Министерство образования и науки Российской Федерации
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

С.Н. ПОСТОВАЛОВ
Е.В. ЧИМИТОВА
В.С. КАРМАНОВ

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Конспект лекций

Утверждено Редакционно-издательским советом университета
в качестве учебного пособия

2-е издание

НОВОСИБИРСК
2017

УДК 519.2(075.8)
П 636

Рецензенты:
д-р техн. наук, доцент *В.С. Тимофеев*
канд. техн. наук, доцент *А.В. Фаддеенков*

Работа подготовлена на кафедре прикладной математики
для студентов III курса ФПМИ

Постовалов С.Н.
П 636 Математическая статистика. Конспект лекций: учеб. пособие / С.Н. Постовалов, Е.В. Чимитова, В.С. Карманов. – 2-е изд. – Новосибирск: Изд-во НГТУ, 2017. – 140 с.

ISBN 978-5-7782-3372-0

Конспект лекций предназначен для проведения лекционных занятий по курсу «Математическая статистика» (направление 010400.62 – «Прикладная математика и информатика»)

УДК 519.2(075.8)

ISBN 978-5-7782-3372-0

Оглавление

Введение

Математическая статистика – наука, разрабатывающая математические методы систематизации и использования статистических данных для научных и практических выводов.

Во многих своих разделах математическая статистика опирается на теорию вероятностей, позволяющую оценить надежность и точность выводов, делаемых на основании ограниченного статистического материала (например, оценить необходимый объем выборки для получения результатов требуемой точности при выборочном обследовании).

Целью математической статистики является разработка методов регистрации, описания и анализа данных наблюдений и экспериментов с целью построения вероятностных моделей массовых случайных явлений. В зависимости от математической природы конкретных результатов наблюдений математическая статистика делится на статистику чисел, многомерный статистический анализ, анализ функций (процессов) и временны́х рядов, статистику объектов нечисловой природы.

Основными разделами статистики являются описательная статистика, теория оценивания и теория проверки гипотез. Описательная статистика есть совокупность эмпирических методов, используемых для визуализации и интерпретации данных (расчет выборочных характеристик, таблицы, диаграммы, графики и т. д.), как правило, не требующих предположений о вероятностной природе данных.

Методы оценивания и проверки гипотез опираются на вероятностные модели происхождения данных. Эти модели делятся на параметрические и непараметрические. В параметрических моделях предполагается, что характеристики изучаемых объектов описываются посредством распределений, зависящих от одного или нескольких числовых параметров. Непараметрические модели не связаны со спецификацией параметрического семейства для распределения изучаемых характеристик. Непараметрические модели в общем случае являются робастными, т. е. устойчивыми, «нечувствительными» к различным отклонениям и неоднородностям в выборке, связанным с теми или иными, в общем случае неизвестными, причинами.

В математической статистике оценивают параметры и функции от них, представляющие важные характеристики распределений (например, математическое ожидание, медиана, стандартное отклонение, квантили и др.), плотности и функции распределения и пр.

Тема 1. Выборочный метод в статистике

1.1. Выборка. Выборочный метод

Пусть $\mathbb{X}_n = \{X_1, ..., X_n\}$ – выборка объема n, полученная в результате наблюдения случайной величины ξ, имеющей распределение (закон распределения) $F_\xi(x)$.

Будем считать, что:

- наблюдения $X_1, ..., X_n$ независимы и имеют одно и то же распределение $F_\xi(x)$;
- $F_{\mathbb{X}_n}(x_1, ..., x_n) = F_\xi(x_1) \cdot F_\xi(x_2) \cdot ... \cdot F_\xi(x_n)$, и нам не важен порядок следования наблюдений;
- множество возможных значений ξ (с распределениями F_ξ) образуют генеральную совокупность $L(\xi)$, которой принадлежит выборка $\mathbb{X}_n$;
- $F_\xi \in F = \{F_\xi(x, \theta), \theta \in \Theta\}$ – параметрическая статистическая модель. Параметр θ может быть как скалярным, так и векторным.

1.2. Порядковые статистики и вариационный ряд

Упорядочим все наблюдения в выборке и произведем их перенумерацию: $X_{(1)} \le X_{(2)} \le X_{(3)} \le ... \le X_{(n)}$ – вариационный ряд.

Определение 1.1. Величина $X_{(i)}$ называется *i*-й порядковой статистикой.

Определение 1.2. Статистикой называется любая измеримая функция от выборки, которая, в свою очередь, также является случайной величиной или случайной функцией.

Найдем распределение *i*-й порядковой статистики.

Введем вспомогательную случайную функцию: $\mu_n(x)$ – количество наблюдений $X_i \le x$.

Найдем $P\{\mu_n(x)=k\}$.

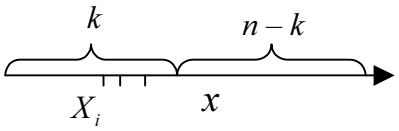


Рис. 1.1. Нахождение вероятности $P\{\mu_n(x)=k\}$

Событие $\mu_n(x)=k$ означает, что в интервал $(-\infty,x]$ попало ровно k наблюдений, а в интервал $(x,+\infty)$ – $(n-k)$ наблюдений.

Число способов, которыми можно выбрать k элементов из n, равно C_n^k, поэтому в результате получаем:

$$P\{\mu_n(x)=k\}=C_n^k F^k(x)\left(1-F(x)\right)^{n-k}.$$

$$P\left\{X_{(i)}\le x\right\}=P\{\mu_n(x)\ge i\}=$$

$$=P\{\mu_n(x)=i\vee\mu_n(x)=i+1\vee\ldots\vee\mu_n(x)=n\}=$$

$$=\sum_{k=i}^{n}P\{\mu_n(x)=k\}=\sum_{k=i}^{n}C_n^k F(x)^k\left(1-F(x)\right)^{n-k}.$$

1.3. Эмпирическая функция распределения

Функция

$$F_n(x)=\frac{\mu_n(x)}{n}$$

называется эмпирической функцией распределения. По определению, эмпирическая функция распределения является случайной функцией;

$\forall x \in R$, $F_n(x)$ – дискретная случайная величина, принимающая значения

$$0 = \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, \frac{n}{n} = 1,$$

при этом $P\left\{F_n(x) = \frac{k}{n}\right\} = P\{\mu_n(x) = k\} = C_n^k F^k(x)\left(1 - F(x)\right)^{n-k}$.

Если все X_i (наблюдения в выборке) различны, то

$$F_n(x) = \begin{cases} 0, & x < X_{(1)}, \\ \frac{k}{n}, & X_{(k)} \le x < X_{(k+1)}, \ k = 1, ..., n-1, \\ 1, & x \ge X_{(n)}, \end{cases}$$

или $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} h(x - x_i)$, где $h(x) = \begin{cases} 0, x < 0 \\ 1, x \ge 0 \end{cases}$ – функция Хевисайда (единичного скачка).

Теорема 1.1

Пусть $F_n(x)$ – эмпирическая функция распределения случайной величины ξ, имеющей функцию распределения $F_\xi(x)$. Тогда

$$\forall |x| < \infty, \forall \varepsilon > 0 \ \lim_{n \to \infty} P\left\{\left|F_n(x) - F(x)\right| < \varepsilon\right\} = 1.$$

Доказательство

Закон больших чисел (теорема Бернулли).

Если η_i – независимые, одинаково распределенные случайные величины, $M\eta_i = a$, то

$$\frac{1}{n}\sum \eta_i \xrightarrow{P} a \ \text{ при } \ n \to \infty$$

$$\left(P\left\{\frac{1}{n}\sum \eta_i - a < \varepsilon\right\} \to 1; \forall \varepsilon > 0, \ n \to \infty \right).$$

Введем случайную величину $\eta_i = h(x - X_i) = \begin{cases} 1, X_i \le x \\ 0, X_i > x \end{cases}$, найдем ее математическое ожидание

$$M\eta_i = Mh(x - X_i) = 1 \cdot P\{X_i \le x\} + 0 \cdot P\{X_i > x\} = F_\xi(x),$$

подставим в Закон больших чисел и получим условия теоремы.

Таким образом, при $n \to \infty$ эмпирическая функция распределения $F_n(x)$ является оценкой теоретической функции распределения $F_\xi(x)$.

Введем статистику $D_n = \sup\limits_{|x|<\infty} |F_n(x) - F(x)|$.

Теорема 1.2 (Гливенко – Кантелли)

$$P\left\{\lim_{n\to\infty} D_n = 0\right\} = 1.$$

Теорема 1.3 (Колмогорова)

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

$K(t)$ – распределение Колмогорова.

Используя теорему Колмогорова, можно построить доверительный интервал для теоретической функции распределения.

$$P\left\{D_n < \frac{t}{\sqrt{n}}\right\} = K(t) = \gamma,$$

$$\forall |x| < \infty : P\left\{F_n(x) - \frac{t_\gamma}{\sqrt{n}} \le F(x) \le F_n(x) + \frac{t_\gamma}{\sqrt{n}}\right\} \approx \gamma \in [0,1],$$

$$n \to \infty,\ n > 20,$$

где $K(t_\gamma) = \gamma$ (γ – квантиль распределения Колмогорова), т. е.

$$t_\gamma = K^{-1}(\gamma).$$

1.4. Непараметрическое оценивание плотности распределения

1.4.1. Гистограмма

Разобьем область определения на k интервалов.

$$t_0 < t_1 < t_2 < \ldots < t_k,$$

n_i – количество наблюдений на $\left[t_i, t_{i+1}\right)$, $\sum_{i=1}^{k} n_i = n$.

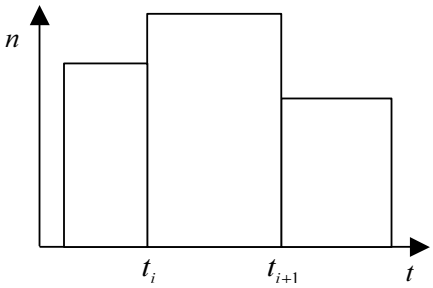


Рис 1.2. Гистограмма

Высота определяется из условия нормировки:

$$\sum_i \frac{n_i}{n\left(t_i - t_{i-1}\right)}\left(t_i - t_{i-1}\right) = \frac{\sum_i n_i}{n} = \frac{n}{n} = 1.$$

Гистограмма – довольно грубый способ оценивания плотности распределения, связанный с неопределенностью выбора числа интервалов k, границ интервалов, потерей информации при группировании.

1.4.2. Ядерные оценки плотности и эмпирической функции распределения

Пусть $g(t)$ – неотрицательная функция, удовлетворяющая условиям:

$$g(t) = g(-t),$$

$$\int_{-\infty}^{+\infty} g(t)dt = 1,$$

$$\int_{-\infty}^{+\infty} t^2 g(t)dt = 1,$$

$$\int_{-\infty}^{+\infty} t^m g(t)dt < \infty; 0 \le m < \infty,$$

$$\lim_{n\to\infty} \lambda_n = 0 \text{ и } \lim_{n\to\infty} n\lambda_n = \infty,$$

тогда функцию плотности можно оценить следующим образом:

$$\hat{f}_n(x) = \frac{1}{n\lambda_n} \sum_{i=1}^{n} g\left(\frac{x - X_i}{\lambda_n}\right),$$

при $n \to \infty$; $\hat{f}_n(x) \to f(x)$,

а функцию распределения как

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} G\left(\frac{x - X_i}{\lambda_n}\right),$$

где λ_n – параметр размытости ядерной функции.

$$G(x) = \int_{\infty}^{x} g(x)dt .$$

Основное преимущество «ядерных» оценок состоит в том, что они непрерывны, в отличие от эмпирической функции распределения и гистограмм.

Тема 2. Точечные оценки и их свойства

2.1. Понятие статистической оценки

Пусть имеется выборка $\mathbb{X}_n = (X_1, ..., X_n)$ из распределения случайной величины $\xi \in F = \{F(x;\theta), \theta \in \Theta\}$. В общем случае задача оценивания заключается в том, чтобы, используя статистическую информацию, доставляемую выборкой $\mathbb{X}_n$, сделать статистические выводы об истинном значении неизвестного параметра θ.

Определение 2.1. *Точечной* оценкой неизвестного параметра θ по выборке $\mathbb{X}_n$ называется значение некоторой статистики $T_n = T(\mathbb{X}_n)$, которое приближенно равно значению параметра θ: $\hat{\theta} = T_n(x)$.

Так как любая статистика является случайной величиной (имеющей некоторое распределение $G_{T_n}(x)$), то для каждой новой реализации выборки $\mathbb{X}_n$ будет получаться другое значение оценки, в общем случае отличное от истинного значения параметра θ.

Определение 2.2. *Интервальной* оценкой параметра θ называют интервал $[T_1(X_n), T_2(X_n)]$, содержащий истинное значение параметра θ с вероятностью γ.

Понятно, что для оценивания θ можно использовать различные оценки, и для того, чтобы выбрать лучшую из них, нужно иметь критерий сравнения качества оценок.

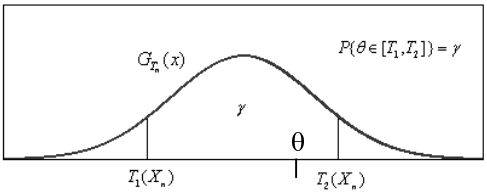


Рис. 2.1. Определение интервальной оценки

В свою очередь, и критерии могут быть разными (в зависимости от целей, для которых строятся оценки) и будут определяться выбором меры близости к истинному значению оцениваемого параметра. Таким образом, если определен *класс* рассматриваемых оценок и выбрана *мера близости*, то, по определению, оценка, минимизирующая меру близости, является оптимальной в этом классе.

2.2. Критерии сравнения оценок

2.2.1. Несмещенность

Определение 2.3. Статистика $T(\mathbb{X}_n)$ называется *несмещенной* оценкой параметра θ, если выполняется условие $M[T(X_n)] = \theta, \forall\theta \in \Theta$.

Для смещенных оценок можно найти величину

$$b(\theta) = M\ [T(X_n)] - \theta,$$

называемую смещением оценки $T(X_n)$.

Определение 2.4. Величину

$$M[T(\mathbb{X}_n) - \theta]^2 = D[T(\mathbb{X}_n)] + b^2(\theta)$$

называют *среднеквадратической ошибкой* оценки *T*.

Для несмещенных оценок среднеквадратическая ошибка совпадет с дисперсией оценки (так как $b(\theta) = 0$).

Отметим, что несмещенные оценки могут не существовать, либо не принадлежать Θ.

Пример 2.1

Пусть наблюдается случайная величина $\xi \in \overline{Bi}(1,\theta)$, и $\mathbb{X}_1 = \{X_1\}$ – выборка из одного наблюдения. Требуется оценить параметр θ.

$$M[T(X_n)] = M[T(X_1)] = \sum_{x=0}^{\infty} T(x)\theta^x(1-\theta) = \theta, \theta \in (0,1)$$

$$\left[\xi \in \overline{Bi}(r,\theta); P\{\xi = x\} = C_{r+x-1}^x \theta^x (1-\theta)^r, x = 0,1,...\right]$$

$$\Rightarrow \sum_{x=0}^{\infty} T(x)\theta^k = \frac{\theta}{1-\theta} = \sum_{r=1}^{\infty} \theta^r, \forall\theta \in (0,1)$$

$$\Rightarrow T(X_1)=\begin{cases}0, X_1=0,\\ 1, X_1\geq 1,\end{cases}$$

но значения 0 и 1 вообще не принадлежат параметрическому множеству и поэтому эта оценка практически бесполезна.

Иногда оценка с малым смещением и малой среднеквадратической ошибкой предпочтительней несмещенной оценки с большой дисперсией.

Пример 2.2

Пусть $\mathbb{X}_n=(X_1,...,X_n)$ – выборка из $\xi\in N(\mu,\theta)$. Найдем оценку дисперсии σ^2. В качестве оценки σ^2 возьмем статистику $S_0^2=\frac{n}{n-1}S_n^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X_n})^2$.

Эта оценка является несмещенной, так как

$$M[S_0^2]=\frac{n}{n-1}M[S_n^2]=\frac{1}{n-1}\frac{n-1}{n}D\xi=\sigma^2.$$

Дисперсия этой оценки равна

$$D[S_0^2]=\frac{2\sigma^4}{n-1}.$$

2.2.2. Несмещенные оценки с равномерно минимальной дисперсией

Определение 2.5. *Несмещенной оценкой с равномерно минимальной дисперсией* (НОРМД) называется такая оценка $T^*(X_n)$, что

$$D[T^*(X_n)]\leq D[T(X_n)]:\forall\theta\in\Theta,\forall T(X_n):M[T(X_n)]=\theta$$

$$\forall x\in X,\left|f(x)-f_0(x)\right|<\varepsilon,\forall\theta\in\Theta.$$

Требование равномерной минимальности дисперсии сильное и не всегда имеет место. Может оказаться, что для двух несмещенных оценок T_1 и T_2 для одних значений θ минимальна дисперсия одной оценки, а для других – дисперсия другой оценки.

Теорема 2.1 (О единственности НОРМД)

Если существует $T^*(\mathbb{X}_n)$, то она единственная.

Доказательство

Предположим, что существуют две НОРМД $T_1(\mathbb{X}_n)$ и $T_2(\mathbb{X}_n)$ функции θ: $M[T_1(\mathbb{X}_n)] = M[T_2(\mathbb{X}_n)] = \theta$ и $D[T_1(\mathbb{X}_n)] = D[T_2(\mathbb{X}_n)] = v = v(\theta) = \inf\limits_{T_n} D[T_n]$.

Рассмотрим оценку

$$T_3 = \frac{T_1 + T_2}{2},\ M[T_3(\mathbb{X}_n)] = \frac{M[T_1] + M[T_2]}{2} = \frac{2\theta}{2} = \theta,$$

$$D[T_3(\mathbb{X}_n)] = D\left[\frac{T_1 + T_2}{2}\right] = \frac{1}{4}(D[T_1] + D[T_2] + 2\operatorname{cov}(T_1, T_2)) =$$

$$= \frac{v + \operatorname{cov}(T_1, T_2)}{2}.$$

Воспользуемся неравенством Коши – Буняковского $\forall \eta_1, \eta_2$: $|\operatorname{cov}(\eta_1, \eta_2)| \le \sqrt{D\eta_1 D\eta_2}$, причем равенство достигается тогда и только тогда, когда $\eta_1 = k\eta_2 + a$.

Но так как v – минимальное значение дисперсии оценки, то T_3 – НОРМД и $\Rightarrow \operatorname{cov}(T_1, T_2) = \sqrt{vv} = v \Rightarrow T_1$ и T_2 линейно связаны, т. е. $T_1 = kT_2 + a$.

Из несмещенности оценки следует

$$MT_1 = kMT_2 + a \Rightarrow \theta = k\theta + a \Rightarrow$$
$$a = \theta(1-k) \Rightarrow T_1 = kT_2 + \theta(1-k) \Rightarrow T_1 - \theta = k(T_2 - \theta).$$

Теперь найдем k:

$$v = \operatorname{cov}(T_1, T_2) = M\left[(T_1 - \theta)(T_2 - \theta)\right] = kM\left[(T_2 - \theta)^2\right] =$$
$$= kDT_2 = kv$$
$$\Rightarrow k \equiv 1 \Rightarrow T_1 \equiv T_2. \blacktriangleleft$$

2.2.3. Состоятельность оценок. Критерий состоятельности

Определение 2.6. Оценка $T_n(\mathbb{X}_n)$ некоторой функции $\tau(\theta)$ называется состоятельной, если при

$$n \to \infty,\ T_n \xrightarrow{P} \tau(\theta), \forall \theta \in \Theta .$$

То есть $\forall \varepsilon > 0 : P\{|\, T_n(\mathbb{X}_n) - \tau(\theta) \,| > \varepsilon\} \to 0, n \to \infty$.

Свойство состоятельности обязательно для любого правила оценивания, однако оно является асимптотическим и не связано со свойствами оценки при фиксированном объеме выборки (в отличие от свойств несмещенности и минимальной дисперсии).

Теорема 2.2. (Критерий состоятельности)

Пусть $M_\theta T_n = \tau(\theta) + \varepsilon_n, D_\theta T_n = \delta_n$ и $\varepsilon_n = \varepsilon_n(\theta) \to 0,\ \delta_n = \delta_n(\theta) \to 0$ при $n \to \infty$. Тогда T_n – состоятельная оценка функции $\tau = \tau(\theta)$.

Доказательство

По определению оценка состоятельна, если

$$\forall \varepsilon > 0 : P\{|\, T_n(\mathbb{X}_n) - \tau(\theta) \,| > \varepsilon\} \to 0, n \to \infty .$$

Имеем

$$|\, T_n(\mathbb{X}_n) - \tau(\theta) \,| = |\, (T_n - MT_n) + (MT_n - \tau(\theta)) \,| \le |\, T_n - MT_n \,| + |\, MT_n - \tau \,| .$$
$$|\, T_n - MT_n \,| \ge |\, T_n - \tau \,| - |\, MT_n - \tau \,| .$$

Пусть $|\, T_n - \tau \,| > \varepsilon$, тогда $|\, T_n - MT_n \,| \ge \varepsilon - |\, \varepsilon_n \,|$. На основании неравенства Чебышева $P\{|\, T_n - \tau \,| \ge \varepsilon\} \le P\{|\, T_n - MT_n \,| \ge \varepsilon - |\, \varepsilon_n \,|\} \le \dfrac{\delta_n}{(\varepsilon - |\, \varepsilon_n \,|)^2} \to 0$ при $n \to \infty, \forall \theta \in \Theta$.

2.3. Функция правдоподобия. Информационное количество Фишера

Пусть $f(x,\theta)$ – плотность случайной величины ξ в непрерывном случае или вероятность в дискретном случае, и $\mathbb{X}_n = (X_1,...,X_n)$ – выборка.

Так как все наблюдения в выборке независимы, то совместная плотность распределения вектора X_n

$$L(\mathbb{X}_n,\theta) = \prod_{i=1}^{n} f(X_i,\theta), \text{ при этом } \int L(X_1,...,X_n,\theta)dx_1...dx_n = 1.$$

Определение 2.7. Функция $L(\mathbb{X}_n,\theta)$, рассматриваемая при фиксированном $\mathbb{X}_n$ как функция θ, называется функцией *правдоподобия*.

Если $L(X_n,\theta) > 0, \forall X_n \in X, \theta \in \Theta$, то

$$\ln L(\mathbb{X}_n,\theta) = \sum_{i=1}^{n} \ln f(X_i,\theta).$$

Определение 2.8. Функция $U(\mathbb{X}_n,\theta) = \dfrac{\partial \ln L(\mathbb{X}_n,\theta)}{\partial\theta}$ называется *вкладом выборки*.

Найдем, чему равно $MU(\mathbb{X}_n,\theta)$.

$$\int L(x_1,...,x_n;\theta)\,dx_1...dx_n = 1.$$

$$0 = \frac{\partial}{\partial\theta}\int_X L(\mathbb{X}_n,\theta)d\mathbb{X}_n = \int_X \frac{\partial \ln L(\mathbb{X}_n,\theta)}{\partial\theta} L(\mathbb{X}_n,\theta)d\mathbb{X}_n =$$

$$= M\left[\frac{\partial \ln L(\mathbb{X}_n,\theta)}{\partial\theta}\right] = M[U(\mathbb{X}_n,\theta)].$$

Значит, $M[U(\mathbb{X}_n,\theta)] = 0$.

Определение 2.9. *Информацией Фишера* о параметре θ, содержащейся в выборке $\mathbb{X}_n$, называется дисперсия вклада выборки

$$I_n(\theta) = D[U(\mathbb{X}_n, \theta)] = D\left[\frac{\partial \ln L(\mathbb{X}_n, \theta)}{\partial \theta}\right] = D\left[\sum_{i=1}^{n} \frac{\partial \ln f(X_i, \theta)}{\partial \theta}\right] = ni(\theta),$$

$$i(\theta) = \int_X \left(\frac{\partial \ln f(x, \theta)}{\partial \theta}\right)^2 f(x, \theta) dx,$$

где $i(\theta)$ – количество информации Фишера, содержащейся в одном наблюдении.

Если продифференцировать выражение $M[U(\mathbb{X}_n, \theta)] = 0$ еще раз, то получим

$$0 = \int \frac{\partial^2 \ln f(x, \theta)}{\partial \theta^2} dx + \int \left(\frac{\partial \ln f(x, \theta)}{\partial \theta}\right)^2 f(x, \theta) dx$$

$$\Rightarrow i(\theta) = -M\left[\frac{\partial^2 \ln f(x, \theta)}{\partial \theta^2}\right].$$

Пример 2.3

Найти $i(\theta)$ для модели $N(\theta, \sigma^2)$.

$$f(x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}; \quad \ln f(x, \theta) = -\frac{(x-\theta)^2}{2\sigma^2} - \ln\left(\sqrt{2\pi\sigma^2}\right);$$

$$\frac{\partial f(x, \theta)}{\partial \theta} = \frac{x-\theta}{\sigma^2}; \quad \frac{\partial^2 f(x, \theta)}{\partial \theta^2} = -\frac{1}{\sigma^2};$$

$$i(\theta) = -M\left[\frac{\partial^2 f(x, \theta)}{\partial \theta^2}\right] = \frac{1}{\sigma^2}.$$

Пример 2.4 (Нерегулярная модель)

Рассмотрим равномерное распределение

$$f(x,\theta)=\begin{cases}\dfrac{1}{\theta}, x\in[0,\theta],\\ 0, x\notin[0,\theta].\end{cases}$$

Здесь из $\int_0^{\theta}\frac{1}{\theta}dx=1$ не следует $\int_0^{\theta}\frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\right)dx=0,$ так как при дифференцировании интеграла по верхнему пределу появляется еще одно слагаемое. В данном случае модель является нерегулярной, так как выборочное пространство зависит от неизвестного параметра θ.

2.4. Неравенство Рао – Крамера и эффективные оценки

Определение 2.10. Модель называется регулярной, если выполняются два условия:

1) область определения не зависит от θ;

2) $f(x,\theta)$ дифференцируема по θ.

Теорема 2.3 (Неравенство Рао – Крамера)

Для любой несмещенной оценки $T(X_n)$ параметрической функции $\tau(\theta)$ справедливо неравенство

$$D[T(X_n)]\geq\frac{[\tau'(\theta)]^2}{ni(\theta)}.$$

Равенство имеет место тогда и только тогда, когда $T(x)=\tau(\theta)+a(\theta)U(X_n,\theta),$ где $a(\theta)$ – некоторая функция от θ.

Доказательство

Так как модель $F(x,\theta)$ регулярная, то, дифференцируя это тождество по $\theta,$ получаем (используя равенство Коши – Буняковского)

$$\tau'(\theta)=\int T(X_n)\frac{\partial\ln L(X_n,\theta)}{\partial\theta}L(X_n,\theta)=M[T(X_n)U(X_n,\theta)]=$$

$$= \operatorname{cov}(T(X_n)U(X_n,\theta)) \le \sqrt{D[T(X_n)]D[U(X_n,\theta)]} = \sqrt{DT \cdot I_n(\theta)} \Rightarrow$$

$$[\tau'(\theta)]^2 \le I_n D[T(X_n)],$$

причем неравенство обращается в равенство тогда и только тогда, когда $T(x)$ и $U(x)$ линейно связаны. ◂

Неравенство Рао – Крамера определяет нижнюю границу дисперсий всех несмещенных оценок, заданных параметрической функцией $\tau(\theta)$ для регулярных моделей.

Определение 2.11. Оценка, при которой достигается нижняя граница неравенства Рао – Крамера, называется *эффективной*.

Так как эффективная оценка является НОРМД, то по теореме 2.1 она является единственной.

Критерий эффективности оценки

Пусть имеется условие линейной связи между $T(X_n)$ и $U(X_n,\theta) = \dfrac{\partial \ln(X_n,\theta)}{\partial\theta}$, т. е. $T(X_n) = a(\theta)U(X_n,\theta) + b(\theta)$.

Найдем

$$b(\theta): M(T(X_n)) = a(\theta)\underbrace{M(U(X_n,\theta))}_{=0} + \underbrace{M(b(\theta))}_{=\tau(\theta)} = \tau(\theta) \Rightarrow$$

$T(X_n) - \tau(\theta) = a(\theta)U(X_n,\theta)$ – *критерий эффективности*, где $a(\theta)$ – некоторая функция от θ.

Определение 2.12. Модель $F = \{F(x,\theta), \theta \in \Theta\}$ – экспоненциальная, если $f(x,\theta) = \exp\{A(\theta)B(x) + C(\theta) + D(x)\}$.

В частности, экспоненциальными являются: $N(\theta,\sigma^2)$, $N(\mu,\theta^2)$, $\Gamma(\theta,\lambda)$, $Bi(X,\theta)$, $\overline{Bi}(r,\theta)$, $\Pi(\theta)$.

Вклад выборки для экспоненциальной модели равен

$$U(\mathbb{X}_n,\theta) = \frac{\partial}{\partial\theta}\sum_{i=1}^{n}\exp(A(\theta)B(X_i) + C(\theta) + D(X_i)) = A'(\theta)\sum_{i=1}^{n}B(X_i) +$$

$$+nC'(\theta) = nA'(\theta)\left(\frac{1}{n}\sum_{i=1}^{n} B(X_i) + \frac{C'(\theta)}{A'(\theta)}\right).$$

Обозначим:

$$T(\mathbb{X}_n) = \frac{1}{n}\sum_{i=1}^{n} B(X_i),$$

$$\tau(\theta) = -\frac{C'(\theta)}{A'(\theta)},$$

$$a(\theta) = \frac{1}{nA'(\theta)}.$$

Таким образом, для регулярной экспоненциальной модели существует эффективная оценка $T(\mathbb{X}_n)$ параметрической функции $\tau(\theta)$.

Верно и обратное утверждение: если эффективная оценка существует, то модель является экспоненциальной.

2.5. Критерий оптимальности в векторном случае

Рассмотрим критерий оптимальности в случае, когда $\bar{\theta} = (\theta_1, \theta_2, ..., \theta_r)$ – векторный параметр и оценка ищется для функции $\tau(\theta)$, дифференцируемой по всем переменным. Модель $F = \{F(x, \bar{\theta}), \bar{\theta} \in \Theta\}$ также считается регулярной.

Определение 2.13. Матрица $I_n(\bar{\theta}) = \| g_{ij}(\bar{\theta}) \|_1^r$, где

$$g_{ij}(\bar{\theta}) = M\left[\frac{\partial \ln L(\mathbb{X}_n, \bar{\theta})}{\partial \theta_i}\frac{\partial \ln L(\mathbb{X}_n, \bar{\theta})}{\partial \theta_j}\right] = nM\left[\frac{\partial \ln f(x, \bar{\theta})}{\partial \theta_i}\frac{\partial \ln f(x, \bar{\theta})}{\partial \theta_j}\right],$$

называется *информационной матрицей Фишера.*

Если функция правдоподобия дважды дифференцируема, то из тождества $\int L(\mathbb{X}_n, \bar{\theta}) d\mathbb{X}_n = 1$ следует

$$g_{ij} = -M\left[\frac{\partial^2 \ln L(\mathbb{X}_n, \bar{\theta})}{\partial\theta_i \partial\theta_j}\right] = -nM\left[\frac{\partial^2 \ln f(X_i, \bar{\theta})}{\partial\theta_i \partial\theta_j}\right].$$

Теорема 2.4 (Неравенство Рао – Крамера в векторном случае)

Если $T = T(\mathbb{X}_n)$ – несмещенная оценка функции $\tau(\bar{\theta})$, то при всех $\bar{\theta} \in \Theta$ $D[T(\mathbb{X}_n)] \geq \sum_{i,j=1}^{r} g_{ij} c_i(\bar{\theta}) c_j(\bar{\theta})$ и коэффициенты $c_i(\theta)$ определяются из уравнений $\sum_{j=1}^{r} g_{ij} c_i(\theta) = \frac{\partial\tau(\bar{\theta})}{\partial\theta_i}$, $i = 1,...,r$.

Если матрица $I_n(\theta)$ не вырождена и $I_n^{-1}(\theta) = \| \widetilde{g_{ij}}(\theta) \|_1^r$, то $D[T(\mathbb{X}_n)] \geq \sum_{i,j=1}^{r} \widetilde{g_{ij}} \frac{\partial\tau(\theta)}{\partial\theta_i} \frac{\partial\tau(\theta)}{\partial\theta_j}$.

Равенство достигается тогда и только тогда, когда $T(\mathbb{X}_n) - \tau(\theta) = \sum_{i=1}^{r} c_i(\theta) \frac{\partial \ln(X_i, \bar{\theta})}{\partial\theta_i}$. Это равенство является критерием эффективности оценки в векторном случае.

Пример 2.5

Пусть $\mathbb{X}_n = (X_1,...,X_n)$ – выборка из $N(\theta_1, \theta_2^2)$. Требуется оценить $\tau(\bar{\theta}) = \tau(\theta_1, \theta_2) = \theta_1$.

Возьмем в качестве оценки $\tau(\theta)$ выборочное среднее $\overline{X}$, тогда

$$L(\mathbb{X}_n, \theta) = \frac{1}{\left(\sqrt{2\pi}\theta_2\right)^n} e^{-\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(X_i - \theta_1)^2}.$$

$$\frac{\partial \ln\left(\mathbb{X}_n, \bar{\theta}\right)}{\partial \theta_1} = \frac{1}{\theta_2{}^2}\sum_{i=1}^{n}(X_i - \theta_1) = \frac{1}{\theta_2{}^2}\left[\sum_{i=1}^{n} X_i - n\theta_1\right] = \frac{1}{\theta_2{}^2} n\left(\overline{X_n} - \theta_1\right).$$

$$\Rightarrow \frac{\theta_2^2}{n}\frac{\partial \ln\left(\mathbb{X}_n, \bar{\theta}\right)}{\partial \theta_1} = \overline{X_n} - \theta_1.$$

и таким образом выполняется критерий эффективности оценки $\overline{X_n}$ параметрической функции $\tau(\theta_1, \theta_2) = \theta_1$.

2.6. Достаточные статистики

Определение 2.14. Статистика $T = T(\mathbb{X}_n)$ называется достаточной для модели $F = \{F(x,\theta), \theta \in \Theta\}$, если условная плотность (или условная вероятность в векторном случае) $L(\mathbb{X}_n \mid t;\theta) = P\{X_1 = x_1, ..., X_n = x_n \mid T(\mathbb{X}_n) = t\}$ случайного вектора $\mathbb{X}_n$ при условии $T(\mathbb{X}_n) = t$ не зависит от параметра θ.

Свойство достаточности статистики $T(\mathbb{X}_n)$ означает, что выборка содержит всю информацию о параметре θ, имеющуюся в выборке, и поэтому все заключения, которые можно сделать при наблюдении $\mathbb{X}_n$, зависят только от $T(\mathbb{X}_n) = t$. Следовательно, достаточная статистика дает оптимальный в определенном смысле способ представления статистических данных, что особенно важно при обработке большого объема статистической информации.

Теорема 2.5. (Критерий факторизации)

Для того, чтобы статистика $T(\mathbb{X}_n)$ была достаточной для θ, необходимо и достаточно, чтобы функция правдоподобия имела вид $L(\mathbb{X}_n, \theta) = g(T(\mathbb{X}_n);\theta)h(\mathbb{X}_n)$, где функция $g(t,\theta)$ зависит от выборки только через $T(\mathbb{X}_n) = t$, а функция $h(\mathbb{X}_n)$ не зависит от θ.

Доказательство (для дискретной модели)

Пусть $T(\mathbb{X}_n)$ – достаточная статистика. Тогда $L(\mathbb{X}_n \mid T(\mathbb{X}_n) = t;\theta) = h(\mathbb{X}_n)$, где h – непрерывная функция, не зависящая от параметра θ.

В дискретном случае. Пусть

$$L(\mathbb{X}_n,\theta)=P\{X_1=x_1,X_2=x_2,...,X_n=x_n \mid X_i\in F\{F(x,\theta)\mid\theta\in\Theta\}\}=$$

$$P\{X_1=x_1,X_2=x_2,...,X_n=x_n,T(\mathbb{X}_n)=t\}=\left|\begin{matrix}P(A\mid B)=\dfrac{P(AB)}{P(B)}\\ P(AB)=P(A\mid B)P(B)\end{matrix}\right|=$$

$$=P\{X_1=x_1,X_2=x_2,...,X_n=x_n \mid T(\mathbb{X}_n)=t\}P\{T(\mathbb{X}_n)=t\}=$$

$$=h(\mathbb{X}_n,t)g(T(\mathbb{X}_n);\theta).$$

Обратно, пусть имеется $L(\mathbb{X}_n,\theta)=g(T(\mathbb{X}_n);\theta)h(\mathbb{X}_n)$, тогда

$$\forall\mathbb{X}_n: T(\mathbb{X}_n)=t$$

$$L(\mathbb{X}_n\mid t;\theta)=P\{X_1=x_1,...,X_n=x_n \mid T(\mathbb{X})=t\}=$$

$$\frac{P\{\mathbb{X}=X_n,T(\mathbb{X})=t\}}{P\{T(\mathbb{X})=t\}}=\frac{L(\mathbb{X}_n;\theta)}{\sum\limits_{X'_nT(X'_n)=t}L(X'_n;\theta)}=\frac{g(t;\theta)h(\mathbb{X}_n)}{\sum\limits_{X'_nT(X'_n)=t}g(t;\theta)h(\mathbb{X}_n)}=$$

$$=\frac{h(\mathbb{X}_n)}{\sum\limits_{X'_nT(X'_n)=t}h(\mathbb{X}_n)},$$

т. е. не зависит от θ. Если $T(\mathbb{X}_n)\neq t$, то очевидно, что $L(\mathbb{X}_n\mid t,\theta)=0$.

Отметим, что всякая эффективная оценка является одновременно достаточной статистикой. Однако обратное не всегда верно – достаточная статистика может существовать, но не быть эффективной.

Если статистика $T(\mathbb{X}_n)$ – достаточная, то достаточной является и любая взаимно однозначная функция $\varphi(T(\mathbb{X}_n))$.

Пример 2.6

Найдем достаточную статистику для $N\left(\theta_1,\theta_2^2\right)$.

$$L\left(\mathbb{X}_n,\bar{\theta}\right)=\frac{1}{\left(\sqrt{2\pi}\theta_2\right)^n}e^{-\frac{1}{2\theta_2^2}\sum\limits_{i=1}^{n}(X_i-\theta_1)^2}=\frac{1}{\left(\sqrt{2\pi}\theta_2\right)^n}e^{-\frac{1}{2\theta_2^2}\sum\limits_{i=1}^{n}\left(X_i-\overline{X_n}+\overline{X_n}-\theta_1\right)^2}.$$

$$\sum_{i=1}^{n}\left(X_i-\theta_1\right)^2=\sum_{i=1}^{n}\left(\left(X_i-\overline{X_n}\right)+\left(\overline{X_n}-\theta_1\right)\right)^2=\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)^2+$$

$$+2\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)\left(\overline{X_n}-\theta_1\right)+\sum_{i=1}^{n}\left(\theta_1-\overline{X_n}\right)^2=$$

$$=\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)^2+2\left(\theta_1-\overline{X_n}\right)\underbrace{\sum_{i=1}^{n}X_i-n\overline{X_n}}_{=0}+n\left(\theta_1-\overline{X_n}\right)^2=$$

$$=\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)^2+n\left(\overline{X_n}-\theta_1\right)^2=\frac{1}{\left(\sqrt{2\pi}\theta_2\right)^n}e^{-\frac{1}{2\theta_2^2}\underbrace{\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)^2}_{T2}-\frac{n\left(\overline{X_n}-\theta_1\right)^2}{2\theta_2^2}}.$$

$$g(T_1,T_2,\theta_1,\theta_2)=\frac{1}{\left(\sqrt{2\pi}\theta_2\right)^n}e^{-\frac{1}{2\theta_2^2}T_2-\frac{n(T_1-\theta_1)^2}{2\theta_2^2}}.$$

$$h(t)=1.$$

Отсюда вытекает по критерию факторизации: $T=(T_1,T_2)$; $T_1=\overline{X_n}$; $T_2=\sum_{i=1}^{n}\left(X_i-\bar{X}_n\right)^2$.

Пример 2.7

Найдем достаточную статистику для равномерного распределения Rav(θ):

$$L(\mathbb{X}_n,\theta)=\begin{cases}\dfrac{1}{\theta^n},X_{(n)}=\max\limits_{1\le i\le n}X_i\le\theta\\ 0,\text{ иначе}\end{cases}=\frac{h(\theta-X_{(n)})}{\theta^n},$$

где $h(x)=\begin{cases}1,x\ge 0,\\ 0,x<0.\end{cases}$

По критерию факторизации имеем, что $T(\mathbb{X}_n) = X_{(n)}$.

Пример 2.8 (Коши)

$$L(\mathbb{X}_n, \theta) = \frac{1}{\pi^n} \prod \frac{1}{1 + (X_i - \theta)^2}.$$

Здесь нельзя найти статистику *T*, обеспечивающую факторизацию.

Теорема 2.6 (Рао – Блекуэлла – Колмогорова)

НОРМД, если она существует, является функцией от достаточной статистики.

Определение 2.15. Достаточная статистика называется полной, если $\forall \varphi(T(X_n)) : (M\varphi(T) = 0, \forall \theta) \Rightarrow \varphi(t) \equiv 0, \ \forall T$.

Пример 2.9

Пусть $\mathbb{X}_n = (X_1, ..., X_n)$ – выборка из $Bi(1, \theta)$.

Статистика $T = \sum_{i=1}^{n} X_i$ является эффективной оценкой для $n\theta$ и, следовательно, является достаточной. Покажем, что она является полной. Распределение статистики T имеет вид $g(t, \theta) = C_n^t \theta^t (1-\theta)^{n-t}$; $t = 0, 1, ..., n$ (по свойству воспроизводимости по параметрам биномиального распределения $X_i \sim Bi(1, \theta); \sum_{i=1}^{n} X_i \sim Bi(n, \theta)$).

Пусть $\varphi(t)$ – произвольная функция, заданная на множестве $\{0, 1, ..., n\}$. Тогда условие $M\varphi(T) = 0; \forall \theta \in [0,1]$ записывается в виде $\sum_{t=0}^{n} \varphi(t) \, C_n^t \theta^t (1-\theta)^{n-t} = 0; \forall \theta \in (0,1)$. Сделаем замену:

$$x = \frac{\theta}{1-\theta} \Rightarrow \theta^t (1-\theta)^{n-t} = \frac{\theta^t}{(1-\theta)^t} (1-\theta)^n = x^t (1-\theta)^n.$$

$$\sum_{t=0}^{n} \varphi(t) C_n^t x^t = 0; \forall x \in [0, \infty).$$

Отсюда следует, что $\varphi(t) \equiv 0, t = 0,1,\ldots$, следовательно, T – полная достаточная статистика.

Теорема 2.7

Если существует полная достаточная статистика, то всякая функция от нее является НОРМД своего математического ожидания.

Доказательство

Пусть $T = T(\mathbb{X}_n)$ – полная достаточная статистика и $H_1(T)$, $H_2(T)$ – две произвольные функции, такие что

$$M\left[H_1(T)\right] = \tau(\theta)$$
$$M\left[H_2(T)\right] = \tau(\theta).$$

Тогда

$$M\left[H_1(T) - H_2(T)\right] = \tau - \tau = 0 \Rightarrow H_1(T) - H_2(T) \equiv 0 \Rightarrow$$
$$H_1(T) \equiv H_2(T) \equiv H(T).$$

По теореме Рао – Блекуэлла – Колмогорова НОРМД $\tau(\theta)$ надо искать в классе функций, зависящих от T, но $H(T)$ – единственная функция, несмещенно оценивающая $\tau(\theta)$, следовательно, $H(T)$ является НОРМД $\tau(\theta)$. ◂

Определение 2.16. Уравнение $M\left[H(T)\right] = \tau(\theta)$, где T – полная достаточная статистика, а $H(T)$ – неизвестная функция, называют уравнением несмещенности.

Для непрерывных моделей уравнение несмещенности имеет вид $\int H(T) g(t;\theta)\, dt = \tau(\theta)$, где $g(t;\theta)$ – плотность распределения достаточной статистики T.

Пример 2.10

Пусть $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ – выборка из $Bi(1,\theta)$. Требуется найти НОРМД функции $\tau(\theta) = \theta^k$.

$T(\mathbb{X}_n) = \sum_{i=1}^{n} X_i$ является полной достаточной статистикой с распределением $g(t;\theta) = C_n^t \theta^t (1-\theta)^{n-t}$. Следовательно, уравнение несмещенности имеет вид $M\left[H(T)\right] = \theta^k$.

Производящая функция случайной величины $T(X_n)$ равна

$$\varphi(z,\theta) = Mz^T = Mz^{x_1}, \ldots, z^{x_n} = Mz^{x_1}, \ldots, Mz^{x_n} = \left(Mz^{\xi}\right)^n.$$

$$Mz^{\xi} = \sum_{x=0}^{1} z^x \theta^x (1-\theta)^{1-x} = z^0\theta^0(1-\theta)^1 + z^1\theta^1(1-\theta)^0 =$$

$$= 1 - \theta + z\theta = 1 + (z-1)\theta.$$

Обозначим $(a)_k = a(a-1)\ldots(a-k+1), k \geq 1,$ по свойству производящей функции $M(T_k) = \left.\frac{\partial^k \varphi(z,\theta)}{\partial z^k}\right|_{z=1} = (n)_k \theta^k \Rightarrow M\frac{(T)_k}{(n)_k} = \theta^k$, и из уравнения несмещенности следует, что статистика $\frac{(T)_k}{(n)_k}$ является НОРМД $\tau(\theta) = \theta^k$ при $k \leq n$.

Тема 3. Построение оценок параметров по полным выборкам

3.1. Метод максимального правдоподобия

Определение 3.1. Оценкой максимального правдоподобия (ОМП) параметра θ называется точка параметрического множества Θ, в которой функция максимального правдоподобия $L(\mathbb{X}_n, \theta)$ достигает максимума:

$$L(\mathbb{X}_n, \theta) = \sup_{\theta \in \Theta} L(\mathbb{X}_n, \theta).$$

Определение 3.2. Если для любой выборки $\mathbb{X}_n$ из выборочного пространства $\mathbb{X}$ максимум $L(\mathbb{X}_n, \theta)$ достигается во внутренней точке θ и $L(\mathbb{X}_n, \theta)$ дифференцируема по θ, то ОМП $\hat{\theta}$ удовлетворяет уравнению $\dfrac{\partial \ln L(\mathbb{X}_n, \theta)}{\partial \theta_i} = 0,\ i = 1, ..., r$, которое называется *уравнением правдоподобия.*

Свойства ОМП

1. Если существует эффективная оценка $T(\mathbb{X}_n)$, то $\hat{\theta} = T(\mathbb{X}_n)$, так как по критерию эффективности Рао – Крамера

$$\frac{\partial \ln L(\mathbb{X}_n, \theta)}{\partial \theta} = \frac{1}{a(\theta)}\left[T(\mathbb{X}_n) - \theta\right] = 0 \Rightarrow T(\mathbb{X}_n) = 0.$$

2. Если существует достаточная статистика $T = T(\mathbb{X}_n)$ и ОМП существует и единственная, то она является функцией от T. Так как по теореме факторизации $L(\mathbb{X}_n, \theta) = g(T(\mathbb{X}_n), \theta) h(\mathbb{X}_n)$ и максимизация функции правдоподобия эквивалентна максимизации $g(T(\mathbb{X}_n); \theta)$ по θ, следовательно, $\hat{\theta}$ зависит от $T(\mathbb{X}_n)$.

3. ОМП является инвариантной относительно преобразования параметров, т. е. $\hat{\tau}(\theta) = \tau(\hat{\theta})$, если $\tau(\theta)$ – взаимно однозначное преобразование.

4. ОМП является асимптотически несмещенной, т. е. $M\hat{\theta} \to \theta$, $n \to \infty$.

5. Если модель F является регулярной, а функция правдоподобия $\forall n \geq 1$ и $\mathbb{X}_n \in \mathbb{X}$ имеет один локальный максимум, то $\hat{\theta}$ является состоятельной оценкой параметра θ, т. е. $\hat{\theta} \xrightarrow{p} \theta, \forall \theta \in \Theta$.

6. Если F регулярная, а $L(\mathbb{X}_n, \theta)$ имеет один максимум, $f(x,\theta)$ трижды дифференцируема и $\exists M(x) : \forall \theta \in \Theta \left| \frac{\partial^3 \ln f(x,\theta)}{\partial\theta_i \partial\theta_j \partial\theta_k} \right| < H(x)$, то при $n \to \infty$ $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{p} \xi \in N(0, I^{-1}(\theta))$, где $I(\theta)$ – информационная матрица Фишера.

7. Если $\tau(\theta)$ – непрерывно дифференцируемая функция от θ и $\hat{\tau}_n = \tau(\hat{\theta})$ – ОМП $\tau(\theta)$, то

$$\sqrt{n}\left(\hat{\tau}_n - \tau(\theta)\right) \xrightarrow{p} N\left(0, \sigma_\tau^2\right)$$

$$\sigma_\tau^2 = b^T(\theta) I^{-1}(\theta) b(\theta), b(\theta) =$$

$$= \left(\frac{\partial\tau(\theta)}{\partial\theta_1}, \ldots \frac{\partial\tau(\theta)}{\partial\theta_r} \right).$$

Свойства 6–7 называются свойствами *асимптотической нормальности.*

Рассмотрим доказательство для скалярного случая.

$\left[\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{p} N\left(0, \frac{1}{i(\theta)}\right) \right]$ разложим в ряд Тейлора $U_n(\theta) = \frac{\partial \ln L(\mathbb{X}_n, \theta)}{\partial\theta}$ в точке θ.

Имеем

$$\theta = U_n\left(\hat{\theta}\right) = U_n(\theta) + \left(\hat{\theta}_n - \theta\right)U'_n(\theta) + \frac{1}{2}\left(\hat{\theta}_n - \theta\right)^2 U''_n(\theta^*),$$

где $\theta^* \in \left[\theta, \hat{\theta}\right]$, отсюда

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \frac{U_n(\theta)}{\sqrt{n}i(\theta)}\left[\underbrace{-\frac{U'_n(\theta)}{ni(\theta)}}_{1} + \underbrace{\varepsilon_n}_{0}\right],$$

где $|\varepsilon_n| = \left|\hat{\theta}_n - \theta\right|\left|U''_n(\theta^*)\right| \Big/ 2ni(\theta) \le \dfrac{\left|\hat{\theta}_n - \theta\right|}{2i(\theta)} \dfrac{1}{n}\sum_{j=1}^{n} M\left(X_j\right)$. Так как $\hat{\theta} \xrightarrow{p} \theta$, то $\varepsilon \xrightarrow{p} 0$. Применим к величине $\dfrac{1}{n}U'_n(\theta) = \dfrac{1}{n}\sum_{j=1}^{n}\dfrac{\partial^2 \ln f(X_j,\theta)}{\partial\theta^2}$ *ЗБЧ* (*закон больших чисел*), согласно которому

$$-\frac{1}{ni(\theta)}U'_n(\theta) \xrightarrow{p} -\frac{1}{i(\theta)}M\left[\frac{\partial^2 \ln f(X_j,\theta)}{\partial\theta^2}\right] = 1.$$

К случайной величине

$$\frac{1}{\sqrt{n}i(\theta)}U_n(\theta) = \frac{1}{\sqrt{n}i(\theta)}\sum_{j=1}^{n}\frac{\partial \ln f(X_j,\theta)}{\partial\theta}$$

применим *ЦПТ* (*центральную предельную теорему*), из которой следует

$$-\frac{1}{\sqrt{n}i(\theta)}\varphi_n(\theta) \xrightarrow{p} \xi \in N\left(0, \frac{1}{i(\theta)}\right).$$

Определение 3.3. *Асимптотической дисперсией* статистики T_n, удовлетворяющей при $n \to \infty$ условию $\sqrt{n}\left(T_n - T(\theta)\right) \xrightarrow{p} \xi \in N\left(0, \sigma^2(\theta)\right)$, называется величина $\dfrac{\sigma^2(\theta)}{n}$.

Определение 3.4. Если оценка T_n параметра θ является асимптотически нормальной $N\left(0,1/ni(\theta)\right)$, то эта оценка называется *асимптотически эффективной.*

Пример 3.1

Дано $N\left(\theta_1, \theta_2^2\right)$.

$$L\left(X_n, \theta\right) = \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{1}{2\theta_2^2}\sum\left(x_i - \bar{x}\right)^2 - \frac{n\left(\bar{x}-\theta_1\right)}{2\theta_2^2}}.$$

Пусть $S^2 = \dfrac{1}{n}\sum\left(x_i - \bar{x}\right)^2$, тогда $L\left(X_n, \theta\right) = \dfrac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{1}{2\theta_2^2}nS^2 - \frac{n(\bar{x}-\theta_1)}{2\theta_2^2}}$.

Равномерное распределение

$$L\left(X_n, \theta\right) = \frac{1}{\theta^n},\ x_{(1)} \geq 0; \quad F(x) = \begin{cases} \dfrac{1}{\theta}, x \in [0, \theta] \\ 0, x \notin [0, \theta] \end{cases} \quad L\left(X_n, \theta\right) = \begin{cases} \dfrac{1}{\theta^n}, x_i \leq \theta, \\ 0, x > \theta. \end{cases}$$

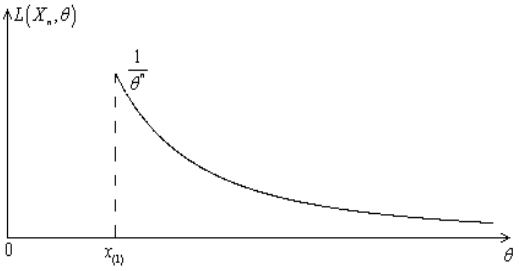


Рис. 3.1. Функция правдоподобия для равномерного распределения

Экспоненциальный закон распределения

$$L\left(X_n,\theta\right)=\theta^n e^{-\theta\sum x_i},$$

$$\ln L\left(X_n,\theta\right)=n\ln\theta-\theta\sum x_i,$$

$$\frac{\partial \ln L(X_n,\theta)}{\partial\theta}=\frac{n}{\theta}-\sum x_i=0,$$

$$\frac{n}{\theta}=\sum x_i,\theta=\frac{n}{\sum x_i}.$$

3.2. Метод моментов

Определение 3.5

$$\mathbb{X}_n=(X_1,...,X_n)\in F(x,\theta),$$
$$\alpha_k=MX^k,\ k=1,2,...$$

Приравнивая теоретические и выборочные моменты, можно найти оценки неизвестных параметров: $\alpha_k(\theta)=\frac{1}{n}\sum_{i=1}^{n}X_i^k,\ k=1,2,...,n.$

Такой метод называется *метод моментов.*

Оценки, полученные по методу моментов, при некоторых условиях являются состоятельными. В общем случае являются эффективными.

Пример 3.2

Имеется выборка $\mathbb{X}_n=\left\{X_1,\ldots,X_n\right\}\in\Gamma\left(\theta_1,\theta_2\right).$

$$f\left(x,\theta_1,\theta_2\right)=\begin{cases}\dfrac{x^{\theta_2-1}e^{-x/\theta_1}}{\Gamma\left(\theta_2\right)\theta_1\theta_2},x\geq 0,\\ 0,x<0,\end{cases}$$

$$\alpha_k = MX^k = \int_0^\infty x^k f\left(x,\theta_1,\theta_2\right) dx = \frac{1}{\Gamma\left(\theta_2\right)\theta_1\theta_2} \int_0^\infty x^k x^{\theta_2-1} e^{-x/\theta_1} dx =$$

$$= \frac{\theta_1^x \Gamma\left(\theta_2 + k\right)}{\Gamma\left(\theta_2\right)},$$

$$\alpha_1 = \frac{\theta_1 \Gamma\left(\theta_2 + 1\right)}{\Gamma\left(\theta_2\right)} = \frac{\theta_1\theta_2 \Gamma\left(\theta_2\right)}{\Gamma\left(\theta_2\right)} = \theta_1\theta_2 = \frac{1}{n}\sum_{i=1}^n X_i \,,$$

$$\alpha_2 = \frac{\theta_1^2 \Gamma\left(\theta_2 + 2\right)}{\Gamma\left(\theta_2\right)} = \frac{\theta_1^2\left(\theta_2+1\right)\Gamma\left(\theta_2+1\right)\Gamma\left(\theta_2+1\right)}{\Gamma\left(\theta_2\right)} = \frac{\theta_1^2\left(\theta_2+1\right)\theta_2\Gamma\left(\theta_2\right)}{\Gamma\left(\theta_2\right)} =$$

$$= \theta_1^2\left(\theta_2+1\right)\theta_2 = \frac{1}{n}\sum_{i=1}^n X_i^2 \,.$$

Получаем $\widetilde{\theta_1} = \dfrac{\dfrac{1}{n}\sum\limits_{i=1}^n \left(X_i - \overline{X}\right)^2}{\overline{X}}$, $\widetilde{\theta_2} = \dfrac{\overline{X}^2}{\dfrac{1}{n}\sum\limits_{i=1}^n \left(X_i - \overline{X}\right)^2}$.

Тема 4. Доверительное оценивание

4.1. Интервальное оценивание

Рассмотрим методы оценивания, позволяющие находить точечные оценки параметров. Для каждой конкретной выборки значения оценки отличаются от истинного значения параметра, поэтому полезно знать возможную погрешность, возникающую при использовании оценки. Например, можно указать такой интервал, внутри которого с вероятностью γ находится истинное значение параметра θ. В этом случае говорят об интервальном оценивании. Иногда интервальное оценивание называют доверительным.

4.2. Понятие доверительного интервала

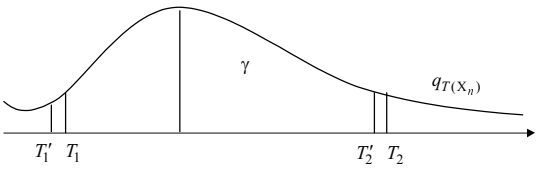
Определение 4.1. γ-доверительным интервалом для параметра θ называют интервал $(T_1(\mathbb{X}_n), T_2(\mathbb{X}_n))$, такой, что

$$P\left\{T_1\left(\mathbb{X}_n\right) < \theta < T_2\left(\mathbb{X}_n\right)\right\} = \gamma, \forall\theta \in \Theta.$$

Статистику $T_1\left(\mathbb{X}_n\right)$ называют нижней границей доверительного интервала, а $T_2\left(\mathbb{X}_n\right)$ – верхней.

В случае многомерного параметра доверительные границы определяют для каждой компоненты векторного параметра. В общем случае доверительных интервалов может быть несколько, поэтому из них выбирают интервал минимальной длины.

Пример 4.1



Оба интервала, $[T_1, T_2]$ и $[T_1', T_2']$, дают одинаковую вероятность, но нам необходимо выбрать интервал с наименьшей длиной. Если распределение является симметричным, то минимальную длину будет давать такой интервал, что для него выполнялось бы:

$a - T_1 = T_2 - a$, где a – середина интервала.

4.3. Построение доверительного интервала с использованием центральных статистик

Определение 4.2. $G(\mathbb{X}_n, \theta)$ – центральная статистика, если распределение $G(\mathbb{X}_n, \theta)$ не зависит от θ. При любом фиксированном θ статистика $G(\mathbb{X}_n, \theta)$ непрерывна и строго монотонна по θ.

С помощью центральной статистики можно построить доверительный интервал.

Пусть $f_G(g)$ – плотность распределения статистики $G(\mathbb{X}_n, \theta)$, так как функция плотности не зависит от θ, то для любого $\gamma \in (0,1)$ можно найти такие $g_1, g_2, g_1 < g_2$, что $\{g_1 < G(\mathbb{X}_n, \theta) < g_2\} = \int_{g_1}^{g_2} f_G(g)\,dg = \gamma$.

Составим два уравнения

$$\begin{cases} G(\mathbb{X}_n, \theta) = g_1, \\ G(\mathbb{X}_n, \theta) = g_2. \end{cases}$$

Из решений выберем минимальное и максимальное:

$$T_1(\mathbb{X}_n) = \min\{\tilde{T}_1(\mathbb{X}_n), \tilde{T}_2(\mathbb{X}_n)\},$$

$$T_2(\mathbb{X}_n) = \min\{\tilde{T}_1(\mathbb{X}_n), \tilde{T}_2(\mathbb{X}_n)\},$$

тогда неравенство $g_1 < G(\mathbb{X}_n) < g_2, T_1(\mathbb{X}_n) < \theta < T_2(\mathbb{X}_n)$.

В силу строгой монотонности функции G следует, что

$P\{g_1 < G(\mathbb{X}_n, \theta) < g_2\} = P\{T_1(\mathbb{X}_n) < \theta < T_2(\mathbb{X}_n)\} = \gamma$ является γ-доверительным интервалом.

Пример 4.2

Дана $\mathbb{X}_n = \{X_1, \ldots, X_n\} \in N(\theta, \sigma^2)$. Построить доверительный интервал для θ. Воспользуемся ЦПТ:

$$MX_i = \theta;\ DX_i = \sigma^2;\ \frac{\sqrt{n}\left(\bar{X} - \theta\right)}{\sigma} \sim N(0,1).$$

Таким образом, можно взять $G\left(X_n, \theta\right) = \dfrac{\sqrt{n}\left(\bar{X} - \theta\right)}{\sigma}$, так как данная статистика имеет распределение, не зависящее от θ, и G монотонно убывает по θ.

Составим два неравенства.

Пусть $g_1 < g_2, P\left(g_1 < G\left(\mathbb{X}_n, \theta\right) < g_2\right) = \Phi\left(g_2\right) - \Phi\left(g_1\right) = \gamma,$

$$\frac{\sqrt{n}(\bar{X} - \tilde{T}_1)}{\sigma} = g_1;\ \frac{\sqrt{n}(\bar{X} - \tilde{T}_2)}{\sigma} = g_2.$$

Отсюда $\tilde{T}_1 = \bar{X} - \dfrac{\sigma}{\sqrt{n}} g_1;\ \ \tilde{T}_2 = \bar{X} - \dfrac{\sigma}{\sqrt{n}} g_2.$

В силу $g_1 < g_2 \Rightarrow \tilde{T}_1 < \tilde{T}_2$, тогда $\left(\bar{X} - \dfrac{\sigma}{\sqrt{n}} g_2;\ \bar{X} - \dfrac{\sigma}{\sqrt{n}} g_1 \right)$ – γ-доверительный интервал для θ.

Пусть $g_1 = -g_2$, тогда

$$\Phi\left(g_2\right) - \Phi\left(g_1\right) = \Phi\left(g_2\right) - \Phi\left(-g_2\right) =$$

$$= \Phi\left(g_2\right) - \left(1 - \Phi\left(g_2\right)\right) = 2\Phi\left(g_2\right) - 1 = \gamma;$$

$$\Phi\left(g_2\right) = \frac{\gamma + 1}{2};\ \ g_2 = \Phi^{-1}\left(\frac{\gamma + 1}{2}\right).$$

Для построения доверительного интервала с помощью центральной статистики основная проблема заключается в нахождении этой центральной статистики. Можно выделить класс моделей, для которых центральная статистика существует и имеет простой вид.

Пусть $F(x,\theta)$ – функция распределения наблюдаемой случайной величины и монотонна по параметру θ, т. е. $F(x,\theta_1) \geq F(x,\theta_2)$; $\forall \theta_1 < \theta_2$ или $F(x,\theta_1) \leq F(x,\theta_2)$; $\forall \theta_1 > \theta_2$.

Можно положить в качестве центральной статистики функцию

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

Действительно:

1. Непрерывность и монотонность $G(\mathbb{X}_n,\theta)$ следует из непрерывности и монотонности $F(x,\theta)$.

2. Найдем распределение случайной величины $G(\mathbb{X}_n,\theta)$. Случайная величина $\eta = F(X_i,\theta)$ подчинена $\mathrm{Rav}(0,1)$:

$$P\{\eta < x\} = P\{F(X_i,\theta) < x\} = P\left\{X_i < F^{-1}(x,\theta)\right\} =$$
$$= F\left(F^{-1}(x,\theta)\right) = x, \quad x \in (0,1).$$

Случайная величина $-\ln\eta$ подчинена $\mathrm{Exp}(1) = \mathrm{Gamma}(1,1)$:

$$P\{-\ln\eta < x\} = P\{\ln\eta > -x\} = P\left\{\eta > e^{-x}\right\} =$$
$$= 1 - P\left\{\eta > e^{-x}\right\} = 1 - e^{-x}.$$

Из воспроизводимости гамма-распределения по параметру следует, что случайная величина $-\sum_{i=1}^{n} \ln F(X_i,\theta)$ подчинена $\mathrm{Gamma}(1,n)$, т. е.

$$F_G(g)P\left\{-\sum_{i=1}^{n} \ln F(X_i,\theta) < g\right\} =$$
$$= \frac{1}{\Gamma(n)} g^{n-1} e^{-g},\ g > 0\ .$$

Таким образом, распределение случайной величины $G(\mathbb{X}_n,\theta)$ не зависит от θ и, следовательно, $G(\mathbb{X}_n,\theta)$ – центральная статистика.

В результате получаем следующий метод построения γ-доверительного интервала:

1) выбираем $g_1 < g_2$ так, чтобы $\dfrac{1}{\Gamma(n)}\int\limits_{g_1}^{g_2} g^{n-1}e^{-g}dg = \gamma$;

2) решая уравнения

$$-\sum_{i=1}^{n}\ln F(X_i,\theta) = g_1,$$

$$-\sum_{i=1}^{n}\ln F(X_i,\theta) = g_2,$$

находим корни $T_1(\mathbb{X}_n) < T_2(\mathbb{X}_n)$ (в общем случае численными методами). Тогда $(T_1(\mathbb{X}_n) < T_2(\mathbb{X}_n))$ – искомый γ-доверительный интервал.

Пример 4.3

Для выборки $\mathbb{X}_n = \{X_1,\ldots,X_n\}$ из $\mathrm{Rav}(0,\theta)$ построить γ-доверительный интервал для параметра θ. В качестве статистики возьмем функцию

$$G(\mathbb{X}_n,\theta) = -\sum_{i=1}^{n}\ln F(X_i,\theta) = g_1,$$

$$-\sum_{i=1}^{n}\ln\frac{X_i}{\theta} = -\sum_{i=1}^{n}\ln X_i + n\ln\theta.$$

Решаем уравнение $g_1 < g_2$:

$$G(\mathbb{X}_n,\theta) = g_1,$$
$$G(\mathbb{X}_n,\theta) = g_2,$$

$$-\sum_{i=1}^{n}\ln X_i + n\ln\tilde{T}_1 = g_1,$$

$$n\ln\tilde{T}_1 = g_1 - \sum_{i=1}^{n}\ln X_i,$$

$$\tilde{T}_1 = e^{\frac{g_1-\sum_{i=1}^{n}\ln X_i}{n}} = e^{\frac{g_1}{n}\left(\prod_{i=1}^{n}X_i\right)^{\frac{1}{n}}},$$

$$\tilde{T}_2 = e^{\frac{g_2}{n}\left(\prod_{i=1}^{n}X_i\right)^{\frac{1}{n}}}.$$

При $g_1 < g_2$ $\tilde{T}_1(\mathbb{X}_n) < \tilde{T}_2(\mathbb{X}_n)$, следовательно, γ-доверительный интервал имеет вид $\left(e^{\frac{g_1}{n}\left(\prod_{i=1}^{n}X_i\right)^{1/n}}, e^{\frac{g_2}{n}\left(\prod_{i=1}^{n}X_i\right)^{1/n}}\right)$ $\frac{1}{\Gamma(n)}\int_{g_1}^{g_2} g^{n-1}e^{-g}dg = \gamma.$

4.4. Построение доверительного интервала с использованием распределения точечной оценки параметров

Если имеется некоторая точечная оценка $T = T(\mathbb{X}_n)$ для параметра θ и известна ее функция распределения $F_T(t,\theta)$, то доверительный интервал можно построить, основываясь на этой функции.

Пусть функция распределения $F_T(t,\theta)$ непрерывна и монотонна по θ и задан доверительный уровень γ.

Найдем $t_1 = t_1(\theta)$ и $t_2 = t_2(\theta)$ такие, что

$$P\{t_1 < T(\mathbb{X}_n) < t_2\} = F_T(t_2,\theta) - F_T(t_1,\theta) = \gamma.$$

Чтобы процедура была однозначной, будем выбирать t_1 и t_2 из условия $F_T(t_1,\theta)=\frac{1-\gamma}{2}$; $1-F_T(t_2,\theta)=\frac{1-\gamma}{2}$, а соответствующий доверительный интервал будем называть *центральным*.

Обозначим через $D\gamma=\{(\theta,\theta'):t_1(\theta)<\theta'<t_2(\theta)\}$, тогда $P\{(\theta,T(X_n))\in D\gamma\}=P\{t_1<T(X_n)<t_2\}=\gamma;\ \forall\theta\in\Theta$.

Рассмотрим множество $D\gamma(\theta')=\{\theta:(\theta,\theta')\in D\gamma\}$. Событие $\theta\in D\gamma(T(\mathbb{X}_n))$ произойдет тогда и только тогда, когда $T(X_n)\in(t_1(\theta),t_2(\theta))$ и, следовательно, имеет вероятность γ. Так как $t_i(\theta)$ являются монотонными одного типа в силу монотонности функции распределения, то $D\gamma(\theta')$ является интервалом $(\theta_1(\theta'),\theta_2(\theta'))$, а γ-доверительный интервал есть $(T_1(\mathbb{X}_n),T_2(\mathbb{X}_n))$; $T_i(\mathbb{X}_n)=\theta_i(T(\mathbb{X}_n))$.

Алгоритм построения γ-доверительных интервалов

1. Находим точечную оценку t – наблюдаемое значение оценки.
2. Решаем относительно θ уравнения

$$F_T(t,\tilde{\theta}_1)=\frac{1-\gamma}{2},\ F_T(t,\tilde{\theta}_2)=\frac{1+\gamma}{2}.$$

3. Принимаем $\theta_1=\min(\tilde{\theta}_1,\tilde{\theta}_2),\ \theta_2=\max(\tilde{\theta}_1,\tilde{\theta}_2)$.

В случае дискретных моделей рассуждения аналогичны. Единственное отличие заключается в том, что в силу дискретности функции распределения можно обеспечить выполнение неравенства

$$P\{t_1<T(X_n)<t_2\}=F(t_2-0;\theta)-F(t_1,\theta)\ge\gamma.$$

При выборе различных точечных оценок $T(X_n)$ будут получаться различные доверительные интервалы. Доверительные интервалы с минимальной длиной будут получаться, если брать точечные оценки с минимальной дисперсией, т. е. эффективные или асимптотически эффективные оценки.

Оценки максимального правдоподобия при достаточно общих условиях являются асимптотически эффективными и асимптотически нормальными, следовательно,

$$P\left\{\left|\tilde{\theta}_n - \theta\right|\sqrt{ni\left(\tilde{\theta}_n\right)} \le c_\gamma\right\} \Rightarrow \Phi\left(c_\gamma\right) - \Phi\left(-c_\gamma\right) = 2\Phi\left(c_\gamma\right) - 1 = \gamma.$$

Если $c_\gamma = \Phi^{-1}\left(\frac{\gamma+1}{2}\right)$, следовательно, $\left(\tilde{\theta}_n - \frac{c_\gamma}{\sqrt{ni(\theta)}};\ \tilde{\theta}_n + \frac{c_\gamma}{\sqrt{ni(\theta)}}\right)$ является асимптотическим кратчайшим γ-доверительным интервалом для θ.

Тема 5. Проверка статистических гипотез

Определение 5.1. *Статистической гипотезой* называется любое утверждение о виде или свойствах распределения наблюдаемых в эксперименте случайных величин.

Проверка статистической гипотезы состоит в том, чтобы сформулировать такое правило, которое позволило бы по результатам проведенных наблюдений принять или отклонить гипотезу.

Определение 5.2. Правило, согласно которому гипотеза принимается или отвергается, называется *критерием* проверки статистической гипотезы.

С проверкой статистических гипотез связывают ошибки двух типов.

Определение 5.3. *Ошибкой первого рода* называют событие, когда верная проверяемая гипотеза отвергается критерием.

Определение 5.4. *Ошибкой второго рода* называют событие, когда неверная проверяемая гипотеза принимается критерием.

Вероятности ошибок первого и второго рода обозначают α и β соответственно. Вероятность ошибки второго рода зависит от выдвигаемой конкурирующей гипотезы.

Определение 5.5. Вероятность отклонения ложной проверяемой гипотезы, т. е. принятия правильного решения в пользу конкурирующей, называется *мощностью*, и она равна $1-\beta$.

Определение 5.6. Вероятность ошибки первого рода также называют *уровнем значимости* критерия.

Гипотезу, которую мы проверяем, будем называть *основной* или *нулевой* гипотезой и будем всегда обозначать H_0. Альтернативные или *конкурирующие* гипотезы будем обозначать $H_1, H_2, \ldots, H_m$.

5.1. Виды статистических гипотез

Обычно статистические гипотезы делят на следующие виды: *однородности*, если имеется две или более выборок случайных величин; *независимости*, если имеется выборка многомерной случайной величины; *случайности*, если есть предположения о независимости и одинаковом распределении наблюдений в выборке; *о виде распределения*, если есть предположения о законе распределения случайной величины (рис. 5.1).

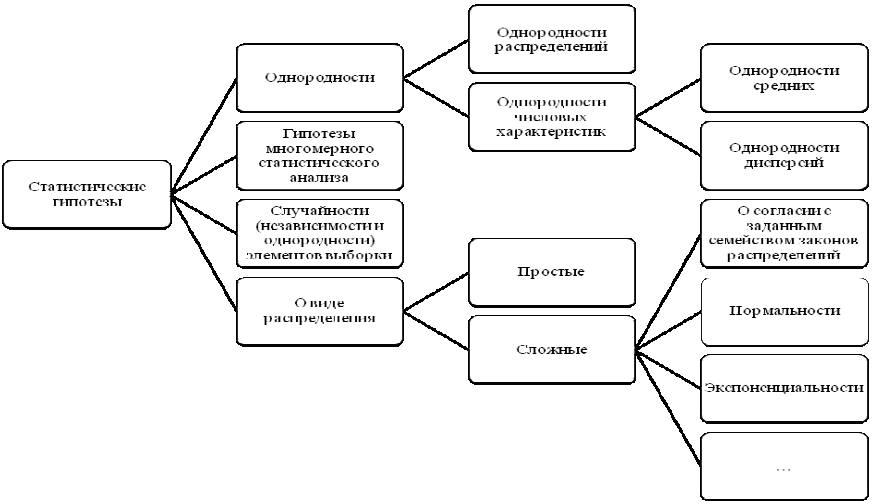


Рис. 5.1. Классификация статистических гипотез

5.1.1. Гипотеза о виде распределения

Пусть имеется выборка $\mathbb{X}_n = \{X_1, ..., X_n\}$ наблюдаемой случайной величины с функцией распределения $F_\xi(x)$.

а) *Простой* гипотезой является утверждение $H_0 : F_\xi(x) = F(x)$, где $F(x)$ полностью задана.

б) *Сложной* гипотезой является утверждение $H_0 : F_\xi(x) \in F = \{F(x, \theta), \theta \in \Theta\}$.

5.1.2. Гипотеза однородности

Пусть произведено k серий независимых наблюдений $\mathbb{X}_{n_1} = \left\{X_1^1,...,X_n^1\right\}$, $\mathbb{X}_{n_2} = \left\{X_1^2,...,X_n^2\right\} ... \mathbb{X}_{n_k} = \left\{X_1^n,...,X_n^n\right\}$ и пусть $F_i(x)$ – функция распределения i-й серии. Чтобы проверить, менялось ли распределение от серии к серии, можно сформулировать гипотезу *однородности* $H_0 : F_1(x) = F_2(x) = ... = F_k(x),$ при этом само распределение $F(x)$ может быть неизвестным.

5.1.3. Гипотеза независимости

В эксперименте наблюдается двумерная случайная величина $\xi = (\xi_1, \xi_2)$ с неизвестной функцией распределения $F_\xi(x, y),$ и есть основания предполагать, что ξ_1 и ξ_2 независимы. В этом случае нужно проверить гипотезу *независимости* $H_0 : F_\xi(x, y) = F_{\xi_1}(x)F_{\xi_2}(y),$ где $F_{\xi_1}(x)$ и $F_{\xi_2}(y)$ – некоторые одномерные функции распределения.

5.1.4. Гипотеза случайности

Результат эксперимента описывается n-мерной случайной величиной $\bar{\xi} = (\xi_1, \xi_2, ..., \xi_n)$ с некоторой функцией распределения $F_{\bar{\xi}}(x_1, ..., x_n).$ Чтобы проверить утверждение о том, что все наблюдения независимы и одинаково распределены, можно сформулировать гипотезу *случайности* $H_0 : F_{\bar{\xi}}(x_1, ..., x_n) = F_\xi(x_1)F_\xi(x_2)...F_\xi(x_n).$

5.2. Выбор критерия проверки статистической гипотезы

Для проверки одной и той же гипотезы, как правило, существует несколько различных статистических критериев $T_1, T_2, ..., T_k.$ Выбор подходящего статистического критерия, вообще говоря, не является тривиальной задачей. Можно сформулировать следующие принципы подбора статистических критериев.

1. Должны выполняться «стандартные» предположения, обусловливающие возможность применения рассматриваемого критерия

(например, о виде распределения случайной величины и о наблюдаемых данных). Так, например, нельзя применять критерий Колмогорова по группированным данным или по наблюдениям дискретной случайной величины.

2. Критерий должен быть *состоятельным*, т. е. его мощность должна стремиться к единице с ростом объема выборки.

3. Критерий должен быть *несмещенным*, т. е. мощность должна быть больше, чем вероятность ошибки первого рода.

4. Критерий должен обладать наибольшей мощностью при заданном объеме выборки и заданном уровне значимости критерия.

Добиться выполнения последнего принципа на практике не представляется возможным, потому что построить наиболее мощный критерий удается только в очень редких случаях, например, когда основная и конкурирующая гипотезы являются *простыми*. Чаще всего для разных конкурирующих гипотез, для разных уровней значимости, для разных объемов выборки более мощными оказываются разные критерии.

В этой ситуации для выбора оптимального критерия можно применить классическую теорию *принятия решений в условиях неопределенности*. Критерии являются стратегиями, конкурирующие гипотезы – состояниями среды, функция полезности $u\left(T_i, H_j\right)$ – это мощность критерия (табл. 5.1). Другим способом определения функции полезности при выборе критерия может быть стоимость проведения эксперимента по различению основной и конкурирующей гипотез с заданными вероятностями ошибок первого и второго рода.

Таблица 5.1

Матрица полезности выбора критерия T_i при конкурирующих гипотезах H_j

$T_i \setminus H_j$	H_1	H_2	...	H_m
T_1	$u(T_1, H_1)$	$u(T_1, H_2)$	...	$u(T_1, H_m)$
T_2	$u(T_2, H_1)$	$u(T_2, H_2)$	...	$u(T_2, H_m)$
...	...	...	...	...
T_k	$u(T_k, H_1)$	$u(T_k, H_2)$	...	$u(T_k, H_m)$

Существуют разные подходы к выбору оптимальной стратегии при принятии решения в условиях неопределенности. В случае, когда нет никакой информации о том, какая конкурирующая гипотеза может быть верна, рациональным выглядит выбор критерия по правилу Вальда (известны также такие названия, как «критерий крайнего пессимиста» или «критерий осторожного наблюдателя»):

$$T^* = \arg\max_{T_i} \min_{H_j} u\left(T_i, H_j\right).$$

Критерий, выбранный по правилу Вальда, максимизирует полезность против самой «неудобной» конкурирующей гипотезы.

Любой критерий проверки статистической гипотезы разбивает выборочное пространство на доверительную область $\boldsymbol{X}_0$ и критическую область $\boldsymbol{X}_1$. Чаще всего такое разбиение производится с помощью одномерной *статистики* – функции от выборки, поэтому критическая и доверительная область формулируются уже как подмножества множества вещественных чисел.

Доверительная область включает такие значения статистики критерия, при которых гипотеза принимается, а критическая область – значения, при которых гипотеза отвергается. Кроме того, вероятность попадания выборки (статистики критерия) в критическую область, когда гипотеза верна, по определению равна вероятности ошибки первого рода, а вероятность попадания выборки (статистики критерия) в доверительную область, когда гипотеза не верна, равна вероятности ошибки второго рода.

Как правило, встречаются три вида критических областей для статистики критерия:

- правосторонняя критическая область $\left(t_\alpha, \infty\right)$;
- левосторонняя $\left(-\infty, t_\alpha\right)$;
- двусторонняя $\left(-\infty, t_{\alpha_1}\right) \cup \left(t_{\alpha_2}, \infty\right),\ \alpha_1 + \alpha_2 = \alpha.$

5.3. Вычисление достигаемого уровня значимости

Достигаемый уровень значимости (p-value) определяется как вероятность попадания статистики критерия:

- в область $\left(S(X_n), \infty\right)$, если критическая область правосторонняя;

- в область $(-\infty, S(X_n))$, если критическая область левосторонняя;

где $S(X_n)$ – вычисленное значение статистики по реализации выборки.

Гипотеза отвергается, если достигаемый уровень значимости оказывается меньше заданной вероятности ошибки первого рода. Достоинство процедуры проверки гипотезы с использованием *p*-value в том, что не нужно заранее фиксировать уровень значимости и определять критическую область для значений статистики критерия. Кроме того, *p*-value характеризует «степень уверенности» в принимаемом решении, т. е. чем меньше *p*-value, тем больше оснований для отвержения основной гипотезы.

Если критическая область двусторонняя, то однозначного способа вычисления достигаемого уровня значимости нет. Например, можно вычислять *p*-value как $2\min(p, 1-p)$, где $p = P\{S \in (S(X_n), \infty) | H_0\}$.

Достигаемый уровень значимости является случайной величиной, определенной на интервале [0,1], на основании которой делается статистический вывод о принятии гипотезы. Чем ближе значение *p*-value к единице, тем больше оснований для принятия гипотезы, чем ближе значение *p*-value к нулю, тем больше оснований для отвержения гипотезы. Однако следует помнить о следующем важном замечании относительно *p*-value.

Когда основная гипотеза ложна, то *p*-value будет стремиться к нулю с ростом объема наблюдаемой выборки. Однако, когда основная гипотеза истинна, *p*-value не стремится к единице, а распределено равномерно на интервале [0,1].

Тема 6. Проверка гипотезы о виде распределения

Для проверки гипотезы о виде распределения используются *критерии согласия*.

6.1. Критерий Колмогорова

В критериях типа Колмогорова измеряемое расстояние между эмпирическим $F_n(x)$ и теоретическим $F(x, \theta)$ распределениями имеет вид

$$D_n = \sup_{|x|<\infty} |F_n(x) - F(x,\theta)|, \tag{6.1}$$

где n – объем выборки.

Предпочтительнее в критерии Колмогорова (Колмогорова – Смирнова) использовать статистику с поправкой Большева вида

$$S_k = \frac{6nD_n + 1}{6\sqrt{n}}, \tag{6.2}$$

где

$$D_n = \max(D_n^+, D_n^-),\ D_n^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F(X_{(i)},\theta)\right\},$$

$$D_n^- = \max_{1\le i\le n}\left\{F(X_{(i)},\theta) - \frac{i-1}{n}\right\}, \tag{6.3}$$

и $X_{(1)}, X_{(2)}, ..., X_{(n)}$ – упорядоченные по возрастанию выборочные значения.

Статистика S_k при справедливости простой проверяемой гипотезы в пределе подчиняется закону Колмогорова, а в случае сложной гипотезы – различным законам, в зависимости от вида распределения и оцениваемых параметров. Статистические модели распределений статистик $G(S_k \mid H_0)$ для наиболее распространенных семейств законов распределений приведены в [12].

Если для вычисленного по выборке значения статистики S_k^* выполняется неравенство $P\left\{S > S_k^*\right\} = 1 - G(S_k^* \mid H_0) > \alpha,$ то нет оснований для отклонения гипотезы H_0.

6.2. Критерии типа ω^2

В критериях типа ω^2 расстояние между гипотетическим и истинным распределениями рассматривают в квадратичной метрике. Статистика критерия выражается соотношением

$$\omega_n^2\left[\psi(F)\right] = \int_{-\infty}^{\infty} \left\{E\left[F_n(x)\right] - F(x)\right\}^2 \psi(F(x))dF(x) =$$

$$= \frac{2}{n}\sum_{i=1}^{n}\left\{g\left[F(x_i)\right] - \frac{2i-1}{2n} f\left[F(x_i)\right]\right\} + \int_0^1 (1-t)^2 \psi(t)dt, \qquad (6.4)$$

где

$$f(t) = \int_0^1 \psi(s)ds,\ g(t) = \int_0^1 s\psi(s)ds.$$

При выборе $\psi(t) \equiv 1$ получается статистика критерия Крамера – Мизеса – Смирнова:

$$S_\omega = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left\{F(X_{(i)}, \theta) - \frac{2i-1}{2n}\right\}^2. \qquad (6.5)$$

При выборе $\psi(t) \equiv 1/t(1-t)$ получается статистика критерия Андерсона – Дарлинга:

$$S_\Omega = n\Omega_n^2 =$$

$$= -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(X_{(i)},\theta) + \left(1 - \frac{2i-1}{2n}\right)\ln(1 - F(X_{(i)},\theta))\right\}. \quad (6.6)$$

Статистики S_ω и S_Ω при простой гипотезе в пределе подчиняются законам $a_1(s)$ и $a_2(s)$ соответственно, а в случае сложной гипотезы – различным законам, в зависимости от вида распределения, числа и типа оцениваемых параметров, значений параметров формы, от метода оценивания. Статистические модели распределений статистик $G(S_\omega|H_0)$ и $G(S_\Omega|H_0)$ для наиболее распространенных семейств законов распределений приведены в [12].

Критерии типа ω^2 имеют правостороннюю критическую область.

6.3. Критерии типа χ^2

Процедура проверки гипотез с применением критерия типа χ^2 предусматривает группирование наблюдений. Область определения случайной величины разбивается на k непересекающихся интервалов граничными точками $t_0 < t_1 < ... < t_{k-1} < t_k$, где t_0 – нижняя граница области определения, t_k – верхняя грань. В соответствии с заданным разбиением подсчитывается число n_i выборочных значений, попавших в i-й интервал, и вычисляется вероятность попадания в интервал $P_i(\theta) = F(t_i) - F(t_{i-1}) = \int_{t_{i-1}}^{t_i} f(x,\theta)dx$. При этом $\sum_{i=1}^{k} n_i = n$; $\sum_{i=1}^{k} P_i(\theta) = 1$.

В основе статистик, используемых в критериях согласия типа χ^2, лежит распределение отклонений $\frac{n_i}{n}$ от $P_i(\theta)$.

Критерии типа χ^2 имеют правостороннюю критическую область.

6.3.1. Критерий χ^2 Пирсона

Статистика χ^2 Пирсона имеет вид

$$S_{\chi^2} = n\sum_{i=1}^{k} \frac{\left(\frac{n_i}{n} - P_i(\theta)\right)^2}{P_i(\theta)}.$$

В случае проверки простой гипотезы при $n \to \infty$ статистика S_{χ^2} имеет распределение χ^2_{k-1} с $k-1$ степенями свободы. В случае проверки сложной гипотезы и при условии, что оценки находятся по методу минимума χ^2, статистика S_{χ^2} имеет распределение χ^2_{k-m-1} степенями свободы при $n \to \infty$, где m – число оцениваемых параметров.

6.3.2. Критерий отношения правдоподобия

В критерии отношения правдоподобия используется статистика

$$S_{\text{ОП}} = -2\sum_{i=1}^{k} n_i \ln\left(\frac{P_i(\theta)}{n_i / n}\right),$$

которая подчиняется в случае простой гипотезы χ^2_{k-1} распределению при $n \to \infty$.

В случае проверки сложной гипотезы и при условии, что оценки находятся по методу максимального правдоподобия по группированной выборке, статистика $S_{\text{ОП}}$ при $n \to \infty$ имеет распределение χ^2_{k-m-1} степенями свободы, где m – число оцениваемых параметров, так же как и для критерия χ^2 Пирсона.

Тема 7. Проверка гипотезы однородности распределений

Определение 7.1. *Гипотеза однородности* – это предположение о том, что две (и более) выборки взяты из одной и той же генеральной совокупности. Для данной гипотезы неважно, какое именно распределение имеет генеральная совокупность, поэтому «идеальный» критерий однородности не должен зависеть от вида закона распределения случайной величины.

В большинстве случаев критерии однородности являются двухвыборочной модификацией критериев согласия. Так, критерий однородности Смирнова соответствует критерию согласия Колмогорова, критерий однородности Лемана – Розенблатта – критерию согласия Крамера – Мизеса – Смирнова, критерий однородности Андерсона – Дарлинга – Петита – критерию согласия Андерсона – Дарлинга, критерий согласия Хи-квадрат – критерию однородности Хи-квадрат.

Иногда в качестве критериев однородности используются критерии однородности средних или дисперсий, однако следует понимать, что их применение как критериев однородности возможно только при наложении ограничений на вид распределения случайной величины.

Гипотеза однородности формулируется следующим образом: пусть имеются две независимые случайные выборки $\mathbb{X}_m = (X_1, X_2, ..., X_m)$ и $\mathbb{Y}_n = (Y_1, Y_2, ..., Y_n)$ объемов m и n соответственно. Выборке $\mathbb{X}_m$ соответствует функция распределения $F(x)$, выборке $\mathbb{Y}_n$ – функция распределения $G(x)$. Проверяемая нулевая гипотеза H_0 имеет вид $F(x) = G(x)$ против конкурирующей $H_1 : F(x) \neq G(x)$. Функции $F(x)$ и $G(x)$ будем считать непрерывными.

7.1. Критерий Смирнова

В критерии Смирнова используется статистика вида

$$S_{\mathrm{C}} = \sqrt{\frac{mn}{m+n}} D_{mn}, \tag{7.1}$$

где

$$D_{mn}(\mathbb{X}_m, \mathbb{Y}_n) = \sup_{-\infty<x<\infty} \left| F_m(x) - G_n(x) \right|, \tag{7.2}$$

$F_m(x)$ и $G_n(x)$ – эмпирические функции распределения, построенные по выборкам $\mathbb{X}_m$ и $\mathbb{Y}_n$ соответственно.

На практике значение $D_{mn}(\mathbb{X}_m, \mathbb{Y}_n)$ рекомендуется вычислять в соответствии с выражением (5.3):

$$D_{mn}(\mathbb{X}_m, \mathbb{Y}_n) = \max\left(D_{mn}^{+}, D_{mn}^{-} \right), \tag{7.3}$$

где D_{mn}^{+} и D_{mn}^{-} вычисляются по формуле

$$D_{mn}^{+} = \max_{1<i<m} \left\{ \frac{i}{m} - F_n\left(X_{(i)} \right) \right\}, \quad D_{mn}^{-} = \max_{1<i<m} \left\{ F_n\left(X_{(i)} \right) - \frac{i-1}{m} \right\}, \tag{7.4}$$

где $X_{(i)}$ – i-й элемент упорядоченной по возрастанию выборки $\mathbb{X}_m$.

Критерий Смирнова не зависит от конкретного вида распределений $F(x)$ и $G(x)$, и при стремлении объемов выборок к бесконечности статистика (7.4) сходится к распределению Колмогорова. Однако при малых значениях объемов m и n распределение статистики (7.4) может значительно отклоняться от предельного закона, в связи с чем предложена модификация критерия Смирнова вида

$$S_{\mathrm{C}} = \sqrt{\frac{mn}{m+n}} \left(D_{m,n} + \frac{m+n}{4{,}6mn} \right). \tag{7.5}$$

При использовании критерия Смирнова рекомендуется брать объемы выборок m и n, представляющие собой взаимно простые числа.

7.2. Критерии типа ω^2

В критериях типа ω^2 расстояние между гипотетическим и истинным распределениями рассматривается в квадратичной метрике, в соответствии с выражением

$$\int_{-\infty}^{\infty} \left\{ E\left[F_n(x)\right] - F(x) \right\}^2 \psi\left(F(x)\right) dF(x), \qquad (7.6)$$

где $E\left[\cdot\right]$ – оператор математического ожидания.

Данный критерий применим для проверки согласия полученного опытного распределения с теоретическим. Если же в данном критерии сделать переход от функции распределения вероятностей $F(x)$, с которой проверяется согласие наблюдаемой выборки, к эмпирической функции распределения $F_m(x)$, то данный критерий можно использовать для проверки гипотезы однородности. В этом случае статистика критерия будет опираться на разность эмпирических функций распределений. Одним из таких критериев является критерий Лемана – Розенблатта.

7.2.1. Критерий Лемана – Розенблатта

Поскольку при проверке однородности в (7.6) имеется два равноправных распределения, то можно рассматривать двухвыборочный аналог, в котором используется статистика вида

$$T = \frac{mn}{m+n} \int_{-\infty}^{\infty} [G_m(x) - F_n(x)]^2 dH_{m+n}(x), \qquad (7.7)$$

где $H_{m+n}(x) = \frac{m}{m+n} G_m(x) + \frac{n}{m+n} F_n(x)$ – эмпирическая функция распределения, построенная по вариационному ряду объединения двух выборок.

Данный критерий был предложен Леманом и исследован Розенблаттом. Как правило, критерий используют со статистикой вида

$$T = \frac{1}{mn(m+n)}\left[n\sum_{i=1}^{n}(r_i - i)^2 + m\sum_{j=1}^{m}(s_j - j)^2 \right] - \frac{4mn-1}{6(m+n)}, \qquad (7.8)$$

где r_i – порядковый номер (ранг) наблюдения Y_i; s_j – порядковый номер (ранг) наблюдения Y_j в общем вариационном ряде, построенном по объединенной выборке $X \bigcup Y$ [1].

Розенблаттом было показано, что статистика (7.7) в пределе распределена как $a1(t)$, т. е. выполняется:

$$\lim_{\substack{m\to\infty \\ n\to\infty}} P\{T < t\} = a1(t). \qquad (7.9)$$

7.2.2. Критерий однородности Андерсона – Дарлинга – Петита

В критерии Лемана – Розенблатта весовая функция $\psi(t)$ предполагается равной единице. Если в данном критерии положить весовую функцию $\psi(t)$ равной $\frac{1}{t(1-t)}$, то получится двухвыборочный аналог критерия Андерсона – Дарлинга

$$A^2 = \frac{mn}{m+n}\int_{-\infty}^{\infty} \frac{[G_m(x) - F_n(x)]^2}{(1 - H_{m+n}(x))H_{m+n}(x)} dH_{m+n}(x) \qquad (7.10)$$

или

$$A^2 = \frac{1}{mn}\sum_{i=1}^{m+n-1} \frac{\left(M_i(m+n) - mi\right)^2}{i(m+n-i)}, \qquad (7.11)$$

где M_i – число элементов из первой выборки, меньших или равных i-му элементу вариационного ряда объединенной выборки.

Статистика (7.11) в пределе распределена как $a2(t)$, т. е. выполняется:

$$\lim_{\substack{m\to\infty \\ n\to\infty}} P\{T < t\} = a2(t). \qquad (7.12)$$

7.3. Критерий однородности χ^2

Пусть осуществляется k последовательных серий независимых наблюдений, состоящих из $n_1, n_2, ..., n_k$ наблюдений. Пусть ν_{ij} – число наблюдений i-го исхода в j-й серии. Пусть p_{ij} – неизвестная вероятность появления i-го исхода в j-й серии $(i = 1,...,s; j = 1,...,k)$.

Тогда гипотеза однородности может быть сформулирована следующим образом:

$$H_0 : (p_{1j}, \ldots, p_{sj}) = (p_1, \ldots, p_s),\ j = 1,...,k.$$

Статистика критерия имеет вид

$$\chi_n^2 = n\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{\left(\nu_{ij} - n_j\nu_i / n\right)^2}{n_j\nu_i} = n\left(\sum_{i=1}^{s}\sum_{j=1}^{k}\frac{{\nu_{ij}}^2}{n_j\nu_i} - 1\right).$$

Эта статистика при $n \to \infty$ имеет распределение $\chi^2_{((s-1)(k-1))}$. Критическая область имеет вид $\{t \ge t_\alpha\}$, $t_\alpha = F^{-1}_{\chi^2_{(s-1)(k-1)}}(1-\alpha)$.

Тема 8. Проверка гипотезы однородности средних и дисперсий

Проверяемая гипотеза о равенстве математических ожиданий (об однородности математических ожиданий) случайных величин в общем случае имеет вид

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m \tag{8.1}$$

при конкурирующей

$$H_1 : \mu_{i_1} \neq \mu_{i_2}.$$

Для проверки гипотезы H_0 может использоваться ряд параметрических критериев: сравнения двух выборочных средних при известных дисперсиях; сравнения двух выборочных средних при неизвестных, но равных дисперсиях (критерий Стьюдента); сравнения двух выборочных средних при неизвестных и неравных дисперсиях; F-критерий. В этих же целях применяется целая совокупность непараметрических критериев: критерий Уилкоксона, критерий Манна – Уитни, критерий Краскела – Уаллиса.

Основным предположением, обусловливающим применение параметрических критериев, является принадлежность анализируемых выборок нормальному закону. Непараметрические критерии свободны от этого требования.

Проверяемая гипотеза о постоянстве дисперсии m выборок имеет вид

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2, \tag{8.2}$$

а конкурирующая с ней гипотеза –

$$H_1 : \sigma_{i_1}^2 \neq \sigma_{i_2}^2,$$

где неравенство выполняется, по крайней мере, для одной пары индексов i_1, i_2. Для проверки такого вида гипотез применяется ряд критериев: критерий Бартлетта, критерий Кокрена, критерий Фишера, критерий Левене.

8.1. Критерии проверки гипотез о математических ожиданиях

8.1.1. *t*-критерий Стьюдента

Наибольшей популярностью при проверке гипотез о равенстве двух генеральных средних ($H_0 : \mu_1 = \mu_2$) пользуется *t*-критерий Стьюдента.

Пусть $X_1, X_2, ..., X_{n_1}$, $Y_1, Y_2, ..., Y_{n_2}$ – две выборки взаимно независимых случайных величин, имеющих одинаковую, но неизвестную дисперсию σ^2. По этим наблюдениям можно вычислить оценки $\overline{\mathbb{X}} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$ и $\overline{\mathbb{Y}} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$ для неизвестных математических ожиданий μ_1 и μ_2, а также оценки $s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_i - \overline{\mathbb{X}}\right)^2$ и $s_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(Y_j - \overline{\mathbb{Y}}\right)^2$ для неизвестной дисперсии σ^2. Тогда статистика *t*-критерия Стьюдента имеет вид

$$t = \frac{\left|\overline{\mathbb{X}} - \overline{\mathbb{Y}}\right|}{\sqrt{\frac{n_1 + n_2}{n_1 n_2} \cdot \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}. \qquad (8.3)$$

Если случайные величины распределены по нормальному закону, то данная статистика подчиняется *t*-распределению Стьюдента с ν-степенями свободы:

$$\nu = n_1 + n_2 - 2. \qquad (8.4)$$

Обязательное требование, которое должно выполняться при использовании данного критерия, это *равенство генеральных дисперсий в сравниваемых группах*.

Статистика *t*-критерия Стьюдента для проверки гипотез о генеральных средних двух групп с неравными дисперсиями имеет вид

$$t = \frac{\left|\overline{\mathbb{X}} - \overline{\mathbb{Y}}\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \qquad (8.4)$$

В случае принадлежности выборок нормальному закону при справедливости гипотезы H_0 данная статистика подчиняется *t*-распределению Стьюдента с ν-степенями свободы:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 + 1}} - 2. \qquad (8.5)$$

Причем, если $n_1 = n_2 = n,$ то данная статистика подчиняется *t*-распределению Стьюдента с соответствующими ν-степенями свободы:

$$\nu = n - 1 + \frac{2n - 2}{\frac{s_1^2}{s_2^2} + \frac{s_2^2}{s_1^2}}. \qquad (8.6)$$

Таким образом, предельным распределением *t*-статистики Стьюдента в случае нормальности распределения исследуемого признака в каждой из сравниваемых групп является распределение Стьюдента с соответствующими ν-степенями свободы.

8.1.2. Критерий Манна и Уитни

Ранговый критерий Манна и Уитни основан на критерии Уилкоксона для независимых выборок. Он является непараметрическим аналогом *t*-критерия для сравнения двух средних значений непрерывных распределений. Для вычисления статистики упорядочивают $n_1 + n_2$ значений объединенной выборки, определяют сумму рангов R_1, соответствующую элементам первой выборки, и сумму рангов второй R_2.

При ранжировании следует придерживаться следующих правил.

1. Меньшему значению начисляется меньший ранг.
2. Наименьшему значению начисляется ранг 1.
3. Наибольшему значению начисляется ранг, соответствующий количеству ранжируемых значений, за возможным исключением для тех случаев, которые предусмотрены следующим правилом.
4. В случае, если несколько значений равны, им начисляется ранг, представляющий собой среднее значение из тех рангов, которые они получили бы, если бы не были равны.

Общая сумма рангов должна совпадать с расчетной, которая определяется по формуле

$$\sum(R_i) = \frac{N(N+1)}{2},$$

где N – общее количество ранжируемых наблюдений (значений).

Далее вычисляются

$$U_1 = n_1 n_2 + \frac{n_1(n_1-1)}{2} - R_1, \tag{8.7}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2-1)}{2} - R_2. \tag{8.8}$$

Статистика критерия имеет вид

$$U = \min\{U_1, U_2\}. \tag{8.9}$$

Вместо *U*-статистики удобнее использовать статистику

$$\tilde{z} = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}, \tag{8.10}$$

дискретное распределение которой в случае справедливости проверяемой гипотезы H_0 при $n_1 + n_2 > 60$ хорошо приближается стандартным нормальным законом, когда объем каждой из выборок не слишком мал: $n_1 \geq 8$, $n_2 \geq 8$. При меньших объемах выборок следует учитывать, что достигаемый уровень значимости, вычисляемый по значению статистики в соответствии с функцией распределения стандартного нормального закона, может заметно отличаться от истинного.

8.2. Критерии проверки гипотез о дисперсиях

8.2.1. Критерий Фишера

Для определения того, относятся ли две выборки к одной и той же генеральной совокупности, проверяется гипотеза вида $H_0 : \sigma_1^2 = \sigma_2^2$. Статистика для проверки гипотезы имеет вид

$$F = \frac{s_1^2}{s_2^2}. \tag{8.11}$$

В случае принадлежности выборок нормальному закону и справедливости H_0 эта статистика подчинятся F_{ν_1,ν_2}-распределению Фишера с числом степеней свободы $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$.

8.2.2. Критерий Бартлетта

Статистика критерия Бартлетта вычисляется в соответствии с соотношением

$$\chi^2 = M\left[1 + \frac{1}{3(m-1)}\left(\sum_{i=1}^{m}\frac{1}{\nu_i} - \frac{1}{N}\right)\right]^{-1}, \tag{8.12}$$

где n_i – объемы выборок, $\nu_i = n_i$, если математическое ожидание известно, и $\nu_i = n_i - 1$, если неизвестно, $N = \sum_{i=1}^{m}\nu_i$,

$$M = N \ln\left(\frac{1}{N} \sum_{i=1}^{m} \nu_i S_i^2 \right) - \sum_{i=1}^{m} \nu_i \ln S_i^2, \qquad (8.13)$$

S_i^2 – оценки выборочных дисперсий. При неизвестном математическом ожидании оценки $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(X_{ji} - \overline{X}_i \right)^2$, где $\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ji}$ и $\nu_i = n_i - 1$. Если гипотеза H_0 верна, все $\nu_i > 3$ и выборки извлекаются из нормальной генеральной совокупности, то статистика (8.12) приближенно подчиняется χ^2_{m-1} -распределению.

Тема 9. Проверка гипотезы независимости

В эксперименте наблюдается двумерная случайная величина $\xi = (\xi_1, \xi_2)$ с неизвестной функцией распределения $F_\xi(x, y)$, и есть основания предполагать, что ξ_1 и ξ_2 независимы. В этом случае нужно проверить гипотезу независимости

$$H_0 : F_\xi(x, y) = F_{\xi_1}(x) F_{\xi_2}(y),$$

где $F_{\xi_1}(x)$ и $F_{\xi_2}(y)$ – некоторые одномерные функции распределения.

Для проверки гипотезы независимости используется критерий χ^2 Пирсона. Если исходные данные негруппированы, то предварительно производится группировка наблюдений.

Пусть случайная величина ξ_1 принимает значения $c_1, ..., c_s$, а ξ_2 – $b_1, ..., b_k$. Обозначим ν_{ij} количество наблюдений (c_i, b_j), $\sum_{i=1}^{s}\sum_{j=1}^{k} \nu_{ij} = n$.

Таблица сопряженности признаков ξ_1 и ξ_2 имеет вид

ξ_1 \ ξ_2	b_1	…	b_k	
c_1	ν_{11}		ν_{1k}	$\nu_{1\bullet}$
…				…
c_s	ν_{s1}		ν_{sk}	$\nu_{s\bullet}$
	$\nu_{\bullet 1}$	…	$\nu_{\bullet k}$	

Статистика критерия независимости χ^2 Пирсона

$$\chi^2_n = n\sum_{i=1}^{s}\sum_{j=1}^{k} \frac{\left(\nu_{ij} - \nu_{i\bullet}\nu_{\bullet j}/n\right)^2}{\nu_{i\bullet}\nu_{\bullet j}} = n\left(\sum_{i=1}^{s}\sum_{j=1}^{k} \frac{\nu_{ij}^{\,2}}{\nu_{i\bullet}\nu_{\bullet j}} - 1\right)$$

имеет распределение $\chi^2_{((s-1)(k-1))}$ при $n \to \infty$.

Пример 9.1

В следующей таблице представлены значения показателя Y и значения показателя X в течение 12 лет.

Год	Y	X	Год	Y	X
1986	152	170	1992	177	200
1987	159	179	1993	179	207
1988	162	187	1994	184	215
1989	165	189	1995	186	216
1990	170	193	1996	190	220
1991	172	199	1997	191	225

Проверить гипотезу о независимости величин X и Y.

Решение

Для проверки гипотезы независимости воспользуемся критерием независимости χ^2. Зададимся уровнем значимости $\alpha = 0,05$. Составим таблицу сопряженности двух признаков: $i = \overline{1,s}$, $j = \overline{1,k}$.

X \ Y	(151,161]	(161,171]	(171,181]	(181,191]	$\nu_{i\bullet}$
(165,180]	2	0	0	0	2
(180,195]	0	3	0	0	3
(195,210]	0	0	3	0	3
(210,225]	0	0	0	4	4
$\nu_{\bullet j}$	2	3	3	4	12

Статистика критерия независимости χ^2: $X_n^2 = n\left(\sum_{i,j}\frac{\nu_{ij}^2}{\nu_{i\bullet}\nu_{\bullet j}} - 1\right)$ имеет χ^2-распределение с числом степеней свободы $(s-1)(k-1)$. Вычислим значение статистики: $X_n^2 = 36$, число степеней свободы $(s-1)(k-1) = 9$. Находим по таблице из приложения 3 критическое значение статистики Пирсона при $\alpha = 0,05$: $S_\alpha = 16,9$. Поскольку $X_n^2 > S_\alpha$, то гипотеза о независимости признаков X и Y отвергается.

Тема 10. Проверка гипотезы случайности

Пусть результат эксперимента описывается n-мерной случайной величиной $\mathbb{X} = \{X_1, ..., X_n\}$ с неизвестной функцией распределения $F_X(x)$, $x = (x_1, ...x_n)$. Для ответа на вопрос, можно ли рассматривать $\mathbb{X}$ как случайную выборку из распределения некоторой случайной величины ξ (т. е. являются ли компоненты X_i независимыми и одинаково распределенными), требуется проверить гипотезу

$$H_0 : F_X(x_1, \ldots, x_n) = F(x_1) \cdot \ldots \cdot F(x_n),$$

где $F_X(x_1, \ldots, x_n)$ – совместная функция распределения для элементов $\mathbb{X}$; $F(x)$ – некоторая функция распределения. Такую гипотезу и называют гипотезой случайности.

Первые значительные работы в этой области были выполнены и опубликованы в 1938 году британскими статистиками М. Кендаллом и Б. Смитом. Ими был предложен набор из четырех критериев проверки гипотезы случайности. Далее мы рассмотрим ряд критериев, встречающихся в современной литературе.

10.1. Критерий инверсий

Если случайность действительно имеет место, то компоненты вектора $\mathbb{X}$ «равноправны» и поэтому данные не должны быть ни в коем смысле упорядочены. Другими словами, ситуацию, соответствующую гипотезе H_0, можно охарактеризовать как «полный хаос» или «полный беспорядок». При отклонениях от H_0 исходные данные имеют тот или иной порядок, проявляются связи. Следовательно, критерий проверки H_0 можно построить на основании статистик, измеряющих степень «беспорядка» исходных данных. Одной из таких статистик является число инверсий в выборке.

Гипотезу случайности по критерию инверсий проверяют так: для заданного уровня значимости α определяют число t_α из условия $\Phi(-t_\alpha)=\frac{\alpha}{2}$; по фактически наблюдавшимся данным $\mathbb{X}=\{X_1,...,X_n\}$ вычисляют значение $t=T_n(\mathbb{X})$ числа инверсий в выборке; если $\left|t-\frac{(n-1)n}{4}\right|\frac{6}{\sqrt{n^3}}>t_\alpha$, то гипотезу H_0 отвергают как противоречащую исходным данным; в противном случае признают, что гипотеза случайности наблюдений согласуется с исследуемыми данными.

Вероятность ошибочно отвергнуть при этом истинную гипотезу H_0 равна

$$P\left\{\left|T_n-\frac{(n-1)n}{4}\right|\frac{6}{\sqrt{n^3}}>t_\alpha\,\middle|\,H_0\right\}\to 2\Phi(-t_\alpha)=\alpha.$$

Это правило можно использовать уже при $n>10$.

10.2. Критерии медиан

Также для проверки гипотезы случайности применяются критерии, основанные на использовании медианы выборки. Оценка медианы вычисляется так: из элементов выборки формируется вариационный ряд $X_{(1)}\le...\le X_{(n)}$, тогда $\widehat{X}_{\text{med}}=X_{\left(\frac{n+1}{2}\right)}$, если n нечетно; $\widehat{X}_{\text{med}}=\frac{1}{2}\left(X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n+1}{2}\right)}\right)$, если n четно. Затем выборка $\mathbb{X}$ преобразуется так: на место элемента X_i ставится нуль, если $X_i<\widehat{X}_{\text{med}}$, или единица, если $X_i>\widehat{X}_{\text{med}}$.

Полученную после этого выборку характеризуют двумя статистиками: количеством серий $\nu(n)$ и длиной самой длинной серии $\tau(n)$. При этом под серией понимается последовательность подряд идущих нулей или единиц. Серия может состоять только из одного нуля или единицы. Длина серии – количество подряд идущих нулей или единиц.

Критические значения для обоих статистик вычисляются следующим образом:

$$\nu_\alpha(n) = \frac{1}{2}\left(n+1-u_{\frac{1-\alpha}{2}}\sqrt{n-1}\right);\ \tau_\alpha(n) = 3{,}31\lg(n+1),$$

где $u_{\frac{1-\alpha}{2}}$ – квантиль нормального распределения.

Причем для критерия, основанного на статистике $\nu(n)$, гипотеза не отвергается в случае $\nu(n) > \nu_\alpha(n)$, а для критерия, основанного на статистике $\tau(n)$, – в случае $\tau(n) < \tau_\alpha(n)$.

10.3. Критерии монотонных серий

Также для проверки гипотезы случайности применяются критерии, основанные на анализе серий монотонности в выборке. В этих критериях, так же как и в рассмотренных в предыдущем разделе, формируется последовательность серий из нулей и единиц. Выборку из нулей и единиц, подлежащую дальнейшему исследованию, формируют так: если $X_i < X_{i+1}$, то в выборку записывается нуль, если же $X_i > X_{i+1}$ – то единица.

Полученная выборка характеризуется теми же статистиками: количеством серий $\nu(n)$ и длиной самой длинной серии $\tau(n)$. Критические значения вычисляются так:

$$\nu_{\text{кр}}(n) = \frac{1}{3}(2n-1) - u_{\frac{1-\alpha}{2}}\sqrt{\frac{16n-29}{90}};$$

$$\tau_\alpha(n) = \begin{cases} 5, & n \le 26, \\ 6, & n \le 153, \\ 7, & 153 < n \le 1170, \end{cases}$$

где $u_{\frac{1-\alpha}{2}}$ – квантиль нормального распределения.

Причем для критерия, основанного на статистике $\nu(n)$, гипотеза не отвергается в случае $\nu(n) > \nu_\alpha(n)$, а для критерия, основанного на статистике $\tau(n)$, – в случае $\tau(n) < \tau_\alpha(n)$.

10.4. Критерий знаков

Критерий знаков основан на предположении о том, что количество положительных и отрицательных $\operatorname{sign}(X_i - X_{i-1})$ в исследуемой выборке должно примерно совпадать в случае случайных данных. Для проверки критерия используется тот факт, что число положительных знаков $\operatorname{sign}(X_i - X_{i-1})$ слабо сходится к распределению $N\left(\frac{m}{2}, \frac{m}{4}\right)$, где m – количество ненулевых значений $(X_i - X_{i-1})$.

10.5. Критерий Манна – Кендалла

Критерий основан на использовании статистики, имеющей схожее строение с коэффициентом корреляции Кендалла $\tau^{(K)}$. Если мы рассмотрим время $\{1, 2, ..., n\}$ наблюдаемой последовательности в качестве $\mathbb{X}$, а множество упорядоченных по времени наблюдений $\{Y_1, Y_2, ..., Y_n\}$ – в качестве $\mathbb{Y}$, то ассоциация (связь) между $\mathbb{X}$ и $\mathbb{Y}$ может считаться указанием на закономерность. В отличие от критериев восходящих/нисходящих серий, в выборочном коэффициенте Кендалла рассматриваются знаки относительных величин каждого наблюдения относительно всех предшествующих наблюдений. Статистика имеет вид

$$T = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \operatorname{sign}(Y_i - Y_j)$$

и при верной нулевой гипотезе случайности сходится к нормально распределенной величине: $T \sim N(0, \sigma_3^2)$, где $\sigma_3^2 = \frac{1}{18} n(n-1)(2n+5)$.

Тема 11. Построение наиболее мощных критериев

Любая процедура проверки статистической гипотезы связана с ошибками первого и второго рода. Если вероятность ошибки экспериментатор вправе задать самостоятельно, то вероятность ошибки второго рода зависит от множества параметров эксперимента:

- вероятности ошибки первого рода;
- объема выборки;
- критерия проверки гипотезы;
- альтернативной гипотезы.

Естественно желание экспериментатора, чтобы вероятность ошибки первого рода была минимальна, но в этом случае, естественно, растет вероятность ошибки второго рода. Поэтому на практике, как правило, вероятность ошибки первого рода берут не менее 0,01.

При фиксированной вероятности ошибки первого рода и при увеличении объема выборки вероятность ошибки второго рода должна уменьшаться и стремиться к нулю. Критерии, удовлетворяющие этому условию, называются *состоятельными.* На практике имеет смысл использовать только состоятельные критерии. Однако увеличение объема выборки сопряжено с очевидными трудностями.

Таким образом, мы подходим к тому, что при фиксированной вероятности ошибки первого рода и фиксированном объеме выборки мы должны выбирать такой критерий проверки гипотезы, который дает минимальную вероятность ошибки второго рода.

11.1. Наиболее мощный критерий

Пусть в эксперименте требуется выбрать одну из двух гипотез:

$$H_0 : f_\xi(x) = f_0(x), \quad H_1 : f_\xi(x) = f_1(x),$$

где $f(x)$ – функция плотности, если случайная величина ξ является непрерывной, и вероятность события $P\{\xi = x\}$, если ξ – дискретная случайная величина.

Определение 11.1. *Наиболее мощным критерием* при заданном уровне значимости α называется критерий, имеющий наибольшую мощность $1-\beta$.

11.2. Построение наиболее мощного критерия в случае простой гипотезы

Определение 11.2. *Статистикой отношения правдоподобия* называется статистика

$$\Lambda = \Lambda\left(\mathbb{X}_n\right) = \frac{L\left(\mathbb{X}_n \middle| H_1\right)}{L\left(\mathbb{X}_n \middle| H_0\right)},$$

где $L\left(\mathbb{X}_n \middle| H_j\right) = \prod_{i=1}^{n} f_j(X_i)$ – функции правдоподобия при верной H_j, $j = 1, 2$.

Естественно, что чем больше Λ, тем большее предпочтение мы должны оказать гипотезе H_1, и наоборот, чем меньше Λ, тем большее предпочтение мы должны отдавать гипотезе H_0.

Таким образом, мы получаем следующее правило:

1) если $\Lambda > t_\alpha$, то гипотеза H_0 отвергается;

2) если $\Lambda \le t_\alpha$, то гипотеза H_0 не отвергается.

Найти критическое значение t_α можно из уравнения

$$P\left\{\Lambda > t_\alpha \middle| H_0\right\} = \alpha.$$

Отсюда, в случае, если ξ является непрерывной,

$$t_\alpha = F^{-1}_{\Lambda|H_0}\left(1-\alpha\right),$$

где $F^{-1}_{\Lambda|H_0}(y)$ – обратная функция распределения случайной величины Λ при верной гипотезе H_0.

Построенный критерий называется *критерием отношения правдоподобия.*

Теорема 11.1 (Лемма Неймана – Пирсона)

Среди всех критериев заданного уровня значимости α, проверяющих две простые гипотезы H_0 и H_1, критерий отношения правдоподобия является наиболее мощным.

Доказательство

Пусть критерий отношения правдоподобия задается критической областью $\boldsymbol{X}_1$, т. е. это подмножество выборочного пространства, при попадании в которое гипотеза H_0 отвергается: $\boldsymbol{X}_1 = \{\mathbb{X}_n \in \boldsymbol{X} | H_0 \text{ отвергается}\}$.

Аналогично определим доверительную область $\boldsymbol{X}_0$, при попадании в которую гипотеза H_0 не отвергается: $\boldsymbol{X}_0 = \{\mathbb{X}_n \in \boldsymbol{X} | H_0 \text{ не отвергается}\}$.

Рассмотрим любой другой критерий того же уровня значимости α для проверки тех же гипотез и обозначим через $\boldsymbol{X}_1'$ его критическую область. Тогда графически можно изобразить множество $\boldsymbol{X}_1$ и $\boldsymbol{X}_1'$ как показано на рис. 11.1.

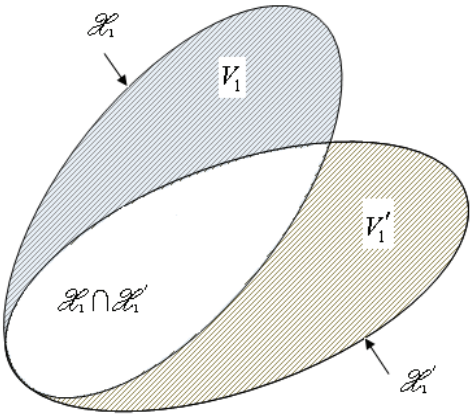


Рис. 11.1. Критические области $\boldsymbol{X}_1$ и $\boldsymbol{X}_1'$

Обозначим через $V = \boldsymbol{X}_1 \setminus \boldsymbol{X}_1'$ такое подмножество выборочного пространства, при попадании в которое выборки $\mathbb{X}_n$ мы должны принять гипотезу H_0 по критерию отношения правдоподобия и отвергнуть по другому критерию.

Аналогично обозначим через $V' = \boldsymbol{X}_1' \setminus \boldsymbol{X}_1$ такое подмножество выборочного пространства, при попадании в которое выборки $\mathbb{X}_n$ мы должны отвергнуть гипотезу H_0 по критерию отношения правдоподобия и принять по другому критерию.

Так как оба критерия имеют одинаковый уровень значимости α, то вероятность попадания выборки $\mathbb{X}_n$ в области V и V' при условии справедливости гипотезы H_0 равны. Обозначим эту вероятность через γ:

$$P\{\mathbb{X}_n \in V \,|\, H_0\} = P\{\mathbb{X}_n \in V' \,|\, H_0\} = \gamma.$$

Теперь определим мощность критерия отношения правдоподобия. По определению, мощность – это вероятность попадания выборки в критическую область при верной альтернативе. В нашем случае критическая область состоит из двух подмножеств – пересечения множеств $\boldsymbol{X}_1$ и $\boldsymbol{X}_1'$ и подмножества V. Таким образом,

$$1-\beta = P\{\mathbb{X}_n \in \boldsymbol{X}_1 \,|\, H_1\} =$$

$$= P\left\{\mathbb{X}_n \in \left(\boldsymbol{X}_1 \cap \boldsymbol{X}_1'\right) \cup V \,|\, H_1\right\} =$$

$$= P\left\{\mathbb{X}_n \in \boldsymbol{X}_1 \cap \boldsymbol{X}_1' \,|\, H_1\right\} + P\{\mathbb{X}_n \in V \,|\, H_1\}.$$

Рассмотрим вероятность попадания выборки в подмножество V при верной гипотезе H_1:

$$P\{\mathbb{X}_n \in V \,|\, H_1\} = \int_V L(x_1, x_2, \ldots, x_n \,|\, H_1) dx_1 dx_2 \ldots dx_n.$$

Заметим, что в случае попадания выборки в подмножество V мы должны отвергнуть гипотезу по критерию отношения правдоподобия, следовательно, выполняется неравенство

$$\Lambda\left(\mathbb{X}_n\right)=\frac{L\left(\mathbb{X}_n\middle|H_1\right)}{L\left(\mathbb{X}_n\middle|H_0\right)}>t_\alpha.$$

Отсюда

$$L\left(\mathbb{X}_n\middle|H_1\right)>t_\alpha L\left(\mathbb{X}_n\middle|H_0\right) \text{ при } \mathbb{X}_n\in V.$$

Тогда

$$P\left\{\mathbb{X}_n\in V\middle|H_1\right\}=\int\limits_V L(x_1,x_2,\ldots,x_n\middle|H_1)dx_1dx_2\ldots dx_n>$$

$$>t_\alpha\int\limits_V L(x_1,x_2,\ldots,x_n\middle|H_0)dx_1dx_2\ldots dx_n=$$

$$=t_\alpha P\left\{\mathbb{X}_n\in V\middle|H_0\right\}=t_\alpha\gamma.$$

Аналогично мощность второго критерия равна

$$1-\beta'=P\left\{\mathbb{X}_n\in\boldsymbol{X}_1\cap\boldsymbol{X}_1'\middle|H_1\right\}+P\left\{\mathbb{X}_n\in V'\middle|H_1\right\}.$$

Но при попадании выборки в подмножество V' гипотеза не отвергается по критерию отношения правдоподобия, и, следовательно,

$$\Lambda\left(\mathbb{X}_n\right)=\frac{L\left(\mathbb{X}_n\middle|H_1\right)}{L\left(\mathbb{X}_n\middle|H_0\right)}\le t_\alpha.$$

Отсюда

$$L\left(\mathbb{X}_n\middle|H_1\right)\le t_\alpha L\left(\mathbb{X}_n\middle|H_0\right) \text{ при } \mathbb{X}_n\in V'.$$

Тогда

$$P\left\{\mathbb{X}_n\in V'\middle|H_1\right\}=\int\limits_{V'} L(x_1,x_2,\ldots,x_n\middle|H_1)dx_1dx_2\ldots dx_n>$$

$$> t_\alpha \int\limits_{V'} L(x_1, x_2, \ldots, x_n | H_0) dx_1 dx_2 \ldots dx_n =$$

$$= t_\alpha P\left\{ \mathbb{X}_n \in V' \middle| H_0 \right\} = t_\alpha \gamma.$$

В результате получили, что

$$1 - \beta \geq P\left\{ \mathbb{X}_n \in \boldsymbol{X}_1 \cap \boldsymbol{X}_1' \middle| H_1 \right\} + t_\alpha \gamma \text{ и } 1 - \beta' \leq P\left\{ \mathbb{X}_n \in \boldsymbol{X}_1 \cap \boldsymbol{X}_1' \middle| H_1 \right\} + t_\alpha \gamma.$$

Следовательно, мощность второго критерия не больше мощности критерия отношения правдоподобия:

$$1 - \beta \geq 1 - \beta'.$$

Замечание. Мощности критерия отношения правдоподобия и второго критерия совпадают тогда и только тогда, когда $\gamma = 0$, т. е. вероятность попадания в области V и V' при условии H_0 равны нулю. Поэтому критерий отношения правдоподобия – единственный с точностью до множеств, вероятность попадания в которые равна нулю.

Пример 11.1

Пусть рассматриваются две простые гипотезы о нормальном распределении с разными параметрами сдвига:

$$H_0 : F(x) \in N\left(\theta_0, \sigma^2\right);$$

$$H_1 : F(x) \in N\left(\theta_1, \sigma^2\right).$$

В простой гипотезе все параметры фиксированы, поэтому для определенности будем считать, что $\theta_1 > \theta_0$. Требуется построить наиболее мощный критерий и определить его мощность.

Запишем функцию правдоподобия для нормального закона с неизвестным параметром сдвига:

$$L\left(\mathbb{X}_n, \theta\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2}\left[(X_1 - \theta)^2 + \ldots + (X_n - \theta)^2 \right]}.$$

Найдем отношение правдоподобия

$$\Lambda(\mathbb{X}_n) = \frac{L(\mathbb{X}_n, \theta_1)}{L(\mathbb{X}_n, \theta_0)} = \exp^{\frac{\theta_1 - \theta_0}{\sigma^2}\left(\sum\limits_{i=1}^{n} X_i - n\frac{\theta_0 + \theta_1}{2}\right)}.$$

Гипотеза H_0 отвергается, если $\Lambda > t_\alpha$. Но чтобы найти t_α, нужно найти распределение статистики отношения правдоподобия $F_{\Lambda|H_0}(t)$. Чтобы упростить нашу задачу, построим критерий, эквивалентный критерию отношения правдоподобия, путем преобразования неравенства $\Lambda > t_\alpha$ таким образом, чтобы получить слева статистику, распределение которой мы знаем:

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

$$\sim \frac{\theta_1 - \theta_0}{\sigma^2}\left(\sum_{i=1}^{n} X_i - n\frac{\theta_0 + \theta_1}{2}\right) > \ln t_\alpha.$$

Так как по условию задачи $\theta_1 > \theta_0$, то можем разделить левую и правую часть неравенства на $\frac{\theta_1 - \theta_0}{\sigma^2}$. Знак неравенства при этом не изменится. В результате получим

$$\sum_{i=1}^{n} X_i > \frac{\sigma^2}{\theta_1 - \theta_0}\ln t_\alpha + n\frac{\theta_0 + \theta_1}{2} = t'_\alpha.$$

Таким образом, вместо критерия $\Lambda > t_\alpha$ мы построили эквивалентный критерий $T(\mathbb{X}_n) = \sum\limits_{i=1}^{n} X_i > t'_\alpha$.

Статистика $T(\mathbb{X}_n) = \sum_{i=1}^{n} X_i$ при верной гипотезе H_0 распределена по нормальному закону $N\left(n\theta_0, n\sigma^2\right)$. Найдем t'_α :

$$\alpha = P\left\{T(\mathbb{X}_n) > t'_\alpha \mid H_0\right\} = 1 - \Phi(t'_\alpha; n\theta_0, n\sigma^2) = 1 - \Phi\left(\frac{t'_\alpha - n\theta_0}{\sqrt{n\sigma^2}}\right).$$

Отсюда

$$t'_\alpha = n\theta_0 + \sqrt{n\sigma^2} \cdot \Phi^{-1}(1-\alpha).$$

Найдем мощность построенного критерия:

$$1 - \beta = P\left\{T(\mathbb{X}_n) > t'_\alpha \mid H_1\right\} = 1 - \Phi(t'_\alpha; n\theta_1, n\sigma^2) = 1 - \Phi\left(\frac{t'_\alpha - n\theta_1}{\sqrt{n\sigma^2}}\right).$$

Пример 11.2

Для критерия, построенного в примере 11.1, требуется найти необходимый объем выборки, чтобы его мощность была не меньше $1-\beta$.

Из найденного выражения для мощности выразим t'_α :

$$t'_\alpha = n\theta_1 + \sqrt{n\sigma^2} \cdot \Phi^{-1}(\beta).$$

Приравняем полученное выражение к t'_α :

$$n\theta_0 + \sqrt{n\sigma^2}\Phi^{-1}(1-\alpha) = n\theta_1 + \sqrt{n\sigma^2} \cdot \Phi^{-1}(\beta);$$

$$n(\theta_1 - \theta_0) = (\Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta))\sqrt{n\sigma^2};$$

$$n = \frac{(\Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta))^2 \sigma^2}{(\theta_1 - \theta_0)^2}.$$

Таким образом, мы нашли минимальный размер выборки для различения заданных гипотез с заданными ошибками первого и второго рода.

11.3. Критерий отношения правдоподобия в случае дискретных распределений

Пусть наблюдается дискретная случайная величина. Тогда отношение правдоподобия

$$\Lambda(\mathbb{X}_n) = \frac{\prod_{i=1}^{n} f_1(X_i)}{\prod_{i=1}^{n} f_0(X_i)}$$

принимает дискретные значения $l_1 < l_2 < \ldots < l_k < l_{k+1} < \ldots$

Проблема построения критерия отношения правдоподобия в дискретном случае заключается в том, что мы не всегда можем найти такое критическое значение t_α, при котором $P\{\Lambda > t_\alpha | H_0\} = \alpha$, так как распределение дискретной случайной величины изменяется скачками, и, например, при некотором k вероятность $P\{\Lambda > l_k | H_0\} = \alpha_0 < \alpha$, но $P\{\Lambda > l_{k+1} | H_0\} = \alpha_0 + p_0 \geq \alpha$ (рис. 11.2).

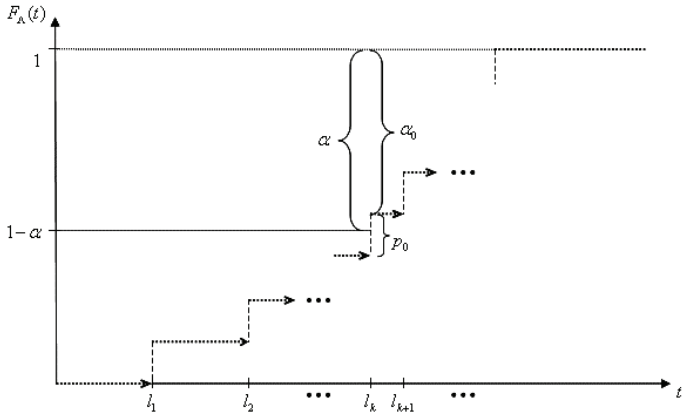


Рис. 11.2. Распределение статистики отношения правдоподобия в дискретном случае

Если $\alpha = \alpha_0$, то критическая область состоит из множества $\left\{\mathbb{X}_n \mid \Lambda(\mathbb{X}_n) > l_{k(\alpha)}\right\}$.

Если $\alpha > \alpha_0$ и $p_0 = P\left\{\Lambda(\mathbb{X}_n) = l_{k(\alpha)}\right\}$, то строят *рандоминизированный* критерий: гипотезу H_0 отвергают с вероятностью $\dfrac{\alpha - \alpha_0}{p_0}$ и принимают с вероятностью $1 - \dfrac{\alpha - \alpha_0}{p_0}$. Тогда

$$P\left\{\Lambda(\mathbb{X}_n) > l_{k(\alpha)} \mid H_0\right\} + \frac{\alpha - \alpha_0}{p_0} P\left\{\Lambda(\mathbb{X}_n) = l_{k(\alpha)}\right\} =$$

$$= \alpha_0 + \frac{\alpha - \alpha_0}{p_0} p_0 = \alpha_0 + \alpha - \alpha_0 = \alpha,$$

и уровень значимости построенного критерия равен α.

11.4. Построение равномерно наиболее мощного критерия

Понятие равномерно наиболее мощного критерия (РНМК) связано с проверкой сложных гипотез, т. е. когда основная и (или) альтернативная гипотезы заданы с точностью до некоторого множества. Рассмотрим частный случай сложных гипотез, когда по основной и альтернативной гипотезам мы предполагаем, что наблюдения принадлежат одному и тому же параметрическому семейству $F(x, \theta), \theta \in \Theta$. Тогда

$$H_0 : \theta \in \Theta_0 \subset \Theta,$$

$$H_1 : \theta \in \Theta_1 \subset \Theta,$$

и, естественно, $\Theta_0 \bigcap \Theta_1 = \varnothing$.

Определение 11.3. *Равномерно наиболее мощным критерием* при заданном уровне значимости α называется критерий, имеющий наибольшую мощность $1-\beta$ для любых пар простых гипотез $H_0:\theta=\theta_0$ и $H_1:\theta=\theta_1$, где $\theta_0\in\Theta_0$ и $\theta_1\in\Theta_1$.

В частности, когда подмножества Θ_0 и Θ_1 состоят из одной точки, мы получаем наиболее мощный критерий.

Надо отметить, что задача построения равномерно наиболее мощного критерия в общем случае неразрешима, т. е., как правило, для одних альтернатив наиболее мощным будет один критерий, а для других – другой.

Построим для каждой фиксированной альтернативы $\theta\in\Theta_1$ критерий отношения правдоподобия $\boldsymbol{X}_{1\alpha}=\boldsymbol{X}_{1\alpha}(\theta_0,\theta_1)=\{\mathbb{X}_n\mid\Lambda(\mathbb{X}_n)>=>t_\alpha(\theta_0,\theta_1)\}$. Если $\boldsymbol{X}_{1\alpha}(\theta_0,\theta_1)$ зависит от альтернативы θ_1, то не существует критической области, которая была бы наилучшей $\forall\theta_1\in\Theta_1$, и РНМК не существует.

Если $\boldsymbol{X}_{1\alpha}(\theta_0,\theta_1)$ не зависит от θ_1, т. е. $\boldsymbol{X}_{1\alpha}(\theta_0,\theta_1)=\boldsymbol{X}_{1\alpha}(\theta_0)$, то эта критическая область максимизирует мощность при любой альтернативе и поэтому $\boldsymbol{X}_{1\alpha}(\theta_0)$ является РНМК.

Сформулируем общее достаточное условие существования РНМК в случае односторонних альтернатив. Пусть θ – скаляр и H_1 – односторонняя гипотеза, т. е. $H_1=H_1^+:\theta>\theta_0$ или $H_1=H_1^-:\theta<\theta_0$. Пусть вероятностная модель $F(x,\theta)$ обладает достаточной статистикой $T(\mathbb{X}_n)$. Тогда из критерия факторизации следует, что

$$\Lambda(\mathbb{X}_n)=\frac{g(T(\mathbb{X}_n),\theta_1)}{g(T(\mathbb{X}_n),\theta_0)}.$$

Пусть класс моделей $F(x,\theta)$ такой, что отношение $\dfrac{g(T(\mathbb{X}_n),\theta_1)}{g(T(\mathbb{X}_n),\theta_0)}$ является монотонной функцией T.

Для таких моделей существует РНМК, который совпадает с критерием отношения правдоподобия для любой альтернативы из H_1.

Действительно, пусть $H_1 = H_1^+$ и $\theta_1 > \theta_0$. Тогда $\dfrac{g(T,\theta_1)}{g(T,\theta_0)}$ возрастает по T.

Тогда $\Lambda(\mathbb{X}_n) \geq t_\alpha \sim T\left(\mathbb{X}_n\right) \geq t_\alpha^+$, причем граница t_α^+ определяется по α и распределению $F(x,\theta_0)$.

В случае двусторонней альтернативы $H_1 = H_1^+ \bigcup H_1^- : \theta \neq \theta_0$ даже для экспоненциальных моделей с монотонным отношением правдоподобия в общем случае не существует РНМК. В этом случае поступают следующим образом: используют достаточную статистику $T\left(\mathbb{X}_n\right)$, с помощью которой строится РНМК против односторонних альтернатив H_1^+ и H_1^-, и задают критическую область в виде объединения $\boldsymbol{X}_{1\alpha} = \left\{T\left(\mathbb{X}_n\right) \leq t_{\alpha_1}^-\right\} \bigcup \left\{T\left(\mathbb{X}_n\right) \geq t_{\alpha_2}^+\right\}$, где $\alpha = \alpha_1 + \alpha_2$.

Пример 11.3

Пусть $\mathbb{X}_n = \left\{X_1, X_2, ..., X_n\right\}$ – выборка из экспоненциального распределения с функцией плотности

$$f(x,\theta) = \begin{cases} \dfrac{1}{\theta} e^{-\frac{x}{\theta}}, x \geq 0, \\ 0, x < 0. \end{cases}$$

Используя достаточную статистику, построить равномерно наиболее мощный критерий для проверки гипотезы $H_0 : \theta = \theta_0$ против альтернативы $H_1 : \theta_1 > \theta_0$.

Решение

Найдем достаточную статистику:

$$L(X_n,\theta) = \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum_{i=1}^{n} X_i}.$$

Отсюда, $g(T(X_n),\theta) = \dfrac{1}{\theta^n} e^{-\frac{1}{\theta}T(X_n)}$, $h(X_n) = 1$.

По критерию факторизации $T(X_n)=\sum_{i=1}^{n} X_i$ – достаточная статистика.

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

так как $-\frac{\theta_0-\theta_1}{\theta_0\theta_1}>0,$ очевидно, что $\Lambda(X_n,\theta)$ монотонна по $T(X_n)=\sum_{i=1}^{n} X_i.$ Таким образом, РНМК существует и совпадает с НМК для некоторого $\theta=\theta_1.$

$$\Lambda(X_n,\theta)=\left(\frac{\theta_0}{\theta_1}\right)^n \exp\left\{-\frac{\theta_0-\theta_1}{\theta_0\theta_1}\sum_{i=1}^{n} X_i\right\}>c_\alpha,$$

$$\exp\left\{-\frac{\theta_0-\theta_1}{\theta_0\theta_1}\sum_{i=1}^{n} X_i\right\}>\left(\frac{\theta_0}{\theta_1}\right)^n c_\alpha,$$

$$\sum_{i=1}^{n} X_i>-\frac{\theta_0-\theta_1}{\theta_0\theta_1}\ln\left[c_\alpha\left(\frac{\theta_0}{\theta_1}\right)^n\right]=\tilde{c}_\alpha,$$

$$P\{H_1\mid H_0\}=P\left\{\sum_{i=1}^{n} X_i>\tilde{c}_\alpha \mid \theta=\theta_0\right\}=\alpha.$$

Так как $X_i\in \mathrm{Exp}(\theta_0)$ или $X_i\in\Gamma(1,\theta_0),$ а $\Gamma(m,\theta_0)$ устойчиво, то

$$\sum_{i=1}^{n} X_i\in\Gamma(n,\theta_0).$$

Тогда

$$\alpha = 1 - F_{\Gamma(n,\theta_0)}(\widetilde{c_\alpha}) \Rightarrow \widetilde{c_\alpha} = F^{-1}_{\Gamma(n,\theta_0)}(1-\alpha).$$

Пример 11.4

Пусть $X_1, X_2, ..., X_n$ – выборка из нормального распределения с функцией плотности

$$f(x,\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}.$$

Используя достаточную статистику, построить равномерно наиболее мощный критерий для проверки гипотезы $H_0 : \theta = \theta_0$ против альтернативы $H_1 : \theta > \theta_0, \theta_0 > 0$.

Решение

Найдем достаточную статистику:

$$L(X_n,\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\theta)^2\right\} =$$

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

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n X_i^2 - 2\theta\sum_{i=1}^n X_i + n\theta^2\right]\right\} =$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2\right\} \exp\left\{\frac{1}{2\sigma^2}\theta\sum_{i=1}^n X_i + n\theta^2\right\},$$

где $h(X_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2\right\}$,

а $g(T(X_n),\theta) = \exp\left\{\frac{1}{2\sigma^2}\theta\sum_{i=1}^n X_i + n\theta^2\right\}$.

Отсюда, по критерию факторизации $T(X_n) = \overline{X}$,

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

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

Отсюда видно, что Λ монотонна по $T(X_n) = \overline{X}$.

$$2(\theta_0 - \theta_1)\sum_{i=1}^{n} X_i + n(\theta_0^2 + \theta_1^2) \le \ln\left(-\frac{c_\alpha}{2\sigma^2}\right),$$

$$\overline{X} \ge \frac{1}{2(\theta_0 - \theta_1)}\left[\ln\left(-\frac{c_\alpha}{2\sigma^2}\right) - (\theta_0^2 + \theta_1^2)\right] = \widetilde{c_\alpha},$$

$$P\{H_1 \mid H_0\} = P\left\{\overline{X} > \tilde{c}_\alpha \mid \theta = \theta_0\right\} = \alpha.$$

11.5. Проверка гипотез и доверительное оценивание

Между задачей проверки простой гипотезы и задачей доверительного оценивания есть связь. Рассмотрим для каждого $\theta_0 \in \Theta$ какой-нибудь критерий $\boldsymbol{X}_{1\alpha} = \{\mathbb{X}_n \mid T(\mathbb{X}_n, \theta_0) > t_\alpha\}$.

Пусть $\boldsymbol{X}_{0\alpha} = \overline{\boldsymbol{X}_{1\alpha}}$ – область принятия гипотезы H_0. Тем самым в выборочном пространстве $\boldsymbol{X}$ задано семейство подмножеств $\{\boldsymbol{X}_{0\alpha}(\theta), \theta \in \Theta\}$. Определим при каждом $\mathbb{X}_n \in \boldsymbol{X}$ подмножество $\Omega(\mathbb{X}_n) \in \Theta$, положив $\Omega(\mathbb{X}_n) = \{\theta : \mathbb{X}_n \in \boldsymbol{X}_{0\alpha}(\theta)\}$. Таким образом, в параметрическом множестве Θ получаем семейство подмножеств $\{\Omega(\mathbb{X}_n), \mathbb{X}_n \in \boldsymbol{X}\}$.

События $\{\theta \in \Omega(\mathbb{X}_n)\}$ и $\{\mathbb{X}_n \in \boldsymbol{X}_{0\alpha}(\theta)\}$ эквивалентны, так как по построению каждое из них влечет за собой другое, поэтому их вероят-

ности при каждом θ совпадают. При верной гипотезе H_0 вероятность второго события равна по построению $1-\alpha$. Следовательно, $P\{\mathbb{X}_n \in \boldsymbol{X}_{0\alpha}(\theta) | H_0\} = P\{\theta \in \Omega(\mathbb{X}_n)\} = 1-\alpha$, т. е. $\Omega(\mathbb{X}_n)$ является доверительной областью для θ с доверительной вероятностью $\gamma = 1-\alpha$.

Верно и обратное, т. е. если имеется семейство γ-доверительных интервалов $\{\Omega_\gamma(\mathbb{X}_n), \mathbb{X}_n \in \boldsymbol{X}\}$ для θ, то множество $\boldsymbol{X}_{0,1-\gamma} = \{\mathbb{X}_n : \theta_0 \in \Omega_\gamma(\mathbb{X}_n)\}$ определяет область принятия гипотезы $H_0 : \theta = \theta_0$ с уровнем значимости $\alpha = 1-\gamma$. При этом РНМК соответствует кратчайшему доверительному интервалу.

Тема 12. Последовательные критерии проверки гипотез

В отличие от классических методов математической статистики, в которых число производимых наблюдений фиксируется заранее, методы последовательного анализа характеризуются тем, что момент прекращения наблюдений является случайным и определяется наблюдателем в зависимости от значений наблюдаемых данных. Преимущество последовательных методов было продемонстрировано А. Вальдом на задаче различия двух простых гипотез, установившим, что такие методы дают выигрыш в среднем числе наблюдений по сравнению с любым другим способом различения с фиксированным объемом выборки и теми же вероятностями ошибочных наблюдений. Вальд указал и тот последовательный метод, названный им критерием последовательных отношений вероятностей, который оказался оптимальным в классе всех последовательных методов.

12.1. Последовательный критерий Вальда

Для проверки гипотезы с заданными вероятностями ошибок первого и второго рода α и β можно вычислить необходимый объем выборочных данных. Кроме правил проверки гипотез, основанных на выборке фиксированного объема, известны последовательные критерии, когда о числе необходимых наблюдений решают в процессе наблюдения. Последовательный критерий отношения правдоподобия или критерий Вальда строят, опираясь на логарифм отношения правдоподобия.

Суть данного метода состоит в следующем. Пусть задача состоит в выборе между гипотезами H_0 и H_1 по результатам независимых наблюдений. Гипотеза H_0 заключается в том, что случайная величина X имеет распределение вероятностей с плотностью $f_0(x)$, а H_1 – в том, что X имеет плотность $f_1(x)$. Для решения этой задачи поступают следующим

образом. Выбирают два числа c_0 и c_1 $(0 < c_0 < c_1)$. После первого наблюдения $(\mathbb{X}_1 = \{X_1\})$ вычисляют значение статистики критерия

$$\Lambda_1 = \left(\frac{L_1(\mathbb{X}_1)}{L_0(\mathbb{X}_1)} \right) = \left(\frac{f_1(X_1)}{f_0(X_1)} \right),$$

где X_1 – результат первого наблюдения.

Если $\Lambda \le c_0$, принимают гипотезу H_0; если $\Lambda \ge c_1$, принимают H_1; если $c_0 < \Lambda < c_1$, производят второе наблюдение $(\mathbb{X}_2 = \{X_1, X_2\})$ и так же исследуют величину

$$\Lambda_2 = \left(\frac{L_1(\mathbb{X}_2)}{L_0(\mathbb{X}_2)} \right) = \left(\frac{f_1(X_1) f_1(X_2)}{f_0(X_1) f_0(X_2)} \right),$$

где X_2 – результат второго наблюдения и т. д.

С вероятностью, равной единице, процесс оканчивается либо выбором H_0, либо выбором H_1, как показано на рис. 12.1.

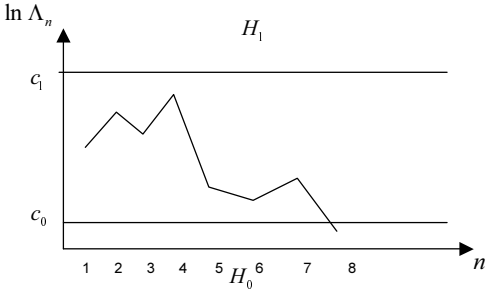


Рис. 12.1. Процедура принятия гипотезы по критерию Вальда

Величины c_0 и c_1 определяются из условия, чтобы вероятности ошибок первого и второго рода имели заданные значения α и β. Обозначим через $c_0(\alpha,\beta)$ и $c_1(\alpha,\beta)$ соответственно величины c_0 и c_1, для которых критерий имеет требуемую силу (α,β). Границы для c_0 и c_1 определяются следующей теоремой.

Теорема 12.1

Критические значения c_0, c_1 критерия Вальда удовлетворяют неравенствам $c_0 \ge \dfrac{\beta}{1-\alpha}$, $c_1 \le \dfrac{1-\beta}{\alpha}$, где α и β – вероятности ошибок первого и второго рода.

Доказательство

Обозначим через $\mathbb{X}_{0n}$ множество тех результатов наблюдений $(X_1,...,X_n)$, для которых процедура заканчивается принятием гипотезы H_0, $\mathbb{X}_{1n}$ – для гипотезы H_1. При этом $\sum\limits_{n=1}^{\infty} P_\theta(\nu = n) = \sum\limits_{n=1}^{\infty} P(\mathbb{X}_{0n}) +$

$+\sum\limits_{n=1}^{\infty} P(\mathbb{X}_{1n}) = 1.$

Далее

$$\alpha = P\{H_1 \mid H_0\} = \sum_{n=1}^{\infty} P(\mathbb{X}_{1n} \mid H_0) \le \frac{1}{c_1} \sum_{n=1}^{\infty} P(\mathbb{X}_{1n} \mid H_1) =$$

$$= \frac{1}{c_1}\left(1 - P\{H_1 \mid H_0\}\right) = \frac{1-\beta}{c_1},$$

так как в точках множества $\mathbb{X}_{1n}$ выполняется неравенство $L_{0n} \le \frac{L_{1n}}{c_1}$.

Аналогично

$$\beta = P\left(H_0 \mid H_1\right) = \sum_{n=1}^{\infty} P\left(\mathbb{X}_{0n} \mid H_1\right) \le c_0 \sum_{n=1}^{\infty} P\left(\mathbb{X}_{0n} \mid H_0\right) =$$

$$= c_0\left(1 - P\left(H_1 \mid H_0\right)\right) = c_0(1-\alpha),$$

поскольку в точках множества $\mathbb{X}_{0n}$ выполняется неравенство $L_{1n} \le c_0 L_{0n}$.

В отличие от критериев с фиксированным объемом выборки, критерий Вальда можно использовать без знания распределения статистики при верной нулевой или альтернативной гипотезе, если в качестве критического значения выбирать оценки снизу и сверху,

$$c_0' \ge \frac{\beta}{1-\alpha},\ c_1' \le \frac{1-\beta}{\alpha}.$$

Пример 12.1

$$H_0 : N(\theta_0, \sigma^2),\ H_1 : N(\theta_1, \sigma^2),\ n = \frac{(\varphi_\alpha - \varphi_\beta)^2 \sigma^2}{(\theta_1 - \theta_0)^2}.$$

Пусть $\alpha = 0{,}05,\ 1-\beta = 0{,}9,\ \sigma^2 = 1,\ \theta_0 = -0{,}5,\ \theta_1 = 0{,}5.$
Тогда $n = 9.$
При использовании критерия Вальда $c_0 = -2{,}25,\ c_1 = 2{,}89,$

$$N_0 = \frac{\alpha c_1 + (1-\alpha)c_0}{M\left[\lambda(X) \mid H_0\right]},\quad N_1 = \frac{\beta c_1 + (1-\beta)c_0}{M\left[\lambda(X) \mid H_1\right]}.$$

Среднее количество наблюдений в случае гипотезы $H_0 : N_0 = 4,$ в случае гипотезы $H_1 : N_1 = 4.$ Следовательно, критерий Вальда требует примерно в два раза меньше наблюдений для различия двух простых гипотез.

Точное определение $c_0(\alpha,\beta)$ и $c_1(\alpha,\beta)$ обычно трудоемкое, и для большинства практических целей достаточно использования в качестве c_0 его нижней границы и в качестве c_1 его верхней границы. Такая замена приведет к изменению вероятностей ошибок первого и второго рода. Если положить c_1 равным величине, большей $c_1(\alpha,\beta),$ и положить $c_0 = c_0(\alpha,\beta),$ то полученная вероятность ошибки первого рода будет меньше α, но вероятность ошибки второго рода будет несколько больше β. Аналогично, если использовать для c_1 точное значение $c_1(\alpha,\beta),$ а для величины c_0 значение ниже точного $c_0(\alpha,\beta),$ то вероятность ошибки второго рода будет меньше $\beta,$ а вероятность ошибки первого рода будет несколько больше $\alpha.$ Но критерий, которому соответствуют $c_0 = \ln\left(\frac{\beta}{1-\alpha}\right)$ и $c_1 = \ln\left(\frac{1-\beta}{\alpha}\right),$ обеспечивает, по крайней мере, такую же гарантию от неправильного решения, как и критерий с $c_0 = c_0(\alpha,\beta)$ и $c_1 = c_1(\alpha,\beta)$ (заданные α и β изменяются незначительно).

Библиографический список

1. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. – М.: Наука, 1983. – 416 с.

2. *Гланц С.* Медико-биологическая статистика. – М.: Практика, 1998. – 459 с.

3. *Гмурман В.Е.* Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 1979. – 400 с.

4. *Губарев В.В.* Вероятностные модели: справочник. В 2-х ч. / Новосиб. электротехн. ин-т. – Новосибирск, 1992. – Ч. 1. – 198 с. Ч. 2 – 188 с.

5. *Ивченко Г.И., Медведев Ю.А.* Математическая статистика: учеб. пособие для втузов. – М.: Высшая школа, 1994. – 248 с.

6. *Ивченко Г.И., Медведев Ю.А., Чистяков А.В.* Сборник задач по математической статистике. – М.: Высшая школа, 1989. – 255 с.

7. *Коршунов Д.А., Чернова Н.И.* Сборник задач и упражнений по математической статистике: учеб. пособие. – Новосибирск: Изд-во Института математики, 2001. – 120 с.

8. *Лемешко Б.Ю., Постовалов С.Н.* Прикладная статистика. Правила проверки опытного распределения с теоретическим: метод. рекомендации. Часть I. Критерии типа χ^2. – Новосибирск: Изд-во НГТУ, 1998. – 126 с.

9. *Лемешко Б.Ю., Постовалов С.Н.* Прикладная статистика. Правила проверки опытного распределения с теоретическим: метод. рекомендации. Часть II. Непараметрические критерии. – Новосибирск: Изд-во НГТУ, 1999. – 85 с.

10. *Лемешко Б.Ю., Постовалов С.Н.* Компьютерные технологии анализа данных и исследования статистических закономерностей: учеб. пособие. – Новосибирск: Изд-во НГТУ, 2004. – 120 с.

11. *Никитина Н.Ш.* Математическая статистика для экономистов: учеб. пособие. – М.: ИНФРА-М; Новосибирск: Изд-во НГТУ, 2001. – 170 с.

12. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход: монография / Б.Ю. Лемешко, С.Б. Лемешко, С.Н. Постовалов, Е.В. Чимитова. – Новосибирск: Изд-во НГТУ, 2011. – 888 с. (серия «Монографии НГТУ»).

Приложение

Основные сведения из курса «Теории вероятностей»

Материал этого раздела является базовым для изучения математической статистики.

Определение. Функцией распределения случайной величины X называется функция действительного переменного x, принимающая при каждом x значение, равное вероятности неравенства $X < x$, т. е. $F(x) = P\{X < x\}$.

Функция распределения обладает следующими свойствами.

1. $F(x) \le F(y)$, если $x < y$.
2. $F(x)$ непрерывна слева при каждом x.
3. $\lim\limits_{x \to -\infty} F(x) = 0$ и $\lim\limits_{x \to \infty} F(x) = 1$.

П1. Виды функций распределения случайных величин

По принимаемым значениям случайные величины делятся на *непрерывные* и *дискретные*. Дискретные случайные величины могут принимать конечное или счетное множество значений. Непрерывные случайные величины в качестве области определения имеют:

- все множество действительных чисел R;
- полуось $x \ge a$ или $x \le a$;
- отрезок $[a, b]$.

Как правило, рассматривают не просто распределения случайных величин, а параметрические семейства распределений $F(x, \theta)$.

П2. Основные числовые характеристики

Числовыми характеристиками называются такие числа, которые характеризуют различные свойства случайной величины.

П2.1. Математическое ожидание

Определение. Математическим ожиданием случайной величины x называется числовая характеристика, вычисляемая по формуле

$$\mathrm{M}\,X = \int_{-\infty}^{\infty} x dF(x),$$

где $F(x)$ – функция распределения случайной величины.

Замечание. В некоторых источниках для обозначения математического ожидания используется символ $\mathrm{E}\,X$.

Математическое ожидание характеризует среднее положение значений случайной величины, однако может быть ситуация, когда величина математического ожидания не принадлежит области определения случайной величины.

П2.2. Дисперсия

Определение. Дисперсией случайной величины X называется числовая характеристика, вычисляемая по формуле

$$\mathrm{D}\,X = \mathrm{M}\left[(X - \mathrm{M}\,X)^2 \right] = \int_{-\infty}^{\infty} \left(x - \mathrm{M}\,X\right)^2 dF(x),$$

где $F(x)$ – функция распределения случайной величины.

Дисперсия является мерой отклонения случайной величины X от ее математического ожидания.

П2.3. Моменты

Определение. *Начальным моментом k-го порядка* случайной величины X называется величина $\mathrm{M}\,X^k$.

В частности, начальным моментом первого порядка случайной величины является математическое ожидание.

Определение. *Центральным моментом k-го порядка* случайной величины X называется величина $\mathrm{M}\left(X - \mathrm{M}\,X\right)^k$.

В частности, центральным моментом второго порядка случайной величины является дисперсия.

П2.4. Ковариация и коэффициент корреляции

Определение. *Ковариацией* случайных величин X и Y называется числовая характеристика, вычисляемая по формуле

$$\operatorname{cov}(X,Y) = \mathrm{M}\big[(X - \mathrm{M}\,X)(Y - \mathrm{M}\,Y)\big].$$

При этом $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$ и $\operatorname{cov}(X,X) = \mathrm{D}\,X$.

С помощью ковариации можно выразить дисперсию суммы случайных величин.

$$\mathrm{D}(X+Y) = \mathrm{D}\,X + \mathrm{D}\,Y + 2\operatorname{cov}(X,Y).$$

Определение. *Коэффициентом корреляции* случайных величин X и Y называется числовая характеристика, вычисляемая по формуле

$$\operatorname{cor}(X,Y) = \operatorname{cov}(X,Y)\big/\sqrt{\mathrm{D}\,X \cdot \mathrm{D}\,Y}\,.$$

Коэффициент корреляции характеризует степень взаимосвязи двух случайных величин.

П2.5. Асимметрия

Определение. *Коэффициентом асимметрии* случайной величины X называется числовая характеристика, вычисляемая по формуле

$$\gamma_1(X) = \frac{\mathrm{M}\left[\left(X - \mathrm{M}\,X\right)^3\right]}{(\mathrm{D}\,X)^{3/2}}.$$

Для симметричного распределения коэффициент асимметрии равен нулю.

П2.6. Эксцесс

Определение. *Эксцессом* называют характеристику унимодального распределения, выражающую островершинность или сглаженность графика плотности в окрестности моды.

Определение. Коэффициентом эксцесса случайной величины x называется числовая характеристика, вычисляемая по формуле

$$\gamma_2(X) = \frac{\mathrm{M}\left[\left(X - \mathrm{M}\,X\right)^4\right]}{(\mathrm{D}\,X)^2} - 3.$$

Для нормального распределения коэффициент эксцесса равен нулю.

П3. Преобразование случайных величин

Пусть задана произвольная случайная величина ξ с функцией распределения $F(x,\Theta)$, где Θ – вектор параметров $(\theta_1,\theta_2,\ldots,\theta_n)$, и функцией плотности распределения $f(x,\Theta)$.

Каждая операция заключается в том, что вместо исходной случайной величины ξ рассматривается случайная величина $\eta = g(\xi)$, где $g(x)$ – некоторая функция. Если функция $y = g(x)$ непрерывна и строго возрастает, то

$$G_\eta(y) = F_\xi(g^{-1}(y)),$$

где $g^{-1}(y)$ – функция, обратная к $g(x)$. Если функция $y = g(x)$ непрерывна и строго убывает, то

$$G_\eta(y) = 1 - F_\xi(g^{-1}(y)).$$

Преобразование $g(x)$ случайной величины ξ в случайную величину η может иметь свои параметры. Для всех операций над распределениями требуется найти аналитический вид для производной функции распределения и область определения случайной величины η.

Замечание. Порядок применения операций влияет на получаемое в результате распределение. Так, например, при выполнении сначала операции сдвига на величину μ, а затем операции масштабирования на величину λ мы получим распределение

$$F\left(\frac{x-\mu}{\lambda}\right),$$

а при обратном порядке их применения мы получим распределение

$$F\left(\frac{x}{\lambda}-\mu\right).$$

П3.1. Сдвиг

Операция сдвига преобразует случайную величину ξ в случайную величину $\eta=\xi+\mu$ с помощью функции

$$y=g(x,\mu)=x+\mu,$$

где μ – параметр сдвига. Тогда

$$x=g^{-1}(y,\mu)=y-\mu,$$

и функция распределения имеет вид

$$G_\eta(x,\theta_1,\theta_2,\ldots,\theta_n,\mu)=F_\xi(x-\mu,\Theta).$$

Функция плотности распределения имеет вид

$$g_\eta(x,\theta_1,\theta_2,\ldots,\theta_n,\mu)=f_\xi(x-\mu,\Theta).$$

Если случайная величина ξ имеет область определения $[l,r]$, то случайная величина η будет иметь область определения $[l+\mu,r+\mu]$.

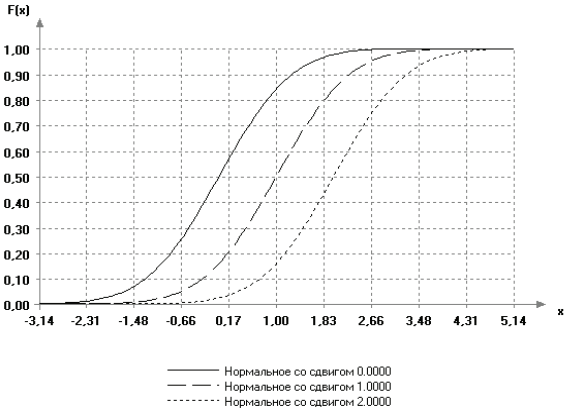


Рис. П1. Функции распределения нормального распределения с параметрами сдвига 0, 1 и 2

П3.2. Масштаб

Операция масштабирования преобразует случайную величину ξ в случайную величину $\eta = \lambda\xi$ с помощью функции

$$y = g(x, \lambda) = \lambda x,$$

где $\lambda > 0$ – параметр масштаба. Тогда

$$x = g^{-1}(y, \lambda) = y / \lambda,$$

и функция распределения имеет вид

$$G_\eta(x, \theta_1, \theta_2, \ldots, \theta_n, \lambda) = F_\xi(x / \lambda, \Theta).$$

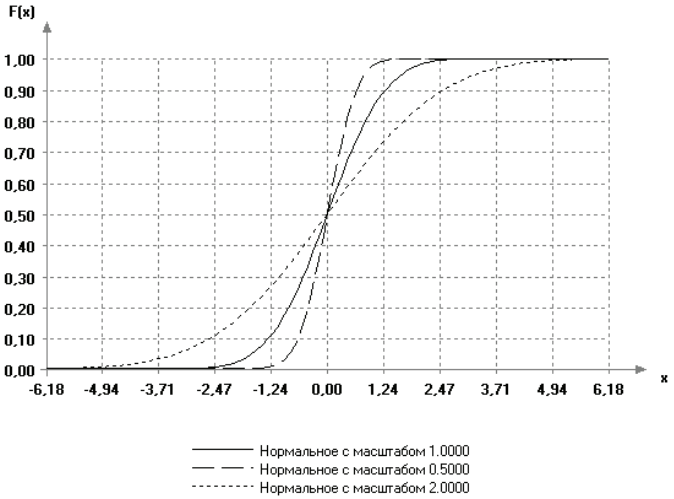


Рис. П2. Нормальное распределение с параметрами масштаба 1, 0,5 и 2

Функция плотности распределения имеет вид

$$g_\eta(x, \theta_1, \theta_2, \ldots, \theta_n, \lambda) = \frac{f(x / \lambda, \theta_1, \theta_2, \ldots, \theta_n)}{\lambda}.$$

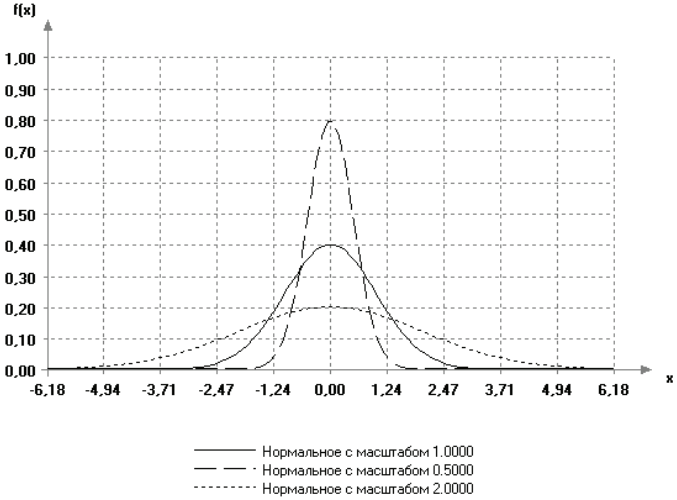


Рис. П3. Функции плотности нормального распределения с параметрами масштаба 1, 0,5 и 2

Если случайная величина ξ имеет область определения $[l, r]$, то случайная величина η будет иметь область определения $[\lambda l, \lambda r]$.

П3.3. Зеркальное отражение

Операция зеркального отражения преобразует случайную величину ξ в случайную величину $\eta = -\xi$ с помощью функции

$$y = g(x) = -x.$$

Тогда

$$x = g^{-1}(y) = -y,$$

и функция распределения имеет вид

$$G_\eta(x, \Theta) = 1 - F_\xi(-x, \Theta).$$

Функция плотности распределения имеет вид

$$g_\eta(x, \Theta) = f_\xi(-x, \Theta).$$

Если случайная величина ξ имеет область определения $[l, r]$, то случайная величина η будет иметь область определения $[-r, -l]$.

Замечание. Распределения, симметричные относительно нуля, при выполнении операции зеркального отражения не меняются. Распределения, симметричные относительно точки a, при выполнении операции зеркального отражения сдвигаются на величину $-2a$ (форма распределения при этом не меняется).

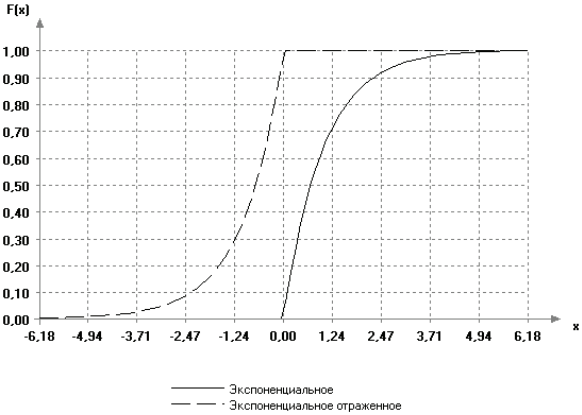


Рис. П4. Функции распределения экспоненциального и зеркального экспоненциального распределений

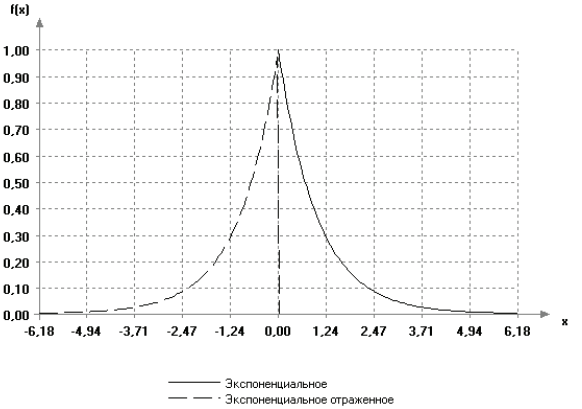


Рис. П5. Функции плотности экспоненциального и зеркального экспоненциального распределений

П3.4. Усечение слева

Операция усечения слева, преобразующая случайную величину ξ в случайную величину $\eta = (\xi \mid \xi > a)$, где a – параметр усечения, задается с помощью условной функции распределения:

$$G_\eta(x, \theta_1, \theta_2, \ldots, \theta_n, a) = F_{\xi|\xi>a}(x, \Theta) = P\{\xi < x \mid \xi > a\} =$$

$$= \frac{P\{\xi < x, \xi > a\}}{P\{\xi > a\}} = \frac{P\{a < \xi < x\}}{P\{\xi > a\}} = \begin{cases} 0, & x \le a; \\ \dfrac{F_\xi(x) - F_\xi(a)}{1 - F_\xi(a)}, & x > a. \end{cases}$$

Функция плотности имеет вид

$$g_\eta(x, \theta_1, \theta_2, \ldots, \theta_n, a) = \begin{cases} 0, & x \le a; \\ \dfrac{f_\xi(x)}{1 - F_\xi(a)}, & x > a. \end{cases}$$

Если случайная величина ξ имеет область определения $[l, r]$, то случайная величина η будет иметь область определения $[\max\{a, l\}, r]$.

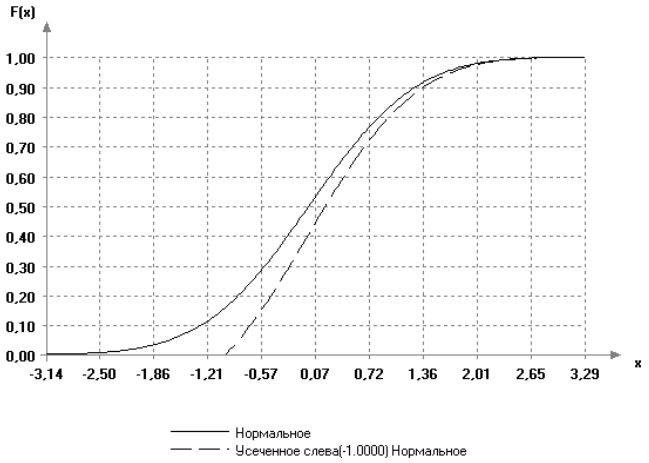


Рис. П6. Функции распределения нормального распределения и нормального, усеченного слева в точке *1*

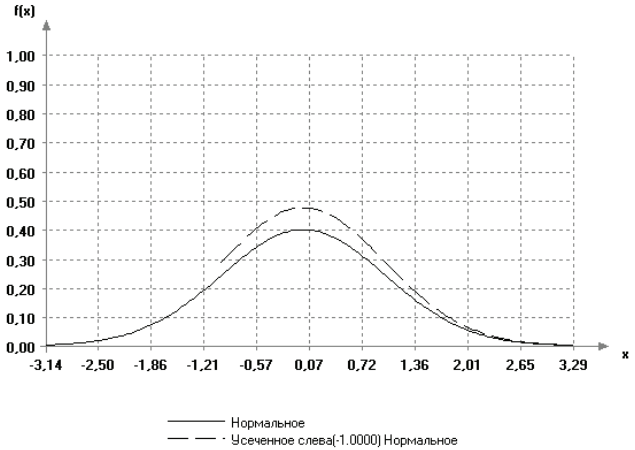


Рис. П7. Функции плотности нормального распределения и нормального, усеченного слева в точке *1*

П3.5. Усечение справа

Операция усечения справа, преобразующая случайную величину ξ в случайную величину $\eta = (\xi \mid \xi < b)$, где b – параметр усечения, задается с помощью условной функции распределения:

$$G_\eta(x, \theta_1, \theta_2, \ldots, \theta_n, b) = F_{\xi|\xi<b}(x, \Theta) = P\{\xi < x \mid \xi < b\} =$$

$$= \frac{P\{\xi < x, \xi < b\}}{P\{\xi < b\}} = \frac{P\{\xi < \min\{x, b\}\}}{P\{\xi < b\}} = \begin{cases} \dfrac{F_\xi(x)}{F_\xi(b)}, & x < b; \\ 1, & x \geq b. \end{cases}$$

Функция плотности распределения имеет вид

$$f_\eta(x, \theta_1, \theta_2, \ldots, \theta_n, b) = \begin{cases} \dfrac{f_\xi(x)}{F_\xi(b)}, & x < b; \\ 0, & x \geq b. \end{cases}$$

Если случайная величина ξ имеет область определения $[l, r]$, то случайная величина η будет иметь область определения $[l, \min\{b, r\}]$.

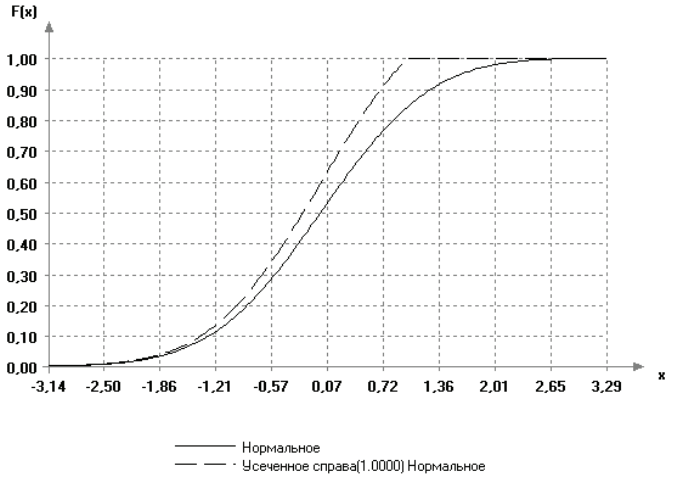


Рис. П8. Функции распределения нормального закона и нормального, усеченного справа в точке *1*

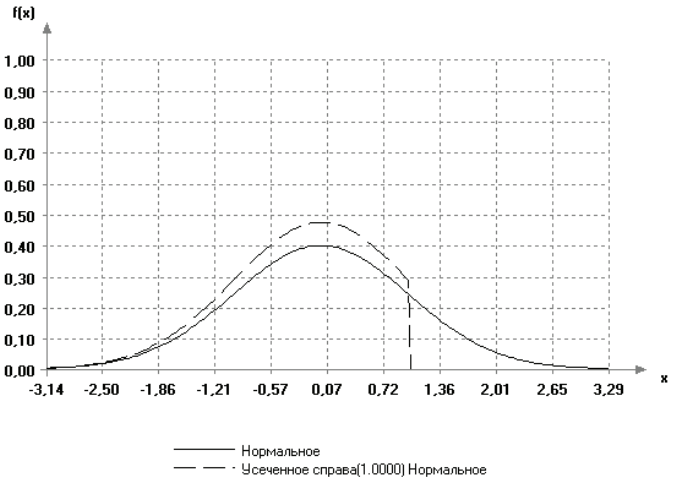


Рис. П9. Функции плотности распределения нормального закона и нормального, усеченного справа в точке *1*

П3.6. Двустороннее усечение

Операция двустороннего усечения, преобразующая случайную величину ξ в случайную величину $\eta = (\xi \mid a < \xi < b)$, где a и b – параметры усечения, задается с помощью условной функции распределения:

$$G_\eta(x, \theta_1, \theta_2, \ldots, \theta_n, a, b) = F_{\xi|a<\xi<b}(x, \Theta) = P\{\xi < x \mid a < \xi < b\} =$$

$$= \frac{P\{\xi < x, a < \xi < b\}}{P\{a < \xi < b\}} = \frac{P\{a < \xi < \min\{x, b\}\}}{P\{a < \xi < b\}} = \begin{cases} 0, & x \le a; \\ \dfrac{F_\xi(x) - F(a)}{F_\xi(b) - F_\xi(a)}, & a < x < b; \\ 1, & x \ge b. \end{cases}$$

Двустороннее усечение получается при последовательном применении операций усечения слева и усечения справа, причем порядок их применения не важен.

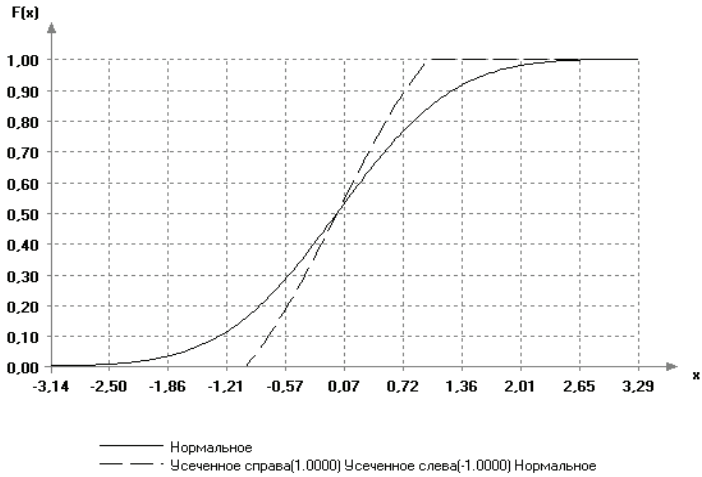


Рис. П10. Функции распределения нормального закона и нормального, усеченного слева в точке *–1* и справа в точке *1*

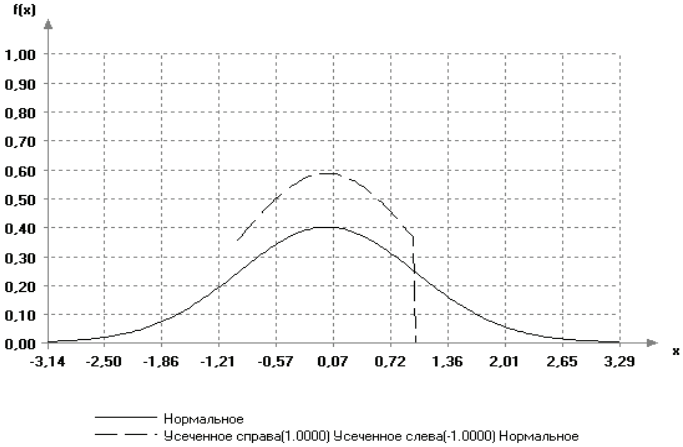


Рис. П11. Функции плотности распределения нормального закона и нормального, усеченного слева в точке *–1* и справа в точке *1*

П3.7. Логарифмирование

Операция логарифмирования преобразует случайную величину ξ в случайную величину $\eta = \exp^{\xi}$ с помощью функции

$$y = g(x) = \exp^{x},$$

тогда

$$x = g^{-1}(y) = \ln y,$$

и функция распределения имеет вид

$$G_{\eta}(x,\Theta) = \begin{cases} F_{\xi}(\ln x,\Theta), & x > 0; \\ 0, & x \le 0. \end{cases}$$

Функция плотности распределения имеет вид

$$g_{\eta}(x,\Theta) = \begin{cases} \dfrac{1}{x} f_{\xi}(\ln x,\Theta), & x > 0; \\ 0, & x \le 0. \end{cases}$$

Если случайная величина ξ имеет область определения $[l,\ r]$, то случайная величина η будет иметь область определения $[\exp e^{l},\ \exp e^{r}]$.

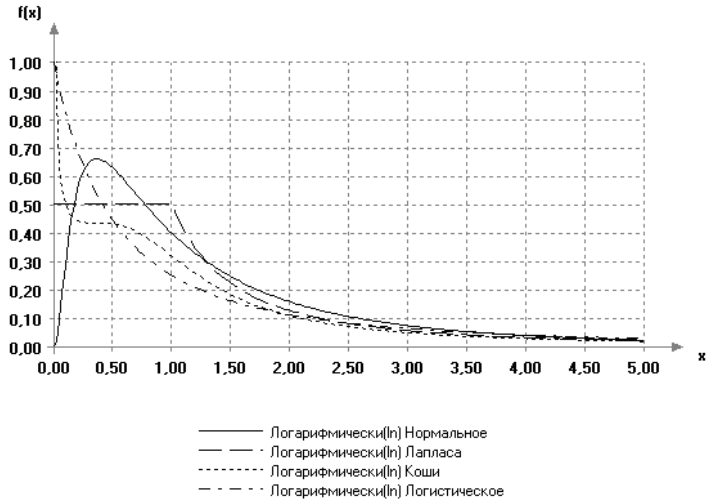


Рис. П12. Функции плотности распределений, полученных при выполнении операции логарифмирования

П3.8. Смесь

Операция смеси преобразует две независимые случайные величины ξ_1 и ξ_2 в случайную величину η, имеющую следующую функцию распределения:

$$G_\eta(x,\Theta_1,\Theta_2,\ \nu) = \nu F_{\xi_1}(x,\Theta_1) + (1-\nu)F_{\xi_2}(x,\Theta_2),$$

где $F_{\xi_1}(x,\Theta_1)$ и $F_{\xi_2}(x,\Theta_2)$ – функции распределения случайных величин ξ_1 и ξ_2 соответственно; ν – параметр смеси, принадлежащий интервалу $[a,b]$,

$$a = \max_{x\in A} \frac{f_{\xi_2}(x)}{f_{\xi_2}(x) - f_{\xi_1}(x)},\ b = \min_{x\in B} \frac{f_{\xi_2}(x)}{f_{\xi_2}(x) - f_{\xi_1}(x)},$$

и $A = \{x \in X : f_{\xi_2}(x) - f_{\xi_1}(x) < 0\}, B = \{x \in X : f_{\xi_2}(x) - f_{\xi_1}(x) > 0\}$.

Функция плотности имеет вид

$$g_\eta(x,\Theta_1,\Theta_2,\ \nu) = \nu f_{\xi_1}(x,\Theta_1) + (1-\nu) f_{\xi_2}(x,\Theta_2).$$

В самом деле, если смешать (объединить) две выборки, подчиненные двум различным законам распределения, то полученная выборка будет подчинена смеси этих законов с параметром смеси $\frac{n_1}{n_1+n_2}$, где n_1 и n_2 – количество наблюдений в первой и второй выборке соответственно.

Если случайная величина ξ_1 имеет область определения $[l_1, r_1]$ и ξ_2 имеет область определения $[l_2, r_2]$, то случайная величина η будет иметь область определения $[l_1, r_1] \cup [l_2, r_2]$.

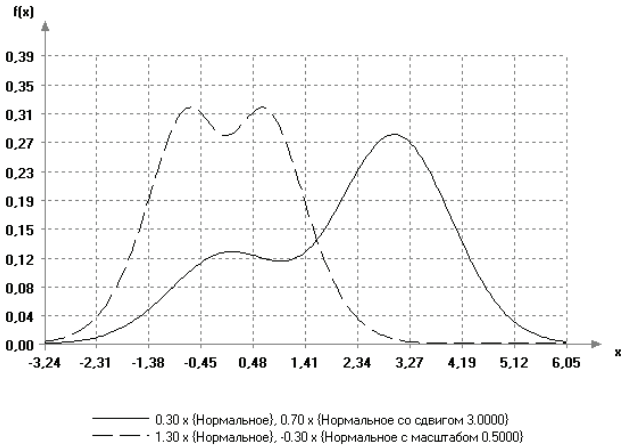


Рис. П13. Смесь двух нормальных распределений с параметром смеси $\nu \in [0,1]$ и $\nu > 1$

П3.9. Произведение

Операция *произведения* преобразует две независимые случайные величины ξ_1 и ξ_2 в случайную величину η, имеющую следующую функцию распределения:

$$G_\eta(x, \Theta_1, \Theta_2) = F_{\xi_1}(x, \Theta_1) F_{\xi_2}(x, \Theta_2),$$

где $F_{\xi_1}(x, \Theta_1)$ и $F_{\xi_2}(x, \Theta_2)$ – функции распределения случайных величин ξ_1 и ξ_2 соответственно.

Функция плотности имеет вид

$$g(x,\Theta_1,\Theta_2) = f_1(x,\Theta_1)F_2(x,\Theta_2) + F_1(x,\Theta_1)f_2(x,\Theta_2).$$

Если случайная величина ξ_1 имеет область определения $[l_1, r_1]$ и ξ_2 имеет область определения $[l_2, r_2]$, то случайная величина η будет иметь область определения $[l_1, r_1] \cap [l_2, r_2]$.

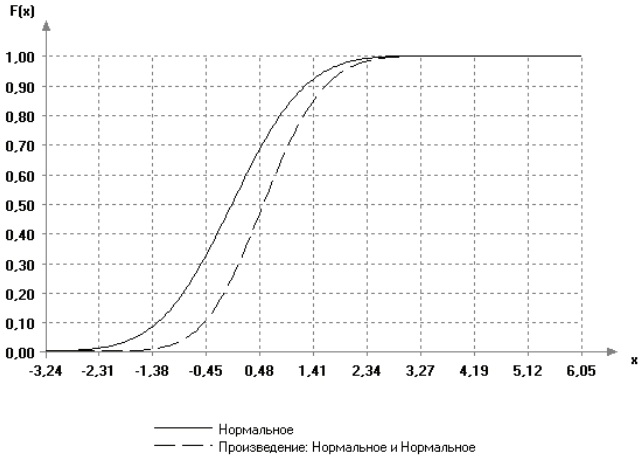


Рис. П14. Произведение двух нормальных распределений

П4. Семейства распределений случайных величин

Многие распределения, широко применяемые на практике, можно свести к нескольким семействам распределений. Семейства распределений можно получать двумя способами. *Групповые семейства* получаются применением операций над распределениями (см. П3), например, *семейство с параметром сдвига* получается при сдвиге распределения, *семейство с параметром масштаба* получается при масштабировании распределения и т. д.

Другой способ задать семейство распределений заключается в том, что функция распределения или функция плотности задаются в некоторой общей форме

$$F(x,\Theta) = H(g(x),\Theta),\ f(x,\Theta) = h(g(x),\Theta)g'(x),$$

где Θ – вектор параметров; $g(x)$ – генерирующая функция, а конкретное распределение из семейства получается при подстановке конкретной функции $g(x)$.

Рассматриваемые далее семейства распределений Джонсона, гамма-распределений и бета-распределений (за исключением бета-распределения III-го рода) принадлежат *экспоненциальному семейству*, функции плотности которого задаются следующим образом:

$$f(x,\Theta) = \exp^{\sum_{i=1}^{s} \eta_i(\Theta)T_i(x) - B(\Theta)} h(x),$$

где η_i и B – вещественнозначные функции от параметров; T_i – вещественнозначные статистики.

П4.1. Семейство распределений Джонсона

Распределения Джонсона задаются формулой

$$F(x) = \Phi(\alpha + \beta g(x)),$$

где $\Phi(x)$ – функция распределения нормального закона распределения; $|\alpha| < \infty$ и $\beta > 0$ – параметры; $g(x)$ – непрерывная, неограниченная, монотонно возрастающая функция, задающая конкретное распределение семейства.

Функция плотности имеет вид

$$f(x) = \frac{\beta}{\sqrt{2\pi}} \exp^{-\frac{1}{2}(\alpha+\beta g(x))^2} g'_x(x).$$

К семейству распределений Джонсона можно отнести следующие законы распределений:

- нормальное, $g(x) = x,\ -\infty < x < +\infty$;

- S_L -Джонсона (логнормальное), $g(x) = \ln x, x > 0$;
- S_B -Джонсона, $g(x) = \ln \frac{x}{1-x}, 0 < x < 1$;
- S_U -Джонсона, $g(x) = \ln(x + \sqrt{x^2 + 1}), -\infty < x < +\infty$.

П4.2. Семейство гамма-распределений

Семейство гамма-распределений задается формулой

$$F(x) = \frac{\gamma(g(x,\alpha),\delta)}{\Gamma(\delta)},$$

где $\Gamma(\delta) = \int_0^\infty y^{\delta-1} \exp^{-y} dy$ и $\gamma(x,\delta) = \int_0^x y^{\delta-1} \exp^{-y} dy$ – полная и неполная гамма-функции, α и $\delta > 0$ – параметры; $g(x,\alpha)$ – непрерывная, монотонно возрастающая от 0 до ∞ функция, задающая конкретное распределение семейства.

К семейству гамма-распределений можно отнести следующие распределения:

- гамма-распределение, $g(x) = x, 0 \le x < +\infty$;
- Г-распределение, $g(x,\alpha) = x^\alpha$, $0 \le x < +\infty$, $\alpha > 0$ (частными случаями этого распределения являются гамма-распределение при $\alpha = 1$ и распределение Вейбулла при $\delta = 1$);
- χ -распределение (распределение модуля n-мерной нормальной случайной величины, $n = 2\delta$);
- $g(x) = \frac{1}{2}x^2, 0 \le x < +\infty$;
- обобщенное распределение минимального значения, $g(x) = \exp^x$, $g(x) = \exp^x, -\infty < x < +\infty$.

Функция плотности распределения имеет вид

$$f(x) = \frac{1}{\Gamma(\delta)} (g(x))^{\delta-1} \exp^{-g(x)} g'_x(x).$$

П4.3. Семейство бета-распределений

Семейство бета-распределений задается формулой

$$F(x) = \frac{B(g(x,\delta),\alpha,\beta)}{B(\alpha,\beta)},$$

где

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy$$

и

$$B(x,\alpha,\beta) = \int_0^x y^{\alpha-1}(1-y)^{\beta-1}\,dy$$

– полная и неполная бета-функции; α, β и δ – параметры; $g(x,\delta)$ – непрерывная, монотонно возрастающая от нуля до единицы функция, задающая конкретное распределение семейства.

К семейству бета-распределений можно отнести следующие распределения:

- бета-распределение I-го рода, $g(x) = x, 0 \le x \le 1$;
- бета-распределение II-го рода,

$$g(x) = \frac{x}{1+x}, 0 \le x < +\infty;$$

- бета-распределение III-го рода,

$$g(x,\delta) = \frac{\delta x}{1+(\delta-1)x}, 0 \le x < 1$$

(частным случаем является бета-распределение I-го рода при $\delta = 1$);

- распределение Парето, $g(x) = \frac{x-1}{x}, 1 \le x < \infty, \alpha = 1;$

- L-распределение (обобщенное логистическое),

$$g(x)=\frac{\exp^x}{1+\exp^x},\quad -\infty<x<+\infty.$$

Функция плотности распределения имеет вид

$$f(x)\ =\ \frac{1}{B(\alpha,\beta)}(g(x,\delta))^{\alpha-1}(1-g(x,\delta))^{\beta-1}g'_x(x,\delta).$$

П5. Стандартные законы распределений

Под *стандартными* распределениями мы будем понимать распределения без параметров сдвига и масштаба (т. е. параметр сдвига равен нулю, а параметр масштаба – единице).

П5.1. Равномерное распределение

Случайная величина, распределенная по равномерному закону, имеет область определения $[0,1]$.

Функция распределения:

$$F(x)=\begin{cases}0, & x<0;\\ x, & 0\le x\le 1;\\ 1, & x>0.\end{cases}$$

Функция плотности распределения:

$$f(x)=\begin{cases}0, & x<0\vee x>1;\\ 1, & 0\le x\le 1.\end{cases}$$

Равномерное распределение является частным случаем бета-распределения I-го рода при $\alpha=\beta=1$ и предельным для двустороннего экспоненциального при $\alpha\to+\infty$.

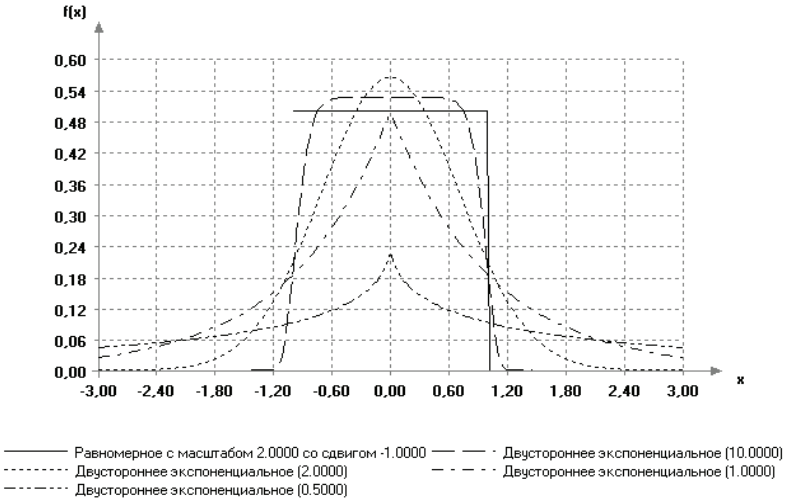


Рис. П15. Функции плотности равномерного и двустороннего экспоненциального распределений

П5.2. Экспоненциальное распределение

Случайная величина, распределенная по экспоненциальному закону, имеет область определения $[0,+\infty)$.

Функция распределения:

$$F(x)=\begin{cases} 0, & x<0; \\ 1-\exp^{-x}, & x\geq 0. \end{cases}$$

Функция плотности распределения:

$$f(x)=\begin{cases} 0, & x<0; \\ \exp^{-x}, & x\geq 0. \end{cases}$$

Экспоненциальное распределение является частным случаем распределения Вейбулла при $\alpha=1$.

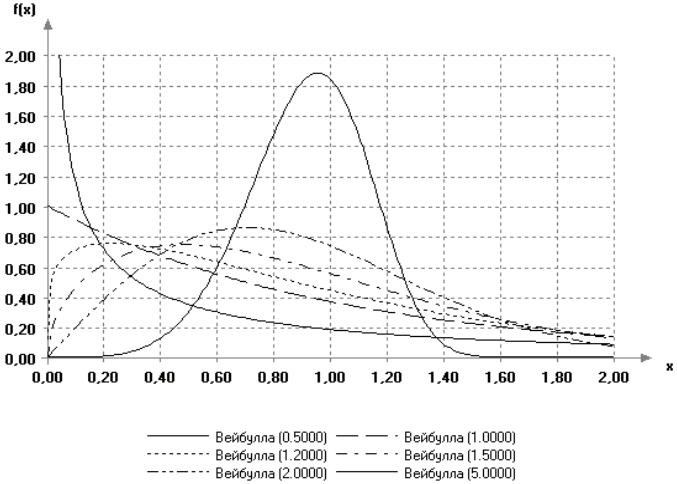


Рис. П16. Функции плотности распределения Вейбулла с параметром формы $\alpha = 1,\ 1,2,\ 1,5,\ 2,\ 5$

П5.3. Полунормальное распределение

Случайная величина, распределенная по полунормальному закону, имеет область определения $[0, +\infty)$.

Функция распределения:

$$F(x) = \begin{cases} 0, & x < 0; \\ 2\Phi_0(x) = \dfrac{1}{\Gamma(1/2)}\gamma\left(\dfrac{x^2}{2}, \dfrac{1}{2}\right), & x \geq 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x < 0; \\ \dfrac{1}{\sqrt{\pi/2}} \exp^{-\frac{x^2}{2}}, & x \geq 0. \end{cases}$$

Полунормальное распределение является частным случаем распределения модуля многомерного нормального вектора при $n = 1$ (см. П5.6), а также усеченным слева в точке 0 нормальным распределением.

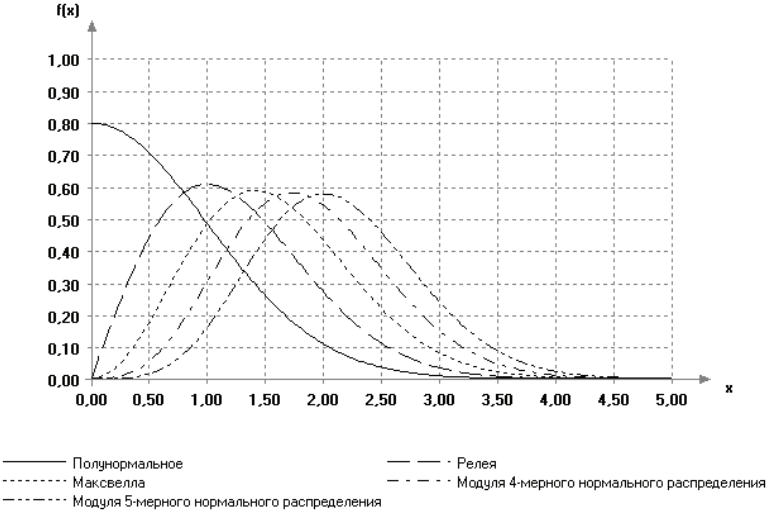


Рис. П17. Функции плотности модуля n-мерного нормального распределения с параметром $n = 1,\ 2,\ 3,\ 4,\ 5$

П5.4. Распределение Рэлея

Случайная величина, распределенная по закону Рэлея, имеет область определения $[0, +\infty)$.

Функция распределения:

$$F(x) = \begin{cases} 0, & x < 0; \\ 1 - \exp^{-\frac{x^2}{2}}, & x \geq 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x < 0; \\ x \exp^{-\frac{x^2}{2}}, & x \geq 0. \end{cases}$$

Распределение Рэлея является частным случаем распределения модуля многомерного нормального вектора при $n = 2$.

П5.5. Распределение Максвелла

Случайная величина, распределенная по закону Максвелла, имеет область определения $[0,+\infty)$.

Функция распределения:

$$F(x)=\begin{cases}0, & x<0;\\ \dfrac{1}{\sqrt{\pi}/2}\gamma\left(\dfrac{x^2}{2},\dfrac{3}{2}\right), & x\geq 0.\end{cases}$$

Функция плотности распределения:

$$f(x)=\begin{cases}0, & x<0;\\ \dfrac{2x^2}{\sqrt{2\pi}}\exp^{-\frac{x^2}{2}}, & x\geq 0.\end{cases}$$

Распределение Максвелла является частным случаем распределения модуля многомерного нормального вектора при $n=3$.

П5.6. Распределение модуля многомерного нормального вектора

Случайная величина, распределенная по закону модуля многомерного нормального вектора, имеет область определения $[0,+\infty)$.

Функция распределения:

$$F(x)=\begin{cases}0, & x<0;\\ \dfrac{1}{\Gamma(n/2)}\gamma\left(\dfrac{x^2}{2},\dfrac{n}{2}\right), & x\geq 0.\end{cases}$$

Функция плотности распределения:

$$f(x)=\begin{cases}0, & x<0;\\ \dfrac{2x^{n-1}}{2^{n/2}\Gamma(n/2)}\exp^{-\frac{x^2}{2}}, & x\geq 0.\end{cases}$$

Распределение модуля n-мерного нормального вектора относится к семейству гамма-распределений (см. П4.2) с генерирующей функцией $g(x)=\dfrac{x^2}{2}$ и целым параметром $n=2\delta$.

П5.7. Распределение Парето

Случайная величина, распределенная по закону Парето, имеет область определения $[1,+\infty)$.

Функция распределения:

$$F(x)=\begin{cases}0, & x<1;\\ 1-\dfrac{1}{x^{\beta}}, & x\geq 1,\end{cases}$$

где $\beta>0$ – параметр.

Функция плотности распределения:

$$f(x)=\begin{cases}0, & x<1;\\ \dfrac{\beta}{x^{\beta+1}}, & x\geq 1.\end{cases}$$

Распределение Парето относится к семейству бета-распределений с генерирующей функцией $g(x)=\dfrac{x-1}{x}$, $x>1$, и параметром $\alpha=1$.

П5.8. Распределение Эрланга

Случайная величина, распределенная по закону Эрланга, имеет область определения $[0,+\infty)$. Распределение Эрланга (или показательно-степенное распределение) является частным случаем гамма-распределения с целым параметром $n=\alpha$.

П5.9. Распределение Лапласа

Случайная величина, распределенная по закону Лапласа, имеет область определения $(-\infty, +\infty)$.

Функция распределения:

$$F(x) = \frac{1}{2} + \frac{\operatorname{sign}(x)}{2} \exp^{-|x|}.$$

Функция плотности распределения:

$$f(x) = \frac{1}{2} \exp^{-|x|}.$$

Распределение Лапласа является частным случаем двустороннего экспоненциального распределения с параметром $\alpha = 1$ (см. П5.1).

П5.10. Нормальное распределение

Случайная величина, распределенная по нормальному закону, имеет область определения $(-\infty, +\infty)$.

Функция распределения:

$$F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp^{-\frac{x^2}{2}} dx = \frac{1}{2} + \frac{\operatorname{sign}(x)}{2\sqrt{\pi}} \gamma\left(\frac{x^2}{2}, \frac{1}{2}\right).$$

Функция плотности распределения имеет вид

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{x^2}{2}}.$$

Нормальное распределение является частным случаем двустороннего экспоненциального распределения с параметром $\alpha = 2$ и масштабом $\sqrt{2}$.

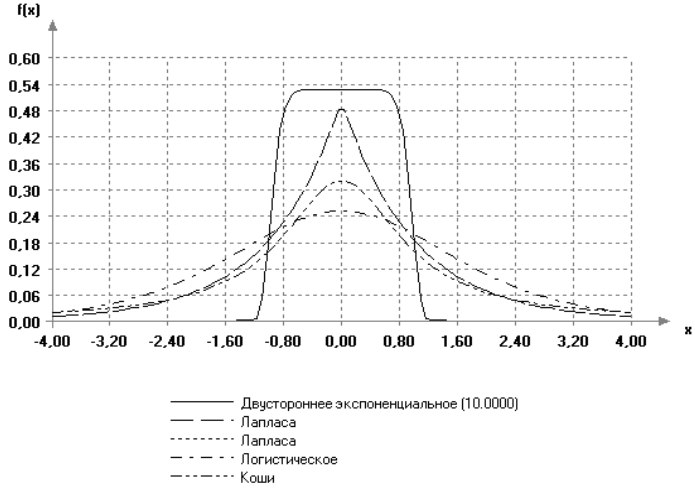


Рис. П18. Множество симметричных распределений: двустороннее экспоненциальное с параметром $\alpha = 10$, Лапласа, нормальное, логистическое, Коши

П5.11. Логарифмически (ln) нормальное распределение

Случайная величина, распределенная по логарифмически (ln) нормальному закону, имеет область определения $(0, +\infty)$.

Функция распределения:

$$F(x) = \begin{cases} 0, & x \le 0; \\ \Phi(\ln x) = \dfrac{1}{2} + \dfrac{\operatorname{sign}(\ln x)}{2\sqrt{\pi}} \gamma\left(\dfrac{(\ln x)^2}{2}, \dfrac{1}{2} \right), & x > 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \le 0; \\ \dfrac{1}{x\sqrt{2\pi}} \exp^{-\frac{(\ln x)^2}{2}} & x > 0. \end{cases}$$

Логарифмически нормальное распределение получается из нормального применением операции логарифмирования и является частным случаем распределения S_L -Джонсона при $\alpha = \beta = 1$.

П5.12. Логарифмически (lg) нормальное распределение

Случайная величина, распределенная по логарифмически (lg) нормальному закону, имеет область определения $(0,+\infty)$.

Функция распределения:

$$F(x)=\begin{cases} 0, & x\le 0; \\ \Phi(\lg x)=\dfrac{1}{2}+\dfrac{\operatorname{sign}(\lg x)}{2\sqrt{\pi}}\gamma\left(\dfrac{(\lg x)^2}{2},\dfrac{1}{2}\right), & x>0. \end{cases}$$

Функция плотности распределения:

$$f(x)=\begin{cases} 0, & x\le 0; \\ \dfrac{1}{x\ln 10\sqrt{2\pi}}\exp^{-\frac{(\lg x)^2}{2}} & x>0. \end{cases}$$

Логарифмически (lg) нормальное распределение получается из логарифмически (ln) нормального применением операции масштабирования с параметром $\lambda=\ln 10$.

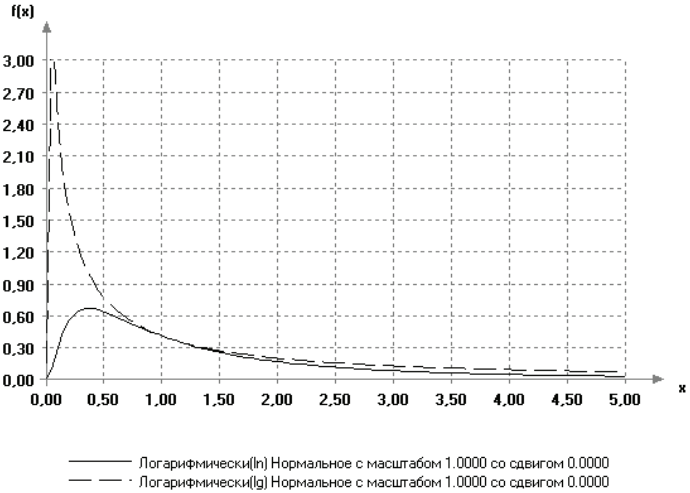


Рис. П19. Функции распределения логарифмически (ln) и логарифмически (lg) нормальных распределений

П5.13. Распределение Коши

Случайная величина, распределенная по закону Коши (Брейта – Вигнера, арктангенса), имеет область определения $(-\infty, +\infty)$.

Функция распределения:

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arctg} x.$$

Функция плотности распределения:

$$f(x) = \frac{1}{\pi(1 + x^2)}.$$

П5.14. Логистическое распределение

Случайная величина, распределенная по логистическому закону, имеет область определения $(-\infty, +\infty)$.

Функция распределения:

$$F(x) = \frac{1}{1 + \exp^{-x}}.$$

Функция плотности распределения:

$$f(x) = \frac{\exp^{-x}}{(1 + \exp^{-x})^2} = \frac{\exp^{x}}{(1 + \exp^{x})^2}.$$

Логистическое распределение является частным случаем L-распределения при $\alpha = \beta = 1$.

П5.15. Распределение Вейбулла

Случайная величина, распределенная по закону Вейбулла (Вейбулла – Гнеденко, экстремальных значений III-го типа), имеет область определения $(0, +\infty)$.

Функция распределения:

$$F(x) = \begin{cases} 0, & x \le 0; \\ 1 - \exp^{-x^{\alpha}}, & x > 0. \end{cases}$$

Функция плотности распределения:

$$f(x)=\begin{cases} 0, & x\le 0; \\ \alpha x^{\alpha-1}\exp^{-x^{\alpha}} & x>0. \end{cases}$$

Распределение Вейбулла является частным случаем Г-распределения при $\delta=1$.

П5.16. Распределение минимального значения

Случайная величина, распределенная по закону минимального значения, имеет область определения $(-\infty,+\infty)$.

Функция распределения:

$$F(x)=1-\exp^{-\exp^{x}}.$$

Функция плотности распределения:

$$f(x)=\exp^{x-\exp^{x}}.$$

Распределение минимального значения является частным случаем обобщенного распределения минимального значения при $\delta=1$.

П5.17. Распределение максимального значения

Случайная величина, распределенная по закону максимального значения (Гумбеля, экстремальных значений I-го типа, Фишера – Типпета, двойной показательный), имеет область определения $(-\infty,+\infty)$.

Функция распределения:

$$F(x)=\exp^{-\exp^{-x}}.$$

Функция плотности распределения:

$$f(x)=\exp^{-x-\exp^{-x}}.$$

Распределение максимального значения получается из распределения минимального значения применением операции *отражения*.

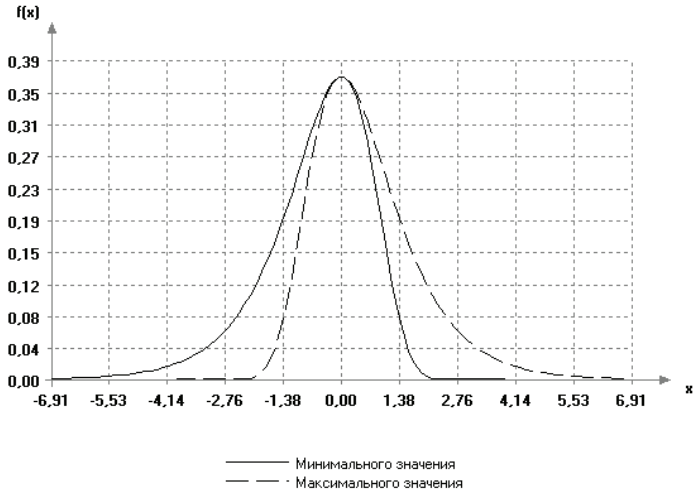


Рис. П20. Функции плотности распределений минимального значения, максимального значения

П5.18. Обобщенное распределение минимального значения

Случайная величина, распределенная по закону обобщенного минимального значения, имеет область определения $(-\infty, +\infty)$.

Функция распределения:

$$F(x) = \frac{1}{\Gamma(\delta)} \gamma(\exp^x, \delta),$$

где параметр $\delta > 0$.

Функция плотности распределения:

$$f(x) = \frac{1}{\Gamma(\delta)} \exp^{\delta x - \exp^x}.$$

Обобщенное распределение минимального значения принадлежит к семейству гамма-распределений с генерирующей функцией $g(x) = \exp^x$.

П5.19. Распределение Накагами

Случайная величина, распределенная по закону Накагами, имеет область определения $(0,+\infty)$.

Функция распределения:

$$F(x)=\begin{cases}0, & x\le 0;\\ \dfrac{1}{\Gamma(\alpha)}\gamma(\alpha x^2,\alpha), & x>0,\end{cases}$$

где параметр $\alpha \ge 1/2$.

Функция плотности распределения:

$$f(x)=\begin{cases}0, & x\le 0;\\ \dfrac{2\alpha^{\alpha}x^{2\alpha-1}}{\Gamma(\alpha)}\exp^{-\alpha x^2}, & x>0.\end{cases}$$

Распределение Накагами также принадлежит к семейству гамма-распределений, но генерирующая функция $g(x)=\alpha x^2$ зависит от основного параметра.

П5.20. Гамма-распределение

Случайная величина, имеющая гамма-распределение (распределение Пирсона III), определена на области $(0,+\infty)$.

Функция распределения:

$$F(x)=\begin{cases}0, & x\le 0;\\ \dfrac{1}{\Gamma(\delta)}\gamma(x,\delta), & x>0,\end{cases}$$

где параметр $\delta>0$.

Функция плотности распределения:

$$f(x)=\begin{cases}0, & x\le 0;\\ \dfrac{1}{\Gamma(\delta)}x^{\delta-1}\exp^{-x}, & x>0.\end{cases}$$

Гамма-распределение является частным случаем Г-распределения при $\alpha=1$.

П5.21. Бета-распределение I-го рода

Случайная величина, имеющая бета-распределение I-го рода, определена на области $[0,1]$.

Функция распределения:

$$F(x)=\begin{cases}0, & x<0;\\ \dfrac{1}{\mathrm{B}(\alpha,\beta)}B(x,\alpha,\beta), & 0\le x\le 1;\\ 1, & x>1,\end{cases}$$

где параметры $\alpha>0$ и $\beta>0$.

Функция плотности распределения:

$$f(x)=\begin{cases}0, & x\le 0\vee x\ge 1;\\ \dfrac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, & 0<x<1.\end{cases}$$

Бета-распределение I-го рода принадлежит к семейству бета-распределений с генерирующей функцией $g(x)=x$.

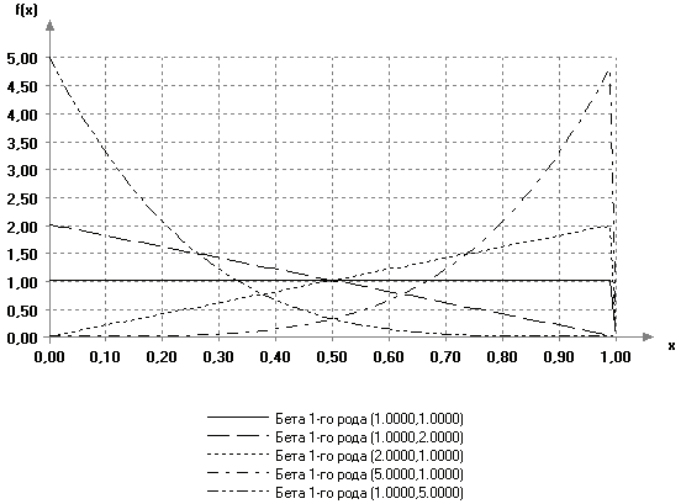


Рис. П21. Функции плотности бета-распределения I-го рода

П5.22. Бета-распределение II-го рода

Случайная величина, имеющая бета-распределение II-го рода, определена на области $[0,+\infty)$.

Функция распределения:

$$F(x)=\begin{cases}0, & x<0;\\ \dfrac{1}{\mathrm{B}(\alpha,\beta)}B\left(\dfrac{x}{1+x},\alpha,\beta\right), & x\le 0,\end{cases}$$

где параметры $\alpha>0$ и $\beta>0$.

Функция плотности распределения:

$$f(x)=\begin{cases}0, & x\le 0;\\ \dfrac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1+x)^{-\alpha-\beta}, & x>0.\end{cases}$$

Бета-распределение II-го рода принадлежит к семейству бета-распределений с генерирующей функцией $g(x)=\dfrac{x}{1+x}$.

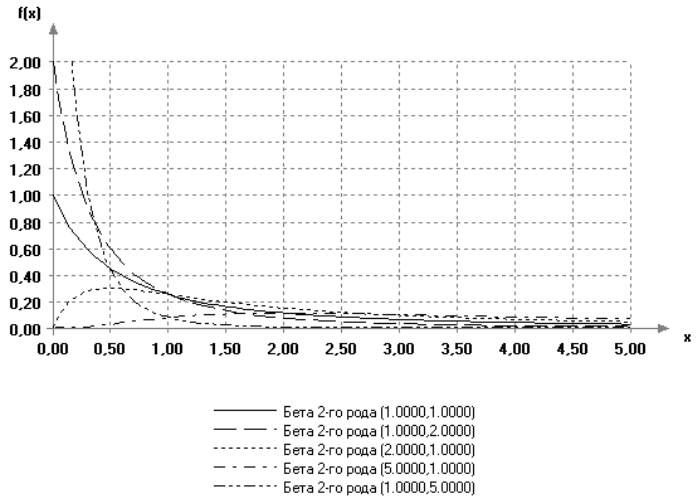


Рис. П22. Функции плотности бета-распределения II-го рода

П5.23. Бета-распределение III-го рода

Случайная величина, имеющая бета-распределение III-го рода, определена на области $[0,1]$.

Функция распределения:

$$F(x)=\begin{cases}0, & x<0;\\ \dfrac{1}{\mathrm{B}(\alpha,\beta)}B\left(\dfrac{\delta x}{1+(\delta-1)x},\alpha,\beta\right), & 0\le x\le 1;\\ 1, & x>1,\end{cases}$$

где параметры $\alpha>0$, $\beta>0$ и $\delta>0$.

Функция плотности распределения:

$$f(x)=\begin{cases}0, & x\le 0\vee x\ge 1;\\ \dfrac{\delta^{\alpha}}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}(1+(\delta-1)x)^{-\alpha-\beta}, & 0<x<1.\end{cases}$$

Бета-распределение III-го рода принадлежит к семейству бета-распределений с генерирующей функцией

$$g(x,\delta)=\frac{\delta x}{1+(\delta-1)x}.$$

П5.24. Распределение S_B-Джонсона

Случайная величина, распределенная по закону S_B-Джонсона, определена на области $(0,1)$.

Функция распределения:

$$F(x)=\begin{cases}0, & x\le 0;\\ \Phi\left(\alpha+\beta\ln\dfrac{x}{1-x}\right), & 0<x<1;\\ 1, & x>1,\end{cases}$$

где параметры $\alpha\in\mathbb{R}$, $\beta>0$.

Функция плотности распределения:

$$f(x)=\begin{cases} 0, & x\le 0\vee x\ge 1;\\ \dfrac{\beta}{x(1-x)\sqrt{2\pi}}\exp^{-\frac{1}{2}(\alpha+\beta\ln\frac{x}{1-x})^2}, & 0<x<1. \end{cases}$$

Распределение S_B-Джонсона принадлежит к семейству распределений Джонсона с генерирующей функцией

$$g(x)=\ln\frac{x}{1-x}.$$

П5.25. Распределение S_L-Джонсона

Случайная величина, распределенная по закону S_L-Джонсона, определена на области $(0,+\infty)$.

Функция распределения:

$$F(x)=\begin{cases} 0, & x\le 0;\\ \Phi(\alpha+\beta\ln x), & x>0, \end{cases}$$

где параметры $\alpha\in\mathbb{R}$, $\beta>0$.

Функция плотности распределения:

$$f(x)=\begin{cases} 0, & x\le 0;\\ \dfrac{\beta}{x\sqrt{2\pi}}\exp^{-\frac{1}{2}(\alpha+\beta\ln x)^2}, & x>0. \end{cases}$$

Распределение S_L-Джонсона принадлежит к семейству распределений Джонсона с генерирующей функцией $g(x)=\ln x$.

П5.26. Распределение S_U -Джонсона

Случайная величина, распределенная по закону S_U -Джонсона, определена на области $(-\infty, +\infty)$.

Функция распределения:

$$F(x) = \Phi(\alpha + \beta \ln(x + \sqrt{x^2 + 1})),$$

где параметры $|\alpha| < \infty$ и $\beta > 0$.

Функция плотности распределения:

$$f(x) = \frac{\beta}{\sqrt{2\pi}\sqrt{x^2 + 1}} \exp^{-\frac{1}{2}(\alpha + \beta \ln(x + \sqrt{x^2 + 1}))^2}.$$

Распределение S_U -Джонсона принадлежит к семейству распределений Джонсона с генерирующей функцией

$$g(x) = \ln(x + \sqrt{x^2 + 1}).$$

П5.27. Двустороннее экспоненциальное распределение

Случайная величина, распределенная по двустороннему экспоненциальному закону (класс экспоненциальных распределений), определена на области $(-\infty, +\infty)$.

Функция распределения:

$$F(x) = \frac{1}{2}\left[1 + \frac{\text{sign}(x)}{\Gamma(1/\alpha)} \gamma(|x|^{\alpha}, 1/\alpha)\right],$$

где параметр $\alpha > 0$.

Функция плотности распределения:

$$f(x) = \frac{\alpha}{2\Gamma(1/\alpha)} \exp^{-|x|^{\alpha}}.$$

Двустороннее экспоненциальное распределение является частным случаем *H*-распределения при $\delta = 1/\alpha$.

П5.28. *H*-распределение

Случайная величина, имеющая *H*-распределение, определена на области $(-\infty, +\infty)$. Функция распределения:

$$F(x) = \frac{1}{2}\left[1 + \frac{\text{sign}(x)}{\Gamma(\delta)}\gamma(|x|^{\alpha}, \delta)\right],$$

где параметры $\alpha > 0$ и $\delta > 0$.

Функция плотности распределения:

$$f(x) = \frac{\alpha}{\Gamma(\delta)} |x|^{\alpha\delta - 1} \exp^{-|x|^{\alpha}}.$$

H-распределение в общем случае (при $\delta \neq 1/\alpha$) является двухмодальным и симметричным относительно нуля.

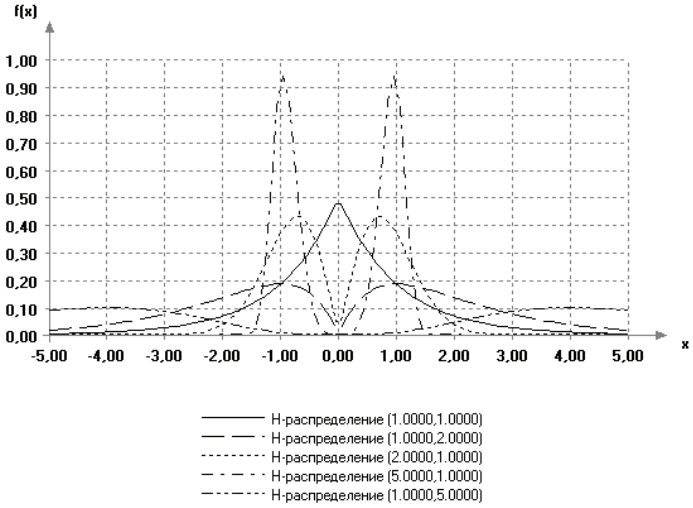


Рис. П23. Функции плотности *H*-распределения

П5.29. Г-распределение

Случайная величина, имеющая Г-распределение, определена на области $(0,+\infty)$.

Функция распределения:

$$F(x)=\begin{cases}0, & x\le 0;\\ \dfrac{1}{\Gamma(\delta)}\gamma(x^{\alpha},\delta), & x>0,\end{cases}$$

где параметры $\alpha>0$ и $\delta>0$.

Функция плотности распределения:

$$f(x)=\begin{cases}0, & x\le 0;\\ \dfrac{\alpha}{\Gamma(\delta)}x^{\alpha\delta-1}\exp^{-x^{\alpha}}, & x>0.\end{cases}$$

Г-распределение принадлежит к семейству гамма-распределений с генерирующей функцией $g(x,\alpha)=x^{\alpha}$.

П5.30. Обобщенное логистическое распределение

Случайная величина, имеющая обобщенное логистическое распределение, определена на области $(-\infty,+\infty)$.

Функция распределения:

$$F(x)=\frac{1}{\mathrm{B}(\alpha,\beta)}B\left(\frac{\exp^{x}}{1+\exp^{x}},\alpha,\beta\right),$$

где параметры $\alpha>0$ и $\beta>0$.

Функция плотности распределения:

$$f(x)=\frac{1}{\mathrm{B}(\alpha,\beta)}\frac{\exp^{\alpha x}}{(1+\exp^{x})^{\alpha+\beta}}.$$

Обобщенное логистическое распределение принадлежит к семейству бета-распределений с генерирующей функцией

$$g(x) = \frac{\exp^x}{1+\exp^x}.$$

П6. Распределение некоторых функций от нормальных случайных величин

Нормальное распределение часто встречается в различных задачах теории вероятностей и математической статистики. Путем различных преобразований совокупности нормальных случайных величин получаются другие распределения: Хи-квадрат, Стьюдента, Фишера.

П6.1. Распределение Хи-квадрат

Распределение Хи-квадрат часто встречается в задачах проверки статистических гипотез.

П6.1.1. Центральное распределение Хи-квадрат

Пусть случайная величина $\xi \in N(0,1)$ имеет функцию распределения $\Phi(x)$. Найдем функцию распределения случайной величины ξ^2.

$$F_{\xi^2}(x) = P\{\xi^2 \le x\} = P\{-\sqrt{x} \le \xi \le \sqrt{x}\} =$$
$$= \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) = \Phi(\sqrt{x}) - (1 - \Phi(\sqrt{x})) =$$
$$2\Phi(\sqrt{x}) - 1, x \ge 0.$$

Тогда плотность распределения мы можем найти путем дифференцирования функции распределения:

$$f_{\xi^2}(x) = \frac{d}{dx} F_{\xi^2} = 2\Phi'(\sqrt{x}) \frac{1}{2\sqrt{x}} = \frac{1}{\sqrt{2\pi}} e^{\frac{-(\sqrt{x})^2}{2}} x^{-1/2} =$$

$$= \frac{2^{-1/2} x^{1/2-1}}{\Gamma(1/2)} e^{-x/2}, \quad x \ge 0.$$

Таким образом, квадрат стандартной нормальной величины подчинен гамма-распределению с параметром формы $1/2$ и параметром масштаба 2: $\xi^2 \in \Gamma(1/2, 2)$.

Рассмотрим теперь сумму из n квадратов стандартных нормальных величин. Пусть случайная величина $\xi_i \in N(0,1)$, $i = \overline{1,n}$. Найдем распределение $\chi_n^2 = \xi_1^2 + \xi_2^2 + ... + \xi_n^2$.

Так как $\xi_i^2 \in \Gamma(1/2, 2)$, а гамма-распределение воспроизводимо по параметру, то

$$\chi_n^2 = \xi_1^2 + \xi_2^2 + ... + \xi_n^2 \in \Gamma(n/2, 2).$$

Функция плотности $f_{\chi_n^2}(x) = \dfrac{2^{-n/2} x^{n/2-1}}{\Gamma(n/2)} e^{-x/2}$, $x \geq 0$.

Это распределение называется *Хи-квадрат распределением* с n степенями свободы и обозначается как $\chi^2(n)$.

П6.1.2. Нецентральное распределение Хи-квадрат

Пусть случайные величины имеют сдвиг: $\xi_i \in N(\mu_i, 1)$, $i = \overline{1,n}$. Тогда их сумма квадратов будет подчинена нецентральному Хи-квадрат распределению:

$$\xi_1^2 + \xi_2^2 + ... + \xi_n^2 \in \chi^2(n, \lambda^2).$$

Параметр λ^2 называется *параметром нецентральности* и равен $\lambda^2 = \mu_1^2 + ... + \mu_n^2$.

Функция плотности нецентрального Хи-квадрат распределения:

$$f_{\chi^2(n,\lambda^2)}(x) = 2^{-n/2} x^{n/2-1} e^{-\frac{x+\lambda^2}{2}} \sum_{j=0}^{\infty} \frac{(x\lambda^2)^j}{j! 2^{2j} \Gamma(n/2 + j)}, x > 0.$$

Если $\lambda^2 = 0$, то $\chi^2(n, \lambda^2) = \chi^2(n)$, т. е. нецентральное Хи-квадрат распределение совпадает с центральным.

П6.2. Распределение Стьюдента

Определение. *Распределением Стьюдента* называется распределение случайной величины $t_n = \xi \Big/ \sqrt{\chi_n^2 / n}$, где случайные величины ξ и χ_n^2 независимы и $\xi \in N(0,1), \chi^2 \in \chi^2(n)$.

Найдем функцию распределения случайной величины t_n. Пусть $\eta = \sqrt{\chi^2 / n}$. Тогда

$$F_{\xi/\eta}(z) = P_{\xi,\eta}\{\xi / \eta \le z\} = \iint\limits_{(x,y:x/y<z)} f_\xi(x) f_\eta(y) dxdy =$$

$$= \int\limits_{-\infty}^{0} (1 - F_\xi(zy)) f_\eta(y) dy + \int\limits_{0}^{+\infty} F_\xi(zy) f_\eta(y) dy.$$

Тогда функция распределения примет вид

$$f_{\xi/\eta}(z) = \frac{d}{dz} F_{\xi/\eta}(z) = \int\limits_{-\infty}^{\infty} f_\xi(zy) f_\eta(y) |y| dy.$$

Так как η является исключительно положительной величиной, то $F_\eta(y) = 0$ при $y < 0$. Рассмотрим случай, когда $y \ge 0$.

$$F_\eta(y) = P\{\sqrt{\chi^2 / n} < y\} = P\{\chi^2 < y^2 n\} = F_{\chi^2}(y^2 n).$$

$$f_\eta(y) = \frac{d}{dy} F_\eta(y) = \frac{d}{dy} F_{\chi^2}(y^2 n) = 2yn f_{\chi^2}(y^2 n) =$$

$$= \frac{2yn 2^{-n/2} (y^2 n)^{n/2-1} e^{-\frac{y^2 n}{2}}}{\Gamma(n/2)} = \frac{2^{1-n/2} y^{n-1} n^{n/2} e^{-\frac{y^2 n}{2}}}{\Gamma(n/2)}.$$

$$f_{\xi/\eta}(z) = \int\limits_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2 y^2}{2}} \frac{2^{1-n/2} y^{n-1} n^{n/2} e^{-\frac{y^2 n}{2}}}{\Gamma(n/2)} dy$$

$$= \frac{2^{1/2-n/2} n^{n/2}}{\sqrt{\pi}\Gamma(n/2)} \int\limits_0^\infty y^{n-1} e^{-\frac{(z^2+n)y^2}{2}} dy.$$

Сделаем замену переменных.

$$x = \frac{(z^2+n)y^2}{2},\ y = \left(\frac{2x}{z^2+n}\right)^{1/2},\ dy = \left(\frac{2}{z^2+n}\right)^{1/2}\frac{1}{2}x^{-1/2}dx.$$

Тогда получим

$$f_{\xi/\eta}(z) = \frac{2^{\frac{1-n}{2}}n^{n/2}}{\sqrt{\pi}\Gamma(n/2)}\frac{2^{-\frac{1}{2}}}{(z^2+n)^{1/2}}\int_0^\infty \frac{2^{\frac{n-1}{2}}x^{\frac{n-1}{2}+1-1}}{(z^2+n)^{\frac{n-1}{2}}}e^{-x}dx =$$

$$= \frac{2^{-\frac{1}{2}}n^{n/2}}{\sqrt{\pi}\Gamma(n/2)(z^2+n)^{(n+1)/2}}\Gamma\left(\frac{n-1}{2}+1\right) =$$

$$= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma(n/2)(z^2/n+1)^{(n+1)/2}},\ z \in R.$$

Распределение Стьюдента с n степенями свободы обозначается как $S(n)$.

Теорема. Пусть $\mathbf{X}_n = (X_1,...,X_n),\ X_i \in N(\mu,\sigma^2)$ и $t = \sqrt{n-1}\cdot\frac{\bar{X}-\mu}{S}$, где $\bar{X} = \sum_{i=1}^n X_i$ – выборочное среднее и $S^2 = \sum_{i=1}^n \left(X_i - \bar{X}\right)^2$ – выборочная дисперсия. Тогда $t \in S(n-1)$.

П6.3. Распределение Снедекора – Фишера

Пусть случайные величины χ_1^2 и χ_2^2 независимы и $\chi_1^2 \in \chi^2(n_1)$, $\chi_2^2 \in \chi^2(n_2)$.

Определение. Распределение случайной величины $\dfrac{\chi_1^2 / n_1}{\chi_2^2 / n_2}$ называют *распределением Снедекора* с n_1 и n_2 степенями свободы. Или F-распределением дисперсионного отношения Фишера.

$$f_{n_1,n_2}(x) = \int\limits_{-\infty}^{\infty} f_{\chi^2/n_1}(xy) f_{\chi^2/n_2}(y)|y|dy = \int\limits_{0}^{\infty} n_1 f_{\chi_1^2}(n_1, xy) f_{\chi_2^2}(n_2, y) y dy =$$

$$= \left(\frac{n_1}{n_2}\right)^{n_1/2} \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \frac{x^{n_1/2-1}}{(1+n_1 x/n_2)^{(n_1+n_2)/2}},\ x>0.$$

Теорема. Пусть

$$X_n \in N(\mu_1, \sigma_1^2),\ Y_m \in N(\mu_2, \sigma_2^2),$$

тогда статистика

$$F = \frac{n(m-1)}{m(n-1)} \frac{\sigma_2^2}{\sigma_1^2} \frac{S^2(x)}{S^2(y)} \in S(n-1, m-1).$$

П7. Метрики в пространстве функций распределения случайных величин

В математической статистике очень часто возникает задача измерения расстояния между распределениями. Многие методы математической статистики отличаются друг от друга фактически выбором другой метрики. Мы не претендуем на полноту изложения, а приведем только наиболее известные способы измерения расстояния между распределениями в математической статистике.

Отметим свойства, которые должны выполняться для метрики $\rho(a,b)$.

1. Положительность: $\rho(a,b) \geq 0$, причем $\rho(a,b) = 0$ тогда и только тогда, когда $a = b$.

2. Симметричность: $\rho(a,b) = \rho(b,a)$.

3. Неравенство треугольника: $\rho(a,b) + \rho(b,c) \geq \rho(a,c)$.

П7.1. Расстояние между функциями распределения

Распределение случайной величины полностью определяется своей функцией распределения. Пусть F_1 и F_2 – функции распределения двух одномерных случайных величин, между которыми мы хотим найти расстояние.

П7.1.1. Расстояние Колмогорова

Расстояние Колмогорова использует равномерную метрику

$$\rho(F_1, F_2) = \sup_{|x|<\infty} \left| F_1(x) - F_2(x) \right|.$$

П7.1.2. Расстояние омега-квадрат

Расстояние омега-квадрат использует квадратичную метрику

$$\rho(F_1, F_2) = \int_{-\infty}^{\infty} \left(F_1(x) - F_2(x) \right)^2 \psi\left(G(x) \right) dG(x),$$

где $G(x) = \nu F_1(x) + (1-\nu) F_2(x)$ – смесь распределений F_1 и F_2 с параметром смеси ν. Весовая функция $\psi(t)$ позволяет настроить чувствительность расстояния к отклонениям на разных участках функций распределения. Например, при $\psi(t) = 1$ мы получим одинаковую чувствительность к отклонениям распределений по всей области распределения, при $\psi(t) = \dfrac{1}{t(1-t)}$ – большую чувствительность на хвостах распределений, а при $\psi(t) = t(1-t)$ – в средней части.

П7.2. Расстояние между функциями плотности распределения

Если случайные величины непрерывны, то можно найти плотность распределения как производную от функции распределения. Тогда расстояние между распределениями можно определять на основании функций плотности распределений. Пусть f_1 и f_2 – функции плотности распределения двух одномерных случайных величин, между которыми мы хотим найти расстояние.

П7.2.1. Расстояние в метрическом пространстве L^1

Если взять в качестве расстояния норму разности функции $f_1 - f_2$ в пространстве L^1, то мера различимости функций f_1 и f_2 будет вычисляться по формуле

$$\rho(f_1, f_2) = \frac{1}{2}\int\left|f_1(x) - f_2(x)\right|dx.$$

Это расстояние обращается в нуль, если плотности f_1 и f_2 совпадают, и принимает свое максимальное значение, равное 2, если они сосредоточены на непересекающихся множествах.

П7.2.2. Расстояние Какутани – Хеллингера

Расстояние Какутани – Хеллингера представляет собой норму функции $\sqrt{f_1(x)} - \sqrt{f_2(x)}$ в пространстве L^2:

$$\rho(f_1, f_2) = \int\left(\sqrt{f_1(x)} - \sqrt{f_2(x)}\right)^2 dx = 2 - 2\int\sqrt{f_1(x)f_2(x)}dx.$$

Расстояние Какутани – Хеллингера обладает способностью правильно отражать «степень трудности» различения распределений по результатам наблюдений.

П7.2.3. Дивергенция Кульбака – Лейблера

Отношение $z_i = \ln\dfrac{f_1(x_i)}{f_0(x_i)}$ называется *информацией о различии* между гипотезами H_0 и H_1, которую несет одно наблюдение x_i. Объяснение именно такой формы информации может быть дано в смысле теоремы Байеса как разность между логарифмами шансов в пользу гипотезы H_1 до и после наблюдения x_i:

$$\ln\frac{f_1(x_i)}{f_0(x_i)} = \ln\frac{P\{H_1 \,|\, \xi = x_i\}}{P\{H_0 \,|\, \xi = x_i\}} - \ln\frac{P\{H_1\}}{P\{H_0\}}.$$

Среднее значение информации о различии между H_0 и H_1 называется *дивергенцией Кульбака – Лейблера*:

$$D(H_0, H_1) = -E_0 z_i = \int_R f_0(x) \ln \frac{f_0(x)}{f_1(x)} dx,$$

$$D(H_1, H_0) = E_1 z_i = \int_R f_1(x) \ln \frac{f_1(x)}{f_0(x)} dx.$$

Дивергенция Кульбака – Лейблера обладает двумя фундаментальными свойствами.

1. Дивергенция Кульбака – Лейблера является неотрицательной, причем $D(H_0, H_1) = 0$ тогда, и только тогда, когда $H_0 = H_1$.

2. Дивергенция Кульбака – Лейблера является асимметричной, т. е. $D(H_0, H_1) \neq D(H_1, H_0)$.

Симметричная дивергенция Кульбака – Лейблера равна

$$\rho(H_0, H_1) = D(H_0, H_1) + D(H_1, H_0) = \int_R \big(f_1(x) - f_0(x)\big) f_0(x) \ln \frac{f_1(x)}{f_0(x)} dx.$$

Хотя $\rho(H_0, H_1)$ и обладает свойством симметричности, она не является расстоянием в пространстве распределений, так как для нее не выполняется неравенство треугольника.

Постовалов Сергей Николаевич
Чимитова Екатерина Владимировна
Карманов Виталий Сергеевич

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

КОНСПЕКТ ЛЕКЦИЙ

Учебное пособие

2-е издание

Редактор *М.А. Кантурова*
Выпускающий редактор *И.П. Бровановa*
Корректор *И.Е. Семенова*
Дизайн обложки *А.В. Ладыжская*
Компьютерная верстка *С.И. Ткачева*

Налоговая льгота – Общероссийский классификатор продукции
Издание соответствует коду 95 3000 ОК 005-93 (ОКП)

Подписано в печать 27.10.2017. Формат 60 × 84 1/16. Бумага офсетная. Тираж 50 экз.
Уч.-изд. л. 8,13. Печ. л. 8,75. Изд. № 297. Заказ № 1344. Цена договорная

Отпечатано в типографии
Новосибирского государственного технического университета
630073, г. Новосибирск, пр. К. Маркса, 20