mid Z = t_i) = X_i - f(t_i)$ между наблюдаемой в i-м эксперименте случайной величиной и её математическим ожиданием.

Итак, $X_i = f(t_i) + \varepsilon_i$, $i = 1, \ldots, n$, где ε_i — ошибки наблюдения, равные в точности разнице между реальным и усредненным значением случайной величины X при значении $Z = t_i$. Про совместное распределение $\varepsilon_1, \ldots, \varepsilon_n$

обычно что-либо известно или предполагается: *например*, что вектор ошибок $\vec{\varepsilon}$ состоит из независимых и одинаково *нормально* распределённых случайных величин с нулевым средним.

Требуется по значениям $t_1, \ldots, t_n$ и $X_1, \ldots, X_n$ оценить как можно точнее функцию $f(t)$. Величины t_i не являются случайными, вся случайность сосредоточена в неизвестных ошибках ε_i и в наблюдаемых X_i. Но пытаться в классе всех возможных функций восстанавливать $f(t)$ по «наилучшим оценкам» для $f(t_i)$ довольно глупо: наиболее точными приближениями к $f(t_i)$ оказываются X_i, и функция $f(t)$ будет просто ломаной, построенной по точкам (t_i, X_i). Поэтому сначала определяют вид функции $f(t)$. Часто в качестве $f(t)$ берут полином небольшой степени с неизвестными коэффициентами.

Будем пока предполагать, что функция $f(t)$ полностью определяется неизвестными параметрами $\theta_1, \ldots, \theta_k$.

§ 2. Метод максимального правдоподобия.

Оценки неизвестных параметров находят с помощью метода максимального правдоподобия. Он предписывает выбирать неизвестные параметры так, чтобы максимизировать функцию правдоподобия случайного вектора $X_1, \ldots, X_n$.

Будем, для простоты, предполагать, что вектор ошибок $\vec{\varepsilon}$ состоит из независимых и одинаково распределённых случайных величин с плотностью распределения $h(x)$ из некоторого семейства распределений с нулевым средним и, вообще говоря, неизвестной дисперсией. Обычно полагают, что ε_i имеют симметричное распределение — нормальное $\mathrm{N}_{0,\,\sigma^2}$, Стьюдента, Лапласа и т. п. Поскольку X_i от ε_i зависят линейно, то распределение X_i окажется таким же, как у ε_i, но с центром уже не в нуле, а в точке $f(t_i)$.

Поэтому X_i имеет плотность $h\big(x - f(t_i)\big)$. Функция правдоподобия вектора $X_1, \ldots, X_n$ в силу независимости координат равна

$$f(\vec{X};\, \theta_1, \ldots, \theta_k) = \prod_{i=1}^{n} h\big(X_i - f(t_i)\big) = h(\varepsilon_1) \cdot \ldots \cdot h(\varepsilon_n). \qquad (20)$$

Если величины ε_i имеют разные распределения, то h следует заменить на соответствующие h_i. Для зависимых ε_i произведение плотностей в формуле (20) заменится плотностью их совместного распределения.

Метод максимального правдоподобия предписывает находить оценки неизвестных параметров θ_i функции $f(t)$ и оценки дисперсии $\sigma^2 = \mathsf{D}\varepsilon_i$, максимизируя по этим параметрам функцию правдоподобия (20).

§ 3. Метод наименьших квадратов.

Рассмотрим, во что превращается метод максимального правдоподобия в наиболее частых на практике предположениях.

Предположим, что вектор ошибок $\vec{\varepsilon}$ состоит из независимых случайных величин с нормальным распределением $\mathrm{N}_{0,\,\sigma^2}$. Функция правдоподобия (20) имеет вид

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

Очевидно, что при любом фиксированном σ^2 максимум функции правдоподобия достигается при наименьшем значении суммы квадратов ошибок

$$\sum (X_i - f(t_i))^2 = \sum \varepsilon_i^2.$$

О п р е д е л е н и е 22. *Оценкой метода наименьших квадратов* (ОМНК) для неизвестных параметров $\theta_1, \ldots, \theta_k$ уравнения регрессии называется набор значений параметров, доставляющий минимум сумме квадратов отклонений

$$\sum_{i=1}^{n}(X_i - f(t_i))^2 = \sum_{i=1}^{n} \varepsilon_i^2.$$

Найдя оценки для θ_i, найдём тем самым оценку $\hat{f}(t)$ для $f(t)$. Обозначим через $\hat{f}(t_i)$ значения этой функции, и через $\hat{\varepsilon}_i = X_i - \hat{f}(t_i)$ соответствующие оценки ошибок. Оценка максимального правдоподобия для σ^2, она же точка максимума по σ^2 функции правдоподобия, равна

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{f}(t_i))^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2. \tag{21}$$

Найдём ОМНК для функций $f(t)$ в ряде частных случаев.

П р и м е р 32. Пусть функция $f(t) = \theta$ — постоянная, θ — неизвестный параметр. Тогда наблюдения равны $X_i = \theta + \varepsilon_i$, $i = 1, \ldots, n$. Легко узнать задачу оценивания неизвестного математического ожидания θ по выборке из независимых и одинаково распределённых случайных величин $X_1, \ldots, X_n$.

Найдём ОМНК $\hat{\theta}$ для параметра θ:

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n}(X_i-\theta)^2 = -2\sum_{i=1}^{n}(X_i-\theta)\Big|_{\theta=\hat{\theta}} = 0 \quad \text{при} \quad \hat{\theta}=\overline{X}.$$

Трудно назвать этот ответ неожиданным. Соответственно, $\hat{\sigma}^2 = S^2$.

Пример 33 (линейная регрессия). Рассмотрим линейную регрессию $X_i = \theta_1 + t_i\theta_2 + \varepsilon_i$, $i = 1,\dots,n$, где θ_1 и θ_2 — неизвестные параметры. Здесь $f(t) = \theta_1 + t\theta_2$ — прямая.

Найдём оценку метода наименьших квадратов $\hat{\theta}_1$, $\hat{\theta}_2$, на которой достигается минимум величины $\sum\varepsilon_i^2 = \sum(X_i - \theta_1 - t_i\theta_2)^2$. Приравняв к нулю частные производные этой суммы по параметрам, найдём точку экстремума.

Упражнение. Убедиться, что решением системы уравнений

$$\frac{\partial}{\partial\theta_1}\sum_{i=1}^{n}\varepsilon_i^2 = 0\,, \qquad \frac{\partial}{\partial\theta_2}\sum_{i=1}^{n}\varepsilon_i^2 = 0$$

является пара

$$\hat{\theta}_2 = \frac{\frac{1}{n}\sum X_i t_i - \overline{X}\cdot\bar{t}}{\frac{1}{n}\sum(t_i-\bar{t})^2}\,, \qquad \hat{\theta}_1 = \overline{X} - \bar{t}\,\hat{\theta}_2.$$

Определение 23. *Выборочным коэффициентом корреляции* называется величина

$$\rho^* = \frac{\frac{1}{n}\sum X_i t_i - \overline{X}\cdot\bar{t}}{\sqrt{\frac{1}{n}\sum(t_i-\bar{t})^2\cdot\frac{1}{n}\sum(X_i-\overline{X})^2}},$$

которая характеризует степень линейной зависимости между наборами чисел $X_1,\dots,X_n$ и $t_1,\dots,t_n$.

Выборочный коэффициент корреляции можно использовать для проверки основной гипотезы H_1, состоящей в отсутствии между случайными величинами линейной корреляционной зависимости (коэффициент корреляции равен нулю). Это нежелательное предположение в регрессионном анализе. Напротив, альтернативой является желательное предположение о наличии корреляционной зависимости.

Если набор данных $(X_1,t_1),\dots,(X_n,t_n)$ есть выборка из двумерного нормального распределения, то для проверки гипотезы об их некоррелированности (отсутствии линейной зависимости) используют статистику

$$t = \frac{\rho^*\sqrt{n-2}}{1-\rho^{*2}}\,.$$

Гипотеза о некоррелированности отвергается, если $|t| > C$, где C есть квантиль уровня $1 - \varepsilon/2$ для распределения Стьюдента T_{n-2}.

П р и м е р 34. Термин «регрессия» ввёл Гальтон (*Francis Galton.* Regression towards mediocrity in hereditary stature // Journal of the Anthropological Institute. — 1886. — v. 15. — p. 246—265).

Гальтон исследовал, в частности, рост детей высоких родителей и установил, что он «регрессирует» в среднем, т. е. в среднем дети высоких родителей не так высоки, как их родители. Пусть X — рост сына, а Z_1 и Z_2 — рост отца и матери. Для линейной модели регрессии

$$\mathsf{E}(X \mid Z_1 = t, Z_2 = u) = f(t, u) = \theta_1 t + \theta_2 u + c$$

Гальтон нашел оценки параметров

$$\mathsf{E}(\text{роста сына} \mid Z_1 = t, Z_2 = u) = 0,27t + 0,2u + \text{const},$$

а средний рост дочери ещё в 1,08 раз меньше. Независимо от добавочной постоянной суммарный вклад высокого роста родителей в рост детей не превышает половины. Остальное — неизменная добавка.

Дальнейшее изучение регрессионных моделей ждёт читателя в курсах эконометрики и многомерного статистического анализа.

ПРИЛОЖЕНИЕ

Таблица 1

Основные дискретные распределения

Название, обозначение, параметры	**Возможные значения k**	$\mathsf{P}(\xi = k)$	$\mathsf{E}\xi$	$\mathsf{D}\xi$
Вырожденное I_c, $c \in \mathbb{R}$	c	$\mathsf{P}(\xi = c) = 1$	c	0
Бернулли B_p $p \in (0,\, 1)$	$k = 0,\, 1$	$\mathsf{P}(\xi = 0) = 1-p$, $\mathsf{P}(\xi = 1) = p$	p	$p(1-p)$
Биномиальное $\mathrm{B}_{n,\,p}$ $p \in (0,\, 1)$ $n = 1, 2, \ldots$	$k = 0, \ldots, n$	$C_n^k p^k (1-p)^{n-k}$	np	$np(1-p)$
Пуассона Π_λ $\lambda > 0$	$k = 0,\, 1,\, 2, \ldots$	$\dfrac{\lambda^k}{k!}\, e^{-\lambda}$	λ	λ
Геометрическое G_p $p \in (0,\, 1)$	$k = 1,\, 2, \ldots$	$p(1-p)^{k-1}$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Гипергеометрическое $n, K, N \in \mathbb{N}$ $0 \leqslant n, K \leqslant N$	целые от $\max(0,\, n+K-N)$ до $\min(n,\, K)$	$\dfrac{C_K^k C_{N_K}^{n-k}}{C_N^n}$	$n\dfrac{K}{N}$	$n\dfrac{K}{N}\left(1-\dfrac{K}{N}\right)\dfrac{N-n}{N-1}$

Таблица 2

Основные абсолютно непрерывные распределения

Название, обозначение, параметры	**Плотность распределения**	$\mathsf{E}\xi$	$\mathsf{D}\xi$	**Асимметрия**	**Эксцесс**
Равномерное на отрезке $[a,\,b]$ $\mathrm{U}_{a,\,b}$, $a<b$	$\begin{cases}\dfrac{1}{b-a}, & x\in[a,\,b],\\ 0, & x\notin[a,\,b]\end{cases}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$	0	$-1{,}2$
Показательное (экспоненциальное) $\mathrm{E}_\alpha=\Gamma_{\alpha,\,1}$, $\alpha>0$	$\begin{cases}\alpha\,e^{-\alpha x}, & x>0,\\ 0, & x\leqslant 0\end{cases}$	$\dfrac{1}{\alpha}$	$\dfrac{1}{\alpha^2}$	2	6
Нормальное (гауссовское) $\mathrm{N}_{a,\,\sigma^2}$, $a\in\mathbb{R}$, $\sigma>0$	$\dfrac{1}{\sigma\sqrt{2\pi}}\,e^{-(x-a)^2/2\sigma^2}$, $-\infty<x<\infty$	a	σ^2	0	0
Коши $\mathrm{C}_{a,\,\sigma}$, $a\in\mathbb{R}$, $\sigma>0$	$\dfrac{1}{\pi}\,\dfrac{\sigma}{\sigma^2+(x-a)^2}$, $-\infty<x<\infty$	—	—	—	—
Гамма $\Gamma_{\alpha,\,\lambda}$, $\alpha>0,\ \lambda>0$	$\begin{cases}\dfrac{\alpha^\lambda}{\Gamma(\lambda)}\,x^{\lambda-1}e^{-\alpha x}, & x>0,\\ 0, & x\leqslant 0\end{cases}$	$\dfrac{\lambda}{\alpha}$	$\dfrac{\lambda}{\alpha^2}$	$\dfrac{2}{\sqrt{\lambda}}$	$\dfrac{6}{\lambda}$
Лапласа $\mathrm{L}_{\alpha,\,\mu}$, $\alpha>0,\ \mu\in\mathbb{R}$	$\dfrac{\alpha}{2}\,e^{-\alpha\lvert x-\mu\rvert}$, $-\infty<x<\infty$	μ	$\dfrac{2}{\alpha^2}$	0	3
Парето, $\alpha>0$	$\begin{cases}\dfrac{\alpha}{x^{\alpha+1}}, & x\geqslant 1,\\ 0, & x<1\end{cases}$	$\dfrac{\alpha}{\alpha-1}$ $(\alpha>1)$	$\dfrac{\alpha}{(\alpha-1)^2(\alpha-2)}$, $\alpha>2$	$\sqrt{\dfrac{\alpha-2}{\alpha}}\,\dfrac{2(\alpha+1)}{\alpha-3}$, $\alpha>3$	$\dfrac{6(\alpha^3+\alpha^2-6\alpha-2)}{\alpha(\alpha-3)(\alpha-4)}$, $\alpha>4$

Таблица 3

Критические точки распределения χ^2

Приведены значения x, при которых $\mathsf{P}(\chi^2_k > x) = \alpha$

k \ α	0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
3	12,84	11,34	9,35	7,81	0,35	0,22	0,12	0,07
4	14,86	13,28	11,14	9,49	0,71	0,48	0,30	0,21
5	16,75	15,09	12,83	11,07	1,15	0,83	0,55	0,41
6	18,55	16,81	14,45	12,59	1,64	1,24	0,87	0,68
7	20,28	18,48	16,01	14,07	2,17	1,69	1,24	0,99
8	21,95	20,09	17,53	15,51	2,73	2,18	1,65	1,34
9	23,59	21,67	19,02	16,92	3,33	2,70	2,09	1,73
10	25,19	23,21	20,48	18,31	3,94	3,25	2,56	2,16
11	26,76	24,73	21,92	19,68	4,57	3,82	3,05	2,60
12	28,30	26,22	23,34	21,03	5,23	4,40	3,57	3,07
13	29,82	27,69	24,74	22,36	5,89	5,01	4,11	3,57
14	31,32	29,14	26,12	23,68	6,57	5,63	4,66	4,07
15	32,80	30,58	27,49	25,00	7,26	6,26	5,23	4,60
16	34,27	32,00	28,85	26,30	7,96	6,91	5,81	5,14
17	35,72	33,41	30,19	27,59	8,67	7,56	6,41	5,70
18	37,16	34,81	31,53	28,87	9,39	8,23	7,01	6,26
19	38,58	36,19	32,85	30,14	10,12	8,91	7,63	6,84
20	40,00	37,57	34,17	31,41	10,85	9,59	8,26	7,43
21	41,40	38,93	35,48	32,67	11,59	10,28	8,90	8,03
22	42,80	40,29	36,78	33,92	12,34	10,98	9,54	8,64
23	44,18	41,64	38,08	35,17	13,09	11,69	10,20	9,26
24	45,56	42,98	39,36	36,42	13,85	12,40	10,86	9,89
25	46,93	44,31	40,65	37,65	14,61	13,12	11,52	10,52
26	48,29	45,64	41,92	38,89	15,38	13,84	12,20	11,16
27	49,65	46,96	43,19	40,11	16,15	14,57	12,88	11,81
28	50,99	48,28	44,46	41,34	16,93	15,31	13,56	12,46
29	52,34	49,59	45,72	42,56	17,71	16,05	14,26	13,12
49	78,23	74,92	70,22	66,34	33,93	31,55	28,94	27,25
99	139,0	134,6	128,4	123,2	77,05	73,36	69,23	66,51
499	584,1	575,4	562,8	552,1	448,2	439,0	428,5	421,4
999	1117,9	1105,9	1088,5	1073,6	926,6	913,3	898,0	887,6

Таблица 4

Критические точки распределения Стьюдента

Приведены значения x, при которых $\mathsf{P}(|t_k| > x) = \alpha$

k \ α	0,001	0,002	0,005	0,01	0,02	0,05	0,1	0,2
3	12,92	10,21	7,45	5,84	4,54	3,18	2,35	1,64
4	8,61	7,17	5,60	4,60	3,75	2,78	2,13	1,53
5	6,87	5,89	4,77	4,03	3,36	2,57	2,02	1,48
6	5,96	5,21	4,32	3,71	3,14	2,45	1,94	1,44
7	5,41	4,79	4,03	3,50	3,00	2,36	1,89	1,41
8	5,04	4,50	3,83	3,36	2,90	2,31	1,86	1,40
9	4,78	4,30	3,69	3,25	2,82	2,26	1,83	1,38
10	4,59	4,14	3,58	3,17	2,76	2,23	1,81	1,37
11	4,44	4,02	3,50	3,11	2,72	2,20	1,80	1,36
12	4,32	3,93	3,43	3,05	2,68	2,18	1,78	1,36
13	4,22	3,85	3,37	3,01	2,65	2,16	1,77	1,35
14	4,14	3,79	3,33	2,98	2,62	2,14	1,76	1,35
15	4,07	3,73	3,29	2,95	2,60	2,13	1,75	1,34
16	4,01	3,69	3,25	2,92	2,58	2,12	1,75	1,34
17	3,97	3,65	3,22	2,90	2,57	2,11	1,74	1,33
18	3,92	3,61	3,20	2,88	2,55	2,10	1,73	1,33
19	3,88	3,58	3,17	2,86	2,54	2,09	1,73	1,33
20	3,85	3,55	3,15	2,85	2,53	2,09	1,72	1,33
21	3,82	3,53	3,14	2,83	2,52	2,08	1,72	1,32
22	3,79	3,50	3,12	2,82	2,51	2,07	1,72	1,32
23	3,77	3,48	3,10	2,81	2,50	2,07	1,71	1,32
24	3,75	3,47	3,09	2,80	2,49	2,06	1,71	1,32
25	3,73	3,45	3,08	2,79	2,49	2,06	1,71	1,32
26	3,71	3,43	3,07	2,78	2,48	2,06	1,71	1,31
27	3,69	3,42	3,06	2,77	2,47	2,05	1,70	1,31
28	3,67	3,41	3,05	2,76	2,47	2,05	1,70	1,31
29	3,66	3,40	3,04	2,76	2,46	2,05	1,70	1,31
49	3,50	3,27	2,94	2,68	2,40	2,01	1,68	1,30
99	3,39	3,17	2,87	2,63	2,36	1,98	1,66	1,29
∞	3,29	3,09	2,81	2,58	2,33	1,96	1,64	1,28

Таблица 5

Критические точки распределения Фишера

Приведены значения x, при которых $\mathsf{P}(f_{k_1,k_2} > x) = 0{,}05$

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10
1	161	199	216	225	230	234	237	239	241	242
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16

Таблица 6

Функция распределения Колмогорова

В таблице приведены значения функции

$$K(y) \;=\; \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2y^2}, \quad y > 0.$$

y	0	1	2	3	4	5	6	7	8	9
0,3	,0000	,0000	,0000	,0001	,0002	,0003	,0005	,0008	,0013	,0019
0,4	,0028	,0040	,0055	,0074	,0097	,0126	,0160	,0200	,0247	,0300
0,5	,0361	,0428	,0503	,0585	,0675	,0772	,0876	,0987	,1104	,1228
0,6	,1357	,1492	,1632	,1778	,1927	,2080	,2236	,2396	,2558	,2722
0,7	,2888	,3055	,3223	,3391	,3560	,3728	,3896	,4064	,4230	,4395
0,8	,4559	,4720	,4880	,5038	,5194	,5347	,5497	,5645	,5791	,5933
0,9	,6073	,6209	,6343	,6473	,6601	,6725	,6846	,6964	,7079	,7191
1,0	,7300	,7406	,7508	,7608	,7704	,7798	,7889	,7976	,8061	,8143
1,1	,8223	,8300	,8374	,8445	,8514	,8580	,8644	,8706	,8765	,8823
1,2	,8878	,8930	,8981	,9030	,9076	,9121	,9164	,9206	,9245	,9283
1,3	,9319	,9354	,9387	,9418	,9449	,9478	,9505	,9531	,9557	,9580
1,4	,9603	,9625	,9646	,9665	,9684	,9702	,9718	,9734	,9750	,9764
1,5	,9778	,9791	,9803	,9815	,9826	,9836	,9846	,9855	,9864	,9873
1,6	,9880	,9888	,9895	,9902	,9908	,9914	,9919	,9924	,9929	,9934
1,7	,9938	,9942	,9946	,9950	,9953	,9956	,9959	,9962	,9965	,9967
1,8	,9969	,9971	,9973	,9975	,9977	,9979	,9980	,9981	,9983	,9984
1,9	,9985	,9986	,9987	,9988	,9989	,9990	,9991	,9991	,9992	,9992
2,0	,9993	,9994	,9994	,9995	,9995	,9996	,9996	,9996	,9997	,9997
2,1	,9997	,9997	,9998	,9998	,9998	,9998	,9998	,9998	,9999	,9999

$$K(2,2) = 0,999874; \quad K(2,25) = 0,999920;$$
$$K(2,3) = 0,999949; \quad K(2,35) = 0,999968;$$
$$K(2,4) = 0,999980; \quad K(2,45) = 0,999988;$$
$$K(2,49) = 0,999992$$

Таблица 7

Функция распределения стандартного нормального закона

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
-5	$0,0000003$	$-2,48$	$0,0066$	$-1,91$	$0,0281$	$-1,58$	$0,0571$
$-4,5$	$0,0000034$	$-2,46$	$0,0069$	$-1,9$	$0,0287$	$-1,57$	$0,0582$
-4	$0,0000317$	$-2,44$	$0,0073$	$-1,89$	$0,0294$	$-1,56$	$0,0594$
$-3,8$	$0,0000724$	$-2,42$	$0,0078$	$-1,88$	$0,0301$	$-1,55$	$0,0606$
$-3,6$	$0,0001591$	$-2,4$	$0,0082$	$-1,87$	$0,0307$	$-1,54$	$0,0618$
$-3,4$	$0,0003370$	$-2,38$	$0,0087$	$-1,86$	$0,0314$	$-1,53$	$0,0630$
$-3,2$	$0,0006872$	$-2,36$	$0,0091$	$-1,85$	$0,0322$	$-1,52$	$0,0643$
-3	$0,00135$	$-2,34$	$0,0096$	$-1,84$	$0,0329$	$-1,51$	$0,0655$
$-2,98$	$0,00144$	$-2,32$	$0,0102$	$-1,83$	$0,0336$	$-1,5$	$0,0668$
$-2,96$	$0,00154$	$-2,3$	$0,0107$	$-1,82$	$0,0344$	$-1,49$	$0,0681$
$-2,94$	$0,00164$	$-2,28$	$0,0113$	$-1,81$	$0,0351$	$-1,48$	$0,0694$
$-2,92$	$0,00175$	$-2,26$	$0,0119$	$-1,8$	$0,0359$	$-1,47$	$0,0708$
$-2,9$	$0,00187$	$-2,24$	$0,0125$	$-1,79$	$0,0367$	$-1,46$	$0,0721$
$-2,88$	$0,00199$	$-2,22$	$0,0132$	$-1,78$	$0,0375$	$-1,45$	$0,0735$
$-2,86$	$0,00212$	$-2,2$	$0,0139$	$-1,77$	$0,0384$	$-1,44$	$0,0749$
$-2,84$	$0,00226$	$-2,18$	$0,0146$	$-1,76$	$0,0392$	$-1,43$	$0,0764$
$-2,82$	$0,00240$	$-2,16$	$0,0154$	$-1,75$	$0,0401$	$-1,42$	$0,0778$
$-2,8$	$0,00256$	$-2,14$	$0,0162$	$-1,74$	$0,0409$	$-1,41$	$0,0793$
$-2,78$	$0,00272$	$-2,12$	$0,0170$	$-1,73$	$0,0418$	$-1,4$	$0,0808$
$-2,76$	$0,00289$	$-2,1$	$0,0179$	$-1,72$	$0,0427$	$-1,39$	$0,0823$
$-2,74$	$0,00307$	$-2,08$	$0,0188$	$-1,71$	$0,0436$	$-1,38$	$0,0838$
$-2,72$	$0,00326$	$-2,06$	$0,0197$	$-1,7$	$0,0446$	$-1,37$	$0,0853$
$-2,7$	$0,00347$	$-2,04$	$0,0207$	$-1,69$	$0,0455$	$-1,36$	$0,0869$
$-2,68$	$0,00368$	$-2,02$	$0,0217$	$-1,68$	$0,0465$	$-1,35$	$0,0885$
$-2,66$	$0,00390$	-2	$0,0228$	$-1,67$	$0,0475$	$-1,34$	$0,0901$
$-2,64$	$0,00415$	$-1,99$	$0,0233$	$-1,66$	$0,0485$	$-1,33$	$0,0918$
$-2,62$	$0,00440$	$-1,98$	$0,0239$	$-1,65$	$0,0495$	$-1,32$	$0,0934$
$-2,6$	$0,00466$	$-1,97$	$0,0244$	$-1,64$	$0,0505$	$-1,31$	$0,0951$
$-2,58$	$0,00494$	$-1,96$	$0,0250$	$-1,63$	$0,0516$	$-1,3$	$0,0968$
$-2,56$	$0,00523$	$-1,95$	$0,0256$	$-1,62$	$0,0526$	$-1,29$	$0,0985$
$-2,54$	$0,00554$	$-1,94$	$0,0262$	$-1,61$	$0,0537$	$-1,28$	$0,1003$
$-2,52$	$0,00590$	$-1,93$	$0,0268$	$-1,6$	$0,0548$	$-1,27$	$0,1020$
$-2,5$	$0,00621$	$-1,92$	$0,0274$	$-1,59$	$0,0559$	$-1,26$	$0,1038$

Окончание табл. 7

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
$-1,25$	$0,1056$	$-0,93$	$0,1762$	$-0,61$	$0,2709$	$-0,29$	$0,3859$
$-1,24$	$0,1075$	$-0,92$	$0,1788$	$-0,6$	$0,2743$	$-0,28$	$0,3897$
$-1,23$	$0,1093$	$-0,91$	$0,1814$	$-0,59$	$0,2776$	$-0,27$	$0,3936$
$-1,22$	$0,1112$	$-0,9$	$0,1841$	$-0,58$	$0,2810$	$-0,26$	$0,3974$
$-1,21$	$0,1131$	$-0,89$	$0,1867$	$-0,57$	$0,2843$	$-0,25$	$0,4013$
$-1,2$	$0,1151$	$-0,88$	$0,1894$	$-0,56$	$0,2877$	$-0,24$	$0,4052$
$-1,19$	$0,1170$	$-0,87$	$0,1922$	$-0,55$	$0,2912$	$-0,23$	$0,4090$
$-1,18$	$0,1190$	$-0,86$	$0,1949$	$-0,54$	$0,2946$	$-0,22$	$0,4129$
$-1,17$	$0,1210$	$-0,85$	$0,1977$	$-0,53$	$0,2981$	$-0,21$	$0,4168$
$-1,16$	$0,1230$	$-0,84$	$0,2005$	$-0,52$	$0,3015$	$-0,2$	$0,4207$
$-1,15$	$0,1251$	$-0,83$	$0,2033$	$-0,51$	$0,3050$	$-0,19$	$0,4247$
$-1,14$	$0,1271$	$-0,82$	$0,2061$	$-0,5$	$0,3085$	$-0,18$	$0,4286$
$-1,13$	$0,1292$	$-0,81$	$0,2090$	$-0,49$	$0,3121$	$-0,17$	$0,4325$
$-1,12$	$0,1314$	$-0,8$	$0,2119$	$-0,48$	$0,3156$	$-0,16$	$0,4364$
$-1,11$	$0,1335$	$-0,79$	$0,2148$	$-0,47$	$0,3192$	$-0,15$	$0,4404$
$-1,1$	$0,1357$	$-0,78$	$0,2177$	$-0,46$	$0,3228$	$-0,14$	$0,4443$
$-1,09$	$0,1379$	$-0,77$	$0,2206$	$-0,45$	$0,3264$	$-0,13$	$0,4483$
$-1,08$	$0,1401$	$-0,76$	$0,2236$	$-0,44$	$0,3300$	$-0,12$	$0,4522$
$-1,07$	$0,1423$	$-0,75$	$0,2266$	$-0,43$	$0,3336$	$-0,11$	$0,4562$
$-1,06$	$0,1446$	$-0,74$	$0,2296$	$-0,42$	$0,3372$	$-0,1$	$0,4602$
$-1,05$	$0,1469$	$-0,73$	$0,2327$	$-0,41$	$0,3409$	$-0,09$	$0,4641$
$-1,04$	$0,1492$	$-0,72$	$0,2358$	$-0,4$	$0,3446$	$-0,08$	$0,4681$
$-1,03$	$0,1515$	$-0,71$	$0,2389$	$-0,39$	$0,3483$	$-0,07$	$0,4721$
$-1,02$	$0,1539$	$-0,7$	$0,2420$	$-0,38$	$0,3520$	$-0,06$	$0,4761$
$-1,01$	$0,1562$	$-0,69$	$0,2451$	$-0,37$	$0,3557$	$-0,05$	$0,4801$
-1	$0,1587$	$-0,68$	$0,2483$	$-0,36$	$0,3594$	$-0,04$	$0,4840$
$-0,99$	$0,1611$	$-0,67$	$0,2514$	$-0,35$	$0,3632$	$-0,03$	$0,4880$
$-0,98$	$0,1635$	$-0,66$	$0,2546$	$-0,34$	$0,3669$	$-0,02$	$0,4920$
$-0,97$	$0,1660$	$-0,65$	$0,2578$	$-0,33$	$0,3707$	$-0,01$	$0,4960$
$-0,96$	$0,1685$	$-0,64$	$0,2611$	$-0,32$	$0,3745$	0	$0,5000$
$-0,95$	$0,1711$	$-0,63$	$0,2643$	$-0,31$	$0,3783$		
$-0,94$	$0,1736$	$-0,62$	$0,2676$	$-0,3$	$0,3821$		

При $x > 0$ значения $\Phi(x)$ находят по такому правилу:

$$\Phi(x) = 1 - \Phi(-x).$$

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

СПИСОК ЛИТЕРАТУРЫ

1. *Бочаров П. П., Печинкин А. В.* Теория вероятностей. Математическая статистика. М.: Гардарика, 1998, 328 с.
2. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М.: Наука, 1965.
3. *Ивченко Г. И., Медведев Ю. И.* Математическая статистика. М.: Высш. шк., 1984, 248 с.
4. *Колемаев В. А., Калинина В. Н.* Теория вероятностей и математическая статистика. М.: ИНФРА-М, 1997, 302 с.
5. *Пугачев В. С.* Теория вероятностей и математическая статистика. М.: ФИЗМАТЛИТ, 2002, 496 с.
6. *Чистяков В. П.* Курс теории вероятностей. М.: Агар, 2000, 255 с.
7. *Гмурман В. Е.* Руководство к решению задач по теории вероятностей и математической статистике. М.: Высшее образование, 2006, 404 с.
8. Сборник задач по теории вероятностей, математической статистике и теории случайных функций / Под редакцией А. А. Свешникова. М.: Наука, 1970, 656 с.

Наталья Исааковна Чернова

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие

Редактор: О. А. Игнатова
Корректор:

Подписано в печать
Формат бумаги $62 \times 84/16$, отпечатано на ризографе, шрифт №10,
изд. л. , зак. № , тир. – экз., СибГУТИ
630102, Новосибирск, ул. Кирова, 86.