МИНОБРНАУКИ РОССИИ

Санкт-Петербургский государственный электротехнический университет „ЛЭТИ“

А. И. ДАУГАВЕТ Е. В. ПОСТНИКОВ А. А. СОЛЫНИН

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие

Санкт-Петербург
Издательство СПбГЭТУ „ЛЭТИ“
2012

УДК 519.22(075)
ББК В172я7
Д21

Д21 Даугавет А. И., Постников Е. В., Солынин А. А. Математическая статистика : Учеб. пособие. СПб.: Изд-во СПбГЭТУ „ЛЭТИ“, 2012. 92 с.

ISBN 978–5–7629–1241–9

Соответствует второй части рабочей программы дисциплины „Теория вероятностей и математическая статистика“ для студентов и бакалавров факультетов электротехники и автоматизации, электроники, экономики и менеджмента и открытого факультета.

Предназначено для студентов всех направлений и специальностей перечисленных выше факультетов.

УДК 519.22(075)
ББК В172я7

Рецензенты: кафедра высшей математики СПбГПУ; д-р физ.-мат. наук, проф. Я. И. Белопольская (СПбГАСУ).

Утверждено
редакционно-издательским советом университета
в качестве учебного пособия

ISBN 978–5–7629–1241–9

Введение

Математическая статистика относится к прикладным математическим дисциплинам. Ее основное назначение – изучение методов обработки экспериментальных данных с целью построения вероятностной модели случайного эксперимента, адекватной реальному эксперименту.

Математическая статистика, являясь разделом математики, оперирует формальными математическими моделями, тогда как случайный эксперимент проводится с реальными физическими объектами. Связь между практикой и математической статистикой можно описать следующей схемой.

Обычно с помощью методов математической статистики исследуется некоторый конечный набор реальных объектов. Все множество таких однородных в некотором смысле объектов называют генеральной совокупностью.

Все объекты генеральной совокупности обладают некоторым случайным признаком, имеющим числовое выражение. Считается, что этот числовой признак описывается случайной величиной. Эта случайная величина и является объектом исследований математической статистики.

В прикладной математической статистике обычно рассматриваются две основные задачи:

1. Указать закон распределения случайной величины или оценить его числовые характеристики.

2. Проверить гипотезу о соответствии указанного закона распределения или значений его числовых характеристик наблюдаемым в эксперименте показателям генеральной совокупности.

Обе эти задачи рассматриваются в настоящем издании.

Для более глубокого изучения математической статистики можно рекомендовать монографии [1] и [2]. Для самостоятельной работы студентов полезно учебное пособие [3].

Для практического применения методов математической статистики могут быть полезны справочники [4] и [5].

Пособие предназначено для студентов технических и экономических специальностей университета, прослушавших соответствующий курс теории вероятностей.

1. ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

1.1. Некоторые сведения из теории вероятностей

Основой математической статистики является математическая теория вероятностей, поэтому приведем некоторые факты из теории вероятностей, необходимые для дальнейшего изложения.

Вероятностное пространство. В соответствии с аксиоматикой теории вероятностей А. Н. Колмогорова вероятностным пространством называют набор из трех объектов: $(\Omega, \mathcal{F}, P)$. Здесь Ω – множество непересекающихся элементарных событий; $\mathcal{F}$ – σ-алгебра случайных событий, т. е. множество подмножеств Ω, содержащее все не более чем счетные объединения и пересечения входящих в него подмножеств; P – вероятностное распределение на $\mathcal{F}$, т. е. аддитивная функция, которая ставит в соответствие каждому событию $A \in \mathcal{F}$ число $P(A) \in [0, 1]$, которое называется вероятностью события A.

Случайная величина. Случайной величиной ξ на вероятностном пространстве $(\Omega, \mathcal{F}, P)$ называется функция $\xi : \Omega \to \mathbb{R}$, такая, что $\forall\, x \in \mathbb{R}$ случайное событие $\{\omega \in \Omega \,|\, \xi(\omega) < x\} \in \mathcal{F}$.

Функция $F_\xi(x) = P(\xi < x)$, определенная для $\forall x \in \mathbb{R}$, называется функцией распределения случайной величины. Функция распределения представляет собой неубывающую непрерывную слева функцию, причем $\lim\limits_{x \to -\infty} F_\xi(x) = 0$ и $\lim\limits_{x \to \infty} F_\xi(x) = 1$.

В зависимости от мощности множества значений случайной величины различают случайные величины с дискретным и абсолютно непрерывным распределениями.

Множество значений дискретной случайной величины конечно или счетно.

Случайная величина имеет абсолютно непрерывное распределение, если множество ее значений несчетно и существует неотрицательная интегрируемая на $\mathbb{R}$ функция f_ξ, называемая плотностью распределения, такая, что $\forall\, x \in \mathbb{R}$ справедливо равенство

$$F_\xi(x) = \int\limits_{-\infty}^{x} f_\xi(t)\, dt.$$

Числовые характеристики случайной величины. Математическим ожиданием случайной величины называется число, представляющее собой ее среднее значение. Это среднее понимается в том смысле, что если

представить вероятность попадания случайной величины в любой интервал как массу этого интервала, то математическое ожидание будет центром тяжести получившегося распределения масс.

Математическое ожидание дискретной случайной величины, которая может принимать значения x_i, вычисляется по формуле

$$M(\xi) = \sum_i x_i P(\xi = x_i).$$

Оно существует, если этот ряд сходится абсолютно.

Математическое ожидание абсолютно непрерывной случайной величины с плотностью распределения f_ξ вычисляется по формуле

$$M(\xi) = \int_{-\infty}^{+\infty} x f_\xi(x)\, dx.$$

Оно существует, если этот интеграл сходится абсолютно.

Математическое ожидание константы, очевидно, равно самой константе. Кроме того, математическое ожидание обладает свойством линейности: для любых случайных величин ξ и η, у которых существуют математические ожидания, и любых вещественных чисел a и b:

$$M(a\xi + b\eta) = aM(\xi) + bM(\eta).$$

Кроме математического ожидания используются и другие числовые характеристики случайной величины, которые в большинстве случаев могут быть представлены в виде математического ожидания функции от случайной величины $M(\varphi(\xi))$. Можно показать, что в дискретном случае

$$M(\varphi(\xi)) = \sum_i \varphi(x_i) P(\xi = x_i);$$

соответственно в непрерывном случае

$$M(\varphi(\xi)) = \int_{-\infty}^{+\infty} \varphi(x) f_\xi(x)\, dx.$$

Чаще всего используются числовые характеристики, которые называются моментами случайной величины.

Начальным моментом порядка k называется величина $\alpha_k = M(\xi^k)$, в частности $\alpha_1 = M(\xi)$. Существование у случайной величины начальных моментов не гарантировано, оно обеспечивается соответствующей скоростью убывания на бесконечности плотности распределения (в дискретном случае – вероятностей значений). Ясно, что из существования начального момента порядка k следует существование начальных моментов порядка меньше k. Центральным моментом порядка k называется величина

$\mu_k = M(\xi - M(\xi))^k$. Заметим, что благодаря свойству линейности математического ожидания $\mu_1 = 0$.

Дисперсией случайной величины называется ее второй центральный момент, т. е. средний квадрат ее отклонения от среднего значения:

$$D(\xi) = \mu_2 = M(\xi - M(\xi))^2.$$

Часто оказывается полезным равенство

$$D(\xi) = M(\xi - a)^2 - (M(\xi - a))^2,$$

которое справедливо при любом вещественном a. Действительно, благодаря свойству линейности математического ожидания,

$$M(\xi - M(\xi))^2 = M((\xi - a) - (M(\xi) - a))^2 =$$
$$= M(\xi - a)^2 - 2M(\xi - a)(M(\xi) - a) + (M(\xi) - a)^2,$$

откуда следует требуемое равенство. В частности, при $a = 0$ получается

$$D(\xi) = M(\xi)^2 - (M(\xi))^2 = \alpha_2 - \alpha_1^2.$$

Дисперсия константы, очевидно, равна нулю. Для линейной функции от случайной величины дисперсия вычисляется следующим образом:

$$D(a\xi + b) = a^2 D(\xi).$$

В качестве характеристики величины разброса значений случайной величины относительно ее среднего значения обычно используют среднее квадратическое отклонение (СКО). Эта величина представляет собой арифметический квадратный корень из дисперсии:

$$\sigma_\xi = \sqrt{D(\xi)}.$$

Центральный момент любого порядка k, подобно дисперсии, может быть выражен через начальные моменты порядка k и ниже. Поэтому существование центрального момента порядка k эквивалентно существованию начального момента того же порядка.

Еще одной числовой характеристикой случайной величины, дающей представление о ее среднем значении, является медиана. Медианой распределения случайной величины ξ называется число med, такое, что

$$P(\xi \le \mathrm{med}) = P(\xi \ge \mathrm{med}).$$

В отличие от математического ожидания медиана существует у любой случайной величины.

Независимость случайных событий и величин. Случайные события A_1, A_2, ..., A_n являются независимыми в совокупности, если тот

факт, что произошли некоторые из них, не влияет на вероятности остальных. Строгое определение независимости в совокупности формулируется следующим образом. Для любого набора из этих событий A_{i_1}, A_{i_2}, ..., A_{i_k} справедливо равенство

$$P(A_{i_1} \cap A_{i_2} \cap ... \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k})$$

(вероятность пересечения нескольких событий, т. е. того, что все они произошли, равна произведению их вероятностей).

Случайные величины ξ_1, ξ_2, ..., ξ_n называются независимыми (в совокупности), если для любых интервалов I_1, I_2, ..., I_n случайные события $(\xi_1 \in I_1)$, $(\xi_2 \in I_2)$, ..., $(\xi_n \in I_n)$ являются независимыми в совокупности.

Если ξ_1 и ξ_2 – независимые случайные величины, то математическое ожидание их произведения (если оно существует) равно произведению математических ожиданий

$$M(\xi_1\xi_2) = M(\xi_1)M(\xi_2).$$

Отсюда, в частности, вытекает правило суммирования дисперсий: для независимых случайных величин ξ_1, ξ_2, ..., ξ_n, у которых существуют дисперсии,

$$D\left(\sum_{i=1}^{n} \xi_i\right) = \sum_{i=1}^{n} D(\xi_i).$$

Совместное распределение случайных величин. Упорядоченный набор из нескольких случайных величин $\boldsymbol{\xi} = [\xi_1, \xi_2, ..., \xi_n]^T$ называют случайным вектором. Задано распределение случайного вектора, которое также называют совместным распределением входящих в этот вектор случайных величин, если для любой области $G \subset \mathbb{R}^n$ определена вероятность случайного события $(\boldsymbol{\xi} \in G)$. Совместное распределение случайных величин называется абсолютно непрерывным, если существует неотрицательная интегрируемая на $\mathbb{R}^n$ функция $f_{\boldsymbol{\xi}}(\boldsymbol{x}) = f_{\boldsymbol{\xi}}(x_1, ..., x_n)$, называемая плотностью совместного распределения, такая, что для любой области $G \subset \mathbb{R}^n$ справедливо равенство

$$P(\boldsymbol{\xi} \in G) = \int\limits_G f_{\boldsymbol{\xi}}(x_1, ..., x_n)dx_1...dx_n.$$

У независимых абсолютно непрерывных случайных величин существует плотность совместного распределения, которая представляет собой произведение отдельных плотностей:

$$f_{\boldsymbol{\xi}}(x_1, ..., x_n) = f_{\xi_1}(x_1)f_{\xi_2}(x_2) \cdots f_{\xi_n}(x_n).$$

Простейшей числовой характеристикой совместного распределения двух случайных величин является их ковариационный момент

$$\mathrm{cov}(\xi_1, \xi_2) = M((\xi_1 - M(\xi_1))(\xi_2 - M(\xi_2))).$$

Эта характеристика отражает степень взаимного влияния двух случайных величин. У независимых случайных величин ковариационный момент равен нулю. Нетрудно показать, что ковариационный момент также может быть вычислен по формуле

$$\mathrm{cov}(\xi_1, \xi_2) = M(\xi_1\xi_2) - M(\xi_1)M(\xi_2).$$

Из приведенных формул видно, что ковариационный момент имеет размерность произведения двух случайных величин.

Безразмерной числовой характеристикой степени взаимного влияния случайных величин является коэффициент корреляции

$$\rho(\xi_1, \xi_2) = \frac{\mathrm{cov}(\xi_1, \xi_2)}{\sqrt{D(\xi_1)D(\xi_2)}}.$$

Коэффициент корреляции может принимать значения из интервала $[-1, 1]$. Если между случайными величинами ξ_1 и ξ_2 есть жесткая линейная связь $\xi_2 = a\xi_1 + b$, то коэффициент корреляции равен 1 или -1, в зависимости от знака a. Случайные величины называются некоррелированными, если их коэффициент корреляции равен нулю. Очевидно, две независимые случайные величины являются некоррелированными. Обратное в общем случае неверно: из некоррелированности не следует независимость.

Ковариационной матрицей R n-мерного случайного вектора $\boldsymbol{\xi}$ называется матрица размера $n \times n$, элементы которой представляют собой ковариационные моменты соответствующих пар компонент этого вектора: $r_{i,j} = \mathrm{cov}(\xi_i, \xi_j)$. Таким образом, диагональными элементами ковариационной матрицы являются дисперсии соответствующих случайных величин. Ковариационная матрица является симметричной и неотрицательно-определенной матрицей. Ковариационная матрица случайного вектора $\boldsymbol{\eta}$ с некоррелированными компонентами является диагональной. Если случайный вектор $\boldsymbol{\xi}$ размерности n имеет ковариационную матрицу $R(\boldsymbol{\xi})$, а случайный вектор $\boldsymbol{\eta} = A\boldsymbol{\xi}$, где A – матрица размера $m \times n$, то ковариационная матрица вектора $\boldsymbol{\eta}$ вычисляется по формуле

$$R(\boldsymbol{\eta}) = AR(\boldsymbol{\xi})A^T$$

(верхний индекс T здесь и далее означает транспонирование матрицы).

Сходимость случайных величин. Пусть имеется последовательность случайных величин

$$\xi_1,\ \xi_2,\ \ldots,\ \xi_n,\ \ldots\ .$$

Говорят, что эта последовательность сходится к числу a с вероятностью 1, если $P(\lim\limits_{n\to\infty} \xi_n = a) = 1$. Это наиболее естественное определение сходимости случайных величин требует, однако, во многих случаях довольно сложных доказательств. Кроме того, обычно бывает важно поведение не самих случайных величин, а их вероятностных распределений. Поэтому чаще используется другое понятие сходимости – сходимость по вероятности.

Последовательность случайных величин ξ_1, ξ_2, ..., ξ_n, ... сходится по вероятности к числу a, если $\forall\, \varepsilon > 0$

$$\lim_{n\to\infty} P(|\xi_n - a| > \varepsilon) = 0.$$

Из сходимости с вероятностью 1 следует сходимость по вероятности, обратное неверно.

Неравенство Чебышева. Это неравенство позволяет оценить вероятность превышения положительной случайной величиной некоторого значения через ее математическое ожидание. Для любой неотрицательной случайной величины η и $\forall\ t > 0$ справедливо неравенство $P(\eta > t) \le \dfrac{M(\eta)}{t}$. Доказательство этого неравенства достаточно просто. Например, для неотрицательной случайной величины с абсолютно непрерывным распределением

$$M(\eta) = \int\limits_0^\infty x f_\xi(x)\,dx = t\int\limits_0^\infty \frac{x}{t} f_\xi(x)\,dx \ge t\int\limits_t^\infty \frac{x}{t} f_\xi(x)\,dx \ge$$

$$\ge t\int\limits_t^\infty f_\xi(x)\,dx = tP(\eta > t),$$

откуда следует требуемая оценка.

Применение этой оценки к случайной величине $|\xi - M(\xi)|$ позволяет легко получить неравенство Чебышева в наиболее часто используемой форме, которая представляет собой оценку вероятности большого отклонения случайной величины от ее математического ожидания через дисперсию, а именно $\forall\, \varepsilon > 0$

$$P(|\xi - M(\xi)| > \varepsilon) \le \frac{D(\xi)}{\varepsilon^2}.$$

Закон больших чисел. Законом больших чисел принято называть утверждения о сходимости последовательности соответствующим образом нормированных сумм независимых случайных величин к некоторому числу. В дальнейшем будет использоваться следующий простейший вариант этого закона, который формулируется следующим образом.

Пусть ξ_1, ξ_2, ..., ξ_n, ... – последовательность независимых случайных величин с одинаковыми распределениями, у которых существует математическое ожидание $M(\xi)$. Тогда последовательность их средних арифметических $\dfrac{1}{n}\sum\limits_{i=1}^{n}\xi_i$ сходится по вероятности к $M(\xi)$. В случае существования у случайных величин дисперсии это утверждение легко получается из неравенства Чебышева и свойств дисперсии.

На самом деле приведенное выше утверждение остается справедливым, если заменить сходимость по вероятности на сходимость с вероятностью 1. В этом случае его обычно называют усиленным законом больших чисел.

Нормальное распределение. Это распределение, называемое также гауссовским, является наиболее используемым в теории вероятностей и математической статистике. Стандартным нормальным распределением называется абсолютно непрерывное распределение с плотностью (рис. 1.1)

$$f(x) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}, \quad x \in \mathbb{R}.$$

Оно имеет нулевое математическое ожидание и дисперсию, равную 1.

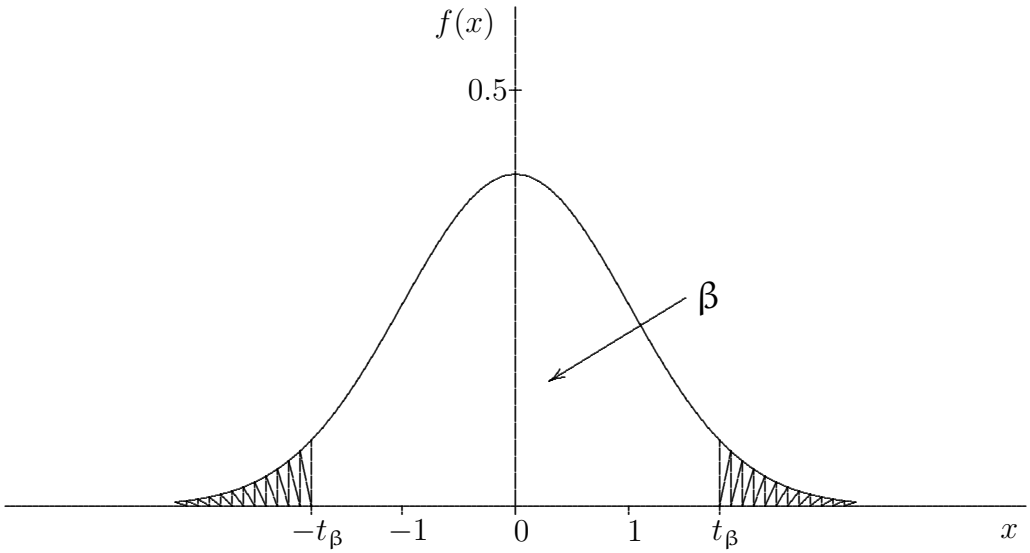

Рис. 1.1. Кривая плотности стандартного нормального распределения

Если случайная величина ξ_0 имеет стандартное нормальное распределение, то $\xi = \sigma\xi_0 + m$, где $\sigma > 0$, имеет математическое ожидание m, дисперсию σ^2 и плотность распределения

$$f(x, m, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}\right\}, \quad x \in \mathbb{R}.$$

Распределение с такой плотностью называется нормальным. Тот факт, что случайная величина ξ имеет такое распределение, кратко записывается

следующим образом: $\xi \in N(m, \sigma^2)$. Соответственно, стандартное нормальное распределение обозначается как $N(0,1)$.

Симметричным нормальным квантилем по уровню β (на рис. 1.1 β – площадь незаштрихованной части) называют число t_β, такое, что

$$P(|\xi_0| < t_\beta) = \beta,$$

где $\xi_0 \in N(0,1)$.

Класс нормальных распределений обладает тем свойством, что любая линейная комбинация случайных величин с нормальным распределением также имеет нормальное распределение. Кроме того, для нормально распределенных случайных величин понятия независимости и некоррелированности становятся равносильными.

Центральная предельная теорема. Важная роль нормального распределения объясняется тем, что распределение суммы независимых случайных величин с конечными дисперсиями приближается при увеличении числа слагаемых к нормальному распределению. Этот факт принято называть центральной предельной теоремой. Простейший вариант центральной предельной теоремы, который в дальнейшем будет использоваться, можно сформулировать следующим образом. Пусть ξ_1, ξ_2 ,..., ξ_n, ... – последовательность независимых случайных величин с одинаковыми распределениями, математическими ожиданиями $M(\xi_i) = m$ и дисперсиями $D(\xi_i) = \sigma^2$. Введем последовательность нормированных сумм

$$Z_n = \frac{\sum_{i=1}^{n} \xi_i - nm}{\sigma\sqrt{n}}.$$

Тогда $\forall\, x \in \mathbb{R}$

$$\lim_{n\to\infty} P(Z_n < x) = \Phi(x),$$

где $\Phi(x)$ – функция распределения стандартного нормального закона $N(0,1)$, которая может быть записана в виде

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{t^2}{2}\right\} dt.$$

1.2. Наблюдаемая случайная величина. Выборка

Рассмотрим случайную величину ξ, относительно которой будем предполагать возможность наблюдения ее значений (реально или мысленно) любое число раз. Эту случайную величину будем называть „наблюдаемая случайная величина“ (НСВ) и обозначать ξ.

Пусть в n повторных наблюдениях НСВ ξ зарегистрированы ее числовые значения x_1, x_2, ..., x_n. Числовой вектор

$$\boldsymbol{x} = [x_1, x_2, ..., x_n]^T$$

называется выборкой из распределения НСВ ξ (числовой выборкой или просто выборкой).

Выборка – это то, что всегда необходимо иметь для использования статистических методов. Выборку можно получить в результате проведения реальных случайных экспериментов с НСВ ξ, а можно предположить факт ее существования. Для этого нужно задать модель генерации выборки любого объема.

В математической статистике делаются следующие теоретико-вероятностные предположения. Математической моделью одного наблюдения с номером i является „копия“ НСВ ξ, т. е. случайная величина X_i, имеющая то же распределение, что и ξ.

Математической моделью выборки $\boldsymbol{x}$ является случайный вектор $\boldsymbol{X} = [X_1, X_2, ..., X_n]^T$. Будем называть этот вектор случайной выборкой.

1.3. Выборочная функция распределения

По выборке $\boldsymbol{x} = [x_1, x_2, ..., x_n]^T$ можно определить дискретное вероятностное распределение, сопоставив каждому значению x_i одну и ту же вероятность $\dfrac{1}{n}$. Такое распределение называется *выборочным распределением*. Соответствующая функция распределения может быть записана в виде

$$F_n(t) = \frac{n_t}{n}, \quad \forall\, t \in \mathbb{R},$$

где n_t – количество элементов выборки, таких, что $x_i < t$. Она называется *выборочной функцией распределения*.Эта функция представляет собой возрастающую от 0 до 1 ступенчатую функцию не более чем с n разрывами первого рода („ступенями“).

Если выборку рассматривать как случайный вектор, то при любом t значение $F_n(t)$ представляет собой случайную величину. Смысл выборочной функции распределения раскрывает следующая теорема Гливенко–Кантелли:

Теорема 1.1. *Если $F_\xi(t)$ – функция распределения НСВ ξ и $n \to +\infty$, то*

$$\sup_t |F_\xi(t) - F_n(t)| \to 0$$

с вероятностью 1.

Таким образом, теорема Гливенко–Кантелли показывает, что если увеличивать объем выборки до бесконечности, то можно приблизить выборочную функцию распределения к теоретической с любой степенью точности.

1.4. Вариационный ряд

Выборка, упорядоченная по значениям элементов, называется вариационным рядом. Таким образом, для выборки $\boldsymbol{x} = [x_1, x_2, ..., x_n]^T$ вариационный ряд строится сортировкой ее элементов в порядке возрастания:

$$x_{n_1} \leq x_{n_2} \leq ... \leq x_{n_n}.$$

Вариационный ряд – очень удобное средство изучения НСВ ξ. Существует целое направление в математической статистике, называемое порядковыми статистиками, основанное на использовании „рангов“, т. е. номеров элементов выборки в вариационном ряду. Приведем в качестве примера простейшие ранговые оценки, получаемые непосредственно из вариационного ряда.

Выборочной медианой называется число

$$\widehat{\text{med}} = \begin{cases} x_{n_{\left(\frac{n+1}{2}\right)}}, & \text{если } n \text{ нечетное,} \\ \dfrac{x_{n_{\left(\frac{n}{2}\right)}} + x_{n_{\left(\frac{n}{2}+1\right)}}}{2}, & \text{если } n \text{ четное.} \end{cases}$$

Размахом выборки называется число

$$R_n = x_{n_n} - x_{n_1}.$$

1.5. Гистограмма

Наглядное представление выборки можно получить с помощью простых геометрических построений. Разобьем область значений НСВ ξ (которую можем получить из вариационного ряда) на k равных интервалов (z_0, z_1), (z_1, z_2), ..., (z_{k-1}, z_k). Пусть длина интервалов равна $z_i - z_{i-1} = l$. Используя каждый интервал как основание, построим на нем прямоугольник с высотой

$$h_i = \frac{m_i}{nl},$$

где m_i – количество элементов выборки, попавших в интервал (z_{i-1}, z_i).

Полученная фигура называется гистограммой (рис. 1.2).

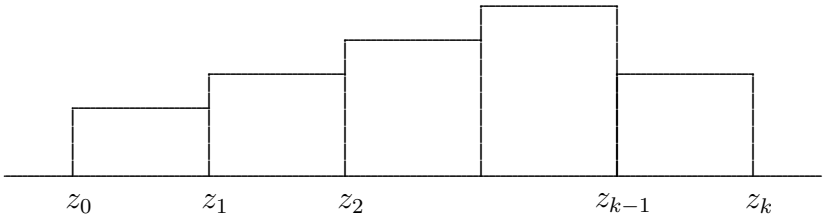

Рис. 1.2

Если НСВ ξ имеет абсолютно непрерывное распределение, то гистограмма дает представление о виде плотности распределения НСВ ξ.

1.6. Распределение „Хи-квадрат“

Пусть случайные величины ξ_1, ξ_2, ..., $\xi_n \in N(0,1)$ и независимы в совокупности. Случайная величина

$$\chi_n^2 = \xi_1^2 + \xi_2^2 + ... + \xi_n^2$$

называется распределенной по закону „Хи-квадрат“ с n степенями свободы.

Это распределение обладает следующими свойствами:

1. Случайная величина $\chi_n^2 \geq 0$, она имеет плотность распределения

$$\mathcal{K}_n(x) = \begin{cases} 0, & \text{если } x \leq 0, \\ \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & \text{если } x > 0, \end{cases}$$

где функция $\Gamma(x) = \int\limits_0^{+\infty} t^{x-1} e^{-t}\, dt$. ($\Gamma(k) = (k-1)!$ для натуральных k.)

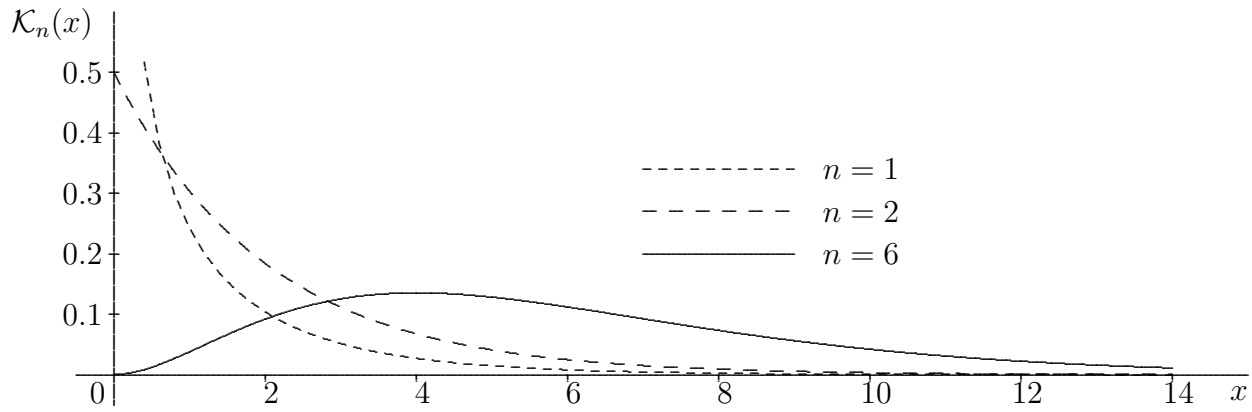

Рис. 1.3. χ^2-распределение. Кривые плотности для n =1, 2, 6

При $n \le 2$ $\mathcal{K}_n(x)$ монотонно убывает, а при $n > 2$ имеет единственный максимум при $x = n - 2$ (рис. 1.3).

2. $M(\chi^2_n) = n$; $D(\chi^2_n) = 2n$.

3. При больших n (практически при $n > 30$) в силу центральной предельной теоремы χ^2_n имеет распределение, близкое к нормальному $N(n; 2n)$.

4. Если $\chi^2_{n_1}$ и $\chi^2_{n_2}$ независимы, то их сумма $\chi^2_{n_1} + \chi^2_{n_2}$ распределена по закону $\chi^2_{n_1+n_2}$.

Теорема 1.2 (Стьюдента). *Пусть случайные величины X_1, X_2, ..., X_n распределены по закону $N(m, \sigma^2)$ и независимы в совокупности. Обозначим*

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

Тогда случайная величина $\dfrac{nS^2}{\sigma^2}$ распределена по закону χ^2_{n-1}, а случайные величины $\overline{X}$ и S^2 независимы.

Распределение χ^2_n табулировано. В дальнейшем интерес будут представлять квантили χ^2_n-распределения. Число $\chi^2_0(n, \beta)$ называется квантилем χ^2_n-распределения по уровню β, если

$$P(\chi^2_n < \chi^2_0(n, \beta)) = \beta.$$

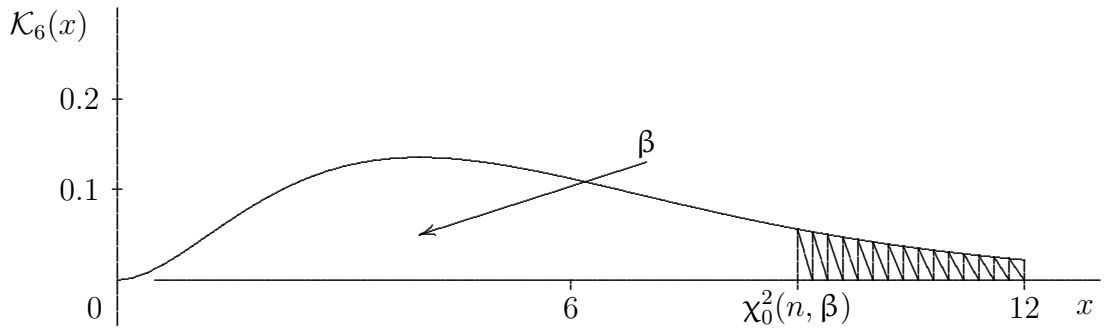

Рис. 1.4. χ^2-распределение (n =6)

На рис. 1.4 β – площадь незаштрихованной части.

1.7. Распределение Стьюдента

Распределение названо в честь английского математика Госсета, публиковавшего свои работы под псевдонимом Student.

Пусть случайные величины ξ, ξ_1, ξ_2, ..., $\xi_n \in N(0,1)$ и независимы в совокупности. Тогда случайная величина

$$\tau_n = \frac{\xi\sqrt{n}}{\sqrt{\sum_{i=1}^{n} \xi_i^2}} = \frac{\xi\sqrt{n}}{\sqrt{\chi_n^2}}$$

распределена по закону Стьюдента с n степенями свободы.

Плотность распределения Стьюдента имеет вид

$$S_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

Плотность $S_n(x)$ – четная функция, график которой (рис. 1.5) симметричен относительно оси ординат. Следовательно, если существует математическое ожидание, то оно равно нулю. При $n = 1$ распределение Стьюдента совпадает с распределением Коши (у которого математическое ожидание не существует). При $n > 1$ $M(\tau_n) = 0$.

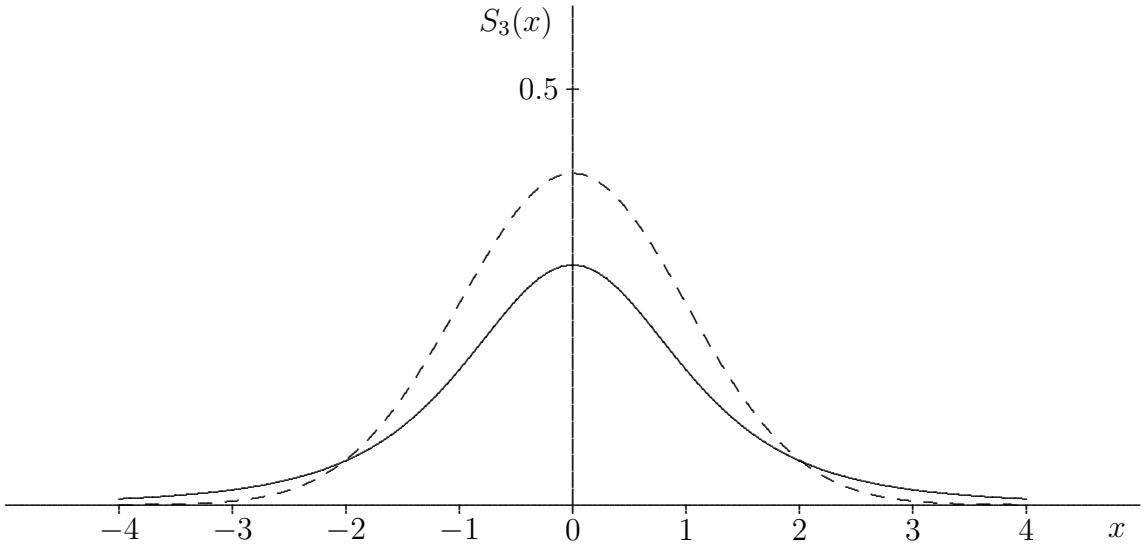

Рис. 1.5. Кривая плотности распределения Стьюдента при $n = 3$
Штрихами показана кривая нормальной плотности, $m = 0,\ \sigma = 1$

Из сформулированной ранее теоремы Стьюдента легко получается следующее утверждение.

Теорема 1.3. *Пусть X_1, X_2, ..., $X_n \in N(m, \sigma^2)$ и независимы в совокупности. Тогда случайная величина $\dfrac{(\overline{X} - m)\sqrt{n-1}}{\sqrt{S^2}}$ распределена по закону Стьюдента с $(n-1)$ степенями свободы τ_{n-1}.*

Распределение Стьюдента табулировано. По аналогии с симметричным нормальным квантилем t_β для распределения Стьюдента вводят понятие симметричного квантиля Стьюдента $\tau_0(n, \beta)$.

Симметричным квантилем Стьюдента по уровню β называют число $\tau_0(n, \beta)$, такое, что $P(|\tau_n| < \tau_0(n, \beta)) = \beta$, (на рис. 1.6 β – площадь незаштрихованной части).

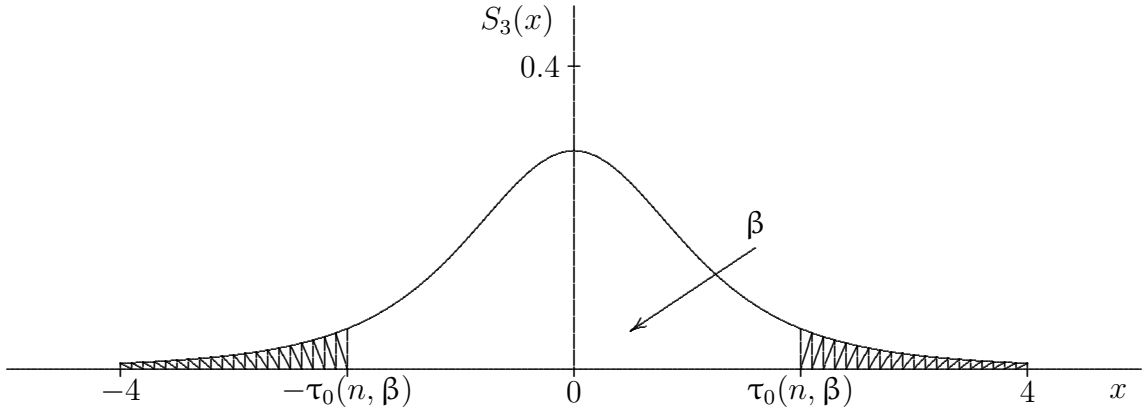

Рис. 1.6. Кривая плотности распределения Стьюдента при $n = 3$

При $n > 30$ для распределения Стьюдента пользуются нормальным приближением. Соответственно, вместо симметричного квантиля Стьюдента $\tau_0(n, \beta)$ используют симметричный нормальный квантиль t_β.

1.8. Распределение Фишера

Пусть χ^2_n и χ^2_m – две независимые случайные величины, имеющие распределение „Хи-квадрат“ соответственно с n и m степенями свободы. Положительная случайная величина $f_{n,m} = \dfrac{m\chi^2_n}{n\chi^2_m}$ имеет распределение Фишера с числом степеней свободы (n, m). Распределение Фишера также иногда называют распределением Снедекора (рис. 1.7).

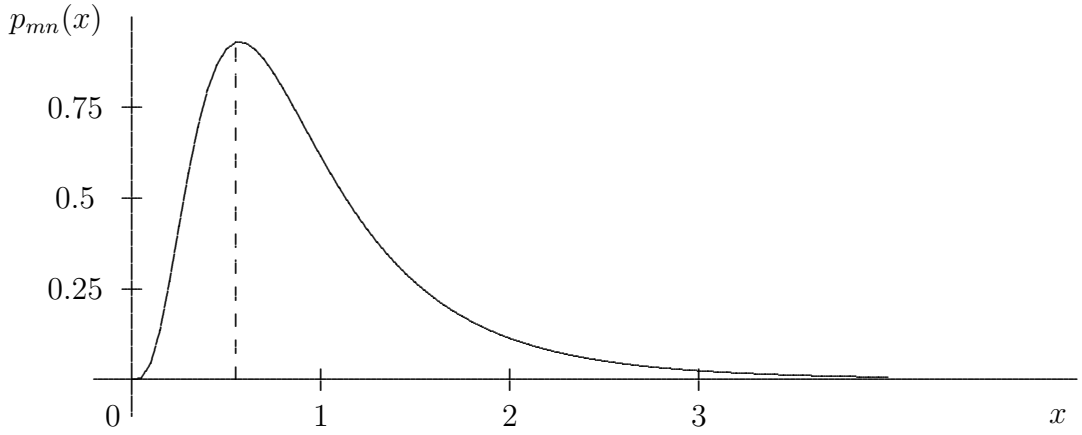

Рис. 1.7. Функция плотности распределения Фишера ($m = 10,\ n = 12$)

Плотность вероятности этого распределения задается формулой

$$p_{mn}(x) = c_{n,m} x^{\frac{n}{2}-1} (m + nx)^{-\frac{(n+m)}{2}}, \quad x > 0,$$

где $c_{n,m}$ – некоторая константа.

2. ОЦЕНИВАНИЕ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ

2.1. Оценки параметров распределения и их свойства. Оценки математического ожидания и вероятности события

Пусть имеется выборка $\boldsymbol{x} = [x_1, x_2, ..., x_n]^T$ из распределения некоторой НСВ ξ. Требуется оценить значение числовой характеристики ξ, т. е. параметра θ ее вероятностного распределения. Решением этой задачи оценивания будем считать интервал, в котором предположительно (с большой вероятностью) должно находиться истинное значение параметра. Такой интервал называют интервальной оценкой параметра, или *доверительным интервалом*.

Итак, *доверительным интервалом* I_β для параметра θ с *доверительной вероятностью* β называется такой построенный по заданной выборке интервал, что

$$P(\theta \in I_\beta) = \beta.$$

Как правило, вероятность β выбирается из диапазона от 0.9 до 0.99. Выбор значения $\beta = 0.9$ означает согласие с тем, что в среднем один из десяти построенных доверительных интервалов I_β не будет накрывать параметр θ. Соответственно, при $\beta = 0.99$ это будет происходить в одном из ста случаев.

Доверительный интервал можно задать его концами, но чаще он задается серединой и половиной длины ε. В качестве самого оцененного значения параметра может быть взято любое число из этого интервала (обычно берется его середина). В любом случае оцененное значение определяется по выборке, т. е. является ее функцией. Всякую функцию выборки принято называть *статистикой*. Как уже отмечалось, выборка может рассматриваться как случайный вектор, точнее, как набор из n независимых одинаково распределенных случайных величин. В этом случае всякая статистика также является случайной величиной.

Основным инструментом построения оценок параметров распределений являются так называемые *точечные оценки*. Точечной оценкой параметра называется статистика, значение которой может использоваться в

качестве значения оцениваемого параметра. Естественным требованием к точечной оценке $\widehat{\theta}_n = \widehat{\theta}(X_1, X_2, ..., X_n)$ является так называемое условие *несмещенности.*

Определение 2.1. *Статистика $\widehat{\theta}_n$ называется несмещенной оценкой параметра θ, если ее математическое ожидание совпадает с истинным значением параметра, т. е. $M(\widehat{\theta}_n) = \theta$ при любом n.*

Требование несмещенности точечной оценки иногда заменяется более слабым требованием *асимптотической несмещенности.*

Определение 2.2. *Статистика $\widehat{\theta}_n$ называется асимптотически несмещенной оценкой параметра θ, если $\lim\limits_{n\to\infty} M(\widehat{\theta}_n) = \theta$.*

Основным требованием к точечной оценке параметра, является *состоятельность*, т. е. приближение (сходимость) к истинному значению параметра при увеличении объема выборки n.

Определение 2.3. *Статистика $\widehat{\theta}_n$ называется состоятельной оценкой параметра θ, если последовательность случайных величин $\widehat{\theta}_n$ сходится по вероятности к числу θ при $n \to \infty$.*

(Определение сходимости по вероятности приводилось в 1.1.)

Доказанное в 1.1 неравенство Чебышева для положительной случайной величины позволяет при любом $\varepsilon > 0$ получить неравенство

$$P(|\widehat{\theta}_n - \theta| > \varepsilon) = P((\widehat{\theta}_n - \theta)^2 > \varepsilon^2) \le \frac{M(\widehat{\theta}_n - \theta)^2}{\varepsilon^2}.$$

Из этого неравенства следует, что для состоятельности оценки достаточно выполнения условия $\lim\limits_{n\to\infty} M(\widehat{\theta}_n - \theta)^2 = 0$. Для несмещенной оценки параметра, согласно определению дисперсии, $M(\widehat{\theta}_n - \theta)^2 = D(\widehat{\theta}_n)$, а для асимптотически несмещенной $\lim\limits_{n\to\infty} (M(\widehat{\theta}_n - \theta)^2 - D(\widehat{\theta}_n)) = 0$. Отсюда получается следующее утверждение.

Теорема 2.1. *Если статистика $\widehat{\theta}_n$ является асимптотически несмещенной оценкой параметра θ и $\lim\limits_{n\to\infty} D(\widehat{\theta}_n) = 0$, то она является состоятельной оценкой θ.*

Для получения оценки параметра в виде доверительного интервала необходимо хотя бы приблизительно знать вероятностное распределение соответствующей точечной оценки. Во многих случаях это распределение оказывается близким к нормальному.

Определение 2.4. *Оценка параметра $\widehat{\theta}_n$ называется асимптотически нормальной, если существует такая числовая последовательность*

c_n, *что* $\forall\, x \quad \lim\limits_{n\to\infty} P(c_n(\widehat{\theta}_n - M(\widehat{\theta}_n)) < x) = \Phi(x)$, *где* $\Phi(x)$ – *функция распределения стандартного нормального закона* $N(0,1)$.

Из центральной предельной теоремы, сформулированной в 1.1, вытекает следующее утверждение.

Теорема 2.2. *Любая точечная оценка параметра* $\widehat{\theta}_n$, *представимая в виде среднего арифметического независимых случайных величин с одинаковыми распределениями и конечными дисперсиями, является асимптотически нормальной.*

В качестве типичного примера рассмотрим задачу оценивания математического ожидания НСВ. Предполагается, что НСВ ξ имеет конечное (неизвестное) математическое ожидание $M(\xi) = \theta$. Для оценивания этого параметра обычно используется статистика, представляющая собой среднее арифметическое элементов выборки:

$$\widehat{\theta}(X_1, X_2, ..., X_n) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Из свойства линейности математического ожидания следует, что

$$M(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = M(\xi),$$

т. е. данная статистика является несмещенной оценкой математического ожидания. Сформулированный в 1.1 закон больших чисел фактически утверждает, что данная статистика является также состоятельной оценкой математического ожидания. Из теоремы 2.2 следует, что при существовании у НСВ конечной дисперсии эта оценка является асимптотически нормальной.

Частным случаем задачи оценивания математического ожидания является задача оценивания вероятности случайного события. Пусть проводится n одинаковых независимых экспериметов, в результате каждого из которых может наступить или не наступить некоторое случайное событие A. Если зафиксировать исходы всех экспериментов, записывая 1 при наступлении события и 0 в противном случае, получится выборка объема n, состоящая из нулей и единиц. Эта выборка является выборкой из распределения двуточечной НСВ ξ, принимающей значение 1 с вероятностью $p = P(A)$ и 0 с вероятностью $1 - p$. Математическое ожидание этой НСВ $M(\xi) = 0(1-p) + 1p = p$. Как было только что показано, несмещенной состоятельной оценкой этого параметра является статистика $\widehat{p} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i = \dfrac{m(A)}{n}$, где $m(A)$ – количество единиц в выборке – факти-

чески представляет собой число наступлений события A в серии из n независимых экспериментов. Статистику $\widehat{p}$ естественно назвать относительной частотой события A. Таким образом, относительная частота события является несмещенной состоятельной и асимптотически нормальной оценкой его вероятности.

2.2. Метод выборочных моментов. Оценка дисперсии

В 1.3 было введено понятие выборочной функции распределения и сформулирована теорема Гливенко–Кантелли о сходимости выборочной функции распределения к истинной с вероятностью 1. Поскольку из сходимости с вероятностью 1 следует сходимость по вероятности, простым следствием теоремы Гливенко–Кантелли, является следующее утверждение.

Теорема 2.3. *При $\forall\, t$ значение выборочной функции распределения $F_n(t)$ является состоятельной оценкой значения истинной функции распределения $F_\xi(t)$.*

Эту теорему несложно также доказать непосредственно. Для этого достаточно заметить, что значение выборочной функции распределения $F_n(t)$ представляет собой относительную частоту случайного события $(\xi < t)$ и поэтому, как было показано в 2.1, является несмещенной состоятельной оценкой для $P(\xi < t) = F_\xi(t)$.

Предположим теперь, что оцениваемый параметр представляет собой начальный или центральный момент НСВ. Исходя из последней теоремы, для оценивания этого параметра естественно использовать статистику, представляющую собой соответствующий момент выборочного распределения. Согласно определению выборочного распределения, каждому значению элемента выборки приписывается вероятность $\dfrac{1}{n}$, поэтому начальные моменты выборочного распределения вычисляются по формуле

$$\widehat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k,$$

а центральные моменты выборочного распределения

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \widehat{\alpha}_1)^k.$$

Эти статистики называются *выборочными моментами* НСВ.

В частности, выборочное математическое ожидание определяется формулой

$$\widehat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

Эта статистика, как было показано в 2.1, является состоятельной несмещенной оценкой математического ожидания $M(\xi) = \alpha_1$. Если аналогичным образом применить закон больших чисел к случайной величине ξ^k, а также теорему 2.2, то получится следующее утверждение.

Теорема 2.4. *Если НСВ ξ имеет конечный момент $\alpha_k = M(\xi^k)$, то соответствующий начальный выборочный момент $\widehat{\alpha}_k$ является состоятельной несмещенной оценкой этого параметра. При существовании момента $\alpha_{2k} = M(\xi^{2k})$ эта оценка является асимптотически нормальной.*

Сложнее обстоит дело с выборочными центральными моментами. Второй выборочный центральный момент (или выборочная дисперсия) определяется формулой

$$S^2 = \widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Требуется доказать, что эта статистика является состоятельной оценкой дисперсии. Для доказательства этого утверждения понадобятся следующие 2 свойства сходимости по вероятности, которые нетрудно получить непосредственно из определения этой сходимости, приведенного в 1.1.

1) Если последовательности случайных величин ξ_n и η_n сходятся по вероятности к числам a и b, то последовательность $(\xi_n - \eta_n)$ сходится по вероятности к числу $(a - b)$.

2) Если последовательность случайных величин ξ_n сходится по вероятности к нулю, то $\forall\, k > 0$ последовательность ξ_n^k сходится по вероятности к нулю.

Пусть случайная величина ζ_n имеет выборочное распределение, построенное по выборке объема n из распределения НСВ с математическим ожиданием m и дисперсией σ^2. Тогда по свойствам дисперсии

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

Первое выражение из получившейся разности представляет собой среднее арифметическое независимых одинаково распределенных случайных величин $(X_i - m)^2$ и, согласно закону больших чисел, сходится по вероятности к математическому ожиданию этих величин, т. е. к σ^2. Далее, согласно закону больших чисел, последовательность $\overline{X}$ сходится по вероятности к m,

следовательно, по свойствам 1 – 2, $(\overline{X} - m)^2$ сходится по вероятности к нулю. Таким образом, по свойству 1, S^2 сходится по вероятности к σ^2, т. е. является состоятельной оценкой дисперсии.

Вместе с тем статистика S^2 не является несмещенной оценкой дисперсии. Действительно, если представить эту статистику в виде разности двух выражений, так же как при доказательстве состоятельности, то математическое ожидание первого выражения, очевидно, равно σ^2. Математическое ожидание второго выражения по свойствам дисперсии можно представить в виде

$$M(\overline{X} - m)^2 = D(\overline{X}) = \frac{1}{n^2} \sum_{i=1}^{n} D(X_i) = \frac{\sigma^2}{n}.$$

Таким образом,

$$M(S^2) = \left(1 - \frac{1}{n}\right) \sigma^2,$$

и статистика S^2 не является несмещенной оценкой дисперсии. Все же она является асимптотически несмещенной оценкой, поскольку последовательность $1 - \dfrac{1}{n}$ стремится к 1 при $n \to \infty$.

При умножении статистики S^2 на величину, обратную множителю $1 - \dfrac{1}{n}$, получится несмещенная состоятельная оценка дисперсии

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

Это следует из несложно проверяемого утверждения, что умножение состоятельной оценки на стремящееся к единице числовое выражение не нарушает свойства состоятельности. Таким образом, доказано следующее утверждение.

Теорема 2.5. *Если НСВ ξ имеет конечную дисперсию, то статистика $\overline{S}^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ является состоятельной несмещенной оценкой этой дисперсии.*

Аналогичным образом обстоит дело с центральными моментами более высокого порядка. Можно сформулировать следующее общее утверждение.

Теорема 2.6. *Если НСВ ξ имеет конечный центральный момент μ_k, то соответствующий центральный выборочный момент $\widehat{\mu}_k$ является состоятельной асимптотически несмещенной оценкой этого параметра.*

Что касается несмещенных состоятельных оценок центральных моментов, то они, как и в случае дисперсии, могут быть получены из соответствующих выборочных моментов введением некоторых дополнительных поправок.

2.3. Метод максимального правдоподобия

Метод максимального правдоподобия может использоваться в тех случаях, когда заранее известно, что НСВ имеет распределение, принадлежащее некоторому параметрическому семейству распределений, т. е. распределение НСВ полностью определяется значением некоторого скалярного или векторного параметра θ, который требуется оценить. В частности, может предполагаться, что распределение НСВ ξ является дискретным и определяется вероятностями $P(\xi = t_i) = p(t_i, \theta)$. Тогда, поскольку элементы выборки рассматриваются как независимые случайные величины с одним и тем же распределением, зависящая от θ вероятность того, что все эти случайные величины примут заданные значения, будет равна:

$$P(X_1 = t_1, X_2 = t_2, ..., X_n = t_n; \theta) = p(t_1, \theta)p(t_2, \theta) \cdots p(t_n, \theta).$$

Если подставить в это выражение в качестве аргументов t_i сами элементы выборки, то при истинном значении θ получившееся значение будет по меньшей мере строго положительным (элементы выборки не могут принимать невозможных значений). Кроме того, сравнительно большие значения каждого сомножителя, а значит, и всего произведения, будут более вероятны, чем малые. Это дает основание использовать в качестве точечной оценки неизвестного параметра то значение θ, при котором данное произведение принимает наибольшее значение, т. е. статистику $\widehat{\theta}_n = \widehat{\theta}(x_1, x_2, ..., x_n)$, представляющую собой точку максимума функции

$$L_n(\theta) = p(X_1, \theta)p(X_2, \theta) \cdots p(X_n, \theta).$$

Эта функция называется *функцией правдоподобия* выборки, а сама статистика называется *оценкой максимального правдоподобия*. Поскольку максимизируемая функция представляет собой произведение большого числа сомножителей, обычно бывает удобнее искать максимум не самой функции, а ее натурального логарифма. На результат такая замена не повлияет, поскольку сама функция положительна, а натуральный логарифм является монотонно возрастающей функцией, определенной при положительных значениях аргумента.

Простым примером оценки максимального правдоподобия в дискретном случае может служить относительная частота случайного события (см. 2.1). Выборка, сформированная для получения этой статистики, соответствует семейству двуточечных распределений, сосредоточенных в точках

0 и 1, где в качестве параметра θ берется вероятность того, что НСВ принимает значение 1, которая как раз и является оцениваемой вероятностью события. Если в серии из n экспериметов событие наступило ровно m раз, то выборка содержит ровно m единиц, которым соответствует вероятность θ, и $n-m$ нулей, которым соответствует вероятность $(1-\theta)$. Таким образом, в этом случае функция правдоподобия

$$L_n(\theta) = \theta^m(1-\theta)^{n-m},$$

соответственно

$$\ln(L_n(\theta)) = m\ln(\theta) + (n-m)\ln(1-\theta).$$

Получившаяся непрерывная на интервале $(0,1)$ функция стремится к $-\infty$ на краях этого интервала и, следовательно, должна иметь точку максимума внутри интервала, где ее производная равна 0. Таким образом, точка максимума является решением уравнения

$$\frac{m}{\theta} - \frac{n-m}{1-\theta} = 0.$$

Единственное решение этого уравнения $\widehat{\theta}_n = \dfrac{m}{n}$ является оценкой максимального правдоподобия для вероятности события. Оно совпадает с оценкой, полученной в 2.1.

Еще одним примером оценки максимального правдоподобия в дискретном случае может служить оценка параметра распределения Пуассона. Случайная величина ξ, имеющая распределение Пуассона, является дискретной, она может принимать все целые неотрицательные значения с вероятностями

$$P(\xi = k) = \frac{\lambda^k}{k!}e^{-\lambda} \quad (k = 0, 1, 2, ...),$$

в частности, $P(\xi = 0) = e^{-\lambda}$. Требуется оценить по выборке параметр $\lambda > 0$.

В этой задаче функция правдоподобия

$$L_n(\lambda) = \frac{\lambda^{X_1}}{X_1!}e^{-\lambda}\frac{\lambda^{X_2}}{X_2!}e^{-\lambda}...\frac{\lambda^{X_n}}{X_n!}e^{-\lambda},$$

соответственно

$$\ln L_n(\lambda) = \ln(\lambda)\sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n}\ln(X_i!).$$

Получившаяся непрерывная на $(0, +\infty)$ функция стремится к $-\infty$ на краях этого интервала и, следовательно, должна иметь точку максимума внутри

интервала, где ее производная равна 0. Таким образом, точка максимума является решением уравнения

$$\frac{1}{\lambda}\sum_{i=1}^{n} X_i - n = 0.$$

Единственное решение этого уравнения $\widehat{\lambda}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ является оценкой максимального правдоподобия для параметра λ. Известно, что для распределения Пуассона $M(\xi) = \lambda$. Таким образом, показано, что в случае распределения Пуассона оценка математического ожидания $\overline{X}$ является оценкой максимального правдоподобия.

Остается рассмотреть другой вариант оценки максимального правдоподобия, когда распределение НСВ ξ является абсолютно непрерывным с функцией плотности $f(t, \theta)$. Тогда совместная плотность распределения всех элементов выборки

$$f(t_1, t_2, ..., t_n, \theta) = f(t_1, \theta) f(t_2, \theta) \cdots f(t_n, \theta).$$

Функция правдоподобия выборки $X_1, X_2, ..., X_n$ определяется в этом случае как

$$L_n(\theta) = f(X_1, \theta) f(X_2, \theta) \cdots f(X_n, \theta),$$

соответственно оценка максимального правдоподобия представляет собой точку максимума этой функции.

Пусть, например, известно, что распределение НСВ является нормальным. Тогда оно полностью определяется двумя параметрами – математическим ожиданием m и дисперсией σ^2. Ставится задача получения оценки максимального правдоподобия для векторного параметра, состоящего из этих параметров:

$$\mathbf{\Theta} = (\theta_1, \theta_2)^T = (m, \sigma^2)^T.$$

Плотность нормального распределения

$$f(t, \mathbf{\Theta}) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left\{-\frac{1}{2}\frac{(t-\theta_1)^2}{\theta_2}\right\}.$$

Соответственно, произведение таких плотностей в точках элементов выборки

$$L_n(\mathbf{\Theta}) = \frac{1}{(2\pi\theta_2)^{n/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(X_i-\theta_1)^2}{\theta_2}\right\}.$$

Натуральный логарифм этого выражения

$$\psi(\theta_1, \theta_2) = \ln L_n(\mathbf{\Theta}) = -\frac{n}{2}(\ln(2\pi) + \ln\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^{n}(X_i - \theta_1)^2.$$

Получившаяся непрерывная функция стремится к $-\infty$ при $\theta_1 \to \pm\infty$, $\theta_2 \to \infty$ и $\theta_2 \to 0$, следовательно, должна иметь точку максимума внутри области допустимых значений, где ее частные производные равны 0. Таким образом, точка максимума является решением системы уравнений

$$\begin{cases} \dfrac{\partial \psi}{\partial \theta_1} = \dfrac{1}{\theta_2} \displaystyle\sum_{i=1}^{n} (X_i - \theta_1) = 0, \\ \dfrac{\partial \psi}{\partial \theta_2} = -\dfrac{n}{2\theta_2} + \dfrac{1}{2{\theta_2}^2} \displaystyle\sum_{i=1}^{n} (X_i - \theta_1)^2 = 0. \end{cases}$$

Решение первого уравнения этой системы дает статистику для оценивания первого параметра

$$\widehat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}.$$

Если подставить это решение во второе уравнение, то из него получится статистика для оценивания второго параметра

$$\widehat{\theta}_2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = S^2.$$

Таким образом, система имеет единственное решение, которое и является требуемой оценкой максимального правдоподобия.

Итак, оценкой максимального правдоподобия для вектора параметров нормального распределения $(m, \sigma^2)^T$ является оценка, состоящая из соответствующих выборочных моментов:

$$\widehat{\boldsymbol{\Theta}} = (\overline{X}, S^2)^T.$$

Задача для самостоятельного решения. Построить оценку максимального правдоподобия для параметра λ экспоненциального распределения, т. е. для сосредоточенного на положительной полуоси $(t \geq 0)$ абсолютно непрерывного распределения с плотностью $f(t, \lambda) = \lambda \exp\{-\lambda t\}$, где $\lambda > 0$.

2.4. Неравенство Рао–Крамера. Условие эффективности оценки

Как и в 2.3, будет оцениваться параметр θ распределения, принадлежащий некоторому параметрическому семейству, причем будет рассматриваться только случай скалярного параметра.

Пусть функция $p(t, \theta)$ при каждом значении θ определяется соответствующим распределением НСВ ξ и в дискретном случае равна зависящей

от θ вероятности $P(\xi = t, \theta)$, а в случае абсолютно непрерывного распределения равна плотности распределения НСВ $f(t, \theta)$. Тогда введенная в 2.3 функция правдоподобия выборки в любом случае может быть записана в виде

$$L_n(\theta) = p(X_1, \theta)p(X_2, \theta) \cdots p(X_n, \theta).$$

При дополнительном предположении, что функция $p(t, \theta)$ дифференцируема по переменной θ, можно ввести величину, которая интерпретируется как количество информации о параметре θ, содержащееся в выборке.

Определение 2.5. *Функция оцениваемого параметра*

$$I_n(\theta) = M\left(\frac{\partial \ln L_n(\theta)}{\partial \theta}\right)^2$$

называется количеством информации о параметре θ, содержащимся в выборке $X_1,\ X_2,\ ...,\ X_n$, или информационным количеством Фишера.

Для исследования свойств $I_n(\theta)$ введем случайную величину

$$\psi(\xi) = \frac{\partial \ln p(\xi, \theta)}{\partial \theta} = \frac{1}{p(\xi, \theta)} \frac{\partial p(\xi, \theta)}{\partial \theta}.$$

Нетрудно убедиться, что эта случайная величина имеет нулевое математическое ожидание. Действительно, в дискретном случае, когда ξ может принимать значения $t_1,\ ...,\ t_m$,

$$M(\psi(\xi)) = \sum_{i=1}^{m} \frac{1}{p(t_i, \theta)} \frac{\partial p(t_i, \theta)}{\partial \theta} p(t_i, \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{m} p(t_i, \theta) = \frac{\partial}{\partial \theta} 1 = 0.$$

В непрерывном случае можно сделать аналогичные преобразования, заменив сумму на интеграл и использовав изменение порядка операций интегрирования и дифференцирования по параметру. (Это изменение порядка допустимо при определенных дополнительных ограничениях на функцию $p(t, \theta)$, которые здесь будут считаться выполненными.)

Используя свойства логарифма, информационное количество Фишера можно записать в виде

$$I_n(\theta) = M\left(\sum_{i=1}^{n} \frac{\partial \ln p(X_i, \theta)}{\partial \theta}\right)^2 = M\left(\sum_{i=1}^{n} \psi(X_i)\right)^2.$$

Поскольку случайные величины $\psi(X_i)$, определяемые разными элементами выборки, являются независимыми случайными величинами с нулевыми математическими ожиданиями,

$$I_n(\theta) = D\left(\sum_{i=1}^{n} \psi(X_i)\right) = \sum_{i=1}^{n} D(\psi(X_i)) = nD(\psi(\xi)).$$

Таким образом, если ввести величину

$$I_1(\theta) = M\left(\frac{\partial \ln p(\xi,\theta)}{\partial\theta}\right)^2 = D(\psi(\xi)),$$

которую по аналогии с $I_n(\theta)$ следует назвать количеством информации о параметре θ, содержащимся в одном элементе выборки, то справедливо равенство $I_n(\theta) = nI_1(\theta)$. Таким образом, введенная функция, определяющая количество информации, обладает естественным свойством аддитивности: количество информации о параметре, содержащееся в выборке, складывается из количеств информации, содержащихся в каждом ее элементе.

Итак, при увеличении объема выборки, с одной стороны, увеличивается количество полученной информации об оцениваемом параметре, а с другой – должно повышаться качество состоятельной оценки параметра, т. е. уменьшаться ее дисперсия. Оказывается, эти две величины могут быть связаны простым неравенством, определяющим нижнюю границу для дисперсии несмещенной оценки. Это неравенство называется *неравенством Рао–Крамера*, соответствующее утверждение можно сформулировать следующим образом.

Теорема 2.7. *Пусть $\widehat{\theta}_n$ – произвольная несмещенная оценка параметра θ. При дополнительных ограничениях на функцию $p(t,\theta)$ справедливо неравенство*

$$D(\widehat{\theta}_n) \geq \frac{1}{I_n(\theta)}.$$

Упомянутые здесь дополнительные ограничения на функцию $p(t,\theta)$ должны позволять изменять порядок операций интегрирования (суммирования) и дифференцирования по параметру θ в явно выписанном выражении для $I_n(\theta)$.

Доказательство неравенства Рао–Крамера основывается на известном неравенстве Коши–Буняковского. Если в некотором линейном пространстве введены скалярное произведение элементов (x,y), имеющее определенный стандартный набор свойств, и норма, определяемая равенством $||x||^2 = (x,x)$, то $(x,y)^2 \leq ||x||^2||y||^2$. Это неравенство обращается в равенство тогда и только тогда, когда элемент y получается из x умножением на число.

В частности, в случае m-мерного евклидова пространства $\mathbb{R}^m$ неравенство Коши–Буняковского представляет собой неравенство для сумм:

$$\left(\sum_{j=1}^m a_j b_j\right)^2 \leq \sum_{j=1}^m a_j^2 \sum_{j=1}^m b_j^2.$$

В случае пространства функций $L_2(G)$ (пространство функций, для которых существует интеграл от их квадрата по области G), где скалярное произведение вводится как интеграл от произведения функций, получается неравенство для интегралов

$$\left(\int\limits_G f(x)g(x)dx\right)^2 \leq \int\limits_G f^2(x)dx \int\limits_G g^2(x)dx.$$

Доказательство неравенства Рао–Крамера. Доказательство будет проводиться для случая дискретного распределения с конечным числом возможных значений, не требующего никаких дополнительных ограничений на функцию $p(t,\theta)$, кроме дифференцируемости по θ. В случае абсолютно непрерывного распределения доказательство проходит точно так же с заменой всех сумм на интегралы. В дальнейших выкладках знак $\sum$ будет означать суммирование по n переменным $t_1,\ t_2,\ ...,\ t_n$, каждая из которых принимает все возможные значения дискретной НСВ. Аналог функции правдоподобия $L_n(\theta)$, где элементы выборки заменены на заданные числовые значения $t_1,\ ...,\ t_n$, будет обозначаться как $L_n(t_1,...,t_n,\theta)$. Напомним, что в дискретном случае эта величина представляет собой вероятность того, что элементы выборки примут заданные значения.

Если продифференцировать по θ очевидное равенство

$$\sum L_n(t_1,...,t_n,\theta) = 1,$$

то получится

$$\sum\left(\frac{\partial L_n(t_1,...,t_n,\theta)}{\partial\theta}\right) = 0.$$

С другой стороны, по определению несмещенной оценки

$$M(\widehat{\theta}_n) = \sum(\widehat{\theta}(t_1,...,t_n)L_n(t_1,...,t_n,\theta)) = \theta.$$

Продифференцировав последнее равенство по θ, получим

$$\sum\left(\widehat{\theta}(t_1,...,t_n)\frac{\partial L_n(t_1,...,t_n,\theta)}{\partial\theta}\right) = 1.$$

Из получившихся равенств следует, что

$$\sum\left((\widehat{\theta}(t_1,...,t_n)-\theta)\frac{\partial L_n(t_1,...,t_n,\theta)}{\partial\theta}\right) = 1.$$

Если умножить и поделить выражение под знаком суммы в последнем равенстве на $L_n(t_1,...,t_n,\theta)$, то это равенство можно переписать в виде

$$\sum(\widehat{\theta}(t_1,...,t_n)-\theta)\sqrt{L_n(t_1,...,t_n,\theta)}\,\frac{\partial \ln L_n(t_1,...,t_n,\theta)}{\partial\theta}\times$$

$$\times \sqrt{L_n(t_1, ..., t_n, \theta)} = 1.$$

Если обе части этого равенства возвести в квадрат и применить неравенство Коши–Буняковского, получится

$$1 \leq \sum (\widehat{\theta}(t_1, ..., t_n) - \theta)^2 L_n(t_1, ..., t_n, \theta) \sum \left(\frac{\partial \ln L_n(t_1, ..., t_n, \theta)}{\partial \theta} \right)^2 \times$$

$$\times L_n(t_1, ..., t_n, \theta) = D(\widehat{\theta}_n) I_n(\theta),$$

откуда следует неравенство Рао–Крамера.

Если в неравенстве Рао–Крамера при всех θ достигается равенство, то соответствующая состоятельная оценка не может быть улучшена и, соответственно, является эффективной. В неравенстве Коши–Буняковского равенство достигается тогда и только тогда, когда 2 вектора (функции) отличаются только числовым множителем. Из проведенного доказательства видно, что это условие, обращающее неравенство Рао–Крамера в равенство, выполнятся, если справедливо тождество

$$\frac{\partial \ln L_n(t_1, ..., t_n, \theta)}{\partial \theta} = A(\theta)(\widehat{\theta}(t_1, ..., t_n) - \theta),$$

где функция $A(\theta)$ зависит только от θ. Если в этом тождестве подставить вместо аргументов элементы выборки, возвести обе части в квадрат и взять математические ожидания, то получится равенство $I_n(\theta) = A^2(\theta) D(\widehat{\theta}_n)$. Далее, если использовать неравенство Рао–Крамера, которое в данном случае превращается в равенство, получится $D(\widehat{\theta}_n) = \dfrac{1}{|A(\theta)|}$.

Можно также переписать левую часть полученного тождества в более удобном виде, представив стоящий в левой части логарифм произведения в виде суммы логарифмов. Таким образом, доказано следующее утверждение.

Теорема 2.8. *Если для некоторой несмещенной оценки параметра* $\widehat{\theta}_n = \widehat{\theta}(X_1, X_2, ..., X_n)$ *существует такая функция* $A(\theta)$, *что*

$$\sum_{i=1}^{n} \frac{\partial \ln p(X_i, \theta)}{\partial \theta} = A(\theta)(\widehat{\theta}_n - \theta),$$

то эта оценка является эффективной, причем ее дисперсия $D(\widehat{\theta}_n) = \dfrac{1}{|A(\theta)|}$.

Можно также показать, что если для параметра θ существует эффективная оценка, то задача нахождения оценки максимального правдоподобия имеет единственное решение, совпадающее с этой оценкой.

Пример. Пусть НСВ имеет нормальное распределение с известной дисперсией σ^2 и неизвестным математическим ожиданием θ. Ранее было установлено, что состоятельной несмещенной оценкой этого параметра является среднее арифметическое элементов выборки $\overline{X}$. Покажем, что в данном случае эта точечная оценка является также эффективной.

Здесь функция $p(t,\theta)$ представляет собой плотность нормального распределения

$$p(t,\theta) = \frac{1}{\sqrt{2\pi}\sigma}\,\exp\left\{-\frac{1}{2}\,\frac{(t-\theta)^2}{\sigma^2}\right\}.$$

Соответственно,

$$\ln p(X_i,\theta) = -\ln(\sqrt{2\pi}\sigma) - \frac{(X_i-\theta)^2}{2\sigma^2},$$

следовательно,

$$\sum_{i=1}^{n}\frac{\partial \ln p(X_i,\theta)}{\partial\theta} = \sum_{i=1}^{n}\frac{X_i-\theta}{\sigma^2} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}X_i - n\theta\right) = \frac{n}{\sigma^2}(\overline{X}-\theta).$$

Таким образом, для несмещенной оценки математического ожидания $\overline{X}$ выполняется условие теоремы 2.8, где $A(\theta) = \dfrac{n}{\sigma^2}$, следовательно, данная оценка является эффективной. Информационное количество Фишера $I_n(\theta) = |A(\theta)| = \dfrac{n}{\sigma^2}$, соответственно $D(\overline{X}) = \dfrac{\sigma^2}{n}$. Последнее равенство нетрудно также проверить непосредственно, исходя из свойств дисперсии.

2.5. Случай неоднородной выборки. Метод наименьших квадратов

Рассмотрим понятие выборки в более широком смысле: предполагается, что она представляет собой набор из n независимых случайных величин X_i не обязательно с одинаковыми распределениями, которые связаны только общими оцениваемыми параметрами.

Пусть, например, одна и та же неизвестная величина θ измеряется n различными способами. Предполагается, что каждое измерение (элемент неоднородной выборки) можно представить в виде $X_i = \theta + \xi_i$, где ξ_i – независимые случайные величины, $M(\xi_i) = 0$, $D(\xi_i) = \sigma_i^2$. Требуется построить линейную несмещенную оценку $\widehat{\theta}_n$ величины θ, минимизирующую $M(\widehat{\theta}_n - \theta)^2$. Линейность оценки означает, что статистика строится в виде $\widehat{\theta}_n = \sum\limits_{i=1}^{n} c_i X_i$, при этом, поскольку $M(X_i) = \theta$, условие несмещенности

требует выполнения дополнительного условия $\sum_{i=1}^{n} c_i = 1$.

Итак, требуется найти минимум по $\boldsymbol{c} = (c_1, c_2, ..., c_n)^T$ (при дополнительном условии $\sum_{i=1}^{n} c_i = 1$) функции

$$\varphi(\boldsymbol{c}) = M\left(\sum_{i=1}^{n} c_i X_i - \theta\right)^2 = M\left(\sum_{i=1}^{n} c_i X_i - \theta \sum_{i=1}^{n} c_i\right)^2 =$$
$$= M\left(\sum_{i=1}^{n} c_i (X_i - \theta)\right)^2 = M\left(\sum_{i=1}^{n} c_i \xi_i\right)^2.$$

Поскольку ξ_i представляют собой независимые случайные величины с нулевыми математическими ожиданиями,

$$\varphi(\boldsymbol{c}) = D\left(\sum_{i=1}^{n} c_i \xi_i\right) = \sum_{i=1}^{n} c_i^2 \sigma_i^2.$$

Условие $\sum_{i=1}^{n} c_i = 1$ позволяет исключить из получившегося выражения последнюю переменную c_n. Таким образом, нужно минимизировать по c_i выражение

$$\sum_{i=1}^{n-1} c_i^2 \sigma_i^2 + \left(1 - \sum_{i=1}^{n-1} c_i\right)^2 \sigma_n^2.$$

Данное выражение является положительным, фактически оно представляет собой положительно-определенную квадратичную форму и поэтому имеет единственную точку минимума. Для нахождения этой точки нужно приравнять к нулю частные производные этого выражения по всем c_i. В результате получится система линейных уравнений

$$2c_i \sigma_i^2 - 2\left(1 - \sum_{j=1}^{n-1} c_j\right) \sigma_n^2 = 0 \quad (i = 1, ..., n-1),$$

или

$$c_i \sigma_i^2 = c_n \sigma_n^2 = b,$$

где b – некоторая константа, не зависящая от i.

Таким образом, получившиеся значения коэффициентов обратно пропорциональны соответствующим дисперсиям $c_i = \dfrac{b}{\sigma_i^2}$. При этом из условия

$\sum_{i=1}^{n} c_i = 1$ следует, что

$$b = \frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}.$$

Итак, требуемая линейная несмещенная оценка имеет вид

$$\widehat{\theta}_n = \frac{\sum_{i=1}^{n} \frac{X_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}.$$

(В случае одинаковых дисперсий σ_i^2 эта статистика совпадает с классической состоятельной оценкой математического ожидания – средним арифметическим $\overline{X}$.)

Нетрудно проверить, что эта же статистика получается как точка минимума суммы квадратов нормированных отклонений элементов выборки от оцениваемой величины $\sum_{i=1}^{n} \left(\frac{X_i - \theta}{\sigma_i} \right)^2$. Таким образом, она является также оценкой *метода наименьших квадратов* (МНК).

В более общем виде задача, решаемая МНК, может быть поставлена следующим образом. Пусть $\mathbf{\Theta} = (\theta_1, \theta_2, ..., \theta_m)^T$ – оцениваемый вектор параметров, а элементы неоднородной выборки (измерения) являются линейными функциями $\mathbf{\Theta}$, на которые накладываются независимые случайные ошибки:

$$X_i = \sum_{j=1}^{m} b_{ij}\theta_j + \xi_i,$$

где $M(\xi_i) = 0$ и $D(\xi_i) = \sigma_i^2$. Оценка векторного параметра $\mathbf{\Theta}$ по МНК получается минимизацией суммы квадратов нормированных невязок измерений

$$\varphi(\mathbf{\Theta}) = \sum_{i=1}^{n} \left(\frac{X_i - \sum_{j=1}^{m} b_{ij}\theta_j}{\sigma_i} \right)^2.$$

Можно заметить, что к минимизации той же функции $\varphi(\mathbf{\Theta})$ сводится задача построения для $\mathbf{\Theta}$ оценки максимального правдоподобия, если

дополнительно предположить, что случайные ошибки имеют нормальное распределение.

Для определения точки минимума $\varphi(\mathbf{\Theta})$ нужно приравнять к нулю все ее частные производные:

$$\frac{\partial\varphi(\mathbf{\Theta})}{\partial\theta_k}=\sum_{i=1}^{n}\frac{2\left(X_i-\sum\limits_{j=1}^{m}b_{ij}\theta_j\right)}{\sigma_i^2}(-b_{ik})\qquad(k=1,...,m).$$

В результате получится система линейных уравнений

$$\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\frac{b_{ij}b_{ik}}{\sigma_i^2}\right)\theta_j=\sum_{i=1}^{n}\frac{b_{ik}X_i}{\sigma_i^2}\qquad(k=1,...,m).$$

Если ввести матрицу H размера $n\times m$ с элементами $h_{ij}=\dfrac{b_{ij}}{\sigma_i}$ и вектор нормированных измерений $\boldsymbol{Z}$ с элементами $z_i=\dfrac{X_i}{\sigma_i}$, то получившуюся систему можно записать в виде

$$(H^TH)\mathbf{\Theta}=H^T\boldsymbol{Z}.$$

Эта система носит название нормальной системы уравнений МНК. Можно показать, что если матрица H содержит m линейно независимых строк, матрица H^TH будет невырожденной и данная система линейных уравнений имеет единственное решение, которое и дает оценку МНК для $\mathbf{\Theta}$. Таким образом, оценка МНК

$$\widehat{\mathbf{\Theta}}_n=(H^TH)^{-1}H^T\boldsymbol{Z}.$$

Нетрудно также найти ковариационную матрицу оценки МНК. В силу независимости элементов выборки вектор нормированных измерений $\boldsymbol{Z}$ имеет единичную ковариационную матрицу I. Из свойств ковариационной матрицы, сформулированных в 1.1, следует, что ковариационная матрица оценки $\widehat{\Theta}_n$ может быть записана в виде

$$R=(H^TH)^{-1}H^TI((H^TH)^{-1}H^T)^T=(H^TH)^{-1}H^TH((H^TH)^{-1})^T=(H^TH)^{-1},$$

поскольку матрица H^TH и, соответственно, $(H^TH)^{-1}$ очевидно являются симметричными.

Покажем теперь, что получившаяся оценка МНК является несмещенной. Согласно принятой модели измерений математическое ожидание каждого измерения $M(X_i)=\sum\limits_{j=1}^{m}b_{ij}\theta_j$, соответственно для каждой компоненты

нормированного вектора измерений

$$M(z_i) = \frac{M(X_i)}{\sigma_i} = \sum_{j=1}^{m} h_{ij}\theta_j.$$

Таким образом, математическое ожидание вектора $\boldsymbol{Z}$ может быть записано в виде $M(\boldsymbol{Z}) = H\boldsymbol{\Theta}$. Поскольку оценка МНК является решением системы нормальных уравнений, справедливо равенство

$$(H^T H)\widehat{\boldsymbol{\Theta}}_n = H^T \boldsymbol{Z}.$$

Взяв математическое ожидание от обеих частей этого равенства и воспользовавшись свойством линейности математического ожидания, получим

$$(H^T H)M(\widehat{\boldsymbol{\Theta}}_n) = H^T H\boldsymbol{\Theta},$$

следовательно,

$$M(\widehat{\boldsymbol{\Theta}}_n) = (H^T H)^{-1}(H^T H)\boldsymbol{\Theta} = \boldsymbol{\Theta},$$

что и означает несмещенность оценки $\widehat{\boldsymbol{\Theta}}_n$.

Итак, на основании вышеизложенного можно сделать следующие выводы.

Теорема 2.9. *Оценка МНК, которая получена решением системы нормальных уравнений, является несмещенной оценкой вектора оцениваемых параметров, а ее ковариационная матрица является обратной матрицей к матрице системы. Сумма диагональных элементов этой ковариационной матрицы является минимально возможной в классе линейных оценок вектора параметров. В случае нормального распределения ошибок измерений оценка МНК является эффективной.*

Одним из применений МНК является решение задачи построения сглаживающей функции для измерений. Пусть в моменты времени $t_1, t_2, ..., t_n$ измеряется некоторая неизвестная зависящая от времени величина $f(t)$. Каждое измерение может быть представлено в виде $X_i = f(t_i) + \xi_i$, где ξ_i – случайные ошибки измерений, $M(\xi_i) = 0$ и $D(\xi_i) = \sigma_i^2$. Если предположить, что неизвестная функция $f(t)$ представляется в виде линейной комбинации некоторых гладких базисных функций: $f(t) = \sum_{j=1}^{m} \theta_j \psi_j(t)$, то задача сводится к определению коэффициентов θ_j, таких, что получившаяся функция наилучшим образом приближается к измерениям, т. е.,

согласно МНК, минимизируется величина

$$\sum_{i=1}^{n} \left(\frac{X_i - \sum_{j=1}^{m} \theta_j \psi_j(t_i)}{\sigma_i} \right)^2 .$$

Таким образом, данная задача является частным случаем описанной ранее задачи МНК, где коэффициенты $b_{ij} = \psi_j(t_i)$, соответственно матрица H состоит из элементов $h_{ij} = \dfrac{\psi_j(t_i)}{\sigma_i}$.

Задача для самостоятельного решения. Получить в явном виде выражения для оценки МНК параметров θ_0 и θ_1 для случая линейной сглаживающей функции измерений $f(t) = \theta_0 + \theta_1(t - t_0)$ и одинаковых дисперсий измерений $\sigma_i^2 = \sigma^2$. Выписать также ковариационную матрицу получившейся оценки вектора коэффициентов. (В качестве значения t_0 удобно выбрать среднее арифметическое моментов t_i.)

2.6. Линейная регрессия

В приложениях статистики часто требуется оценить характер зависимости между двумя совместно наблюдаемыми случайными величинами ξ и η. Обычно предполагается, что эта зависимость может быть описана в виде функции, линейно зависящей от подлежащего оценке неизвестного вектора параметров $\boldsymbol{\Theta} = (\theta_1, \theta_2, ..., \theta_m)^T$. Такая зависимость может быть записана в виде

$$p(\eta) = \sum_{j=1}^{m} \theta_j \psi_j(\xi),$$

где $p(x)$ и $\psi_j(x)$ – любые функции.

Пусть $[(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)]^T$ – двумерная выборка из распределения наблюдаемого случайного вектора $(\xi, \eta)^T$. Для отдельных элементов выборки описанная зависимость содержит случайные ошибки, точнее, предполагается, что

$$p(Y_i) = \sum_{j=1}^{m} \theta_j \psi_j(X_i) + \nu_i,$$

где ν_i – независимые случайные величины с $M(\nu_i) = 0$ и одинаковыми конечными дисперсиями $D(\nu_i) = \sigma^2$. Такая модель называется *линейной моделью регрессии.*

Для оценивания параметров регрессии используется описанный в 2.5 метод наименьших квадратов, т. е. вектор параметров находится минимизацией выражения

$$\sum_{i=1}^{n}\left(p(Y_i)-\sum_{j=1}^{m}\theta_j\psi_j(X_i)\right)^2.$$

Эта задача аналогична описанной в 2.5 задаче сглаживания измерений. Отличие состоит лишь в том, что вместо измерений X_i берутся значения $p(Y_i)$, времена измерений t_i заменяются на выборочные значения X_i и отсутствует нормировка (деление на σ_i), поскольку при одинаковых дисперсиях ошибок в этом нет необходимости. Таким образом, требуемая оценка вектора параметров может быть найдена как решение системы линейных уравнений

$$(H^TH)\widehat{\boldsymbol{\Theta}}_n = H^T\boldsymbol{p}(\boldsymbol{Y}),$$

где вектор $\boldsymbol{p}(\boldsymbol{Y}) = (p(Y_1), p(Y_2), ..., p(Y_n))^T$, а матрица H состоит из элементов $h_{ij} = \psi_j(X_i)$.

В описанной модели в качестве функции $p(y)$ чаще всего используются y, $\dfrac{1}{y}$, $\ln y$. В качестве функций $\psi_j(x)$ чаще всего используются степени x. Случай, когда $p(y) = y$, а в качестве $\psi_j(x)$ используются степени x либо полиномы от x, называется *полиномиальной регрессией*.

Простейшим и наиболее часто используемым случаем полиномиальной регрессии является *простая линейная регрессия*, когда предполагается линейная связь НСВ:

$$\eta = \theta_0 + \theta_1\xi.$$

Параметр θ_1 в этой модели называется *коэффициентом линейной регрессии*. В данном случае нетрудно получить статистики для оценивания параметров θ_0 и θ_1 в явном виде.

Действительно, в случае простой линейной регрессии матрица H имеет размер $n \times 2$, ее элементы $h_{i1} = 1$, $h_{i2} = X_i$. Таким образом, матрица системы линейных уравнений

$$H^TH = \begin{pmatrix} n & \sum\limits_{i=1}^{n} X_i \\ \sum\limits_{i=1}^{n} X_i & \sum\limits_{i=1}^{n} X_i^2 \end{pmatrix},$$

а вектор правых частей

$$H^T \boldsymbol{Y} = \begin{pmatrix} \sum\limits_{i=1}^{n} Y_i \\ \sum\limits_{i=1}^{n} X_i Y_i \end{pmatrix}.$$

Если поделить обе части получившейся системы на n и использовать обозначения $\widehat{\alpha}_{X2} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i^2$ и $\widehat{\alpha}_{XY} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i Y_i$, то систему можно записать в виде

$$\begin{pmatrix} 1 & \overline{X} \\ \overline{X} & \widehat{\alpha}_{X2} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} = \begin{pmatrix} \overline{Y} \\ \widehat{\alpha}_{XY} \end{pmatrix}.$$

Определитель матрицы этой системы $\Delta = \widehat{\alpha}_{X2} - \overline{X}^2 = S_X^2$ представляет собой выборочную дисперсию выборки $X_1, X_2, ..., X_n$. Таким образом, если выписать решение данной системы по формулам Крамера, получится, что для оценивания свободного члена служит статистика

$$\widehat{\theta}_0 = \frac{\overline{Y}\widehat{\alpha}_{X2} - \overline{X}\widehat{\alpha}_{XY}}{S_X^2},$$

а для оценивания коэффициента регрессии – статистика

$$\widehat{\theta}_1 = \frac{\widehat{\alpha}_{XY} - \overline{XY}}{S_X^2} = \frac{\widehat{\mu}_{XY}}{S_X^2}.$$

Из формулы для вычисления ковариационного момента, приведенной в 1.1, видно, что статистика $\widehat{\mu}_{XY}$ представляет собой выборочный ковариационный момент двумерной выборки из распределения наблюдаемого случайного вектора $(\xi, \eta)^T$. Изначально эта величина определяется формулой

$$\widehat{\mu}_{XY} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}).$$

Таким образом, оценка коэффициента линейной регрессии представляет собой отношение выборочного ковариационного момента к выборочной дисперсии первой составляющей выборки.

2.7. Применение МНК к нелинейной модели измерений

Рассмотрим применение принципа МНК к случаю нелинейной зависимости измерений от вектора оцениваемых параметров:

$$X_i = h_i(\boldsymbol{\Theta}) + \xi_i,$$

где $h_i(\boldsymbol{x})$ – некоторые дифференцируемые функции m переменных, а независимые случайные величины ξ_i по-прежнему имеют $M(\xi_i) = 0$ и $D(\xi_i) = \sigma_i^2$. Согласно принципу МНК минимизируется сумма квадратов нормированных невязок измерений

$$\varphi(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \left(\frac{X_i - h_i(\boldsymbol{\Theta})}{\sigma_i} \right)^2.$$

Для определения точки минимума этой функции следовало бы приравнять к нулю все ее частные производные:

$$\frac{\partial \varphi(\boldsymbol{\Theta})}{\partial \theta_k} = \sum_{i=1}^{n} \frac{2(X_i - h_i(\boldsymbol{\Theta}))}{\sigma_i^2} \left(-\frac{\partial h_i}{\partial \theta_k}(\boldsymbol{\Theta}) \right) \quad (k = 1, ..., m).$$

Однако получившаяся система будет нелинейной, ее решение может оказаться трудной задачей.

Предположим, что для вектора оцениваемых параметров $\boldsymbol{\Theta}$ имеется некоторая приближенная оценка $\boldsymbol{\Theta}_0$. Тогда приращения функций $h_i(\boldsymbol{\Theta})$ за счет изменения компонент $\boldsymbol{\Theta}$ можно приближенно представить в виде линейных выражений

$$h_i(\boldsymbol{\Theta}) - h_i(\boldsymbol{\Theta}_0) \approx \sum_{j=1}^{m} \frac{\partial h_i}{\partial \theta_j}(\boldsymbol{\Theta}_0)(\theta_j - \theta_{0j}).$$

Если предположить, что производные функций $h_i(\boldsymbol{\Theta})$ не слишком быстро меняются в окрестности $\boldsymbol{\Theta}_0$, можно считать, что $\dfrac{\partial h_i}{\partial \theta_k}(\boldsymbol{\Theta}) \approx \dfrac{\partial h_i}{\partial \theta_k}(\boldsymbol{\Theta}_0)$. Тогда при всех $\boldsymbol{\Theta}$, достаточно близких к $\boldsymbol{\Theta}_0$, справедливы приближенные равенства

$$\frac{\partial \varphi(\boldsymbol{\Theta})}{\partial \theta_k} \approx \sum_{i=1}^{n} \frac{2 \left((X_i - h_i(\boldsymbol{\Theta}_0)) - \sum_{j=1}^{m} \dfrac{\partial h_i}{\partial \theta_j}(\boldsymbol{\Theta}_0)(\theta_j - \theta_{0j}) \right)}{\sigma_i^2} \left(-\frac{\partial h_i}{\partial \theta_k}(\boldsymbol{\Theta}_0) \right).$$

Соответственно, приблизительное значение вектора параметров $\boldsymbol{\Theta}$, более близкое к точке минимума $\varphi(\boldsymbol{\Theta})$, чем $\boldsymbol{\Theta}_0$, может быть найдено как решение системы линейных уравнений

$$\sum_{j=1}^{m} \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \frac{\partial h_i}{\partial \theta_j}(\boldsymbol{\Theta}_0) \frac{\partial h_i}{\partial \theta_k}(\boldsymbol{\Theta}_0) \right) (\theta_j - \theta_{0j}) = \sum_{i=1}^{n} \frac{\partial h_i}{\partial \theta_k}(\boldsymbol{\Theta}_0) \frac{X_i - h_i(\boldsymbol{\Theta}_0)}{\sigma_i^2}$$

$$(k = 1, ..., m).$$

Если ввести матрицу $H(\mathbf{\Theta})$ размером $n \times m$ с элементами

$$h_{ij}(\mathbf{\Theta}) = \frac{1}{\sigma_i} \frac{\partial h_i}{\partial \theta_j}(\mathbf{\Theta})$$

и вектор нормированных невязок измерений $\boldsymbol{Z}(\mathbf{\Theta})$ с элементами

$$z_i(\mathbf{\Theta}) = \frac{X_i - h_i(\mathbf{\Theta})}{\sigma_i},$$

то получившуюся систему можно записать в виде

$$(H^T(\mathbf{\Theta}_0)H(\mathbf{\Theta}_0))(\mathbf{\Theta} - \mathbf{\Theta}_0) = H^T(\mathbf{\Theta}_0)\boldsymbol{Z}(\mathbf{\Theta}_0).$$

Решив эту систему, можно найти вектор параметров $\mathbf{\Theta}$, являющийся более точным приближением к оценке МНК, чем $\mathbf{\Theta}_0$.

Таким образом, для получения практически точного значения оценки МНК можно организовать итерационный процесс

$$\mathbf{\Theta}_{k+1} = \mathbf{\Theta}_k + (H^T(\mathbf{\Theta}_k)H(\mathbf{\Theta}_k))^{-1}H^T(\mathbf{\Theta}_k)\boldsymbol{Z}(\mathbf{\Theta}_k).$$

Аналогично линейному случаю, ковариационную матрицу получившейся оценки можно найти по приближенной формуле

$$R \approx (H^T(\theta_k)H(\mathbf{\Theta}_k))^{-1}.$$

Итерационный процесс имеет смысл завершить, если для каждой компоненты вектора оцениваемых параметров абсолютная величина приращения $|\theta_{kj} - \theta_{(k-1)j}|$ становится намного меньше, чем СКО, определяемое по соответствующему диагональному элементу ковариационной матрицы $\sigma_j = \sqrt{r_{jj}}$.

Пример. Требуется определить с максимальной точностью координаты объекта $\mathbf{\Theta} = (\theta_1, \theta_2, \theta_3)^T$ в некоторой прямоугольной системе координат по измерениям расстояний от n измерителей с координатами

$$\boldsymbol{r}_i = (r_{i1}, r_{i2}, r_{i3})^T \quad (i = 1, 2, ..., n).$$

(Подобная задача решается, в частности, в приборах спутниковой навигации.) Если пренебречь ошибками измерений, задача записывается в виде системы уравнений

$$h_i(\mathbf{\Theta}) = \sqrt{\sum_{k=1}^{3} (\theta_k - r_{ik})^2} = X_i \quad (i = 1, ..., n).$$

При $n < 3$ для решения данной задачи не хватает данных (число уравнений меньше числа неизвестных).

При $n = 3$ задача записывается в виде системы из трех квадратных уравнений с тремя неизвестными, которая сводится к одному квадратному уравнению и соответственно имеет 2 решения. Обычно не составляет труда выбрать из этих решений правильное, если известна приблизительная область нахождения объекта.

При $n > 3$ уравнений становится больше, чем неизвестных (система переопределенная), а из-за наличия ошибок в измерениях она становится несовместной. Именно в этом случае для получения наиболее точной оценки координат актуально использование МНК.

Если обозначить через σ_i СКО измерения расстояния до соответствующего измерителя, то матрица $H(\mathbf{\Theta})$ будет иметь элементы

$$h_{ij}(\mathbf{\Theta}) = \frac{1}{\sigma_i}\frac{\partial h_i}{\partial \theta_j}(\mathbf{\Theta}) = \frac{\theta_j - r_{ij}}{\sigma_i \sqrt{\sum_{k=1}^{3}(\theta_k - r_{ik})^2}}.$$

Описанный выше итерационный процесс уточнения оценки можно начать с оценки координат $\mathbf{\Theta}_0$, полученной по любым трем измерениям, как было описано для случая $n = 3$.

Отметим, что оценка МНК может быть найдена лишь в том случае, когда матрица системы нормальных уравнений не слишком близка к вырожденной. В данной задаче это означает, что измерители не находятся вблизи одной прямой и что имеется хотя бы 3 измерителя, расстояния между которыми не слишком малы по сравнению с расстояниями от объекта.

2.8. Робастные оценки

В дословном переводе термин „робастный“ означает крепкий, надежный. Статистику принято называть робастной оценкой параметра, если ее значения обладают определенной устойчивостью (нечувствительностью) к некоторым отклонениям истинного распределения НСВ от предполагаемого семейства распределений. На практике такие отклонения часто можно наблюдать в виде „выбросов“ в значениях НСВ, зарегистрированных в выборке. Такие выбросы можно представить в виде следующей модели: с некоторой небольшой вероятностью НСВ имеет распределение, отличающееся от предполагаемого гораздо большей дисперсией.

Можно выделить 2 типа робастных оценок: L-оценки и M-оценки.

L-оценки представляют собой линейные комбинации членов вариационного ряда (см. 1.4):

$$\sum_{i=1}^{n} \omega_i X_{n_i},$$

где весовые коэффициенты ω_i задаются с помощью специальной функции $\lambda(t)$, определенной на $(0,1)$:

$$\omega_i = \frac{1}{n}\lambda\left(\frac{i}{n+1}\right) \qquad (i = 1, ..., n).$$

Снижение чувствительности оценки к отдельным выбросам, которые всегда оказываются крайними членами вариационного ряда, достигается за счет того, что $\lambda(t)$ принимает на краях промежутка $(0,1)$ меньшие значения, чем в его середине.

Примером L-оценки может служить оценка математического ожидания с помощью усеченного среднего. Пусть число $\alpha \in (0, 0.5)$ представляет собой предполагаемую долю аномальных элементов (выбросов) в общем объеме выборки и пусть натуральное число $k = [n\alpha]$ (целая часть числа). Если ввести кусочно-постоянную функцию, определяемую равенством $\lambda(t) = \dfrac{n}{n-2k}$ на интервале $(\alpha, 1-\alpha)$ и равную нулю вне этого интервала, то применение формулы L-оценки даст среднее арифметическое элементов выборки с отброшенными крайними членами:

$$\overline{X}_\alpha = \frac{1}{n-2k}(X_{n_{k+1}} + ... + X_{n_{n-k}}).$$

M-оценки, подобно оценкам МНК, получаются минимизацией суммы значений некоторой положительной функции, вычисленной в точках невязок. В частности, для выборки, представляющей собой n независимых однотипных измерений неизвестной величины θ, M-оценка этой величины получается минимизацией выражения

$$\sum_{i=1}^{n} \rho(X_i - \theta),$$

где $\rho(x)$ – неотрицательная четная функция, возрастающая на положительной полуоси медленнее, чем x^2, и такая, что $\rho(0) = 0$. Если положить $\rho(x) = x^2$, то минимизация данного выражения дает классическую оценку МНК – среднее арифметическое $\overline{X}$. Условие более медленного возрастания $\rho(x)$ дает большую устойчивость к отдельным выбросам. Пусть, например, $\rho(x) = |x|$. Можно показать, что в этом случае требуемой статистикой является *выборочная медиана* $\widehat{\text{med}}$ (см. 1.4). Эта статистика вообще нечувствительна к отдельным выбросам.

Обычно на $\rho(x)$ накладывается еще условие строгой выпуклости, которое обеспечивает единственность решения задачи минимизации. Если при этом $\rho(x)$ имеет производную $\rho'(x)$, то требуемая робастная оценка $\widehat{\theta}_n$ яв-

ляется решением уравнения

$$\sum_{i=1}^{n} \rho'(X_i - \widehat{\theta}_n) = 0.$$

Например, в качестве $\rho(x)$ можно использовать *функцию Хубера*

$$\rho(x) = \begin{cases} \dfrac{x^2}{2}, & |x| \le a; \\ a|x| - \dfrac{a^2}{2}, & |x| > a. \end{cases}$$

Легко проверить, что эта функция непрерывна и имеет непрерывную производную

$$\rho'(x) = \begin{cases} -a, & x < -a; \\ x, & |x| \le a; \\ a, & x > a. \end{cases}$$

Однако уравнение для определения $\widehat{\theta}_n$ здесь не удается решить в явном виде, как в случае классического МНК, поскольку явный вид выражения для $\rho'(x - \theta)$ зависит от θ. Для решения уравнения приходится применять метод последовательных приближений, в качестве начального приближения можно использовать классическую оценку МНК $\overline{X}$.

2.9. Доверительные интервалы для параметров распределений

Для полного решения задачи оценивания неизвестного параметра распределения нужно не только найти состоятельную точечную оценку этого параметра, но и определить ее точность, построив с использованием этой точечной оценки доверительный интервал с заданной доверительной вероятностью (см. 2.1). Естественным свойством такого доверительного интервала является неограниченное уменьшение его длины при увеличении объема выборки.

Простейшим примером доверительного интервала может служить доверительный интервал для математического ожидания при известном значениии дисперсии. Пусть для НСВ ξ с известным значением $D(\xi) = \sigma^2$ требуется построить доверительный интервал для параметра $\theta = M(\xi)$ с заданной доверительной вероятностью β. Как было показано ранее, точечная оценка этого параметра $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ является состоятельной несмещенной и асимптотически нормальной. Последнее свойство можно использовать для построения доверительного интервала. Действительно, согласно

центральной предельной теореме (см. 1.1) при больших n случайная величина

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\theta}{\sigma\sqrt{n}}$$

имеет распределение, близкое к стандартному нормальному. Пусть t_β – симметричный нормальный квантиль по уровню β, т. е. такое число, что $P(|\eta| < t_\beta) = \beta$ для случайной величины $\eta \in N(0;1)$. Тогда

$$\beta \approx P(|Z_n| < t_\beta) = P\left(\left|\sum_{i=1}^{n} X_i - n\theta\right| < t_\beta\sigma\sqrt{n}\right) = P\left(|\overline{X} - \theta| < \frac{t_\beta\sigma}{\sqrt{n}}\right) =$$

$$= P\left(-\frac{t_\beta\sigma}{\sqrt{n}} < \theta - \overline{X} < \frac{t_\beta\sigma}{\sqrt{n}}\right) = P\left(\overline{X} - \frac{t_\beta\sigma}{\sqrt{n}} < \theta < \overline{X} + \frac{t_\beta\sigma}{\sqrt{n}}\right).$$

Таким образом,

$$I_\beta = \left(\overline{X} - \frac{t_\beta\sigma}{\sqrt{n}}, \overline{X} + \frac{t_\beta\sigma}{\sqrt{n}}\right)$$

при больших объемах выборки n является доверительным интервалом для математического ожидания с доверительной вероятностью β.

СКО σ, используемое в последней формуле, часто бывает неизвестно. В этом случае оно может быть заменено на свою состоятельную оценку $\widehat{\sigma}_n$, построенную по той же выборке. Такой оценкой, естественно, является квадратный корень из несмещенной состоятельной оценки дисперсии:

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

Пример. Пусть проведено 100 независимых экспериментов, в результате каждого из которых может наступить или не наступить некоторое случайное событие A, и пусть событие A наступило 50 раз. Требуется построить доверительный интервал для вероятности этого события с доверительной вероятностью $\beta = 0.99$.

Как было показано в 2.1, несмещенной состоятельной оценкой вероятности события является относительная частота его наступления $\widehat{p} = \frac{50}{100} = \frac{1}{2}$, которая представляет собой среднее арифметическое элементов выборки из распределения двуточечной НСВ ξ, имеющей математическое ожидание $M(\xi) = P(\xi = 1) = p = P(A)$. Таким образом, задача сводится к построению доверительного интервала для $M(\xi)$.

Дисперсия НСВ $D(\xi) = M(\xi - M(\xi))^2 = (0-p)^2(1-p) + (1-p)^2 p = $ $= p(1-p)(p+1-p) = p(1-p)$. Это выражение содержит неизвестный параметр p, но поскольку выборка достаточно велика, дисперсию можно заменить ее состоятельной оценкой $S^2 = \widehat{p}(1-\widehat{p}) = \dfrac{1}{4}$.

Для нахождения симметричного нормального квантиля t_β можно, например, воспользоваться таблицей значений функции распределения $\Phi(x)$ стандартной нормальной случайной величины ξ_0. Поскольку это распределение является симметричным, справедливо равенство $\Phi(-x) = 1 - \Phi(x)$. Из равенств

$$\beta = P(|\xi_0| < t_\beta) = P(-t_\beta < \xi_0 < t_\beta) = \Phi(t_\beta) - \Phi(-t_\beta) = 2\Phi(t_\beta) - 1$$

следует, что симметричный нормальный квантиль является решением уравнения $\Phi(t_\beta) = \dfrac{\beta+1}{2} = 0.995$. Из таблицы значений $\Phi(x)$ находим $t_\beta \approx 2.58$. Половина длины доверительного интервала $\varepsilon = \dfrac{t_\beta \sqrt{S^2}}{\sqrt{n}} = \dfrac{2.58 \cdot 0.5}{10} = $ $= 0.129$. Таким образом, с вероятностью $\beta = 0.99$ истинное значение вероятности события находится в интервале

$$I_\beta = (\widehat{p} - \varepsilon, \widehat{p} + \varepsilon) = (0.371, 0.629).$$

Построение доверительного интервала для дисперсии. Предполагается, что НСВ ξ имеет конечный начальный момент 4-го порядка, что, в частности, обеспечивает существование центрального момента $\mu_4 = M(\xi - M(\xi))^4$. Первоначально также будем предполагать, что $m = M(\xi)$ является известным числом. Тогда несмещенной состоятельной и асимптотически нормальной оценкой дисперсии является статистика

$$S_m^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2.$$

Отношение этой статистики к неизвестному истинному значению дисперсии $\theta = \sigma^2$ может быть записано в виде

$$\frac{S_m^2}{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\eta_i,$$

где $\eta_i = \left(\dfrac{X_i - m}{\sigma}\right)^2$ представляют собой независимые случайные величины, причем $M(\eta_i) = 1$. Начальный второй момент этих величин

$$M(\eta_i^2) = M\left(\frac{X_i - m}{\sigma}\right)^4 = \frac{\mu_4}{\sigma^4}.$$

Соответственно, дисперсия

$$D(\eta_i^2) = M(\eta_i^2) - (M(\eta_i))^2 = \frac{\mu_4}{\sigma^4} - 1,$$

т. е. случайные величины η_i имеют СКО

$$\sigma_\eta = \sqrt{\frac{\mu_4}{\sigma^4} - 1}.$$

Согласно центральной предельной теореме, при большом объеме выборки случайная величина

$$Z_n = \frac{\sum_{i=1}^{n} \eta_i - n}{\sigma_\eta \sqrt{n}}$$

имеет распределение, близкое к стандартному нормальному. Поэтому для симметричного нормального квантиля t_β справедливо приближенное равенство

$$\beta \approx P(|Z_n| < t_\beta) = P\left(\left|\sum_{i=1}^{n} \eta_i - n\right| < t_\beta \sigma_\eta \sqrt{n}\right) =$$

$$= P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \eta_i - 1\right| < \frac{t_\beta \sigma_\eta}{\sqrt{n}}\right) = P\left(\left|\frac{S_m^2}{\sigma^2} - 1\right| < \frac{t_\beta \sigma_\eta}{\sqrt{n}}\right).$$

Пусть $\dfrac{t_\beta \sigma_\eta}{\sqrt{n}} = \varepsilon$. Тогда последнее неравенство, которое выполняется с вероятностью, близкой к β, можно переписать в виде

$$1 - \varepsilon < \frac{S_m^2}{\sigma^2} < 1 + \varepsilon.$$

Поскольку значение ε стремится к нулю при увеличении объема выборки, можно считать, что $\varepsilon < 1$. Тогда последнее неравенство равносильно неравенству

$$\frac{S_m^2}{1+\varepsilon} < \sigma^2 < \frac{S_m^2}{1-\varepsilon}.$$

Таким образом, в качестве доверительного интервала для дисперсии может быть выбран

$$I_\beta = \left(\frac{S_m^2}{1+\varepsilon}, \frac{S_m^2}{1-\varepsilon}\right).$$

Статистику $S_m^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - m)^2$ при неизвестном математическом ожидании m можно заменить на выборочную дисперсию S^2, подставив вместо

параметра m его состоятельную оценку $\overline{X}$. Величина

$$\varepsilon = \frac{t_\beta}{\sqrt{n}}\sqrt{\frac{\mu_4}{\sigma^4} - 1}.$$

Если НСВ имеет нормальное распределение, отношение $\frac{\mu_4}{\sigma^4} = 3$, соответственно $\varepsilon = t_\beta\sqrt{\frac{2}{n}}$. В общем случае неизвестные центральные моменты μ_4 и σ^2 можно заменить соответствующими выборочными моментами, тогда $\varepsilon = \frac{t_\beta}{\sqrt{n}}\sqrt{\frac{\widehat{\mu}_4}{(S^2)^2} - 1}$.

2.10. Доверительные интервалы для параметров нормального распределения в случае малой выборки

Построение доверительных интервалов, основанное на асимптотической нормальности точечных оценок (см. 2.9), правомерно только в случае большой выборки, содержащей по меньшей мере несколько десятков элементов. Случай малой выборки требует других подходов, использующих дополнительную априорную информацию о принадлежности распределения НСВ некоторому семейству распределений. Здесь будет сделано наиболее часто используемое предположение о том, что НСВ имеет нормальное распределение. Основанием для такого предположения может служить та же центральная предельная теорема, благодаря которой можно утверждать, что если НСВ формируется в результате суммарного влияния большого количества независимых факторов, то она имеет распределение, близкое к нормальному.

Итак, пусть НСВ $\xi \in N(m, \sigma^2)$. Требуется построить по выборке доверительные интервалы для математического ожидания m и дисперсии σ^2. Для их построения используется теорема Стьюдента 1.2 и ее следствие – теорема 1.3. Состоятельной оценкой для σ^2 может служить выборочная дисперсия $S^2 = \widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2$. Согласно теореме 1.3 случайная величина $\frac{(\overline{X} - m)\sqrt{n-1}}{\sqrt{S^2}}$ распределена по закону Стьюдента с $(n-1)$ степенями свободы. Пусть $\tau_0(n-1, \beta)$ – симметричный квантиль для этого распределения по уровню β. Тогда

$$\beta = P\left(\left|\frac{(\overline{X} - m)\sqrt{n-1}}{\sqrt{S^2}}\right| < \tau_0(n-1, \beta)\right) =$$

$$= P\left(|m - \overline{X}| < \sqrt{\frac{S^2}{n-1}}\,\tau_0(n-1,\beta)\right).$$

Таким образом, доверительным интервалом для математического ожидания m с доверительной вероятностью β является интервал

$$I_\beta = \left(\overline{X} - \sqrt{\frac{S^2}{n-1}}\,\tau_0(n-1,\beta),\ \ \overline{X} + \sqrt{\frac{S^2}{n-1}}\,\tau_0(n-1,\beta)\right).$$

Что касается доверительного интервала для дисперсии, то, согласно теореме Стьюдента, случайная величина $\dfrac{nS^2}{\sigma^2}$ распределена по закону "Хи-квадрат" с $(n-1)$ степенями свободы. Для построения доверительного интервала нужно выделить на положительной полуоси интервал, имеющий для этого распределения заданную доверительную вероятность β. Естественно выбрать этот интервал таким образом, чтобы оставшиеся части положительной полуоси справа и слева от этого интервала имели одинаковые вероятности $\dfrac{1-\beta}{2}$. Тогда получится интервал, ограниченный квантилями распределения „Хи-квадрат" $\chi_0^2\left(n-1, \dfrac{1-\beta}{2}\right)$ и $\chi_0^2\left(n-1, \dfrac{1+\beta}{2}\right)$. Таким образом,

$$\beta = P\left(\chi_0^2\left(n-1,\frac{1-\beta}{2}\right) < \frac{nS^2}{\sigma^2} < \chi_0^2\left(n-1,\frac{1+\beta}{2}\right)\right) =$$

$$= P\left(\frac{\chi_0^2\left(n-1,\dfrac{1-\beta}{2}\right)}{nS^2} < \frac{1}{\sigma^2} < \frac{\chi_0^2\left(n-1,\dfrac{1+\beta}{2}\right)}{nS^2}\right) =$$

$$= P\left(\frac{nS^2}{\chi_0^2\left(n-1,\dfrac{1+\beta}{2}\right)} < \sigma^2 < \frac{nS^2}{\chi_0^2\left(n-1,\dfrac{1-\beta}{2}\right)}\right).$$

Следовательно, в качестве доверительного интервала с доверительной вероятностью β для дисперсии σ^2 можно выбрать интервал

$$I_\beta = \left(\frac{nS^2}{\chi_0^2\left(n-1,\dfrac{1+\beta}{2}\right)},\ \ \frac{nS^2}{\chi_0^2\left(n-1,\dfrac{1-\beta}{2}\right)}\right).$$

Формулы для доверительных интервалов, приведенные в данном подразделе, фактически можно использовать только для выборок, не превышающих нескольких десятков элементов, поскольку таблицы распределений „Хи-квадрат“ и Стьюдента составлены только для такого числа степеней свободы. Однако при достаточно большом объеме выборки можно пользоваться формулами из 2.9, тем более, что в случае нормального распределения НСВ эти формулы более точны, чем в общем случае.

Пример. Пусть имеется выборка из нормального распределения НСВ объема $n = 17$, и пусть для этой выборки вычислены статистики: среднее арифметическое $\overline{X} = 2.0$ и средний квадрат $\widehat{\alpha_2} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i^2 = 5.0$. Требуется построить доверительные интервалы для математического ожидания и дисперсии НСВ с доверительной вероятностью $\beta = 0.9$.

Для построения обоих доверительных интервалов используется значение выборочной дисперсии $S^2 = \widehat{\alpha_2} - \overline{X}^2 = 5.0 - 4.0 = 1.0$. Из таблицы симметричных квантилей Стьюдента находим $\tau_0(n-1, \beta) = \tau_0(16, 0.9) = 1.746$. Половина длины доверительного интервала для математического ожидания

$$\varepsilon = \sqrt{\frac{S^2}{n-1}}\,\tau_0(n-1,\beta) = \sqrt{\frac{1}{16}}\,\tau_0(16, 0.9) = 0.25 \cdot 1.746 \approx 0.44.$$

Таким образом, с вероятностью $\beta = 0.9$ истинное значение математического ожидания НСВ находится в интервале

$$I_\beta = (\overline{X} - \varepsilon, \overline{X} + \varepsilon) = (1.56, 2.44).$$

Из таблицы квантилей распределения „Хи-квадрат“ находим

$$\chi_0^2\left(n-1, \frac{1-\beta}{2}\right) = \chi_0^2(16, 0.05) = 7.96,$$

$$\chi_0^2\left(n-1, \frac{1+\beta}{2}\right) = \chi_0^2(16, 0.95) = 26.3.$$

Таким образом, с вероятностью $\beta = 0.9$ истинное значение дисперсии НСВ находится в содержащем значение $S^2 = 1.0$ интервале

$$I_\beta = \left(\frac{nS^2}{26.3}, \frac{nS^2}{7.96}\right) = \left(\frac{17.0}{26.3}, \frac{17.0}{7.96}\right) \approx (0.64, 2.14).$$

3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

3.1. Статистический критерий

Пусть в результате статистического эксперимента получена выборка $\boldsymbol{x}$ из распределения НСВ ξ. По выборке $\boldsymbol{x}$ относительно НСВ ξ могут быть высказаны некоторые предположения, или, как их называют, гипотезы. Например, можно выдвинуть гипотезу о том, что НСВ ξ имеет нормальное распределение; или можно предположить, что НСВ ξ имеет математическое ожидание, равное m и дисперсию, равную σ^2.

Статистической гипотезой (или просто гипотезой) будем называть утверждение о законе распределения НСВ или значениях его параметров.

Если гипотеза выдвинута относительно значений параметров распределения НСВ, то она называется параметрической. Если распределение НСВ имеет всего l параметров, а гипотеза утверждает, что k из них имеют заданные значения, то при $k = l$ гипотеза называется простой, а при $k < l$ сложной.

Задачей проверки статистической гипотезы является ответ на вопрос: согласуется ли гипотеза с результатами эксперимента (полученной выборкой) или нет. В общем случае любая гипотеза может быть либо ложной, либо истинной. Результатом проверки гипотезы не является ее принятие или непринятие. Проверяя статистическую гипотезу, выясняем лишь то, противоречит ли она данным эксперимента или нет.

Проверяемую гипотезу принято обозначать H_0, тогда как альтернативную гипотезу обозначают H_1.

В общем случае процедура проверки статистической гипотезы выглядит следующим образом. Пусть выдвинута гипотеза H_0. Строится некоторая функция выборки $\mathcal{K}(\boldsymbol{X}, H_0)$, называемая критериальной, или, коротко, критерием. В предположении истинности гипотезы H_0 все множество значений функции $\mathcal{K}(\boldsymbol{X}, H_0)$ делят на две непересекающиеся области W и Q. Область W называют допустимой, а область Q критической. Если теперь значение критериальной функции на полученной выборке $\mathcal{K}(\boldsymbol{X}, H_0)$ попадает в область Q, то гипотеза H_0 отклоняется, если же $\mathcal{K}(\boldsymbol{X}, H_0)$ попадает в область W, то гипотеза H_0 сохраняется как не противоречащая экспериментальным данным.

Таким образом задача проверки статистической гипотезы сводится к задаче выбора критериальной функции $\mathcal{K}(\boldsymbol{X}, H_0)$ и критической области Q (допустимая область W определяется при этом автоматически). Для одной и той же гипотезы H_0 такие построения можно провести по-разному и получить разные статистические критерии. Поясним это следующим примером.

Пусть выдвинута гипотеза H_0: вероятность события A $P(A) = 0.7$.

В результате проведенных 100 экспериментов событие A наступило m раз. Если обозначить через χ_i индикатор события A, т. е. случайную величину, принимающую значение 1, если в i-м испытании A наступило, и 0, если A не наступило, то

$$m = \sum_{i=1}^{100} \chi_i$$

и выборка $\boldsymbol{X} = \{\chi_i\}_{i=1}^{100}$.

В качестве критериальной функции $\mathcal{K}(\boldsymbol{X}, H_0)$ возьмем

$$\mathcal{K}(\boldsymbol{X}, H_0) = \left|0.7 - \frac{m}{100}\right|.$$

Множество значений этой функции есть отрезок $[0, 0.7]$ (рис. 3.1).

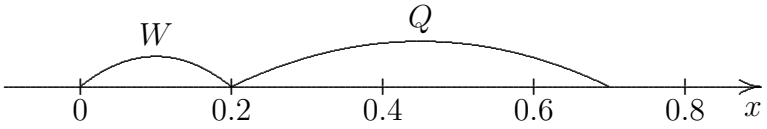

Рис. 3.1

Естественно сохранить H_0, если частота события A $\boldsymbol{\nu} = \dfrac{m}{100}$ не слишком сильно отличается от 0.7, и отвергнуть H_0, если разность велика, например, превышает 0.2. Таким образом, гипотеза H_0 отклоняется, если $\mathcal{K}(\boldsymbol{X}, H_0) > 0.2$, если же $\mathcal{K}(\boldsymbol{X}, H_0) \leq 0.2$, то отклонять H_0 нет оснований в соответствии с принятой договоренностью.

Например, $m = 53$, тогда $\mathcal{K}(\boldsymbol{X}, H_0) = 0.17 < 0.2$ и отклонять гипотезу нет оснований, если же $m = 47$, то $\mathcal{K}(\boldsymbol{X}, H_0) = 0.23 > 0.2$ и гипотезу следует отклонить.

Можно построить область Q по-другому и получить другой статистический критерий. Так, если $Q = (0.1; 0.7]$, то при $m = 53$ гипотезу H_0 следует отклонить, тогда как ранее ее сохраняли.

Получается, что для одной и той же выборки $\boldsymbol{X}$ и функции $\mathcal{K}(\boldsymbol{X}, H_0)$ можем как сохранить гипотезу H_0, так и отклонить ее, назначая по-разному область Q. Становится ясным, что для принятия статистически обоснованных решений необходимо строить область Q некоторым оптимальным образом, основываясь на статистических свойствах НСВ ξ.

Поскольку выборка $\boldsymbol{X}$ случайна, значения критериальной функции тоже случайны, поэтому при проверке статистической гипотезы возможны случайные ошибки следующих двух типов:

1) гипотеза верна, но отклоняется (ошибка I рода);

2) гипотеза ложна, но сохраняется (ошибка II рода).

Случайные ошибки характеризуются вероятностями:

$\boldsymbol{\alpha} = P(\mathcal{K}(\boldsymbol{X}, H_0) \in Q / H_0)$ – вероятность ошибки первого рода;

$\gamma = P(\mathcal{K}(\boldsymbol{X}, H_0) \in W/H_1)$ – вероятность ошибки второго рода.

Вероятность ошибки первого рода α также называют уровнем значимости критерия.

Естественно, что хотелось бы сделать вероятности обеих ошибок нулевыми, но уменьшение вероятности одной ошибки при фиксированной функции $\mathcal{K}$, достигаемое соответствующим изменением области Q или W, ведет к увеличению вероятности другой ошибки. Поэтому при построении статистического критерия применяют принцип Неймана–Пирсона, в соответствии с которым фиксируется вероятность ошибки первого рода α, а за счет выбора критической области Q максимизируется мощность критерия, т. е. вероятность β отклонить гипотезу H_0, когда она неверна.

Таким образом, мощность $\beta = P(\mathcal{K}(\boldsymbol{X}, H_0) \in Q/H_1)$ и $\beta = 1 - \gamma$. Как следует из определения, мощность критерия зависит от альтернативной гипотезы.

Рассмотрим следующий пример. Пусть НСВ $\xi \in N(m, \sigma^2)$. Гипотеза H_0: $m = 0$.

1. Пусть σ^2 известна. Возьмем в качестве критериальной функции $\mathcal{K}(\boldsymbol{X}, H_0) = \overline{X}$. Если гипотеза H_0 верна, то элементы случайной выборки $X_i \in N(0; \sigma^2)$ и $\overline{X} \in N\left(0; \dfrac{\sigma^2}{n}\right)$. Плотность распределения вероятностей $\overline{X}$ есть $f(t/H_0) = f\left(t, 0, \dfrac{\sigma^2}{n}\right)$.

Пусть альтернативная гипотеза H_1: $m \neq 0$. При этой гипотезе $\overline{X} \in N\left(m, \dfrac{\sigma^2}{n}\right)$ и $f(t/H_1) = f\left(t, m, \dfrac{\sigma^2}{n}\right)$. Критическую область Q по заданному уровню значимости α можно построить четырьмя основными способами (рис. 3.2):

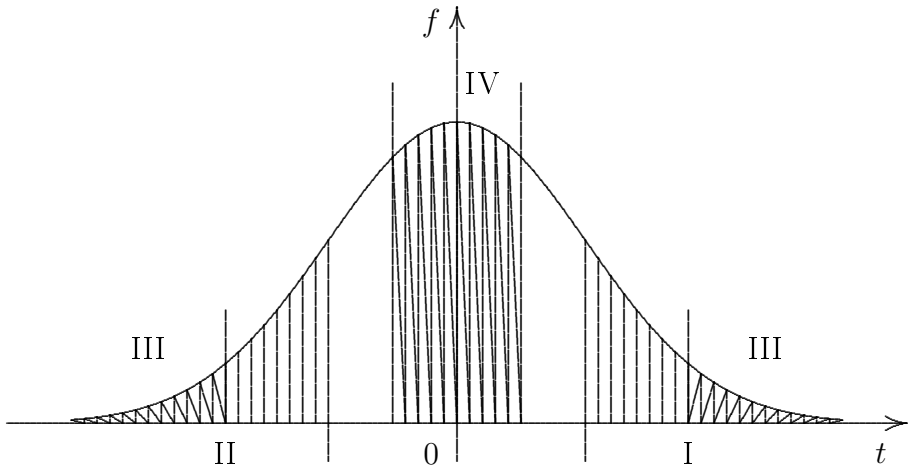

Рис. 3.2

I тип – область больших положительных t;

II тип – область больших отрицательных t;

III тип – область больших по модулю t;

IV тип – область малых по модулю t.

Вероятность попадания $\mathcal{K}(\boldsymbol{X}, H_0)$ в каждую из этих областей при истинности гипотезы H_0 равна α.

Для принятия решения о выборе критической области Q рассмотрим мощность критерия как функцию от параметра m альтернативной гипотезы H_1. Будем менять m от $-\infty$ до $+\infty$ и рассчитывать мощность соответствующего критерия, как показано на рис. 3.3 для критической области первого типа.

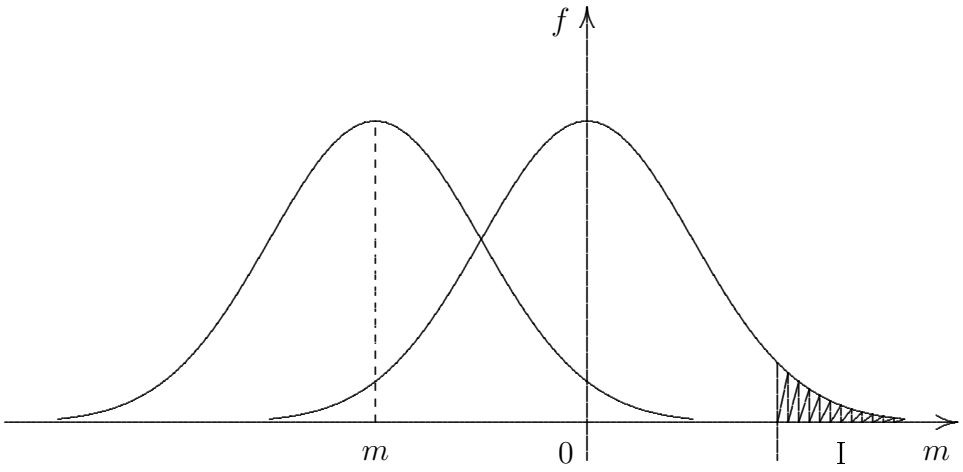

Рис. 3.3

График мощности как функции m для критических областей всех четырех типов показан на рис. 3.4, *а – г*.

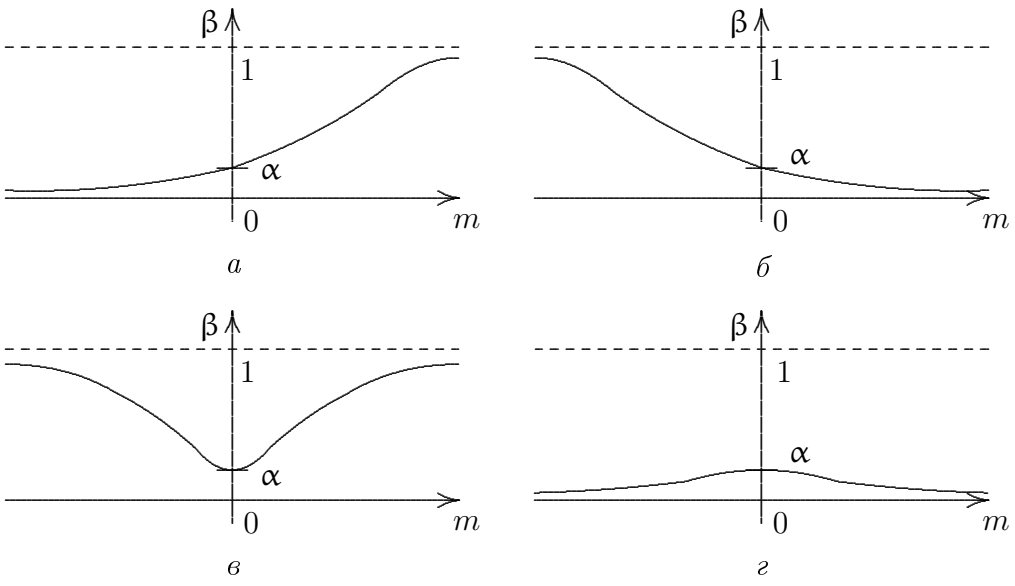

Рис. 3.4

Вид графиков мощности показывает, что если известен знак альтер-

нативы ($m > 0$ или $m < 0$), то следует использовать соответственно критические области первого и второго типов. Если знак альтернативы неизвестен, то следует использовать универсальную критическую область типа III. Критическую область четвертого типа использовать не имеет смысла, так как для нее мощность критерия не поднимается выше, чем α.

2. Пусть σ^2 неизвестно. Возьмем

$$\mathcal{K}(\boldsymbol{X}, H_0) = \frac{\overline{X}}{\frac{S}{\sqrt{n-1}}}.$$

Как известно (теорема 1.3), такая статистика $\mathcal{K}(\boldsymbol{X}, H_0) \in \tau_{n-1}$. Критическая область III типа может быть построена следующим образом. Пусть $\tau_0(n-1, 1-\alpha)$ – симметричный квантиль Стьюдента по уровню $1-\alpha$. Тогда решающее правило будет выглядеть так:

если $|\mathcal{K}(\boldsymbol{X}, H_0)| \geq \tau_0(n-1, 1-\alpha)$, то гипотеза H_0 отклоняется.

Критерий, мощность которого не меньше мощности любого другого критерия проверки гипотезы H_0 против альтернативы H_1 при фиксированном уровне значимости α, называется наиболее мощным критерием.

Критерий, мощность которого не меньше мощности любого другого критерия проверки гипотезы H_0 против любой альтернативы H_1 при фиксированном уровне значимости α, называется равномерно наиболее мощным критерием.

Из определения следует, что равномерно наиболее мощный критерий (РНМК) является наилучшим. К сожалению, РНМК существуют лишь в редких случаях и являются, скорее, исключениями из общего правила.

В связи с тем, что уровень значимости α критерия назначается, принято выбирать α из некоторого стандартного ряда: 0.1, 0.05, 0.02, 0.01. Выбор конкретного значения α определяется потерями за неправильное решение при проверке гипотезы. Так, выбор $\alpha = 0.1$ означает, что экспериментатор готов ошибаться (отклонять верную гипотезу H_0) в среднем 1 раз из 10. При $\alpha = 0.01$ такая ошибка допускается в среднем уже 1 раз из 100. В связи с этим в реальной задаче проверки гипотезы анализируются потери от неправильного решения: что важнее – не отклонить верную гипотезу H_0 (уменьшаем α) или принять ложную H_1 (α увеличиваем).

3.2. Проверка гипотезы о равенстве математических ожиданий двух нормальных случайных величин

Пусть в эксперименте наблюдаются две нормальные случайные величины: $\xi_1 \in N(m_1, \sigma^2)$, $\xi_2 \in N(m_2, \sigma^2)$. Параметры распределений m_1 , m_2 и σ^2 неизвестны. Предполагается, что ξ_1 и ξ_2 независимы.

По гипотезе $H_0 : m_1 = m_2 = m$.

Для проверки гипотезы H_0 имеем две выборки: $\boldsymbol{X} = [X_1, X_2, \ldots, X_{n_1}]^T$ и $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_{n_2}]^T$ из распределений ξ_1 и ξ_2 соответственно.

Построим критериальную функцию для проверки гипотезы H_0. Возьмем выборочные средние: $\overline{X} \in N\left(m, \dfrac{\sigma^2}{n_1}\right)$; $\overline{Y} \in N\left(m, \dfrac{\sigma^2}{n_2}\right)$. Если гипотеза H_0 верна, то

$$\overline{X} - \overline{Y} \in N\left(0, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right).$$

После нормировки получим:

$$\mathcal{K}_1(\boldsymbol{X}, \boldsymbol{Y}, H_0) = \frac{\overline{X} - \overline{Y}}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \in N(0,1).$$

Если взять функцию $\mathcal{K}_1(\boldsymbol{X}, \boldsymbol{Y}, H_0)$ с известным распределением в качестве критериальной, то незнание параметра σ не позволит вычислить ее значение на числовых выборках $\boldsymbol{x}$ и $\boldsymbol{y}$. Поэтому поступим следующим образом. Из теоремы Стьюдента известно, что

$$\frac{n_1 S_x^2}{\sigma^2} \in \chi^2_{n_1-1},$$

где S_x^2 – выборочная дисперсия для НСВ ξ_1,

$$\frac{n_2 S_y^2}{\sigma^2} \in \chi^2_{n_2-1},$$

где S_y^2 – выборочная дисперсия для НСВ ξ_2.

При этом обе указанные случайные величины независимы, что следует из независимости ξ_1 и ξ_2. Тогда их сумма

$$\frac{n_1 S_x^2}{\sigma^2} + \frac{n_2 S_y^2}{\sigma^2}$$

распределена по закону $\chi^2_{n_1+n_2-2}$.

Можно показать, что эта случайная величина независима со статистикой $\mathcal{K}_1(\boldsymbol{X}, \boldsymbol{Y}, H_0)$. Тогда случайная величина

$$\frac{\dfrac{\overline{X} - \overline{Y}}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}\sqrt{n_1 + n_2 - 2}}{\sqrt{\dfrac{n_1 S_x^2}{\sigma^2} + \dfrac{n_2 S_y^2}{\sigma^2}}} \in \tau_{n_1+n_2-2}$$

распределена по закону Стьюдента с $n_1 + n_2 - 2$ степенями свободы, что следует из определения распределения Стьюдента. Так как эта случайная величина не содержит параметр σ , то возьмем в качестве критериальной функции

$$\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0) = \frac{(\overline{X} - \overline{Y})\sqrt{n_1 + n_2 - 2}\sqrt{n_1 n_2}}{\sqrt{n_1 + n_2}\sqrt{n_1 S_x^2 + n_2 S_y^2}} \in \tau_{n_1+n_2-2}.$$

Задав уровень значимости критерия α, найдем симметричный квантиль Стьюдента $\tau_0(n_1 + n_2 - 2, 1 - \alpha)$. Критическая область критерия – это область третьего типа – больших по модулю значений функции $\mathcal{K}$. Таким образом, решающее правило имеет вид

$$\text{если} \quad |\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0)| \geq \tau_0(n_1 + n_2 - 2, 1 - \alpha),$$

то гипотезу H_0 отклоняем ($m_1 \neq m_2$).

Пример. Для двух участков сельскохозяйственных угодий получены следующие результаты: первый участок имеет площадь 10 га, средняя урожайность $\overline{X} = 26$ ц/га при $S_x = 2.0$; второй участок имеет площадь 12 га, средняя урожайность $\overline{Y} = 24$ ц/га при $S_y = 1.9$.

Требуется проверить гипотезу: урожайность участков одинакова при уровне значимости $\alpha = 0.1$. Таким образом, гипотеза $H_0 : M(\xi_1) = M(\xi_2)$.

Считая распределения урожайности участков нормальными, воспользуемся критерием проверки гипотезы о равенстве математических ожиданий двух нормальных случайных величин. Количество гектаров в каждом участке будем считать объемом выборки. Тогда $n_1 = 10$, $n_2 = 12$.

При $\alpha = 0.1$ симметричный квантиль Стьюдента

$$\tau_0(10 + 12 - 2, 0.9) = 1.725.$$

Тогда значение критериальной функции

$$\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0) = \frac{(26 - 24)\sqrt{20}\sqrt{10 \cdot 12}}{\sqrt{22}\sqrt{10 \cdot (2.0)^2 + 12 \cdot (1.9)^2}} \approx 2.29.$$

Так как $|\mathcal{K}| = 2.29 > 1.725$, гипотезу H_0 следует отклонить при уровне значимости 0.1.

Для проверки устойчивости найденного решения уменьшим уровень значимости α в 2 раза. Возьмем $\alpha = 0.05$. При этом $\tau_0(20, 0.95) = 2.09$. Поскольку $2.29 > 2.09$, решение об отклонении гипотезы H_0 сохраняется. Решение устойчиво.

Таким образом, считать, что урожайность участков одинакова, нет оснований.

3.3. Критерий Фишера проверки гипотезы о равенстве дисперсий двух нормальных случайных величин

Пусть наблюдаются две независимые случайные величины ξ и η, имеющие нормальные распределения:

$$\xi \in N(m_\xi, \sigma_\xi^2), \qquad \eta \in N(m_\eta, \sigma_\eta^2).$$

Из распределений ξ и η получены выборки $\boldsymbol{X}$ объемом n_1 и $\boldsymbol{Y}$ объемом n_2 соответственно.

Выдвигается гипотеза $H_0 : \sigma_\xi^2 = \sigma_\eta^2 = \sigma$. Для проверки гипотезы H_0 рассмотрим критерий, называемый критерием Фишера (Фишера–Снедекора).

Из теоремы Стьюдента известно, что случайные величины

$$\frac{n_1 S_1^2}{\sigma_\xi^2} \in \chi^2_{n_1-1}, \qquad \frac{n_2 S_2^2}{\sigma_\eta^2} \in \chi^2_{n_2-1},$$

где $S_1^2 = \dfrac{1}{n_1}\sum\limits_{i=1}^{n_1}(X_i - \overline{X})^2$, $S_2^2 = \dfrac{1}{n_2}\sum\limits_{i=1}^{n_2}(Y_i - \overline{Y})^2$, имеют χ^2-распределения и независимы. Тогда случайные величины

$$\frac{(n_1-1)\tilde{S}_1^2}{\sigma^2} \in \chi^2_{n_1-1}, \qquad \frac{(n_2-1)\tilde{S}_2^2}{\sigma^2} \in \chi^2_{n_2-1},$$

где $\tilde{S}_1^2 = \dfrac{1}{n_1-1}\sum\limits_{i=1}^{n_1}(X_i - \overline{X})^2$, $\tilde{S}_2^2 = \dfrac{1}{n_2-1}\sum\limits_{i=1}^{n_2}(Y_i - \overline{Y})^2$ также имеют χ^2-распределение и независимы.

Взяв в качестве критериальной функции

$$\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0) = \frac{\dfrac{(n_1-1)\widetilde{S}_1^2}{(n_1-1)\sigma^2}}{\dfrac{(n_2-1)\widetilde{S}_2^2}{(n_2-1)\sigma^2}} = \frac{\widetilde{S}_1^2}{\widetilde{S}_2^2} \in F(k_1, k_2),$$

получим случайную величину, имеющую распределение Фишера с числом степеней свободы (k_1, k_2), где $k_1 = n_1 - 1$, $k_2 = n_2 - 1$.

Если гипотеза H_0 верна, то $\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0)$ должна принимать значения, близкие к единице. Поэтому в качестве критической области возьмем область малых и больших значений $\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0)$: $0 \le \mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0) \le F_1$ и $\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0) \ge F_2$, причем выберем значения F_1 и F_2 так, чтобы при заданном уровне значимости критерия α

$$P(\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0) \le F_1) = P(\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0) \ge F_2) = \frac{\alpha}{2}.$$

Плотность распределения Фишера представлена на рис. 3.5, где $1-\alpha$ – площадь незаштрихованной части.

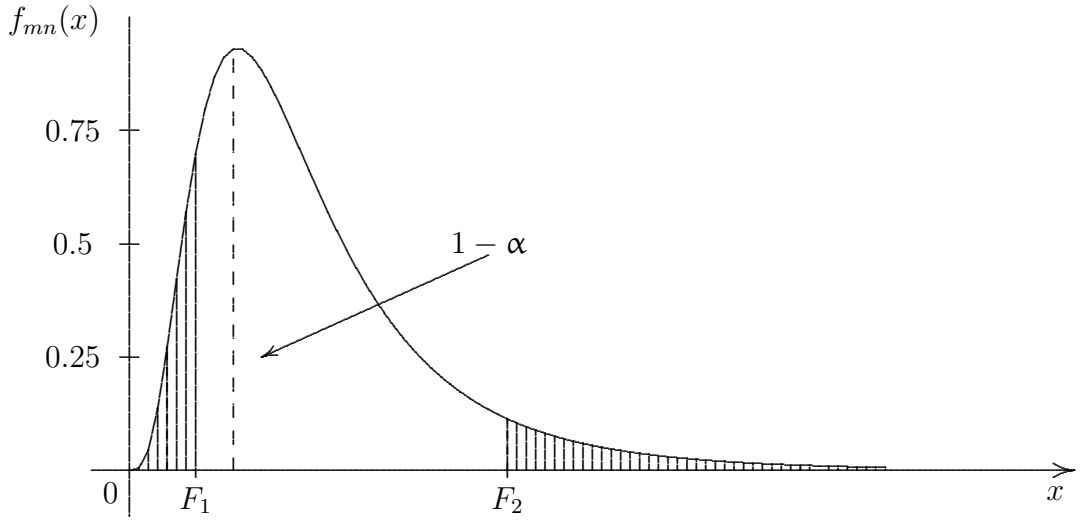

Рис. 3.5. Функция плотности распределения Фишера $m=10,\ n=12$

Пусть $\widetilde{\mathcal{K}}(\boldsymbol{X},\boldsymbol{Y},H_0)=\dfrac{1}{\mathcal{K}(\boldsymbol{X},\boldsymbol{Y},H_0)}=\dfrac{\widetilde{S}_2^2}{\widetilde{S}_1^2}$. Тогда, так как

$$\mathcal{K}(\boldsymbol{X},\boldsymbol{Y},H_0)\in F(k_1,k_2),\quad \text{то}\quad \widetilde{\mathcal{K}}(\boldsymbol{X},\boldsymbol{Y},H_0)\in F(k_2,k_1).$$

При этом

$$\frac{\alpha}{2}=P(\mathcal{K}<F_1)=P(\frac{1}{\mathcal{K}}>\frac{1}{F_1})=P(\widetilde{\mathcal{K}}>\frac{1}{F_1}),$$

т. е. для распределения $\widetilde{\mathcal{K}}$ $\widetilde{F}_2=\dfrac{1}{F_1}$ и $\widetilde{F}_1=\dfrac{1}{F_2}$.

Для упрощения критерия будем рассматривать другую критериальную функцию

$$F(\boldsymbol{X},\boldsymbol{Y},H_0)=\max\left\{\frac{\widetilde{S}_1^2}{\widetilde{S}_2^2},\frac{\widetilde{S}_2^2}{\widetilde{S}_1^2}\right\}.$$

У такого критерия критическая область будет уже односторонней, если $F(\boldsymbol{X},\boldsymbol{Y},H_0)\geq C_\alpha$, то гипотеза H_0 отклоняется.

C_α есть квантиль распределения Фишера с числом степеней свободы k_1 и k_2 по уровную $1-\dfrac{\alpha}{2}$:

$$C_\alpha=F_0\left(k_1,k_2,1-\frac{\alpha}{2}\right),$$

где k_1 – число степеней свободы для выборки с большим значением $\widetilde{S}^2$, т. е. из двух несмещенных выборочных дисперсий $\widetilde{S}_1^2$ и $\widetilde{S}_2^2$ в числитель критериальной функции $F(\boldsymbol{X},\boldsymbol{Y},H_0)$ ставится та, которая больше.

Пример. Получены оценки дисперсий $\widetilde{S}_1^2 = 9.6$ и $\widetilde{S}_2^2 = 5.7$ по выборкам объемами $n_1 = 10$ и $n_2 = 15$ соответственно. Проверим гипотезу о равенстве дисперсий по критерию Фишера при уровне значимости 0.1:

$$F(\boldsymbol{X}, \boldsymbol{Y}, H_0) = \frac{\widetilde{S}_1^2}{\widetilde{S}_2^2} = \frac{9.6}{5.7} = 1.68.$$

Число степеней свободы $k_1 = 10 - 1 = 9$, $k_2 = 15 - 1 = 14$. Пороговое значение $C_\alpha = F_0(9.14, 0.95) = 2.65$. Так как $1.68 < 2.65$, отклонение в выборочных дисперсиях не значимо и гипотезу H_0 можно сохранить.

3.4. Критерии согласия

Критериями согласия в математической статистике принято называть статистические критерии для проверки гипотезы о виде закона распределения.

Пусть $\mathbf{X} = [X_1 \ldots, X_n]^T$ – выборка из распределения НСВ ξ с неизвестной функцией распределения $F_\xi(t)$, о которой выдвинута простая гипотеза $H_0 : F_\xi(t) = F(t)$.

Критерий согласия Колмогорова. Этот критерий можно применять в тех случаях, когда функция распределения $F(t)$ непрерывна и монотонна. В качестве критериальной функции используется величина

$$D_n(X) = \sup_{-\infty<t<+\infty} |F_n(t) - F(t)|,$$

представляющая собой максимальное отклонение эмпирической функции распределения $F_n(t)$ от предполагаемой функции распределение $F(t)$. При каждом t значение $F_n(t)$ является наилучшей оценкой для $F(t)$, причем с увеличением объема выборки и происходит сближение $F_n(t)$ и $F(t)$. Поэтому при больших n, если гипотеза H_0 верна, значение $D_n(X)$ не должно существенно отличаться от нуля. Отсюда следует, что критическую область критерия следует выбирать типа I, т. е. $Q = \{D_n(X) \geq C_\alpha\}$.

Особенностью критериальной функции D_n является то, что ее распределение при гипотезе H_0 не зависит от вида функции $F(t)$.

Пусть U_i – случайная величина, имеющая равномерное $R(0, 1)$ распределение. Рассмотрим новую случайную величину $Y_i = F^{-1}(U_i)$, где F^{-1} – функция, обратная F.

Из монотонности функции F следует монотонность (возрастание) функции F^{-1} и $F^{-1}(U_i) < y$ эквивалентно $U_i < F(y)$.

Тогда

$$F_{Y_i}(y) - P(Y_i < y) - P(F^{-1}(U_i) < y) = P(U_i < F(y)) =$$

$$= \int\limits_{-\infty}^{F(y)} f_{U_i}(\tau)d\tau = \int\limits_{0}^{F(y)} 1d\tau = F(y),$$

т. е. распределение случайной величины Y_i совпадает с заданным $F(t)$.

Другими словами, любая абсолютно непрерывная случайная величина с монотонной непрерывной функцией распределения может быть получена из равномерно распределенной на $(0,1)$ случайной величины.

Если положить $t = F^{-1}(u)$, $0 \le u \le 1$, получим

$$D_n(X) = \sup_{0\le u\le 1} \left|F_n(F^{-1}(u)) - u\right|.$$

Перейдем к новым случайным величинам $U_i = F(X_i), \quad i = 1, 2, \ldots, n$. Построим из них вариационный ряд

$$U_{i_1} \le \cdots \le U_{i_n}.$$

Функция $F(t)$ монотонна, поэтому $U_{i_k} = F(X_{i_k}), \quad k = 1, 2, \ldots, n$.

Следовательно, неравенства $F^{-1}(u) \ge X_{i_k}$ эквивалентны неравенствам $u \ge U_{i_k}$. Если $\delta_0(z)$ – единичная функция Хевисайда, то

$$F_n(F^{-1}(u)) = \frac{1}{n}\sum_{k=1}^{n} \delta_0(F^{-1}(u) - X_{i_k}) = \frac{1}{n}\sum_{k=1}^{n} \delta_0(u - U_{i_k}) = \Phi_n(u),$$

где $\Phi_n(u)$ – эмпирическая функция распределения для выборки из равномерного $R(0,1)$ распределения.

Таким образом, статистика D_n совпадает с $\sup\limits_{0\le u\le 1} |\Phi_n(u) - u|$ и, следовательно, ее распределение не зависит от $F(t)$, если справедлива гипотеза H_0. Тем самым, достаточно вычислить и протабулировать распределение D_n один раз, а именно для выборки из равномерного $R(0,1)$ распределения, и использовать его для проверки гипотезы относительно произвольной монотонной непрерывной функции распределения $F(t)$.

Другим замечательным свойством критерия Колмогорова является то, что распределение статистики D_n при достаточно больших n (уже при $n \ge 20$) практически от n не зависит. При $n \ge 20$ пороговое значение $C_\alpha = C_\alpha(u)$ критической области можно полагать равным $C_\alpha = \lambda_\alpha/\sqrt{n}$. Действительно, в этом случае

$$P(D_n \in Q/H_0) = P(\sqrt{n}D_n \ge \lambda_\alpha/H_0) \cong \alpha.$$

Так, при $\alpha = 0,05$ $\lambda_\alpha = 1,3581$; при $\alpha = 0,01$ $\lambda_\alpha = 1,6276$.

Таким образом, при заданном уровне значимости α число λ_α определяют из соотношения

$$\mathcal{K}(\lambda_\alpha) = 1 - \alpha,$$

где

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

Решающее правило проверки гипотезы H_0 имеет (при $n \geq 20$) следующий вид: если $\sqrt{n}D_n(X) \geq \lambda_\alpha$, то H_0 следует отвергнуть.

Для небольших значений n точное распределение D_n табулировано и для расчета точных значений c_α можно воспользоваться соответствующими таблицами.

Проверка гипотезы о полиномиальном распределении. Если в результате случайного эксперимента может наступить одно из k попарно несовместных событий $A_i,\ i = 1, 2, \ldots, k$, то множество вероятностей $P(A_i)$ этих событий называют полиномиальным распределением. Биномиальное распределение, таким образом, является частным случаем полиномиального распределения при $k = 2$.

Пусть по гипотезе H_0 каждому событию A_i приписана вероятность $p_i = P(A_i), i = 1, 2, \ldots, k$. В результате проведения n случайных экспериментов событие A_i наступило n_1 раз, событие A_2 наступило n_2 раз, событие A_k наступило n_k раз, $\sum\limits_{i=1}^{k} n_i = n$. Таким образом, выборкой является числовой вектор

$$\mathbf{X} = (n_1, n_2, \ldots, n_k)^T.$$

В качестве критериальной функции возьмем функцию вида

$$\mathcal{K}(\boldsymbol{X}, H_0) = \sum_{i=1}^{k} \frac{(n_i - np_i^0)^2}{np_i^0}.$$

Для правильного выбора критической области Q необходимо знать распределение критериальной функции. Можно показать, что если гипотеза H_0 верна, то критериальная функция $\mathcal{K}(\mathbf{X}, H_0/H_0)$ имеет асимптотически (при $n \to +\infty$) χ^2_{k-1}-распределение с $k - 1$ степенями свободы.

Убедимся в этом при $k = 2$. В данном случае

$$\mathcal{K}(\boldsymbol{X}, H_0/H_0) = \frac{(n_1 - np_1^0)^2}{np_1^0} + \frac{(n_2 - np_2^0)^2}{np_2^0}.$$

Очевидно, что $n_2 = n - n_1$, $p_2^0 = 1 - p_1^0$ и $n_2 - np_2^0 = n - n_1 - np_2^0 =$

$= n(1-p_2^0) - n_1 = np_1^0 - n_1$. Значит,

$$\mathcal{K}(\boldsymbol{X}, H_0/H_0) = (n_1 - np_1^0)^2 \left(\frac{1}{np_1^0} + \frac{1}{np_2^0}\right) =$$

$$= (n_1 - np_1^0)^2 \frac{(p_1^0 + p_2^0)}{np_1^0 p_2^0} = \frac{(n_1 - np_1^0)^2}{np_1^0(1-p_1^0)} = \left(\frac{\dfrac{n_1}{n} - p_1^0}{\sqrt{\dfrac{p_1^0(1-p_1^0)}{n}}}\right)^2.$$

Случайную величину n_1 можно трактовать как число успехов в n испытаниях Бернулли, тогда $\dfrac{n_1}{n}$ – частота успеха.

Если гипотеза H_0 верна, то по интегральной теореме Муавра–Лапласа частота успеха имеет асимптотически (при $n \to +\infty$) нормальное распределение:

$$\frac{n_1}{n} \to N\left(p_1^0, \frac{p_1^0(1-p_1^0)}{n}\right).$$

Тогда случайная величина

$$\frac{\dfrac{n_1}{n} - p_1^0}{\sqrt{\dfrac{p_1^0(1-p_1^0)}{n}}} \to N(0;1).$$

Наконец, согласно определению „Хи-квадрат“-распределения, квадрат стандартной нормальной случайной величины будет иметь „Хи-квадрат“-распределение с одной степенью свободы:

$$\left(\frac{\left(\dfrac{n_1}{n} - p_1^0\right)}{\sqrt{\dfrac{p_1^0(1-p_1^0)}{n}}}\right)^2 \to \chi_1^2.$$

Таким образом, этим частным примером проиллюстрировано утверждение об асимптотическом распределении критериальной функции.

В качестве критической области Q возьмем область первого типа – область больших положительных значений $\mathcal{K}(\mathbf{X}, H_0)$. Действительно, если гипотеза H_0 верна, то значение $\mathcal{K}(\boldsymbol{X}, H_0/H_0)$ должно быть небольшим.

Большие значения $\mathcal{K}(\boldsymbol{X}, H_0)$ должны свидетельствовать против гипотезы H_0.

Задавшись теперь уровнем значимости α, в качестве порогового значения c_α возьмем квантиль χ^2_{k-1}-распределения по уровню $1-\alpha$ (рис. 3.6):

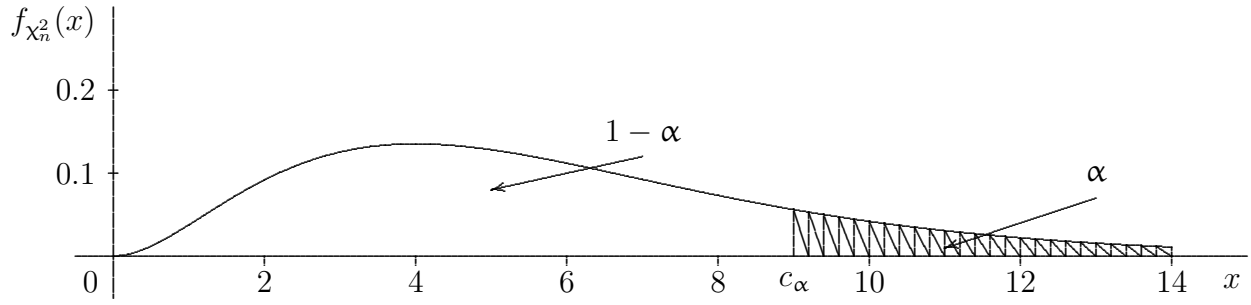

Рис. 3.6. χ^2-распределение. Кривая плотности для $n=6$

$$c_\alpha = \chi^2_{k-1}\ (k-1, 1-\alpha).$$

Решающее правило принимает вид:

если $\mathcal{K}(\boldsymbol{X}, H_0) \geq c_\alpha$, то гипотеза H_0 отклоняется.

Пример. В 100 испытаниях события A_1, A_2 и A_3 наступили соответственно $n_1 = 60,\ n_2 = 21,\ n_3 = 19$ раз. Проверить гипотезу $H_0 : p_1^0 = \frac{1}{2},\ p_2^0 = \frac{1}{4},\ p_3^0 = \frac{1}{4}$ при уровне значимости $\alpha = 0.1$.

В данном случае $k = 3$.

По таблицам χ^2-распределения находим $\chi^2_0(2,\ 0.9) = 4.61$. Вычислим критериальную функцию

$$\mathcal{K}(\boldsymbol{X}, H_0) = \sum_{i=1}^{3} \frac{(n - np_i^0)^2}{np_i^0} = \frac{(60-50)^2}{50} +$$

$$+\frac{(21-25)^2}{25} + \frac{(19-25)^2}{25} = 2.00 + 0.64 + 1.44 = 4.08.$$

Так как $4.08 < 4.61 = c_\alpha$, то отклонить гипотезу H_0 по уровню значимости $\alpha = 0.1$ нет оснований.

Пример. При $n = 4040$ бросаниях монеты Бюффон получил $n_1 = 2048$ выпадений герба и $n_2 = 1992$ выпадений цифры. Проверим гипотезу о симметричности монеты (т. е. гипотезу о биномиальном распределении с вероятностью успеха $p = 0.5$ по уровню значимости $\alpha = 0.05$).

Для биномиального распределения $k = 2$, $c_\alpha = \chi^2_0(1, 0.95) = 3.841$ и

$$\mathcal{K}(\boldsymbol{X}, H_0) = \frac{(n_1 - np_1^0)^2}{np_1^0 p_2^0} = \frac{(2048-2020)^2}{1010} = 0.776.$$

Так как $0.776 < 3.841$, то гипотеза о симметрии монеты не противоречит полученным данным.

Критерий согласия χ^2 (К. Пирсона). В отличие от критерия согласия Колмогорова критерий согласия χ^2 может использоваться для НСВ с произвольной функцией распределения.

Пусть H_0 – простая гипотеза о виде функции распределения и ее параметрах. Сведем задачу проверки гипотезы H_0 к задаче проверки гипотезы о полиномиальном распределении. Для этого разобьем вещественную ось на k промежутков:

$$(y_0, y_1), \ldots, (y_{k-1}, y_k),$$

где $y_0 = -\infty$, $y_k = +\infty$.

Рассмотрим событие $A_j = \{\xi \in (y_{j-1}, y_j)\}$.

Будем считать, что в i-м испытании событие A_j наступило, если $X_i \in \in (y_{j-1}, y_j)$.

Если гипотеза H_0 верна, то распределение ξ полностью определено и можно вычислить

$$P(A_j) = F_\xi(y_j) - F_\xi(y_{j-1}) = p_j^0.$$

Тогда выборка X может быть преобразована к виду (табл. 3.1).

Таблица 3.1

A_1	A_2	$\cdots \cdots \cdots$	A_k
p_1^0	p_2^0	$\cdots \cdots \cdots$	p_k^0
n_1	n_2	$\cdots \cdots \cdots$	n_k

Здесь n_j – количество элементов, попавших в промежуток (y_{j-1}, y_j).

В соответствии с критерием проверки гипотезы о полиномиальном распределении будет справедливо следующее утверждение. При $n \to +\infty$ (см. рис. 3.6)

$$\mathcal{K}(\boldsymbol{X}, H_0) = \sum_{j=1}^{k} \frac{(n_j - np_j^0)^2}{np_j^0} \to \chi^2_{k-r-1},$$

где r – количество утверждений гипотезы H_0 о значениях параметров функции распределения НСВ ξ, оцениваемых по той же выборке X.

Критическая область – область I больших положительных значений (см. рис. 3.2), определяется по заданному уровню значимости α через квантиль χ^2-распределения:

$$c_\alpha = \chi_0^2(k - r - 1, 1 - \alpha).$$

Решающее правило χ^2-критерия согласия выглядит следующим образом:

если $\mathcal{K}(\boldsymbol{X}, H_0) \geq c_\alpha$, то гипотеза H_0 отклоняется.

Приближение к предельному распределению становится достаточно хорошим уже при $n \geq 50$ и $k \geq 5$.

Выбор разбиения вещественной оси на интервалы – задача, слабо поддающаяся теоретическому обоснованию. Обычно считают, что распределение случайной величины $\dfrac{n_j - np_j^0}{\sqrt{np_j^0}}$ приближенно нормально, поэтому при любом j должно выполняться условие $np_j^0 \geq 10$, т. е. в соответствии с гипотезой H_0 в среднем на каждый интервал должно попасть не менее 10 элементов выборки. Если на некоторых интервалах это условие не выполняется, то такие интервалы следует укрупнить, объединяя их с соседними. При этом число оставшихся интервалов не должно быть меньше пяти.

Некоторую стандартизацию может дать следующая процедура. При объеме выборки $n \geq 100$ берут 10 равновероятных по гипотезе H_0 интервалов (y_{j-1}, y_j).Соответственно, все $p_j^0 = 0.1$.

Пример. Дана группированная выборка объемом $n = 200$ (табл. 3.2).

Таблица 3.2

x_i	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5
m_i	2	2	12	30	50	62	28	11	2	1

Требуется проверить гипотезу о нормальности НСВ ξ при уровне значимости $\alpha = 0.1$ с параметрами, найденными из выборки. Находим:

$$m_\xi = \bar{x} = \sqrt{\frac{1}{200}} \sum_{i=1}^{10} m_i x_i = 2.05; \quad \sigma = s = \sqrt{\frac{1}{200} \sum_{i=1}^{10} m_i x_i^2 - \bar{x}^2} = 0.14.$$

Количество параметров распределения, оцененных по той же выборке, $r = 2$. Расчетный материал в соответствии с критерием χ^2 представлен в (табл. 3.3).

Таблица 3.3

y_{j-1}, y_j	$-\infty$, 1.85	1.85, 1.95	1.95, 2.05	2.05, 2.15	2.15, 2.25	2.25, $+\infty$
n_j	16	30	50	62	28	14
p_j^0	0.076	0.163	0.261	0.261	0.163	0.076

Тогда $c_\alpha = \chi_0^2(k - r - 1, 1 - \alpha) = \chi_0^2(3, 0.9) = 6.25$ и $\mathcal{K}(\boldsymbol{X}, H_0) = 2.96$, значит, гипотезу H_0 сохраняем.

3.5. Гипотеза независимости

Гипотезу независимости можно сформулировать следующим образом. Пусть имеются две НСВ ξ и η с неизвестными функциями распределения F_ξ и F_η и совместной функцией распределения $F_{\xi\eta}$. Гипотеза H_0 независимости ξ и η теперь записывается как

$$H_0 \ : \ F_{\xi\eta}(t_1, t_2) = F_\xi(t_1) F_\eta(t_2).$$

В качестве исходных данных для проверки гипотезы H_0 имеются две связанные выборки $\boldsymbol{X}$ и $\boldsymbol{Y}$, полученные при проведении одного случайного эксперимента.

Проверка гипотезы о независимости двух признаков. Вначале рассмотрим критерий проверки гипотезы о независимости двух групп признаков, которыми обладают объекты генеральной совокупности. Пусть каждый объект генеральной совокупности обладает признаками двух возможных типов. Количество признаков первого типа m, а второго типа – k. При этом объект обладает только одним признаком из m возможных первого типа, и только одним признаком из k возможных второго типа. В результате проведенного обследования n случайно выбранных объектов у них обнаружились признаки двух данных типов в разных сочетаниях, разбивающих все множество обследованных объектов на непересекающиеся классы.

Если обозначить через n_{ij} количество объектов, обладающих i-м признаком первого типа ($i = \overline{1,m}$) и j-м признаком второго типа ($j = \overline{1,k}$), то всю полученную выборку можно представить в виде табл. 3.4.

Таблица 3.4

	1	2	$\cdots$	j	$\cdots$	k	
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	ν_1
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2k}	ν_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	ν_i
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
m	n_{m1}	n_{m2}	$\cdots$	n_{mj}	$\cdots$	n_{mk}	ν_m
	μ_1	μ_2	$\cdots$	μ_j	$\cdots$	μ_k	n

В таблице признаки второго типа откладываются горизонтально, а признаки первого типа – вертикально.

Очевидно условие $\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{k} n_{i_j} = n.$

Дополнительно в табл. 3.4 обозначены

$$\nu_i = \sum_{j=1}^{k} n_{ij}, \quad i = \overline{1,m}, \quad \mu_j = \sum_{i=1}^{m} n_{ij} \quad j = \overline{1,m}.$$

Пусть p_{ij} – вероятность события, что наудачу выбранный объект обладает i-м признаком первого типа и j-м признаком второго типа. Соответственно, p_{I_i} – вероятность события, что наудачу выбранный объект обладает i-м признаком первого типа, а p_{II_j} – j-м признаком второго типа.

Тогда гипотеза H_0 независимости признаков первого и второго типов может быть записана так:

$$H_0: \quad p_{ij} = p_i p_j.$$

В сформулированном виде задача проверки гипотезы H_0 может быть сведена к задаче проверки гипотезы о полиномиальном распределении. Действительно, процедура обследования n объектов позволяет сразу выделить $m \times k$ попарно несовместных событий A_{ij} – обладание наудачу выбранным объектом i-м признаком первого типа и j-м признаком второго типа с вероятностями $P(A_{ij}) = p_{ij}$, назначенными в гипотезе H_0.

Тогда можно воспользоваться критерием χ^2, критериальная функция которого примет вид

$$\mathcal{K}(\boldsymbol{X}, H_0) = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(n_{ij} - n p_i p_j)^2}{n p_i p_j} \to \chi^2.$$

Таким образом, гипотеза о независимости эквивалентна гипотезе о том, что существует $m + k$ вероятностей p_i и p_j, таких, что

$$p_{ij} = p_i p_j, \quad \sum_i p_i = \sum_j p_j = 1.$$

Так как вероятности p_i и p_j неизвестны, то естественно взять вместо этих вероятностей их точечные оценки:

$$\widehat{p}_i = \frac{\nu_i}{n}, \quad \widehat{p}_j = \frac{\mu_j}{n}.$$

Тогда

$$\mathcal{K}(\boldsymbol{X}, H_0) = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(n_{ij} - \dfrac{\nu_i \mu_j}{n}\right)^2}{\dfrac{\nu_i \mu_j}{n}} = n \left(\sum_{i=1}^{m} \sum_{j=1}^{k} \frac{n_{ij}^2}{\nu_i \mu_j} - 1 \right).$$

Для подсчета числа степеней свободы предельного χ^2-распределения критериальной функции учтем наличие $m\,k$ событий, т. е. $m\,k - 1$ назначаемых вероятностей и $m + k - 2$ параметров, оцениваемых по той же выборке.

Значит, число степеней свободы предельного χ^2-распределения есть

$$m\,k - (m + k - 2) - 1 = (m - 1)(k - 1).$$

Критическая область задается пороговым значением

$$c_\alpha = \chi^2_0((m - 1)(k - 1), 1 - \alpha),$$

равным квантилю χ^2-распределения с $(m - 1)(k - 1)$ числом степеней свободы по уровню $1 - \alpha$, где α – заданный уровень значимости критерия.

Пример [2]. В результате обследования мужской части населения одного английского города были выявлены пристрастия к двум традиционным видам спорта: крикету и теннису, в зависимости от возраста респондентов (до 45 лет и после 45 лет).

Результаты приведены в табл. 3.5.

Таблица 3.5

Вид спорта	Возраст		Всего
	Не старше 45 лет	Старше 45 лет	
Крикет	58	74	132
Теннис	24	38	62
Всего	82	112	194

Проверим по критерию независимости двух групп признаков χ^2, влияет ли возраст на выбор вида спорта по уровню значимости $\alpha = 0.1$.

Так как $m = k = 2$, то $c_\alpha = \chi^2_0(1, 0.9) = 2.706$ и

$$\mathcal{K}(\boldsymbol{X}, H_0) = n\left(\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{n_{ij}^2}{\nu_i \mu_j} - 1\right) = 0.423.$$

Так как $\mathcal{K}(\boldsymbol{X}, H_0) < c_\alpha$, отвергать гипотезу независимости оснований нет.

Проверка гипотезы независимости в общем случае. Приведенный критерий χ^2 проверки гипотезы независимости для двух групп признаков может быть приспособлен и для проверки общей гипотезы независимости.

В случае дискретных НСВ ξ и η их можно трактовать как случайные признаки, характеризующие объекты генеральной совокупности. Получаемые при проведении единичного случайного эксперимента значения ξ и η дают комбинацию признаков, которая регистрируется в выборке. Таким образом, задача проверки гипотезы независимости ξ и η сводится к задаче проверки гипотезы о независимости двух групп признаков.

Если НСВ ξ и η имеют абсолютно непрерывные распределения, то для использования критерия χ^2 необходимо сформировать две группы признаков. Для этих целей можно применить процедуру разбиения вещественной оси на интервалы, описанную в 3.4. Попадание значений ξ и η, полученных в результате единичного случайного эксперимента, в соответствующие

интервалы, будем интерпретировать как проявление соответствующей комбинации признаков у обследуемого объекта.

3.6. Проверка гипотез однородности

На практике встречаются ситуации, когда требуется проверить отсутствие изменений в законе распределения НСВ ξ. Такие изменения могут возникнуть, когда меняются условия проведения случайного эксперимента. Различного вида гипотезы об отсутствии изменений в законе распределения НСВ ξ называют гипотезой однородности.

Критерий знаков. Одним из наиболее простых критериев проверки гипотезы однородности является критерий знаков.

Пусть $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$ и $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_n]^T$ – выборки из распределений двух НСВ ξ и η с функциями распределения F_ξ и F_η . По гипотезе $H_0 : F_\xi = F_\eta$.

Будем полагать,что выборки $\boldsymbol{X}$ и $\boldsymbol{Y}$ не содержат одинаковых соответствующих по номеру элементов. В противном случае эти совпадающие элементы из выборок исключаем.

Построим новую выборку $\boldsymbol{Z} = [Z_1, Z_2, \ldots, Z_n]^T$,

$$Z_i = \operatorname{sign}(X_i - Y_i) = \begin{cases} -1, & \text{если } X_i < Y_i, \\ +1, & \text{если } X_i > Y_i, \end{cases} i = \overline{1, n}.$$

Найдем распределение Z_k, когда верна гипотеза H_0.

Так как случайные величины X_k и Y_k независимы и их распределение одинаково, то распределение Z_k имеет вид табл. 3.6.

Таблица 3.6

a_i	-1	1
$P(Z_k = a_i)$	1/2	1/2

Таким образом, выборку Z можно рассматривать как последовательность успехов ($Z_i = 1$) и неудач ($Z_i = -1$) в n испытаниях Бернулли с вероятностью успеха $p = \dfrac{1}{2}$ в одном испытании. Если μ_n^+ – число успехов в n испытаниях, то распределение μ_n^+ – биномиальное с вероятностью успеха $p = \dfrac{1}{2}$:

$$P(\mu_n^+ = k) = C_n^k p^k (1-p)^{n-k} = C_n^k \frac{1}{2^n}.$$

Если взять μ_n^+ в качестве критериальной функции, то критическая область $Q(\alpha)$ строится следующим образом. Выберем число r_α так, чтобы выполнялись условия:

$$P\left(\left|\mu_n^+ - \frac{n}{2}\right| \geq r_\alpha\right) = P\left(\mu_n^+ \geq \frac{n}{2} + r_\alpha\right) + P\left(\mu_n^+ \leq \frac{n}{2} - r_\alpha\right) = \alpha.$$

При этом в силу симметрии

$$P(\mu_n^+ \geq \frac{n}{2} + r_\alpha) = \sum_{i=[\frac{n}{2}]+r_\alpha}^{n} C_n^i \frac{1}{2^n} = \frac{\alpha}{2},$$

$$P(\mu_n^+ \leq \frac{n}{2} - r_\alpha) = \sum_{i=0}^{[\frac{n}{2}]-r_\alpha} C_n^i \frac{1}{2^n} = \frac{\alpha}{2}.$$

Если $|\mu_n^+ - \frac{n}{2}| \geq r_\alpha$, то гипотеза H_0 отклоняется.

Для перехода к односторонней критической области обозначим $\mu_n = \min\{\mu_n^+, \mu_n^-\}$, где $\mu_n^- = n - \mu_n^+$. Тогда критерий принимает вид:

если $\mu_n \leq \frac{n}{2} - r_\alpha$, то H_0 отклоняется.

Для выбора значения r_α имеются специальные таблицы, рассчитываемые по формулам биномиального распределения.

Пример. Пусть нужно сравнить две партии транзисторов по распределению их коэффициента усиления по току β. По гипотезе H_0 распределение β в обеих партиях одинаково при уровне значимости $\alpha = 0.1$. У каждой пары транзисторов измеряется коэффициент усиления по току β. Результаты измерений: $\mu_n^+ = 20$, $\mu_n^- = 18$.

Из таблиц для значений r_α находим, что $\frac{n}{2} - r_\alpha = 13$. Критическая область $Q = [0, 13]$, $\mu_n = \min\{20, 18\} = 18$. Так как $13 < 18$, то отклонить гипотезу H_0 нет оснований при уровне значимости $\alpha = 0.1$.

Критерий знаков довольно грубый, так как происходит потеря информации при переходе от случайных величин X_i и Y_i к Z_i. Кроме того, он требует выборок одинакового объема. Однако этот критерий очень прост в использовании и хорошо подходит для проведения экспресс-анализа.

Критерий Вилкоксона. Критерий проверки гипотезы однородности Вилкоксона относится к группе ранговых критериев, использующих ранги элементов выборки.

Пусть дана выборка $\boldsymbol{X}$. Расположим выборку в вариационный ряд:

$$X_{n_1} \leq X_{n_2} \leq \cdots \leq X_{n_n}.$$

Рангом элемента X_i называется число R_i, равное номеру элемента X_i в вариационном ряду. Выборке $\boldsymbol{X}$, таким образом, соответствует вектор рангов $\boldsymbol{R} = [R_1, \ldots, R_n]^T$.

Пусть для абсолютно непрерывных НСВ ξ и η получены две выборки $\boldsymbol{X} = [X_1, \ldots, X_n]^T$ и $\boldsymbol{Y} = [Y_1, \ldots, Y_m]^T$.

В соответствии с гипотезой H_0 распределения ξ и η одинаковы, т. е. $F_\xi = F_\eta$.

Расположим обе выборки в единый вариационный ряд.

Пусть $R_1, \ldots, R_n$ – ранги величин $X_1, \ldots, X_n$ в едином вариационном ряду. Пусть $T = R_1 + R_2 + \cdots + R_n$ – сумма рангов в едином вариационном ряду элементов выборки X.

Введем случайные величины

$$Z_{ij} = \begin{cases} 0, \text{ если } X_i \geq Y_j, \\ 1, \text{ если } X_i < Y_j \quad (i = \overline{1,n},\ j = \overline{1,m}), \end{cases}$$

и пусть

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}.$$

Здесь U – общее число случаев, когда элементы выборки $\boldsymbol{X}$ предшествуют элементам выборки $\boldsymbol{Y}$ в общем вариационном ряду.

Статистики T и U связаны линейным соотношением

$$T + U = nm + n(n+1)/2.$$

Вычислим первые 2 момента статистики U, когда верна гипотеза H_0, т. е. элементы выборок $\boldsymbol{X}$ и $\boldsymbol{Y}$ в общем вариационном ряду перемешаны равномерно:

$$M(U) = \sum_{i=1}^{n} \sum_{j=1}^{m} M(Z_{ij}) = n\,m P(X_1 < Y_1) = n\,m \frac{1}{2},$$

$$D(U) = nm(n+m+1)/12.$$

При $n, m \longrightarrow +\infty$ предельным распределением для U является

$$N\left(\frac{n\,m}{2},\ \frac{n\,m\,(n+m+1)}{12}\right).$$

Этим распределением можно пользоваться уже при $n, m \geq 4$, $n + m \geq 20$.

Критическую область для альтернативной гипотезы $H_1 : F_\xi \neq F_\eta$ строим для заданного уровня значимости α следующим образом:

$$Q \quad : \quad P\left(\left|U - \frac{nm}{2}\right| \geq t_{1-\alpha}\sqrt{\frac{nm(n+m+1)}{12}}\right) = \alpha,$$

где $t_{1-\alpha}$ – симметричный нормальный квантиль по уровню $1 - \alpha$. Окончательно критерий Вилкоксона выглядит так: критериальная функция:

$$\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0) = \left|\frac{nm}{2} + \frac{n(n+1)}{2} - T\right|;$$

решающее правило: если $\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0) \geq c_\alpha$, где

$$c_\alpha = t_{1-\alpha}\sqrt{\frac{nm(n+m+1)}{12}},$$

то гипотезу H_0 отклоняем.

Пример. Даны выборки

$$\mathbf{X} = [15, 23, 25, 26, 28, 29, 31, 32]^T, \quad n = 8;$$

$$\mathbf{Y} = [12, 14, 16, 18, 20, 21, 22, 24, 27, 30, 34, 35]^T, \quad m = 12.$$

Проверим гипотезу об однородности выборок по критерию Вилкоксона при уровне значимости $\alpha = 0.1$.

Составим общий вариационный ряд, выделив элементы выборки $\boldsymbol{X}$:

12	14	[15]	16	18	20	21	22	[23]	24
1	2	3	4	5	6	7	8	9	10
[25]	[26]	27	[28]	[29]	30	[31]	[32]	34	35
11	12	13	14	15	16	17	18	19	20

$$T = 3 + 9 + 11 + 12 + 14 + 15 + 17 + 18 = 99;$$

$$c_\alpha = 1.64\sqrt{\frac{8 \cdot 12 \cdot 21}{12}} = 21.26;$$

$$\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0) = \left|\frac{8 \cdot 12}{2} + \frac{8 \cdot 9}{2} - 99\right| = 15.$$

Так как $\mathcal{K}(\boldsymbol{X}, \boldsymbol{Y}, H_0) < c_\alpha$, то гипотезу H_0 сохраняем.

3.7. Выбор из двух простых параметрических гипотез. Критерий Неймана–Пирсона

Как отмечалось в 3.1, наилучшим критерием с точки зрения его мощности является РНМ-критерий. К сожалению, РНМ-критерии существуют, скорее, как исключения лишь в некоторых простых случаях. Вместе с тем, если рассматривать задачу выбора из двух простых параметрических гипотез (при заданных основной гипотезе H_0 и альтернативной H_1), то можно доказать существование наиболее мощного критерия. Такой критерий носит название критерия Неймана–Пирсона.

Пусть распределение $F_\xi(t)$ НСВ ξ зависит от параметра θ. Рассмотрим две простые гипотезы:

$$H_0 \ : \ \theta = \theta_0, \quad F_\xi(t) = F_\xi(t, \theta_0) = F_0(t);$$

$$H_1 \ : \ \theta = \theta_1, \quad F_\xi(t) = F_\xi(t, \theta_1) = F_1(t).$$

Ставится задача по выборке $\boldsymbol{X}$ выбрать одну из этих гипотез.

Пусть распределениям $F_0(t)$ и $F_1(t)$ соответствуют плотности $f_0(t)$ и $f_1(t)$, удовлетворяющие условию

$$f_j(t) > 0, \quad j = 1, 2; \ t \in \mathbb{R}.$$

Обозначим через $G(X, \theta)$ функцию правдоподобия выборки X и рассмотрим отношение правдоподобия

$$L_n(\boldsymbol{X}) = \frac{G(\boldsymbol{X}, \theta_1)}{G(\boldsymbol{X}, \theta_0)} = \prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)}.$$

Определим функцию

$$\varphi(c) = P(L_n(\boldsymbol{X}) \geq c / H_0).$$

С ростом c эта функция убывает. При $c = 0$, очевидно, $\varphi(0) = 1$. Тогда

$$P(L_n(\boldsymbol{X}) \geq c/H_1) = \int\limits_{\boldsymbol{X}:\, L_n(\boldsymbol{X}) \geq c} G(\boldsymbol{X}, \theta_1) dx \geq$$

$$\geq c \int\limits_{\boldsymbol{X}:\, L_n(\boldsymbol{X}) \geq c} G(\boldsymbol{X}, \theta_0) dx = cP(L_n(\boldsymbol{X}) \geq c/H_0) = c\varphi(c).$$

Так как вероятность не превосходит единицы, то $1 \geq c\varphi(c)$, или $\varphi(c) \leq \leq \dfrac{1}{c}$. Если $c \to +\infty$, то $\varphi(c) \to 0$.

Пусть существует такое значение c_α, что для $0 \leq \alpha \leq 1 \quad \varphi(c_\alpha) = \alpha$. Например, такое условие всегда выполнено, если $\varphi(c)$ – непрерывна.

Докажем теперь следующее утверждение, известное как лемма Неймана – Пирсона.

Лемма 3.1. *При сделанных ранее предположениях существует наиболее мощный критерий проверки гипотезы H_0 против альтернативы H_1 с критической областью*

$$Q^*(\alpha) = \{\boldsymbol{X} : L_n(\boldsymbol{X}) \geq c_\alpha\},$$

где порог c_α определяется условием $\varphi(c_\alpha) = \alpha$.

Доказательство. Рассмотрим любой другой критерий уровня значимости α, определяемый критической областью $Q(\alpha)$ (рис. 3.7). Его мощность

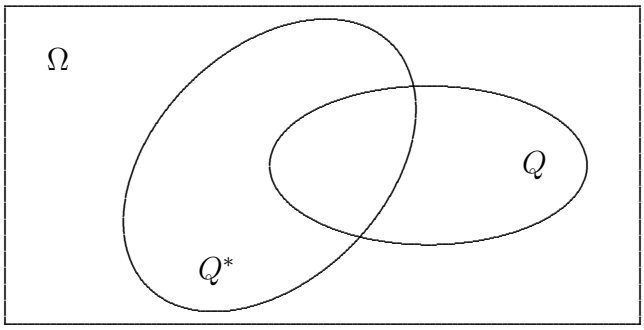

Рис. 3.7

$$\beta = P(\boldsymbol{X} \in Q(\alpha)/H_1) = \int\limits_{Q(\alpha)} G(\boldsymbol{X}, \theta_1)dx =$$

$$= \int\limits_{Q(\alpha)\cap Q^*(\alpha)} G(\boldsymbol{X}, \theta_1)dx + \int\limits_{Q(\alpha)\cap\overline{Q^*(\alpha)}} G(\boldsymbol{X}, \theta_1)dx.$$

Аналогично

$$\beta^* = P(\boldsymbol{X} \in Q^*(\alpha)/H_1) = \int\limits_{Q^*(\alpha)} G(\boldsymbol{X}, \theta_1)dx =$$

$$= \int\limits_{Q^*(\alpha)\cap Q(\alpha)} G(\boldsymbol{X}, \theta_1)dx + \int\limits_{Q^*(\alpha)\cap\overline{Q(\alpha)}} G(\boldsymbol{X}, \theta_1)dx.$$

Тогда

$$\beta = \beta^* - \int\limits_{Q^*(\alpha)\cap\overline{Q(\alpha)}} G(\boldsymbol{X}, \theta_1)dx + \int\limits_{Q(\alpha)\cap\overline{Q^*(\alpha)}} G(\boldsymbol{X}, \theta_1)dx.$$

В точках множества $Q^*(\alpha)$ $L_n(\boldsymbol{X}) \geq c_\alpha$, в точках множества $\overline{Q^*(\alpha)}$ $L_n(\boldsymbol{X}) < c$. Значит,

$$\beta = \beta^* + \int\limits_{Q^*(\alpha)\cap\overline{Q(\alpha)}} (-\alpha_n(X))G(X, \theta_0)dx + \int\limits_{\overline{Q^*(\alpha)}\cap Q(\alpha)} (\alpha_n(\boldsymbol{X}))G(\boldsymbol{X}, \theta_0)dx \leq$$

$$\leq \beta^* - c_\alpha \int\limits_{Q^*(\alpha)\cap\overline{Q(\alpha)}} G(\boldsymbol{X}, \theta_0)dx + c_\alpha \int\limits_{\overline{Q^*(\alpha)}\cap Q(\alpha)} G(\boldsymbol{X}, \theta_0)dx =$$

$$= \beta^* + c_\alpha \left(\int\limits_{\overline{Q^*(\alpha)}\cap Q(\alpha)} G(\boldsymbol{X}, \theta_0)dx - \int\limits_{Q^*(\alpha)\cap\overline{Q(\alpha)}} G(\boldsymbol{X}, \theta_0)dx \right). \qquad (3.1)$$

По условию $P(\boldsymbol{X} \in Q^*(\alpha)/H_0) = P(\boldsymbol{X} \in Q(\alpha)H_0) = \alpha$, или

$$\int\limits_{Q^*(\alpha)} G(\boldsymbol{X}, \theta_0)dx = \int\limits_{Q(\alpha)} G(\boldsymbol{X}, \theta_0)dx = \alpha. \tag{3.2}$$

Область интегрирования в (3.1) в одном интеграле представляет собой (рис. 3.7) $Q(\alpha) \setminus (Q(\alpha) \cap Q^*(\alpha))$, а во втором – $Q^*(\alpha) \setminus (Q(\alpha) \cap Q^*(\alpha))$, поэтому каждый из интегралов в (3.1), с учетом (3.2), отличается от α на величину

$$\int\limits_{Q(\alpha)\cap Q^*(\alpha)} G(X, \theta_0)dx.$$

Значит, сумма интегралов в (3.1) равна нулю и $\beta \leq \beta^*$, т. е. критерий Q^* – наиболее мощный. ■

Критерий Неймана–Пирсона часто называют также критерием отношения правдоподобия.

Пример. Пусть НСВ ξ имеет нормальное распределение с известной дисперсией σ^2 и неизвестным средним θ. По гипотезе H_0 $\theta = \theta_0$, по гипотезе H_1 $\theta = \theta_1$.

Построим критерий Неймана–Пирсона. В этом случае отношение правдоподобия имеет вид

$$L_n(X) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(X_i-\theta_1)^2 - (X_i-\theta_0)^2\right]\right\} =$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i^2 - 2\theta_1 X_i + \theta_1^2 - X_i^2 + 2\theta_0 X_i - \theta_0^2\right)\right\} =$$

$$= \exp\left\{\frac{n}{\sigma^2}(\theta_1-\theta_0)\overline{X} - \frac{n}{2\sigma^2}(\theta_1^2-\theta_0^2)\right\}.$$

В соответствии с формулировкой критерия Неймана–Пирсона критическая область задается неравенством $L_n(\boldsymbol{X}) \geq c_\alpha$. Это неравенство эквивалентно неравенству

$$\frac{n}{\sigma^2}(\theta_1-\theta_0)\overline{X} - \frac{n}{2\sigma^2}(\theta_1^2-\theta_0^2) \geq \ln c_\alpha,$$

или

$$\overline{X} \geq \frac{\sigma^2}{n}\ \ln c_\alpha\ \frac{1}{\theta_1-\theta_0} + \frac{\theta_1+\theta_0}{2}.$$

Так как по условию примера НСВ имеет нормальное распределение с известной дисперсией σ^2, то в силу линейности математического ожидания

$\overline{X} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i \in N(\theta, \dfrac{\sigma^2}{n})$. Тогда при истинной гипотезе H_0 центрируем и нормируем $\overline{X}$:

$$\overline{X} - \theta_0 \geq \frac{\sigma^2}{n}\ \ln c_\alpha\ \frac{1}{\theta_1 - \theta_0} + \frac{\theta_1 - \theta_0}{2},$$

$$\frac{\overline{X} - \theta_0}{\dfrac{\sigma}{\sqrt{n}}} \geq \frac{\sigma}{\sqrt{n}}\ \ln c_\alpha\ \frac{1}{\theta_1 - \theta_0} + \frac{\sqrt{n}}{2\sigma^2}(\theta_1 - \theta_0).$$

Обозначим правую часть неравенства через $t(c_\alpha)$. Тогда

$$\frac{\overline{X} - \theta_0}{\dfrac{\sigma}{\sqrt{n}}} \geq t(c_\alpha).$$

Таким образом,

$$\varphi(c_\alpha) = P(L_n(\boldsymbol{X}) \geq c_\alpha / H_0) = P\left(\frac{\overline{X} - \theta_0}{\dfrac{\sigma}{\sqrt{n}}} \geq t(c_\alpha)\right) = 1 - \Phi(t(c_\alpha)),$$

где $\Phi(t)$ – функция стандартного нормального распределения. При $c_\alpha > 0$ функция $t(c_\alpha)$ непрерывна, поэтому $\varphi(c_\alpha)$ – непрерывная функция c_α.

Для любого уровня значимости α однозначно определяется пороговое значение c_α из следующих соотношений:

$$\Phi(t_\alpha) = 1 - \alpha, \quad 1 - \Phi(t_\alpha) = \alpha, \quad t(c_\alpha) = t_\alpha.$$

Критическая область Q^*_α определяется неравенством

$$Q^*_\alpha = \left\{ \boldsymbol{X} : \quad \frac{\sqrt{n}(\overline{X} - \theta_0)}{\sigma} \geq t_\alpha \right\}.$$

Как видно из данного соотношения, Q^*_α не зависит от альтернативы H_1 ($\theta = = \theta_1$), значит, построенный критерий Неймана–Пирсона – равномерно наиболее мощный.

4. ПОНЯТИЕ О ПОСЛЕДОВАТЕЛЬНОМ АНАЛИЗЕ

При рассмотрении задач проверки гипотез объем выборки n обычно предполагается фиксированным. Однако в некоторых задачах объем выборки представляет собой случайную величину, зависящую от значений наблюдений.

Например, отбирается 10 конденсаторов с определенным номиналом. Поочередный выбор продолжается до тех пор, пока не наберутся нужные 10 штук. При этом общий объем просмотренных элементов может существенно колебаться. Он не может быть меньше 10 (если повезет), но может быть и бесконечным (с нулевой вероятностью). Подобного рода процедуры называются последовательными. Их отличительной чертой является правило, которое регламентирует процедуру отбора на каждом шаге: прекратить или продолжить выбор. Если последовательная процедура останавливается с вероятностью 1 – ее называют замкнутой, если выбор может продолжаться до бесконечности с нулевой вероятностью – процедуру называют открытой.

Применим последовательный анализ к задаче различения двух простых гипотез относительно параметра θ. По гипотезе $H_0 : \theta = \theta_0$, по гипотезе $H_1 : \theta = \theta_1$.

4.1. Критерий отношения Вальда

Пусть выбираются одно за другим n значений $X_1, \ldots, X_n$ из распределения НСВ ξ с плотностью $f(x, \theta)$ (функция правдоподобия выборки).

Если верна гипотеза H_0, то совместная плотность выборки (функция правдоподобия) есть

$$G(X_1, \ldots, X_n/H_0) = \prod_{i=1}^{n} f(X_i, \theta_0).$$

Если верна гипотеза H_1, то функция правдоподобия

$$G(X_1, \ldots, X_n/H_1) = \prod_{i=1}^{n} f(X_i, \theta_1).$$

Рассмотрим отношение функций правдоподобия

$$L_n = \frac{\prod\limits_{i=1}^{n} f(X_i, \theta_1)}{\prod\limits_{i=1}^{n} f(X_i, \theta_0)}.$$

Предположим, что выбрали 2 числа A и B, соответствующие желаемым вероятностям α и γ ошибок I и II рода. Тогда последовательный критерий отношения правдоподобия (ПКОП) имеет вид: выбор продолжается, пока $B < L_n < A$; когда впервые $L_n \geq A$, принимаем H_1, когда впервые $L_n \leq B$ – принимаем H_0.

Более удобна для вычислений форма

$$\ln B < \ln L_n < \ln A.$$

Если обозначить через $z_i = \ln f(X_i, \theta_1) - \ln f(X_i, \theta_0)$, то тогда

$$\ln B < \sum_{i=1}^{n} z_i < \ln A.$$

Докажем, что ПКОП замкнут, т. е. заканчивается с вероятностью, равной 1. Случайные величины z_i независимы, так как независимы элементы выборки. Пусть дисперсии z_i есть σ^2. Тогда $\sum\limits_{i=1}^{n} z_i$ имеет дисперсию $n\sigma^2$. С ростом n дисперсия увеличивается. Распределение выборочного среднего $\overline{z} = \dfrac{1}{n}\sum\limits_{i=1}^{n} z_i$ стремится в силу центральной предельной теоремы к нормальному с дисперсией $\dfrac{\sigma^2}{n}$. Значит, вероятность того, что

$$\frac{\ln B}{n} < \overline{z} < \frac{\ln A}{n},$$

будет стремиться к нулю, так как к нулю стремятся границы. Значит, рано или поздно решение будет принято.

Рассмотрим теперь вопрос определения A и B. Пусть выборка такова, что для первых $n-1$ испытаний L_n лежит между A и B, а в n-м испытании становится больше A, и принимается H_1. По определению, вероятность получения такой выборки по крайней мере в A раз больше при гипотезе H_1, чем при гипотезе H_0. Вероятность принятия H_1, когда верна H_0, равна α, а вероятность принятия H_1, когда верна H_1, равна $\beta = 1-\gamma$. Следовательно, $1 - \gamma \geq A\alpha$, или $A \leq \dfrac{1-\gamma}{\alpha}$.

Рассуждая аналогично, приходим к выводу, что вероятность получения выборки для принятия в n-м испытании H_0 должна быть по крайней мере в $\dfrac{1}{B}$ раз больше при истинной гипотезе H_0, чем при H_1. Вероятность принятия H_0, когда верна H_0, равна $1 - \alpha$, а вероятность принятия H_0, когда верна $H_1 - \gamma$. Значит, $1 - \alpha \geq \dfrac{1}{B}\gamma$, или $B \geq \dfrac{\gamma}{1-\alpha}$.

Обычно предполагается, что

$$A = \frac{1-\gamma}{\alpha}, \qquad B = \frac{\gamma}{1-\alpha}.$$

4.2. Средний объем выборки ПКОП

Подсчитаем средний объем выборки ПКОП. Рассмотрим последовательность из n одинаково распределенных случайных величин z_i. Если бы n было фиксированным, то

$$M\left(\sum_{i=1}^{n} z_i\right) = nM(z_1).$$

Для последовательного выбора это неверно, так как n – случайная величина.

Справедлива следующая теорема.

Теорема 4.1. $M\left\{\sum\limits_{i=1}^{n} z_i\right\} = M\{n\}M\{z_1\}$.

Доказательство. Пусть каждое z_i имеет среднее μ,

$$M\{|z_i|\} \le c < +\infty,$$

и пусть вероятность того, что n примет значение k, есть p_k, $k = 1, 2, \ldots$. Пусть $\varkappa_i$ – индикатор события Ω_i: $\varkappa_i = 1$, если z_i наблюдалось, т. е. $n \ge i$, и 0 в противном случае. Тогда

$$P(\varkappa_i = 1) = P(n \ge i) = \sum_{j=i}^{+\infty} p_j.$$

Пусть $\Lambda_n = \sum\limits_{i=1}^{n} z_i$.

Тогда $\Lambda_n = \sum\limits_{i=1}^{+\infty} \varkappa_i z_i$,

$$M\{\Lambda_n\} = M\left\{\sum_{i=1}^{+\infty} \varkappa_i z_i\right\} = \sum_{i=1}^{+\infty} M\{\varkappa_i z_i\}.$$

$M\{\Lambda_n\}$ конечно, если конечно $M\{n\}$. Кроме того, $\varkappa_i$ зависит только от $z_1, \ldots, z_{i-1}$ и не зависит от z_i, поэтому

$$M\{\varkappa_i z_i\} = M\{\varkappa_i\}M\{z_i\}.$$

Следовательно,

$$M\{\Lambda_n\} = \sum_{i=1}^{+\infty} M\{\varkappa_i\}M\{z_i\} = \mu \sum_{i=1}^{+\infty} M\{\varkappa_i\},$$

$$M\{\varkappa_i\} = 1P(\varkappa_i = 1) = \sum_{j=i}^{+\infty} p_j,$$

$$\mu \sum_{i=1}^{+\infty} \sum_{j=i}^{+\infty} p_j = \mu \sum_{i=1}^{+\infty} (p_i + p_{i+1} + \dots) = \mu \sum_{i=1}^{+\infty} i p_i = \mu M\{n\}.$$

Тогда средний объем выборки

$$M\{n\} = \frac{M\{\Lambda_n\}}{M\{z_1\}}. \blacksquare$$

Вероятность сохранить неверную гипотезу H_0 обозначена γ (Λ_n в этом случае принимает значение, равное $\ln B$), а гипотеза H_1 принимается с вероятностью $1 - \gamma$ ($\Lambda_n = \ln A$), когда она верна. Таким образом, если верна гипотеза H_1, то средний объем выборки

$$M\{n\} = \frac{\gamma \ln B + (1 - \gamma) \ln A}{M\{z_i\}}.$$

Если верна гипотеза H_0, то она принимается с вероятностью $1 - \alpha$ ($\Lambda_n = \ln B$) и отклоняется с вероятностью α ($\Lambda_n = \ln A$). Таким образом,

$$M\{n\} = \frac{(1 - \alpha) \ln B + \alpha \ln A}{M\{z_1\}}.$$

Пример. Построим ПКОП для проверки гипотезы о среднем НСВ, распределенных по закону $N(m, 1)$.

$$H_0: \quad m = m_0;$$

$$H_1: \quad m = m_1, \qquad m_0 \neq m_1.$$

Плотность распределения НСВ

$$f(x, m) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x - m)^2\right\}.$$

Тогда

$$z_i = \ln f(X_i, m_1) - \ln f(X_i, m_0) = -\frac{1}{2}(X_i - m_1)^2 + \frac{1}{2}(X_i - m_0)^2 =$$
$$= (m_1 - m_0)X_i - \frac{1}{2}(m_1^2 - m_0^2)$$

и

$$\Lambda_n = \sum_{i=1}^{n} z_i = n(m_1 - m_0)\overline{X} - \frac{1}{2} n(m_1^2 - m_0^2).$$

Решающее правило $\ln B < \Lambda_n < \ln A$. При этом $\ln B = \ln \dfrac{\gamma}{1-\alpha}$; $\ln A = \ln \dfrac{1-\gamma}{\alpha}$. Далее находим:

$$M\{z_i\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \left\{ (m_1 - m_0)x - \frac{1}{2}(m_1^2 - m_0^2) \right\} \exp\left\{ -\frac{1}{2}(x-m)^2 \right\} dx =$$

$$= \frac{m_1 - m_0}{2\pi} \int_{-\infty}^{+\infty} x \exp\left\{ -\frac{1}{2}(x-m)^2 \right\} dx -$$

$$- \frac{\frac{1}{2}(m_1^2 - m_0^2)}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left\{ -\frac{1}{2}(x-m)^2 \right\} dx =$$

$$= -\frac{1}{2}(m_1^2 - m_0^2) + \frac{m_1 - m_0}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (x-m) \exp\left\{ -\frac{1}{2}(x-m)^2 \right\} dx +$$

$$+ \frac{m_1 - m_0}{\sqrt{2\pi}} m \int_{-\infty}^{+\infty} \exp\left\{ -\frac{1}{2}(x-m)^2 \right\} dx =$$

$$= -\frac{1}{2}(m_1^2 - m_0^2) + m(m_1 - m_0).$$

Подставляя сюда различные значения m, можно определить средний объем выборки.

Так, если верна H_1, то $m = m_1$ и

$$M\{z_i\} = \frac{1}{2}(m_1 - m_0)^2,$$

следовательно,

$$M\{n\} = \frac{\gamma \ln \dfrac{\gamma}{1-\alpha} + (1-\gamma) \ln \dfrac{1-\gamma}{\alpha}}{\dfrac{1}{2}(m_1 - m_0)^2}.$$

Если, например, $m_0 = 0$, $m_1 = 1$, $\alpha = \gamma = 0.1$, то $M\{n\} = 3.52$ (если верна H_1). Если $\alpha = \gamma = 0.01$, то $M\{n\} = 9.00$.

5. ПОНЯТИЕ О КОРРЕЛЯЦИОННОМ АНАЛИЗЕ

5.1. Корреляция

При решении различных задач бывает необходимо знать, связаны или не связаны между собой две или более случайные величины. Решение задач

такого рода в инженерной практике обычно сводится к выявлению зависимости между некоторой предполагаемой возмущающей силой и наблюдаемой реакцией исследуемой физической системы. Существование этих связей и их тесноту можно выразить коэффициентом корреляции $\rho(\xi, \eta)$.

Когда рассматриваются две случайные величины ξ и η, $\rho(\xi, \eta)$ определяется как

$$\rho(\xi, \eta) = \frac{\operatorname{cov}(\xi, \eta)}{\sigma_\xi \sigma_\eta},$$

где $\operatorname{cov}(\xi, \eta))$ – корреляционный момент (ковариация) ξ и η, определяемый как

$$\operatorname{cov}(\xi, \eta) = M\{(\xi - M(\xi))(\eta - M(\eta))\}.$$

Предположим, что производится выборка из распределений двух НСВ ξ и η, в результате чего получается n пар наблюденных значений $\boldsymbol{X} = (x_1, \ldots, x_n)^T$, $\boldsymbol{Y} = y_1, \ldots, y_n)^T$. Если $\overline{x}$ и $\overline{y}$, S_x и S_y – выборочные средние и среднеквадратичные отклонения, а

$$S_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}),$$

то коэффициент корреляции можно оценить по методу выборочных моментов следующим образом:

$$r(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\rho}(\xi, \eta) = \frac{S_{xy}}{S_x S_y} = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2 \sum\limits_{i=1}^{n} (y_i - \overline{y})^2}} =$$

$$= \frac{\sum\limits_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\left(\sum\limits_{i=1}^{n} x_i^2 - n\overline{x}^2\right)\left(\sum\limits_{i=1}^{n} y_i^2 - n\overline{y}^2\right)}}. \qquad (5.1)$$

Как и коэффициент корреляции $\rho(\xi, \eta)$, выборочный коэффициент корреляции $r(\boldsymbol{X}, \boldsymbol{Y})$ лежит в пределах от -1 до 1. Граничные значения достигаются, когда наблюдения обнаруживают идеальную линейную зависимость. Если же зависимость отлична от линейной или же наблюдается разброс измеренных значений, то независимо от того, обусловлено ли это обстоятельство ошибками измерений или нелинейным характером связи исследуемых величин ξ и η, коэффициент $r(\boldsymbol{X}, \boldsymbol{Y})$ уменьшается.

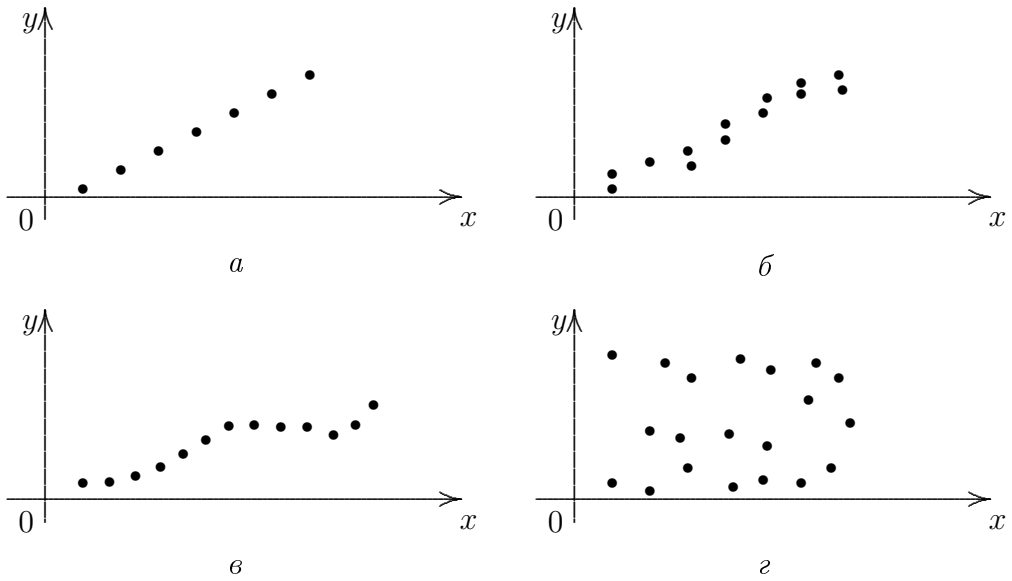

Рис. 5.1. Символом • отмечены на плоскости точки (x_i, y_i)

На рис. 5.1: *а* – $r(\boldsymbol{X}, \boldsymbol{Y}) = 1$ (идеальная линейная корреляция); *б* – $0 < r(\boldsymbol{X}, \boldsymbol{Y}) < 1$ (линейная корреляция с умеренным рассеянием); *в* – $0 < r(\boldsymbol{X}, \boldsymbol{Y}) < 1$ (нелинейная корреляция); *г* – $r(\boldsymbol{X}, \boldsymbol{Y}) \approx 0$ (отсутствие корреляции).

Для того чтобы оценить точность полученной оценки $r(\boldsymbol{X}, \boldsymbol{Y})$ (5.1), рассмотрим случайную величину

$$\omega = \frac{1}{2} \ln \left(\frac{1 + r(\boldsymbol{X}, \boldsymbol{Y})}{1 - r(\boldsymbol{X}, \boldsymbol{Y})} \right) . \tag{5.2}$$

При нормальном распределении наблюдаемых случайных величин $\boldsymbol{X}$ и $\boldsymbol{Y}$ величина ω (5.2) имеет приближенно нормальное распределение со средним значением и дисперсией:

$$m_\omega = \frac{1}{2} \ln \left(\frac{1 + \rho(\xi, \eta)}{1 - \rho(\xi, \eta)} \right) + \frac{\rho(\xi, \eta)}{2(n-1)}, \quad \sigma_\omega^2 = \frac{1}{n-3}. \tag{5.3}$$

Зная оценку $r(\boldsymbol{X}, \boldsymbol{Y})$, нетрудно найти доверительные интервалы для коэффициента $\rho(\xi, \eta)$.

Ввиду выборочной изменчивости оценок коэффициента корреляции желательно убедиться в том, что отличное от нуля значение оценки действительно отражает наличие статистически значимой корреляционной зависимости между исследуемыми величинами.

Для этого достаточно проверить гипотезу $H_0 : \ \rho(\xi, \eta) = 0$. Если гипотеза отвергается, то корреляционную зависимость признают статистически значимой.

Согласно (5.3) распределение случайной величины ω при гипотезе $H_0 \ : \ \rho(\xi, \eta) = 0$ является нормальным с нулевым средним и дисперсией

$\sigma_\omega^2 = \dfrac{1}{n-3}$. Следовательно, область принятия гипотезы H_0 определяется соотношением

$$P(|\sqrt{n-3}\omega| < t_\alpha) = 1 - \alpha,$$

где t_α – симметричный нормальный квантиль по уровню α. Если же

$$|\sqrt{n-3}\omega| \geq t_\alpha,$$

то это означает наличие корреляционной зависимости при уровне значимости α.

Пример. Получены результаты измерений роста x в сантиметрах и массы y в килограммах $n = 25$ студентов, выбранных случайным образом. Есть ли основание считать, что при уровне значимости $\alpha = 0.05$ рост и масса студентов связаны корреляционной зависимостью?

x	175	185	175	163	173	182	180	173	180	190	185	180	170
y	56	84	59	58	73	66	62	68	70	62	74	74	66
x	175	177	170	182	163	182	185	160	180	180	167	182	
y	88	74	72	68	54	70	72	60	68	66	58	68	

Находим значения:

$$\sum_{i=1}^{n} x_i y_i = 299\ 056; \quad \sum_{i=1}^{n} x_i^2 = 780\ 760; \quad \sum_{i-1}^{n} y_i^2 = 115\ 838;$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 176.6, \qquad S_x = 7.54;$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 67.58, \qquad S_y = 7.98.$$

Определяем

$$r(\boldsymbol{X}, \boldsymbol{Y}) = \frac{299\ 056 - 25 \cdot 176.6 \cdot 67.58}{\sqrt{(781\ 163 - 25 \cdot (176.6)^2)(115\ 777 - 25 \cdot (67.58)^2)}} = 0.45,$$

$$\omega = \frac{1}{2}\ln\left(\frac{1 + r(\boldsymbol{X}, \boldsymbol{Y})}{1 - r(\boldsymbol{X}, \boldsymbol{Y})}\right) = 0.485 \text{ и } \sqrt{25-3}\omega = 2.27.$$

Гипотеза $H_0 : \rho(\xi, \eta) = 0$ отклоняется при $\alpha = 0.05$, так как значение $\sqrt{n-3}\omega = 2.27$ лежит вне области принятия гипотезы H_0 при $t_\alpha = 1.96$. Другими словами, можно полагать, что между ростом и массой имеется существенная корреляционная зависимость.

5.2. Применение линейной регрессии

С помощью корреляционного анализа можно установить, насколько тесна связь между двумя или более случайными величинами. Однако в дополнение к этому желательно располагать моделью зависимости, которая

позволяла бы предсказать значение некоторой величины по заданным значениям других величин. Например, в результате корреляционного анализа в примере удалось установить значимую в статистическом смысле зависимость между ростом и массой студентов. Хорошо было бы теперь получить оценку такой зависимости, которая позволила бы предсказывать массу студентов в зависимости от их роста. Методы решения подобных задач составляют предмет регрессионного анализа. Понятие линейной регрессии было введено в 2.6.

Рассмотрим простейший случай двух коррелированных случайных величин ξ и η. Наличие линейной зависимости между двумя этими величинами означает, что η можно вычислить, зная ξ:

$$\eta = A + B\xi. \tag{5.4}$$

При идеальной линейной корреляции ($r(\boldsymbol{X}, \boldsymbol{Y}) = 1$) предсказанное значение $\widetilde{y_i}$ при любом x_i всегда равно измеренному значению y_i. На практике это не так. Обычно наблюдается некоторый разброс точек, обусловленный наличием посторонних случайных помех и, возможно, искажениями за счет нелинейных эффектов. Тем не менее, если допустить существование линейной зависимости между величинами и располагать неограниченно большим количеством измерений, то можно найти такие подходящие значения коэффициентов A и B, которые позволяют предсказать ожидаемые значения y_i при любых заданных x_i. При этом значение $\widetilde{y_i}$ не обязательно совпадает с измеренным значением y_i, соответствующим данному x_i; оно представляет собой некоторое среднее для всех таких измеренных значений.

На практике для определения коэффициентов A и B в (5.4) пользуются методом наименьших квадратов, согласно которому в качестве значений A и B выбираются такие числа, при которых сумма квадратов отклонений измеренных значений от предсказанных была бы минимальной.

Отклонение измеренного значения от предсказанного

$$y_i - \widetilde{y_i} = y_i - (A + Bx_i),$$

поэтому сумма квадратов отклонений

$$Q = \sum_{i=1}^{n} (y_i - A - Bx_i)^2.$$

Наименьшее значение Q достигается в том случае, когда коэффициенты A и B удовлетворяют условию

$$\frac{\partial Q}{\partial A} = \frac{\partial Q}{\partial B} = 0. \tag{5.5}$$

Данные измерений представляют собой ограниченную выборку, состоящую из n пар измеренных значений ξ и η, поэтому условие (5.5) позволяет получить лишь оценки коэффициентов A и B – a и b. Из (5.5) находим:

$$a = \overline{y} - b\overline{x}, \tag{5.6}$$

где

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}. \tag{5.7}$$

Используя оценки (5.6) и (5.7) можно записать формулу для прогноза y при заданном x:

$$\widehat{y} = a + bx = \overline{y} - b\overline{x} + bx = \overline{y} + b(x - \overline{x}). \tag{5.8}$$

Прямая линия, описываемая уравнением (5.8), называется линией регрессии y по x. Меняя местами зависимую и независимую переменные в (5.8), получим линию регрессии x по y:

$$\widehat{x} = \overline{x} + b'(y - \overline{y}),$$

где

$$b' = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} y_i^2 - n(\overline{y})^2}. \tag{5.9}$$

Сравнивая выражения (5.7) и (5.9), нетрудно убедиться, что

$$r(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{bb'}. \tag{5.10}$$

Рассмотрим точность определения оценок a и b. Если величина r (см. (5.10)) при заданном x подчиняется нормальному распределению, то можно показать, что a и b представляют собой несмещенные оценки соответственно коэффициентов A и B. Выборочные распределения этих оценок связаны с τ-распределением Стьюдента следующим образом:

$$\frac{a - A}{\sqrt{\dfrac{1}{n} + \dfrac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} = S_{y/x}\tau_{n-2},$$

$$\frac{b - B}{\sqrt{\dfrac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} = S_{y/x}\tau_{n-2}.$$

Выборочное распределение $\widehat{y}$, соответствующей некоторому фиксированному значению $x = x_0$:

$$\frac{\widehat{y} - \widetilde{y}}{\sqrt{\dfrac{1}{n} + \dfrac{(x_0 - \overline{x})^2}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}}} = S_{y/x}\tau_{n-2}.$$

Во всех формулах $S_{y/x} = \sqrt{\dfrac{\sum\limits_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-2}}$ – выборочное стандартное отклонение измеренных значений y_i от прогноза $\widehat{y}_i = a + bx_i$.

Все эти соотношения позволяют найти доверительные интервалы для A, B и $\widetilde{y}$ на основании оценок a, b и $\widehat{y}$.

Пример. Используя данные предыдущего примера, найдем линию регрессии, с помощью которой можно давать линейный прогноз средней массы студентов в зависимости от их роста.

По формулам (5.6) и (5.7) находим оценки a и b:

$$b = \frac{299\ 056 - 25 \cdot 176.6 \cdot 67.58}{781\ 163 - 25(176.6)^2} = 0.47,$$

$$a = \overline{y} - b\overline{x} = 67.58 - 0.47 \cdot 176.6 = -15.4.$$

Следовательно, уравнение регрессии имеет вид

$$\widehat{y} = -15.4 + 0.47x.$$

Определим теперь 95 %-й доверительный интервал для средней массы студентов, рост которых составляет 180 см. Оценка массы $\hat{y} = 69.2$ кг. Вычислим $S_{y/x}$. Это значение удобнее вычислять по формуле

$$S_{y/x} = \left(\frac{1}{n-2} \left(\sum_{i=1}^{n}(y_i - \overline{y})^2 - \frac{\left(\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \right)^2}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} \right) \right)^{1/2} =$$

$$= \left(\frac{1}{23} \left(1600 - \frac{(690.3)^2}{1474} \right) \right)^{1/2} = 7.45.$$

Тогда 95 %-й доверительный интервал для средней массы студентов с ростом 180 см составляет

$$\widehat{y} \pm S_{y/x} t(n-2, \alpha) \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{1/2} =$$

$$= 69.2 \mp 7.75 t(23, 0.05) \left(\frac{1}{25} + \frac{(180 - 176.6)^2}{1474} \right)^{1/2} =$$

$$= 69.2 \mp 7.45 \cdot 2.069 \cdot 0.219 = 69.2 \mp 3.4 = 65.8 \div 72.6 \text{ кг.}$$

Список литературы

1. Крамер Г. Математические методы статистики.–М.: Мир, 1975.

2. Кендалл М. Дж., Стьюарт А. Статистические выводы и связи.–М.: Гл. ред. физ.-мат. лит., 1973.

3. Коваленко И. Н., Филиппова А. А. Теория вероятностей и математическая статистика.–М.: Высш. шк., 1982.

4. Справочник по прикладной статистике. Т. 1.–М.: Финансы и статистика, 1989.

5. Справочник по прикладной статистике. Т. 2.–М.: Финансы и статистика, 1990.

ОГЛАВЛЕНИЕ

Даугавет Александр Игоревич
Постников Евгений Валентинович
Солынин Андрей Александрович

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие

Редактор Э. К. Долгатов

Подписано в печать 16.10.12. Формат 60×84 1/16. Бумага офсетная.
Печать офсетная. Гарнитура „Times New Roman“. Печ. л. 5,75.
Тираж 200 экз. Заказ

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