МИНИСТЕРСТВО ОБЩЕГО И ПРОФЕССИОНАЛЬНОГО
ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

На правах рукописи

Постовалов Сергей Николаевич

СТАТИСТИЧЕСКИЙ АНАЛИЗ ИНТЕРВАЛЬНЫХ НАБЛЮДЕНИЙ ОДНОМЕРНЫХ НЕПРЕРЫВНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

Специальность 05.13.16 — применение вычислительной техники, математического моделирования и математических методов в научных исследованиях (в области технических наук)

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель
к.т.н., доцент Лемешко Б.Ю.

Новосибирск, 1997

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Современное состояние и актуальность темы исследований. Любая математическая теория, родившаяся из практических задач, в своей основе содержит ряд явных или неявных предположений об исследуемом ею объекте. Когда выводы теории перестают соответствовать практике, требуется либо пересмотреть ее основные предположения, либо ограничить область ее применения. Математическая статистика в своем развитии прошла несколько этапов, на каждом из которых идеализированное представление о статистическом эксперименте становилось все более приближенным к реальности. Так, например, робастная статистика [1, 2, 3] внесла предположение о наличии аномальных наблюдений в выборке; непараметрическая статистика [4] отбросила предположение о том, что экспериментатору известно параметрическое семейство распределений, которому подчинены выборочные наблюдения.

Дальнейшее развитие статистики связано с непосредственным учетом факта погрешностей наблюдений в статистических процедурах [5, 6, 7]. Если в статистическом эксперименте фиксируется положение монеты "герб" или "решка", то погрешности наблюдений отсутствуют, а полученную выборку можно назвать "точной". Однако, если измеряется сила тока, уровень воды в реке, рост человека и т.п., то статистическая выборка уже не может быть "точной", так как приборы или органы восприятия дают нам лишь некоторое приближенное значение. Естественно, что если затем по этим значениям делаются статистические выводы, игнорирующие погрешности измерений, то результат оказывается не соответствующим действительности.

Почему же столь очевидный факт игнорировался при разработке математической теории статистики? Ответ, скорее всего, заключается в следующем: малый объем рассматриваемых выборок и отсутствие мощной вычислительной техники [6]. Действительно, когда объем выборки невелик, то статистическая погрешность в определении оценки превосходит погрешность, связанную с ошибками измерений. Решение же асимптотических статистических задач было продиктовано необходимостью построения общей математической теории, так как, например, теория малых выборок из нормального распределения будет отличаться от теории малых выборок из закона Пуассона [8]. С другой стороны, корректный учет погрешностей резко усложняет статистическую модель, и без применения вычислительной техники решение практических задач было бы невозможно.

Существует несколько различных способов описания реального наблюдения. Пожалуй, наиболее общее описание наблюдения дает *теория нечетких множеств* [9, 10, 11]. Каждое наблюдение представляется в виде функции $\varphi(x) : \mathrm{R} \to [0, 1]$, которая задает достоверность его нахождения в той или иной точке. В области, где наблюдение быть не может, достоверность равна нулю. Множество $\{x \mid \varphi(x) > \alpha\}$ задает область, в которой наблюдение находится с уровнем достоверности α. Вид функции $\varphi(x)$ определяет степень нечеткости нашего знания о наблюдении. Достоинством такого представления знаний является то, что можно анализировать статистическими методами выборки, заданные лингвистически: “много”, “мало”, “приблизительно один”, “около трех” и т.д., если, конечно, заданы соответсвующие функции принадлежности $\varphi(x)$. Недостатком нечеткого описания наблюдения является прозвольный (экспертный) выбор функций $\varphi(x)$.

Теория нечетких множеств получила развитие в статистике в двух модификациях — статистический анализ *размытых* наблюдений [12] и статистический анализ *нечетких* наблюдений [13, 14, 15].

Более простое описание дается в духе *интервальной математики* [16]. Каждое наблюдение представляется в виде интервала $[\underline{x}, \overline{x}]$, задающего верхнюю и нижнюю границу возможного расположения точного значения наблюдения. Выбор границ интервалов не является произвольным, а определяется из условий эксперимента.

Далее, возможен вероятностный подход — представление наблюдения в виде одного числового значения с априорно заданным распределением ошибок измерения. Аксиоматический выбор нормального в качестве априорного на основании центральной предельной теоремы теории вероятностей не согласуется с практикой, — так, например, в результате специальных исследований, проведенных в 1965–1975 гг., было установлено, что законы рапределения ошибок измерений весьма разнообразны и очень часто отличаются от нормального [17, 18]. Более того, распределения ошибок измерений приборов с течением времени могут изменяться [19]. Интересно, что многие фактические распределения ошибок, отличающиеся по форме, пересекаются в области 0.05-й и 0.95-й квантили в очень узком интервале значений $x/\sigma = 1.6 \pm 0.05$, что позволяет с допустимой в технических расчетах точностью 0.05σ определить 90%-й доверительный интервал, содержащий наблюдение [20].

Наконец, самый простой способ, — это использование обычных выборок, но не превосходящих по объему некоторого критического значения, называемого в [5] *рациональным объемом выборки*. Рациональный объем выборки получается из принципа "уравнивания статистической и измерительной погрешностей". Заметим, что в этом случае рациональный объем выборки зависит не только от погрешностей исходных данных, но и от вида решаемой статистической задачи.

Наряду с модернизацией понятия статистического наблюдения, существуют работы, в которых предлагается обобщение теории вероятностей на базе новой аксиоматики и замена вероятностных моделей нечеткими [10] и интервальными [21]. Этот подход заслуживает внимания, так как, предположив, что наблюдение может быть интервальным, нетрудно допустить, что и вероятность события может задаваться интервалом. К сожалению, в [21] отсутствуют ясные рекомендации по практическому применению интервальных моделей, что ограничивает практическое использование таких моделей.

Из рассмотренных подходов достаточно полным, практически обоснованным, и в то же время не слишком далеким от классических статистических схем представляется интервальное описание наблюдения. Развитие интервальной математики, родившейся первоначально из задач вычислительной математики по корректному учету ошибок округления [22, 23, 24], привело к ее постепенному проникновению в другие разделы математики. Не осталась в стороне и статистика. И хотя термин "интервальная статистика" еще не является общепризнанным, наличие целого ряда публикаций за последние 10 лет [5, 25, 6, 7, 16, 26, 27, 28, 21, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39] позволяет утверждать, что рождение этой теории на стыке интервальной математики и математической статистики уже произошло, и сейчас необходимо развивать и расширять это направление.

Цель и задачи исследований. Целью исследования диссертационной работы является повышение надежности статистических процедур за счет использования интервальных моделей представления исходных данных и расширения множества применяемых вероятностных моделей описания случайной величины смесями и усеченными распределениями.

Для достижения поставленной цели решаются следующие задачи:

- осуществляется перенос классических процедур проверки согласия на случай интервальных представления наблюдений;
- исследуются асимптотические свойства критериев согласия по интервальным наблюдениям;
- разрабатываются методы точечного и интервального оценивания параметров распределений по интервальным наблюдениям;
- создается программная система параметрического и непараметрического статистического анализа интервальных наблюдений, с возможностью преобразования исходной выборки и/или распределения случайной величины операторами сдвига, масштаба, отражения, усечения, логарифмирования, смеси, произведения на базе объектно-ориентированного программирования.

Методы исследования. Для решения поставленных задач используется аппарат теории вероятностей, математической статистики, интервального анализа, теории нечетких множеств, вычислительной математики, статистического моделирования.

Научная новизна диссертационной работы заключается

- в построении оценок границ вероятности согласия по критериям χ^2 Пирсона, отношения правдоподобия, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса для интервальных выборок;
- в построении интервальных L-, M- и MD- оценок параметров распределений по интервальным наблюдениям;

- в построении алгоритма проверки гипотез о согласии непараметрическими критериями в случае предварительного оценивания параметров произвольным методом;

- в построении алгоритма отбраковки аномальных наблюдений с использованием оценок, получаемых при минимизации статистики Колмогорова.

Основные положения, выносимые на защиту.

1. Метод проверки согласия по критериям χ^2 Пирсона, отношения правдоподобия, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса теоретического распределения с интервальной выборкой.

2. Теорема об асимптотических свойствах критерия Колмогорова по интервальной выборке.

3. Нестатистические интервальные оценки параметров по интервальной выборке.

4. Объектно-ориентированная программная система статистического анализа интервальных наблюдений одномерных непрерывных случайных величин.

Обоснованность и достоверность полученных результатов обеспечивается

- применением аналитических методов исследования свойств оценок и критериев;

- подтверждением аналитических выводов результатами статистического моделирования.

Практическая ценность и реализация результатов. Работа над системой статистического анализа ведется в рамках госбюджетной НИР по теме "Объектно-ориентированная программная система статистического анализа" [40]. Программа успешно используется для статистического анализа геодезических наблюдений [41, 42, 43].

Аппробация работы. Результаты исследований докладывались на Российской научно-технической конференции “Информатика и проблемы телекоммуникаций” (Новосибирск, 1996); Международной научно-технической конференции “Информатика и проблемы телекоммуникаций” (Новосибирск, 1995, 1997); VIII-м международном симпозиуме по непараметрическим и робастным методам в кибернетике (Красноярск, 1995); Международной конференции “Актуальные проблемы электронного приборостроения АПЭП-96” (Новосибирск, 1996); III-й международной научно-технической конференции “Микропроцессорные системы автоматики” (Новосибирск, 1996); Втором Сибирском Конгрессе по Прикладной и Индустриальной Математике ИНПРИМ-96 (Новосибирск, 1996); Межреспубликанском совещании по интервальной математике (Новосибирск, 1996); Международной конференции “Информационные технологии в моделировании и управлении” (Санкт-Петербург, 1996); Международной научно-методической конференции “Новые информационные технологии в университетском образовании” (Новосибирск, 1997); Международной научной конференции “Всесибирские чтения по математике и механике” (Томск, 1997); Международном совещании по интервальной математике (Красноярск, 1997); IX-м международном симпозиуме по непараметрическим и робастным методам в кибернетике (Железногорск, 1997); First Korea-Russia Internnational Symposium of Science and Technology (Ulsan, 1997).

Публикации. Основные результаты исследований по теме диссертации опубликованы в 23 печатных работах и 4 зарегистрированных отчетах по НИР.

Структура работы. Диссертация состоит из введения, трех глав основного содержания, заключения, списка литературы и приложений.

Краткое содержание работы. В первой главе сделан обзор литературы по способам нетрадиционного представления наблюдений одномерных непрерывных случайных величин, построена их классификация и установлена взаимосвязь размытых и нечетких наблюдений (теоремы 1.3 и 1.4).

Во второй главе исследованы теоретические аспекты статистического анализа интервальных наблюдений. В первом параграфе второй главы вводятся понятия интервальной эмпирической функции распределения и интервальной гистограммы. Во втором параграфе второй главы разработана методика проверки статистических гипотез по интервальным наблюдениям. Для критериев согласия отношения правдоподобия, χ^2 Пирсона, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса выведены оценки границ соответствующих статистик. Получена и доказана теорема об асимптотических свойствах статистики Колмогорова по интервальным наблюдениям. В третьем параграфе второй главы разработана методика получения L-, M- и MD- оценок параметров распределений по интервальным наблюдениям, проанализированы свойства этих оценок.

В третьей главе описывается объектно-ориентированная программная система статистического анализа интервальных наблюдений. В первом параграфе третьей главы рассматриваются преимущества объектно-ориентированного программирования при разработке программного обеспечения задач статистического анализа. Со второго по шестой параграф третьей главы описываются классы системы: “Наблюдение”, “Выборка”, “Распределение”, “Статистика”, “Идентификация”; представление данных, алгоритмы и методы работы с ними. В седьмом параграфе третьей главы приведен пример использования системы для статистического анализа геодезических наблюдений.

В приложении 1 содержится руководство пользователя системы. В приложении 2 приведены акты о внедрении программной системы.

1. СПОСОБЫ НЕТРАДИЦИОННОГО ПРЕДСТАВЛЕНИЯ СТАТИСТИЧЕСКИХ ДАННЫХ

Пусть в статистическом эксперименте наблюдается одномерная непрерывная случайная величина ξ. Непрерывная случайная величина может принимать бесконечное число значений, однако практически зафиксировать можно только их конечное количество, так как наблюдения всегда производятся с конечной точностью, т.е.

$$\tilde{\xi} = \xi + \varepsilon,$$

где ξ — "истинное" значение случайной величины, $\tilde{\xi}$ — измеренное значение, ε — ошибка измерения.

$\tilde{\xi}$ также является случайной величиной, но уже дискретной. Исследователь, как правило, располагает выборкой из реализаций $\tilde{\xi}$, но его интересует распределение случайной величины ξ. Можно ли по наблюдениям $\tilde{\xi}$ восстановить распределение ξ? Ответ на этот вопрос зависит от того, являются ли ошибки измерений случайными или систематическими, и известно ли распределение случайных ошибок измерений.

Предположим, что систематическая ошибка измерения отсутствует (это является вполне реалистичным, когда средство измерения новое, и это предположение не соответствует реальности после его длительной эксплуатации), а распределение случайной ошибки получено с помощью эталонного прибора. Тогда по значению случайной величины $\tilde{\xi}$ можно построить доверительный интервал, который с вероятностью α (практически берут α равное $0.9 - 0.95$ [19]) содержит неизвестное точно значение ξ. Варьируя доверительную вероятность от 0 до 1 мы получим множество вложенных интервалов или *размытое* число [12]. Таким образом, получаются следующие способы представления наблюдения: размытое число — интервал — точка. Аналогом размытых чисел с точки зрения теории нечетких множеств [10] являются *нечеткие* числа. В зависимости от способа представления наблюдения будем различать *нечеткие, размытые, интервальные и точечные* наблюдения и выборки. На рис. 1.1. приведена классификация одномерных наблюдений.

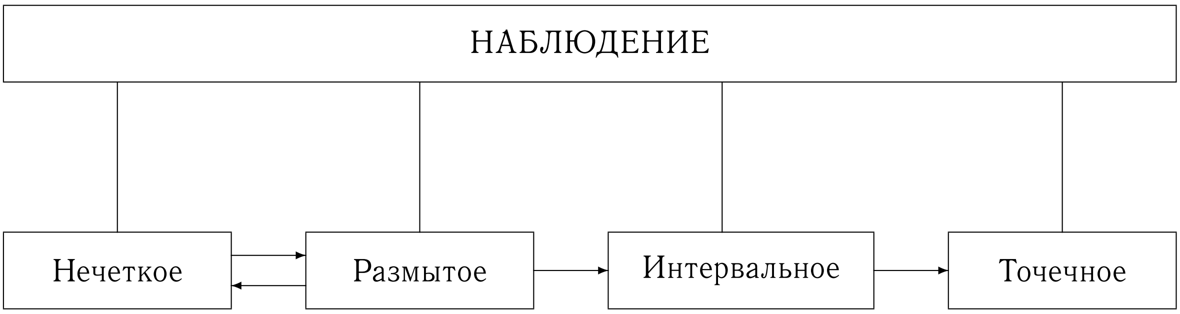


Рис. 1.1. Классификация наблюдений

Самым общим способом представления наблюдения являются нечеткие и размытые числа. Интервальное наблюдение можно получить из размытого, если зафиксировать доверительную вероятность. Из интервального наблюдения можно получить точечное наблюдение, если взять произвольную точку из интервала, в котором находится значение случайной величины ξ.

1.1. Выборка из нечетких наблюдений

Теория нечетких множеств [10] получила широкое распространение в тех областях науки и техники, где знания невозможно представить точно. Статистические наблюдения также можно рассматривать в терминах этой теории. Для этого вводятся понятия нечеткого числа и нечеткого вектора, на основе которых получаются соответственно нечеткое наблюдение и нечеткая выборка. Основные результаты, изложенные в этом параграфе, получены в [14, 13, 15].

Определение 1.1. *Кусочно-непрерывная функция* $\varphi_{a^*} : \mathrm{R} \to [0,1]$*, отображающая* R *в интервал* $[0,1]$*, определяет нечеткое число* a^* *на* R*, если семейство множеств* $(C(a^*)_\alpha)_{\alpha\in[0,1]}$*, полученное из* $\varphi_{a^*}(\cdot)$ *по формуле*

$$C(a^*)_\alpha = \{a \in \mathrm{R} : \varphi_{a^*}(a) \geq \alpha\}, \forall\alpha \in [0,1], \tag{1.1}$$

обладает следующими свойствами:

$$\begin{array}{ll} C(a^*)_\alpha & \alpha = 1, \\ C(a^*)_\alpha & [C_L(a^*)_\alpha, C_U(a^*)_\alpha] \subset \mathrm{R}. \end{array} \tag{1.2}$$

Функция $\varphi_{a^*}(\cdot)$ *называется характеризующей функцией* a^**, множество* $(C(a^*)_\alpha)_{\alpha\in[0,1]}$ *называется* α*-отсечением* a^**.*

Понятно, что выбор функции $\varphi_{a^*}(\cdot)$ определяет свойства нечеткого числа. Так, если в качестве характеризующей функции взять

$$\varphi_{x^*}(x) = I_{x_0}(x) = \begin{cases} 1, & x = x_0, \\ 0, & x \neq x_0; \end{cases}$$

то получится точечное наблюдение; если взять

$$\varphi_{x^*}(x) = I_{[\underline{x},\overline{x}]}(x) = \begin{cases} 1, & x \in [\underline{x},\overline{x}], \\ 0, & x \notin [\underline{x},\overline{x}]; \end{cases}$$

то получится интервальное наблюдение. На практике в качестве характеризующих функций нечеткого числа можно рекомендовать функции следующего вида:

$$\varphi_{x^*}(x) = \begin{cases} L(x), & x < m_1, \\ \quad 1, & m_1 \leq x \leq m_2, \\ R(x), & x > m_2, \end{cases} \tag{1.3}$$

где $L(x)$ — монотонно возрастающая функция при $x < m_1$, и $R(x)$ — монотонно убывающая функция при $x > m_2$. Например, для трапецидального нечеткого числа $T^*(m_1, m_2, a_1, a_2)$ функции $L(x)$ и $R(x)$ выглядят следующим образом:

$$L(x) = \max\left\{\frac{x - m_1 + a_1}{a_1}, 0\right\},\ R(x) = \max\left\{\frac{m_2 + a_2 - x}{a_2}, 0\right\}; \tag{1.4}$$

для экспоненциального нечеткого числа $E^*(m_1, m_2, a_1, a_2, p_1, p_2)$ (см. рис. 1.2):

$$L(x) = \mathrm{e}^{-|\frac{x-m_1}{a_1}|^{p_1}},\ R(x) = \mathrm{e}^{-|\frac{x-m_2}{a_2}|^{p_2}}. \tag{1.5}$$

Обобщая определение 1 на случай n-мерного пространства, получим следующее определение:

Определение 1.2. *Кусочно-непрерывная функция* $\varphi_{\mathbf{a}^*} : \mathrm{R}^n \to [0,1]$*, отображающая* R^n *в интервал* $[0,1]$*, определяет нечеткий вектор* $\mathbf{a}^*$ *на* R^n*, если семейство множеств* $(C(\mathbf{a}^*)_\alpha)_{\alpha\in[0,1]}$*, полученное из* $\varphi_{\mathbf{a}^*}(\cdot)$ *по формуле*

$$C(\mathbf{a}^*)_\alpha = \{\mathbf{a} \in \mathrm{R}^n : \varphi_{\mathbf{a}^*}(\mathbf{a}) \geq \alpha\}, \forall\alpha \in [0,1], \tag{1.6}$$

обладает следующими свойствами:

$$\begin{aligned} &C(\mathbf{a}^*)_\alpha \quad \alpha = 1, \\ &C(\mathbf{a}^*)_\alpha \ \ , \quad \mathrm{R}^n. \end{aligned} \tag{1.7}$$

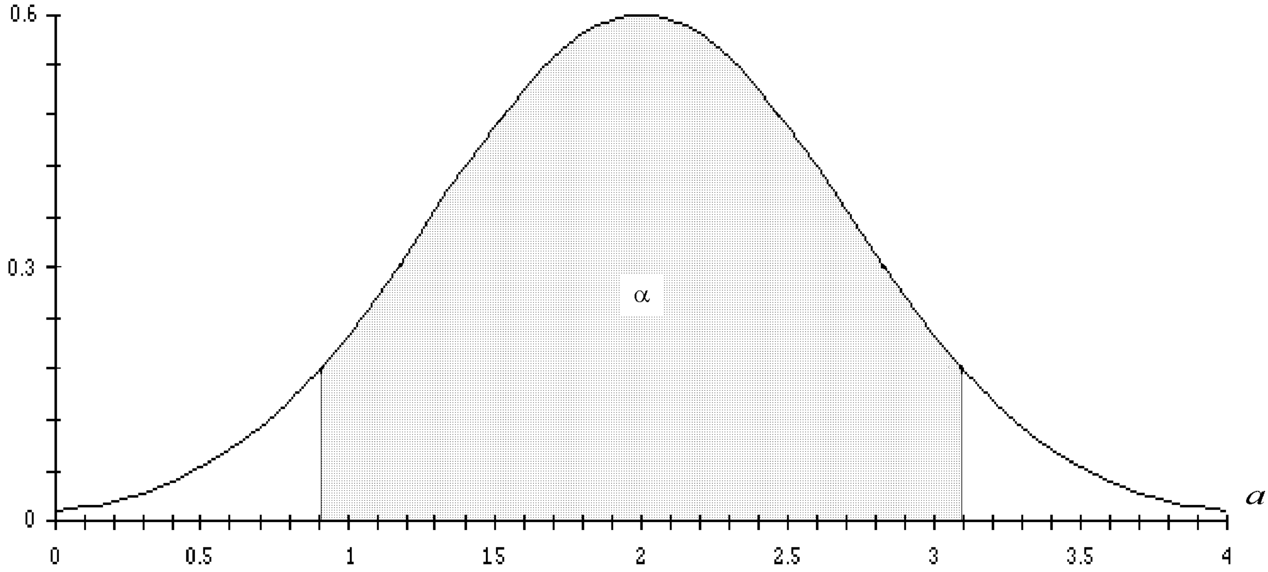


Рис. 1.2. α-отсечения экспоненциального нечеткого числа

Если задано n нечетких чисел, то построить характеризующую функцию нечеткого вектора можно несколькими способами:

- правило минимума [14]

$$\varphi_{\mathbf{x}^*}(\mathbf{x}) = \varphi_{(x_1^*, x_2^*, \ldots, x_n^*)}(x_1, x_2, \ldots, x_n) = \min\{\varphi_{x_1^*}(x_1), \ldots, \varphi_{x_n^*}(x_n)\}; \quad (1.8)$$

- правило произведения [15]

$$\varphi_{\mathbf{x}^*}(\mathbf{x}) = \varphi_{(x_1^*, x_2^*, \ldots, x_n^*)}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \varphi_{x_i^*}(x_i). \quad (1.9)$$

Если задано отображение n-мерного вектора в число, то можно построить аналогичное отображение нечеткого n-мерного вектора в нечеткое число. Следующее определение и теоремы показывают каким образом это можно сделать.

Определение 1.3. *Пусть* $\mathbf{a}^*$ *— нечеткий* n*-мерный вектор с характеризующей функцией* $\varphi_{\mathbf{a}^*}(\cdot)$ *и* α*-отсечением* $C(\mathbf{a}^*)_\alpha, \alpha \in [0,1]$*, и пусть* $g : \mathrm{R}^n \to Y$*,* $Y \subseteq \mathrm{R}$*, — непрерывное отображение и* $X_u = \{\mathbf{a} \in \mathrm{R}^n : g(\mathbf{a}) = u\}, \forall u \in \mathrm{R}$*. Нечеткий образ* $u^* = g(\mathbf{a}^*)$ *нечеткого вектора* $\mathbf{a}^*$ *определяется следующей характеризующей функцией:*

$$\varphi_{u^*}(u) = \begin{cases} \sup\limits_{\mathbf{a} \in X_u} \varphi_{\mathbf{a}^*}(\mathbf{a}), & X_u \ ; \\ 0, & X_u \ . \end{cases} \quad (1.10)$$

Теорема 1.1. *[15] Пусть* $\mathbf{a}^*$ *— нечеткий* n*-мерный вектор с характеризующей функцией* $\varphi_{\mathbf{a}^*}(\cdot)$ *и* α*-отсечением* $C(\mathbf{a}^*)_\alpha, \alpha \in [0,1]$*, и пусть* $g : \mathrm{R}^n \to Y$*,* $Y \subseteq \mathrm{R}$*, — непрерывное отображение. Если функция* $\varphi_{u^*}(u)$ *определена согласно (1.10), то*

1. $u^* = g(\mathbf{a}^*)$ *— нечеткое число в смысле определения 1.1.*
2. $C(u^*)_\alpha = \left[\min\limits_{\mathbf{a}\in C(\mathbf{a}^*)_\alpha} g(\mathbf{a}), \max\limits_{\mathbf{a}\in C(\mathbf{a}^*)_\alpha} g(\mathbf{a})\right]$

Теорема 1.2. *[15] Пусть* $\mathbf{a}^*$ *— нечеткий* n*-мерный вектор.* α*-отсечение* $\mathbf{a}^*$ *является Декартовым произведением* α*-отсечений компонент вектора тогда и только тогда, когда характеризующая функция* $\varphi_{\mathbf{a}^*}(\cdot)$ *построена по правилу минимума (1.8):*

$$C(\mathbf{a}^*)_\alpha = C(a_1^*)_\alpha \times C(a_2^*)_\alpha \times \cdots \times C(a_n^*)_\alpha,\ \forall\alpha \in [0,1] \Leftrightarrow$$

$$\Leftrightarrow \varphi_{\mathbf{a}^*}(\mathbf{a}) = \min_i \varphi_{a_i^*}(a_i),\ \forall a \in \mathrm{R}^n.$$

На базе теорем (1.1) и (1.2) нетрудно получить нечеткие аналоги для обычных статистик, типа выборочного среднего и выборочной дисперсии.

1. Пусть задана нечеткая выборка $\mathbf{x}^* = (x_1^*, x_1^*, \ldots, x_n^*)$. Тогда выборочное среднее есть нечеткое число $\overline{x}^*$ с характеризующей функцией

$$\varphi_{\overline{x}^*} = \sup_{x\in X_y} \varphi_{\mathbf{x}^*}(\mathbf{x}),\ X_y = \left\{\mathbf{x} \in \mathrm{R}^n : \frac{1}{n}\sum_{i=1}^n x_i = y\right\}, \forall y \in \mathrm{R},$$

и α-отсечением

$$C(\overline{x}^*)_\alpha = \left[\min_{\mathbf{x}\in C(\mathbf{x}^*)_\alpha} \frac{1}{n}\sum_{i=1}^n x_i, \max_{\mathbf{x}\in C(\mathbf{x}^*)_\alpha} \frac{1}{n}\sum_{i=1}^n x_i\right], \forall\alpha \in [0,1].$$

Если выборочный вектор $\mathbf{x}^*$ построен по правилу минимума, то α-отсечение выборочного среднего выглядит следующим образом:

$$C(\overline{x}^*)_\alpha = \left[\frac{1}{n}\sum_{i=1}^n C_L(x_i^*)_\alpha, \frac{1}{n}\sum_{i=1}^n C_U(x_i^*)_\alpha,\right], \forall\alpha \in [0,1]. \tag{1.11}$$

где $C(x_i^*)_\alpha = [C_L(x_i^*)_\alpha, C_U(x_i^*)_\alpha]$.

2. Пусть задана нечеткая выборка $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$. Тогда выборочная дисперсия есть нечеткое число $(S^2)^*$ с характеризующей функцией

$$\varphi_{(S^2)^*} = \sup_{x\in X_y} \varphi_{\mathbf{x}^*}(\mathbf{x}),\ X_y = \left\{\mathbf{x} \in \mathrm{R}^n : \frac{1}{n-1}\sum_{i=1}^n \left(x_i - \frac{1}{n}\sum_{i=1}^n x_i\right)^2 = y\right\}, \forall y \geq 0,$$

и α-отсечением

$$C((S^2)^*)_\alpha = \left[\min_{\mathbf{x}\in C(\mathbf{x}^*)_\alpha} g(\mathbf{x}), \max_{\mathbf{x}\in C(\mathbf{x}^*)_\alpha} g(\mathbf{x})\right], \forall\alpha \in (0,1],$$

где $g(\mathbf{x}) = g(x_1, x_2, \ldots, x_n) = \frac{1}{n-1}\sum\limits_{i=1}^{n}\left(x_i - \frac{1}{n}\sum\limits_{i=1}^{n} x_i\right)^2$. Границы α-отсечения для выборочной дисперсии можно найти численно.

По выборке из нечетких чисел можно построить эмпирическую функцию распределения $F_n(x)$:

$$(F_n)^L_\alpha(z) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,z)}(C_L(x_i^*)_\alpha), (F_n)^U_\alpha(z) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,z)}(C_U(x_i^*)_\alpha), \quad (1.12)$$

где $C_U(x_i^*)_\alpha$ и $C_L(x_i^*)_\alpha$ — верхняя и нижняя граница α-отсечения i-го наблюдения.

1.2. Выборка из размытых наблюдений

Понятие размытых чисел было введено в [12] для описания субъективных величин и также базируется на теории нечетких множеств.

Определение 1.4. *Характеризующая функция $\psi_{a^*}(x)$, определяет размытое число a^*, если она обладает следующими свойствами:*

1. $\psi_{a^*}(x) \geq 0, \forall x \in \mathrm{R}$;
2. $\int\limits_{\mathrm{R}} \psi_{a^*}(x)dx < \infty$.

Если $\int\limits_{\mathrm{R}} \psi_{a^*}(x)dx = 1$, то размытое число называется нормированным, иначе — размытое число называется ненормированным.

Нетрудно заметить, что характеризующая функция $\psi_{a^*}(x)$ нормированного размытого числа является плотностью рапределения. Точечное наблюдение соответствует размытому числу с характеризующей функцией $\psi_{a^*}(x) = \delta(x - x_0)$, где $\delta(x)$ — дельта-функция (сингулярное распределение). Интервальное наблюдение соответствует размытому числу с характеризующей функцией $\psi_{a^*}(x) = \frac{1}{\overline{x}-\underline{x}} I_{[\underline{x},\overline{x}]}(x)$ (равномерное распределение).

Для размытых чисел можно формально ввести операции, которые применяются к обычным числам [12].

Определение 1.5. *n-арная операция над точечными числами $f(x_1, x_2, \ldots, x_n)$ порождает n-арную операцию над размытыми числами*

$f(x_1^*, x_2^*, \dots, x_n^*)$ по правилу максмина, если характеризующая функция $\psi_{f(x_1^*, x_2^*, \dots, x_n^*)}(z)$ имеет следующий вид:

$$\psi_{f(x_1^*, x_2^*, \dots, x_n^*)}(z) = \max_{x_1, x_2, \dots, x_n | f(x_1, x_2, \dots, x_n) = z} \min_i \psi_{x_i^*}(x_i).$$

Пусть a^*, b^*, c^* — размытые числа, $x_0 \in \mathrm{R}$. Операции суммы и произведения, введенные по определению 1.5, обладают следующими свойствами [12]:

1. $\psi_{a^*+x_0}(z) = \psi_{a^*}(z - x_0)$.
2. $a^* + 0 = a^*$.
3. $\psi_{a^* \cdot x_0}(z) = \psi_{a^*}(z/x_0), x_0 \neq 0$.
4. $a^* \cdot 1 = a^*$.
5. $a^* \cdot 0 = 0$.
6. $a^* + b^* = b^* + a^*$.
7. $a^* \cdot b^* = b^* \cdot a^*$.
8. $a^* \cdot (b^* + c^*)$ не всегда равно $a^* \cdot b^* + a^* \cdot c^*$.
9. $a^* - a^* \neq 0$.
10. $a^* + a^* \neq 2 \cdot a^*$.

Арифметические операции с размытыми числами существенно отличаются от обычных арифметических операций, поэтому вычисления с размытыми числами представляют значительные трудности.

Между нечеткими (определение 1.1) и размытыми (определение 1.4) числами существует тесная связь, то есть при определенных условиях нечеткое число является размытым и наоборот.

Теорема 1.3. *Нечеткое число a^* с характеризующей функцией $\varphi_{a^*}(x)$ является размытым числом, тогда и только тогда, когда*

$$S = \int_{\mathrm{R}} \varphi_{a^*}(x) dx < \infty.$$

При этом характеризующая функция размытого числа имеет вид:

$$\psi_{a^*}(x) = \frac{1}{S}\varphi_{a^*}(x).$$

Теорема 1.4. *Размытое число a^* с характеризующей функцией $\psi_{a^*}(x)$ является нечетким числом, тогда и только тогда, когда функция $\psi_{a^*}(x)$ является унимодальной и*

$$M = \max_{x \in \mathrm{R}} \psi_{a^*}(x) < \infty.$$

При этом характеризующая функция нечеткого числа имеет вид:

$$\varphi_{a^*}(x) = \frac{1}{M}\psi_{a^*}(x).$$

Доказательство этих теорем очевидно и опирается только на определения 1.1 и 1.4.

Заметим, что даже одно нормированное размытое наблюдение своей характеризующей функцией задает непараметричечскую оценку распределения случайной величины. В случае n наблюдений в качестве непараметрической оценки плотности можно использовать смесь характеризующих функций:

$$\tilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \psi_{x_i^*}(x). \tag{1.13}$$

Пусть все наблюдения подчинены одному и тому же распределению с плотностью $f(x)$. Тогда оценка (1.13) является асимптотически несмещенной в каждой точке непрерывности $f(x)$, т.е.

$$M\{\tilde{f}_n(x)\} \to f(x), \quad n \to \infty,$$

в случае, если характеризующие функции размытых наблюдений построены из точечной выборки $\{x_1, x_2, \ldots, x_n\}$ с помощью ядерных функций со следующими свойствами [44]:

$$\psi_{x_i^*}(x) = \frac{1}{h_n}K\left(\frac{x - x_i}{h_n}\right), \ \sup_x |K(x)| < \infty, \lim_{x\to\infty} |xK(x)| = 0, \ \int\limits_{-\infty}^{+\infty} K(x)dx = 1,$$

и

$$h_n \to 0, \quad n \to \infty. \tag{1.14}$$

Понятно, что свойство асимптотической несмещенности оценки функции плотности выполняется тогда, когда дисперсия каждого размытого наблюдения стремится к нулю при увеличении объема выборки. Реально же добиться выполнения этого условия достаточно сложно. Поэтому имеет смысл модифицировать оценку (1.13) к следующему виду:

$$\tilde{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \psi_{x_i^*}\left(\frac{x}{h_n}\right), \tag{1.15}$$

где h_n удовлетворяет условию (1.14).

1.3. Выборка из интервальных наблюдений

Представление числа в виде интервала с целью учета погрешностей измерений довольно очевидно и берет свое начало с монографии Р.Е.Мура [22]. Однако в статистике эта идея развивалась почти независимо от интервального анализа [38]. Методологию нестатистического оценивания сформировал Л.В. Канторович [45], предложив оценивать интервалы (области) неопределенности методами математического программирования. В ряде работ интервальный нестатистический анализ [46, 31, 38] противопоставляется статистическому анализу, но обсуждение этой проблемы на круглом столе, проведенном редакцией журнала "Заводская лаборатория", [32, 33, 34, 35, 6, 28] показало, что более плодотворным является их совместное использование. Отметим работы Орлова, который первым ввел термин “интервальная статистика” [16] и исследовал интервальные выборки в предположении, что длины интервалов достаточно малы [5, 25, 7, 26].

1.3.1. Интервальная арифметика

Основная идея интервального анализа [22, 23, 24] заключатся в том, что вещественное число представляется не одним, а двумя числами — оценкой снизу и оценкой сверху, образующими интервальное число. Арифметические операции над числами выполняются так, что если $[a_1, a_2] = [b_1, b_2] \circ [c_1, c_2], b \in [b_1, b_2], c \in [c_1, c_2]$, то $b \circ c \in [a_1, a_2]$, где $\circ \in \{+, -, \times, /\}$. Это позволяет автоматически учитывать погрешности в задании исходных данных и погрешности, вызываемые округлениями при вычислениях на ЭВМ.

Множество всех интервалов на R обозначется через IR. Арифметические операции над интервальными числами вводятся следующим образом [22]:

Определение 1.6. *Если $r(x)$ – непрерывная унарная операция на R, то*

$$r(X) = [\min_{x\in X} r(x), \max_{x\in X} r(x)],\ X \in \mathrm{IR}, \tag{1.16}$$

определяют соответствующую ей операцию на множестве IR.

Определение 1.7. *Пусть $A = [a_1, a_2]$ и $B = [b_1, b_2]$. Тогда*

$$\begin{array}{rcl} A + B &=& [a_1 + b_1, a_2 + b_2], \\ A - B &=& [a_1 - b_1, a_2 - b_2] = A + [-1, -1] \cdot B \\ A \cdot B &=& [\min\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}, \max\{a_1b_1, a_1b_2, a_2b_1, a_2b_2\}] \\ A : B &=& [a_1, a_2] \cdot [1/b_2, 1/b_1] \end{array} \tag{1.17}$$

Для такого определения интервальных чисел особенностью является то, что произвольный невырожденный элемент из IR не имеет обратного ни по сложению, ни по умножению, отсутствует свойство дистрибутивности. Вычисления над такими числами достаточно трудоемки. Имеются различные расширения и обобщения интервальных чисел [24].

1.3.2. Интервальная выборка без пересечений

Частный случай интервальной выборки, когда интервалы не пересекаются, хорошо известен в классической статистике. Это *группированные, цензурированные и частично-группированные выборки.*

Определение 1.8. *Выборка называется группированной, если область определения случайной величины разбита на непересекающиеся интервалы, и известно только количество точечных наблюдений, попавших в эти интервалы.*

Определение 1.9. *Выборка называется цензурированной слева, если область определения случайной величины разбита на два непересекающихся интервала: в первом интервале известно только количество точечных наблюдений, а во втором интервале известны все точечные наблюдения.*

Определение 1.10. *Выборка называется цензурированной справа, если область определения случайной величины разбита на два непересекающихся интервала: в первом интервале известны все точечные наблюдения, а во втором интервале известно только количество точечных наблюдений.*

Определение 1.11. *Выборка называется цензурированной с двух сторон, если область определения случайной величины разбита на три непересекающихся интервала: в первом и последнем интервале известно только количество точечных наблюдений, а в среднем интервале известны все точечные наблюдения.*

Определение 1.12. *Выборка называется частично-группированной, если область определения случайной величины разбита на непересекающиеся интервалы двух типов: в интервалах первого типа известно только количество точечных наблюдений, а в интервалах второго типа известны все точечные наблюдения.*

Статистический анализ данных в группированном виде хорошо разработан, так как группированную выборку можно рассматривать как реализацию дискретной случайной величины [47]. Анализ частично-группированных выборок разработан в работах [48, 49, 50, 51]. Оценивание параметров по группированной выборке с неполным покрытием рассмотрено в работе [36].

1.3.3. Интервальная выборка с пересечениями

Интервальная выборка с пересечениями интервалов получается из точечной с учетом максимальной погрешности измерения. Пусть в результате эксперимента получена точечная выборка

$$Y_n = \{y_1, y_2, \ldots, y_n\}. \tag{1.18}$$

Однако исследователя фактически интересует выборка

$$X_n = \{x_1, x_2, \ldots, x_n\},\ y_i = x_i + \Delta_i, \Delta_i = \delta_i|y_i| + \varepsilon_i \tag{1.19}$$

где ε_i — абсолютная и δ_i — относительная погрешность измерения. Об относительной и абсолютной погрешности из технического паспорта прибора известно только, что

$$|\varepsilon_i| \le \Delta_\varepsilon,\ \ |\delta_i| \le \Delta_\delta. \tag{1.20}$$

Таким образом, границы интервала, содержащего точное значение наблюдения определяются по формулам:

$$a_i = y_i + \Delta_\delta |y_i| + \Delta_\varepsilon, \quad b_i = y_i - \Delta_\delta |y_i| - \Delta_\varepsilon. \tag{1.21}$$

В результате получаем интервальную выборку:

$$\mathbf{X}_n = \{[a_i, b_i], i = 1, \dots, n\}, \tag{1.22}$$

где a_i и b_i определяются по формуле (1.21).

Точечная выборка в пространстве R^n представляет собой точку, а интервальная выборка определяет в пространстве R^n n-мерный параллелепипед. Запись $X_n \in \mathbf{X}_n$ обозначает, что точка $(x_1, x_2, \dots, x_n)$ принадлежит n-мерный параллелепипеду, определенному интервальной выборкой $\mathbf{X}_n$, то есть $\forall i, a_i \le x_i \le b_i$. Выборка Y_n определяет точку, которая является его центром тяжести.

Естественным является предположение, что величины Δ_δ и Δ_ε достаточно малы. На основе этого предположения Орловым разработана теория *реалистической статистики* [5, 25, 7, 26], и введено понятие *нотны*.

Пусть задана статистика $g : \mathrm{R}^n \to \mathrm{R}$.

Определение 1.13. *Нотной N_g статистики g называется максимально возможное абсолютное отклонение $g(Y_n)$ от $g(X_n)$, т.е.*

$$N_g = N_g(\mathbf{X}_n) = \max_{X_n \in \mathbf{X}_n} |g(Y_n) - g(X_n)|. \tag{1.23}$$

При достаточно малых ошибках Δ_i измерения, с точностью до $o(|\Delta_i|)$ выполняется

$$g(Y) - g(X) = \sum_{i=1}^{n} \frac{\partial g}{\partial x_i} \Delta_i. \tag{1.24}$$

Тогда из (1.20) следует, что

$$N_g = \Delta_\varepsilon \sum_{i=1}^{n} |\frac{\partial g}{\partial x_i}| + \Delta_\delta \sum_{i=1}^{n} |x_i \frac{\partial g}{\partial x_i}|. \tag{1.25}$$

Таким образом, нотна статистики пропорциональна максимально допустимым погрешностям.

1.3.3.1. Оценивание параметров и характеристик

Пусть статистика $g(X)$ является асимптотически нормальной (при $n \to \infty$) с математическим ожиданием μ и дисперсией σ^2/n, и

$$Mg - \mu = o(n^{-1/2}).$$

Доверительный интервал для μ, построенный по точечной выборке Y_n с доверительной вероятностью γ, имеет вид

$$\left[g(Y) - u_\gamma \frac{\hat{\sigma}}{\sqrt{n}}, g(Y) + u_\gamma \frac{\hat{\sigma}}{\sqrt{n}}\right], \tag{1.26}$$

где u_γ — квантиль стандартного нормального распределения порядка $(1+\gamma)/2$, а $\hat{\sigma}$ — оценка σ по выборке Y_n.

С учетом нотны статистики g, доверительный интервал расширяется, и принимает вид

$$\left[g(Y) - u_\gamma \frac{\hat{\sigma}}{\sqrt{n}} - N_g, g(Y) + u_\gamma \frac{\hat{\sigma}}{\sqrt{n}} + N_g\right]. \tag{1.27}$$

Длина этого интервала при $n \to \infty$ стремится к $2N_g$, и если нотна статистики не равна нулю, то состоятельное оценивание соответствующего параметра или характеристики невозможно [25].

В качестве суммарной погрешности оценки можно использовать полудлину доверительного интервала при $u_\gamma = 1$ [7], т.е.

$$\frac{\hat{\sigma}}{\sqrt{n}} + N_g. \tag{1.28}$$

Первое слагаемое соответствует статистической погрешности, а второе — измерительной. Если следовать “принципу уравнивания погрешностей” [9], то, приравняв первое и второе слагаемое в (1.28), можно получить рациональный объем выборки

$$n = \frac{\hat{\sigma}^2}{N_g^2}. \tag{1.29}$$

В [5] вычислены доверительные интервалы и рациональный объем выборки для основного параметра гамма-распределения, в [25] — для аддитивных статистик, в [6] — для оценки математического ожидания и дисперсии.

1.3.3.2. Проверка гипотез

При проверке гипотезы вычисляется статистика $g(Y)$. Гипотеза отвергается, если $g(Y) \geq C_\alpha$, где C_α — критическое значение статистики, соответствующее уровню значимости α. С учетом нотны статистики g, критическое значение принадлежит интервалу

$$[C_\alpha - N_g, C_\alpha + N_g,], \tag{1.30}$$

а уровень значимости лежит между $1 - P\{g \leq C_\alpha + N_g\}$ и $1 - P\{g \leq C_\alpha - N_g\}$. Следовательно, для того чтобы надежно (с учетом возможных измерительных погрешностей) отвергнуть гипотезу, следует взять правую границу интервала (1.30), то есть $C_\alpha + N_g$. Эта методика приведена в [7] на примере одновыборочного критерия Стьюдента и двухвыборочного критерия Смирнова.

1.4. Основные задачи интервальной статистики

В данной главе предложена классификация одномерных наблюдений непрерывной случайной величины и рассмотрены самые общие способы представления наблюдений в виде нечетких и размытых чисел. Результаты, полученные для размытых и нечетких выборок, можно легко перенести на интервальные выборки:

1. Нечеткая выборка при фиксированном α-уровне представляет собой множество α-отсечений, т.е. интервальную выборку.

2. Размытая выборка при фиксированной доверительной вероятности α представляет собой множество доверительных интервалов, т.е. интервальную выборку.

Отсюда следует, что если будет разработан аппарат для работы с интервальными выборками, то его можно будет применять и для нечетких выборок при фиксированном α-уровне, и для размытых выборок при фиксированной доверительной вероятности α.

Интервальные выборки очень просто получаются в целом ряде практических задач: при группировании, цензурировании, в моделях порождения данных с погрешностями измерения. Однако, общая теория интервальных выборок (возможно, с использованием интервальной арифметики) еще не построена.

Методологический подход к построению общей теории интервальных выборок состоит в следующем: **неопределённость в задании исходных данных порождает неопределённость в статистических выводах**. Если статистические выводы базируются на некоторой выборочной статистике S, то для интервальной выборки мы получим интервальную статистику $[\underline{S}, \overline{S}]$. С помощью этого подхода требуется решить основные задачи статистического анализа одномерных наблюдений.

Основными задачами интервальной статистики одномерных наблюдений непрерывных случайных величин являются:

- непараметрическое оценивание функции распределения и функции плотности по интервальной выборке;
- проверка гипотез о согласии теоретического распределения с интервальной выборкой;
- точечное и интервальное оценивание параметров теоретического распределения по интервальной выборке.

Решение этих задач позволит учитывать метрологическую погрешность в статистических выводах, что сделает их более надёжными и устойчивыми.

2. ИНТЕРВАЛЬНОЕ ПРЕДСТАВЛЕНИЕ ВЫБОРОЧНЫХ ДАННЫХ. СТАТИСТИЧЕСКИЙ АНАЛИЗ ИНТЕРВАЛЬНЫХ НАБЛЮДЕНИЙ

Рассмотрим следующую модель порождения исходных данных. Пусть в результате эксперимента наблюдаются значения y_i одномерной непрерывной случайной величины ξ:

$$y_i = x_i + \varepsilon_i, \tag{2.1}$$

где x_i — точное значение, а ε_i — погрешность наблюдения. Если погрешность ε_i не превосходит по модулю некоторого числа d_i, то об истинном значении x_i можно сказать, что оно принадлежит интервалу $[a_i, b_i]$, где

$$a_i = y_i - d_i \qquad \text{и} \qquad b_i = y_i + d_i.$$

Таким образом, интервал $[a_i, b_i]$ содержит всю информацию об i-й реализации случайной величины ξ.

Определение 2.1. *Интервальным наблюдением называется интервал, содержащий неизвестное точно значение реализации случайной величины.*

Определение 2.2. *Интервальной выборкой объёма* n *называется множество из* n *интервальных наблюдений:*

$$\mathbf{X_n} = \{[\mathbf{a_i}, \mathbf{b_i}] \in \mathrm{IR} \mid \mathbf{a_i} \le \mathbf{x_i} \le \mathbf{b_i},\ \mathbf{a_i} \in \mathrm{R},\ \mathbf{b_i} \in \mathrm{R},\ \mathbf{i} = \mathbf{1}, \ldots, \mathbf{n}\}. \tag{2.2}$$

Замечание 2.1. К подобной математической модели могут привести процедуры группирования и цензурирования данных, хорошо известные в классической статистике. Отличие заключается в том, что интервалы группирования задаются априори, а в модели (2.1) границы интервалов связаны с наблюдениями. Тем не менее, несмотря на различные порождающие механизмы, все выводы, полученные для интервальной выборки (2.2), можно перенести на случай группированных, цензурированных и частично группированных выборок [52, 53].

Замечание 2.2. Интервалы $[a_i, b_i]$ в модели (2.2) могут быть бесконечными. Эта ситуация может возникнуть, например, в случае, когда стрелка измерительного прибора зашкаливает, и поэтому установить точное значение границы не представляется возможным.

Интервальную выборку (2.2) можно рассматривать как n-мерный параллелепипед в пространстве R^n. Тогда выборку, рассматриваемую в классической статистике

$$\mathrm{X}_n = \{x_1, x_2, \ldots, x_n\}, \tag{2.3}$$

можно интерпретировать как точку, принадлежащую этому параллелепипеду. Частный случай выборки (2.2), в которой наблюдались значения x_i, фиксируемые с точностью до интервала $[a_i, b_i]$, так, что $a_i \leq x_i \leq b_i, i = 1, \ldots, n$, будем называть *точечной* и обозначать $\mathrm{X}_n \in \mathbf{X}_n$.

Классические методы статистического анализа применимы к точечным выборкам. Для адаптации известных методов к интервальным выборкам обычным приёмом может служить построение интервала неопределённости интересующей исследователя статистики [45]. В самом деле, если исходные данные известны с точностью до интервала, то естественным является описание статистики также с помощью интервала. При этом статистические выводы становятся менее определёнными, но более надёжными.

2.1. Эмпирическая функция распределения и гистограмма

Основную информацию о распределении случайной величины ξ исследователь получает по эмпирической функции распределения и/или гистограмме, на которые опираются статистические методы анализа. Однако, для интервальной выборки построение этих функций, в общем случае, неоднозначно. Действительно, для построения гистограммы область определения случайной величины разбивается на k непересекающихся интервалов точками $X_0 < X_1 < \ldots < X_k$ и подсчитывается количество наблюдений, попавших в интервалы $(X_j, X_{j+1}]$, $j = 0, \ldots, k-1$. Если интервальное наблюдение $[a_i, b_i]$ покрывает точку разбиения X_j, то точечное значение наблюдения можно отнести как к интервалу $[X_{j-1}, X_j]$, так и к интервалу $[X_j, X_{j+1}]$. Множество всех допустимых гистограмм можно получить простым перебором. Мощность этого множества равна 2^p, где p — число наблюдений, попавших на границу разбиения. Чтобы наглядно представить это множество, предлагается по интервальной выборке строить *интервальную гистограмму* (см. рис. 2.1).

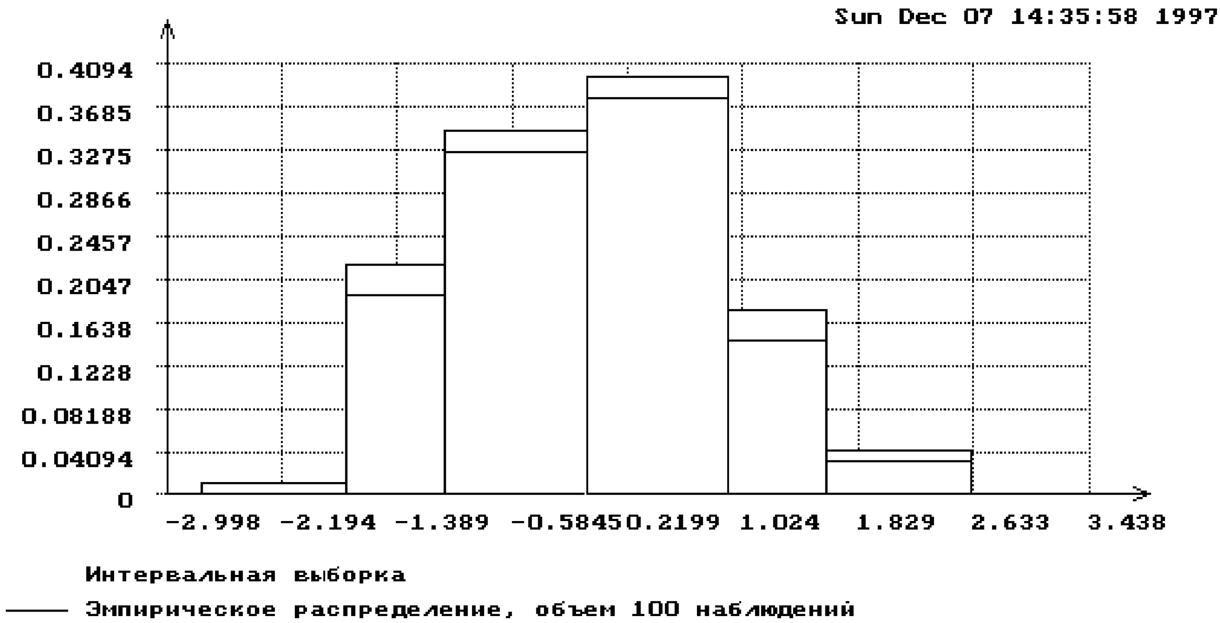


Рис. 2.1. Интервальная гистограмма для интервальной выборки

Высота каждого столбца интервальной гистограммы является интервалом: его нижняя граница определяется минимально возможным числом точечных наблюдений в интервале группирования, а верхняя граница определяется максимально возможным числом точечных наблюдений в интервале группирования. В то же время ни нижняя, ни верхняя границы интервальной гистограммы не удовлетворяют условию нормировки, а лишь задают миноранту и мажоранту для гистограммы с теми же граничными точками, построенную по любой точечной выборке $X_n \in \mathbf{X}_n$.

Более простым оказывается построение множества всех допустимых эмпирических функций распределения. Упорядочим граничные точки интервалов:

$$a_{(1)} \leq a_{(2)} \leq \ldots \leq a_{(n)}, \qquad b_{(1)} \leq b_{(2)} \leq \ldots \leq b_{(n)}.$$

Предположим, что все точечные значения наблюдений x_i совпали с левыми границами интервалов. Тогда эмпирическая функция распределения будет иметь следующий вид:

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

Аналогично, если все точечные значения совпали с правыми границами интервалов, эмпирическая функция распределения примет вид:

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

В общем случае эмпирическая функция распределения будет принадлежать множеству, ограниченному сверху $\overline{F_n}(x)$ и снизу $\underline{F_n}(x)$:

$$\underline{F_n}(x) \ \le\ F_n(x) \ \le\ \overline{F_n}(x) \qquad \forall x \in \mathrm{R}. \tag{2.4}$$

Следующий пример иллюстрирует вид $\overline{F_n}(x)$ и $\underline{F_n}(x)$ в зависимости от формы представления данных.

Пример 2.1. *Была сгенерирована точечная выборка объёмом 100 наблюдений. Её эмпирическая функция распределения приведена на рис. 2.2(а). Рис. 2.2(б) соответствует предположению, что наблюдения фиксировались с абсолютной погрешностью, а рис. 2.2(в) — с относительной погрешностью в исходных данных. Наконец, в последнем случае (рис. 2.2(г)) исходная выборка сгруппирована в 10 интервалов. На рисунках 2.2(б)–2.2(г) показаны графики функций* $\underline{F_n}(x)$ *и* $\overline{F_n}(x)$.

2.2. Проверка гипотез о согласии по интервальным выборкам

В данном параграфе раздела рассматриваются процедуры проверки гипотез о согласии теоретического закона распределения случайной величины с интервальной выборкой. Gastaldi в [54] нашёл верхнюю и нижнюю границы статистики Колмогорова в случае, когда выборка задана с пропусками данных, но при этом известно количество пропущенных наблюдений на

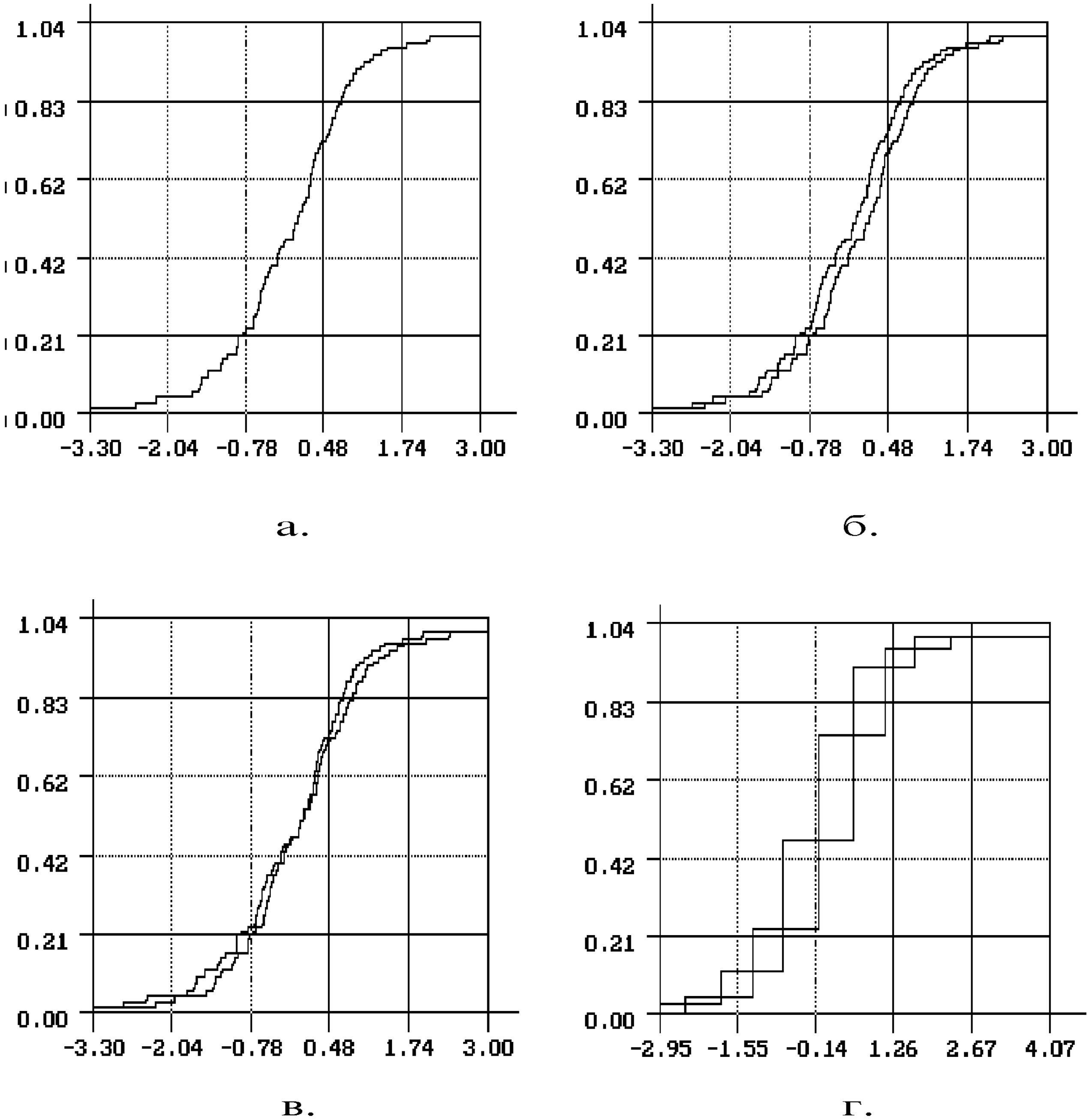


Рис. 2.2. Эмпирическая функция распределения точечной (а) и интервальных (б)–(г) выборок

интервалах между членами вариационного ряда (аналог частично группированной выборки). Этот результат обобщается на случай произвольной интервальной выборки и на статистики критериев согласия χ^2 Пирсона, отношения правдоподобия, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса [55, 56, 57].

При проверке гипотез о согласии теоретического распределения с точечной выборкой для найденного значения соответствующей статистики $S^\star$ вычисляется вероятность

$$p = P\{S > S^\star\} = \int\limits_{S^\star}^{\infty} g(s)ds,$$

где $g(s)$ — плотность распределения статистики при условии истинности нулевой гипотезы. При заданном уровне значимости α гипотеза о согласии не отвергается, если $p > \alpha$. В дальнейшем, вероятность $P\{S > S^\star\}$ будем называть ***вероятностью согласия***. Когда задана интервальная выборка (2.2), то статистика принадлежит интервалу $[\underline{S^\star}, \overline{S^\star}]$, границы которого определяются следующим неравенством:

$$\underline{S^\star}(\mathbf{X}_n) = \inf_{\mathrm{X}_n \in \mathbf{X}_n} S^\star(\mathrm{X}_n, F) \le S^\star(\mathrm{X}_n, F) \le \sup_{\mathrm{X}_n \in \mathbf{X}_n} S^\star(\mathrm{X}_n, F) = \overline{S^\star}(\mathbf{X}_n). \quad (2.5)$$

Вероятность $P\{S > S^\star\}$ будет принадлежать интервалу $[p_{min}, p_{max}]$, где

$$p_{min} = \int\limits_{\overline{S^\star}}^{\infty} g(s)ds, \qquad p_{max} = \int\limits_{\underline{S^\star}}^{\infty} g(s)ds.$$

Тогда, при заданном уровне значимости α, гипотезу о согласии следует отклонить, если $p_{max} \le \alpha$; гипотезу о согласии не следует отвергать, если $p_{min} > \alpha$. Если $p_{min} \le \alpha < p_{max}$, то однозначного вывода сделать невозможно. В этом случае для принятия решения об отклонении гипотезы о согласии необходима дополнительная информация.

Следующий пример иллюстрирует применение рассмотренного подхода.

Пример 2.2. *Была сгенерирована выборка из 100 наблюдений по стандартному нормальному закону с абсолютной погрешностью $\varepsilon = 0.05$ и относительной погрешностью $\delta = 0.01$ и проверено согласие с нормальным распределением с параметрами $\mu = 0.0$ и $\sigma = 1.0$ (см. рис. 2.3 и 2.4).*

На диаграмме в правом верхнем углу цифрами обозначена вероятность согласия по критериям: 1 — отношения правдоподобия, 2 — χ^2 Пирсона, 3

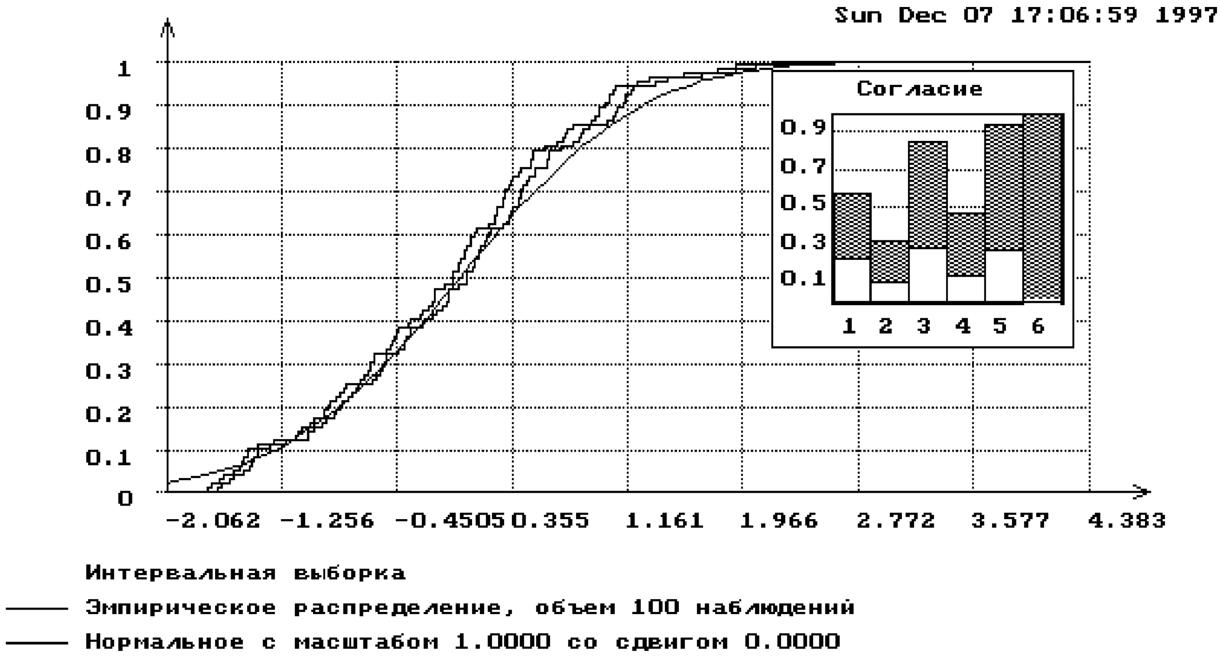


Рис. 2.3. Проверка согласия интервальной выборки с нормальным распределением

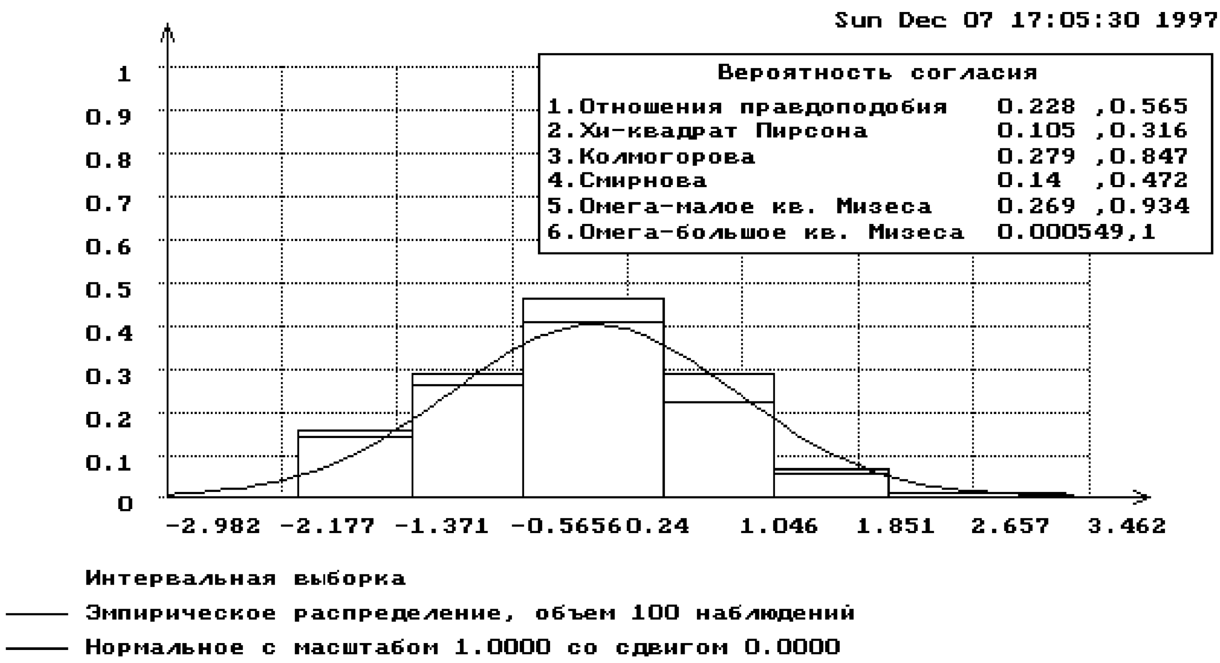


Рис. 2.4. Проверка согласия интервальной выборки с нормальным распределением

— Колмогорова, 4 — Смирнова, 5, 6 — ω^2 и Ω^2 Мизеса. Заштрихованные области показывают интервалы неопределённости вероятности согласия.

На основании проверки гипотез можно сделать следующие выводы:

- *При уровне значимости $\alpha = 0.1$ гипотеза о согласии* **не отвергается** *по критериям отношения правдоподобия, χ^2 Пирсона, Колмогорова, Смирнова, ω^2 Мизеса.*

- *При уровне значимости $\alpha = 0.2$ гипотеза о согласии* **не отвергается** *по критериям отношения правдоподобия, Колмогорова, ω^2 Мизеса.*

- *При уровне значимости $\alpha = 0.5$ гипотеза о согласии* **отвергается** *по критериям χ^2 Пирсона, Смирнова.*

По остальным критериям однозначного вывода сделать невозможно.

2.2.1. Критерии согласия χ^2 Пирсона и отношения правдоподобия

Перед использованием критерия χ^2 Пирсона необходимо сгруппировать исходную выборку. Область определения случайной величины разбивается на k непересекающихся интервалов граничными точками

$$X_0 < X_1 < \ldots < X_k,$$

после чего подсчитывается число наблюдений n_i, попавших в интервалы $(X_i, X_{i+1}], i = 0, 1, \ldots, k-1$. В случае интервальной выборки (2.2) процедура подсчета не является однозначной, и возможные значения n_i в соответствии с (2.4) удовлетворяют ограничениям:

$$n\underline{F_n}(X_i) \leq \sum_{j=0}^{i} n_j \leq n\overline{F_n}(X_i), n_i \geq 0, i = 0, 1, \ldots, k-1; \tag{2.6}$$

$$\sum_{i=0}^{k-1} n_i = n. \tag{2.7}$$

Статистика критерия χ^2 Пирсона имеет вид:

$$\chi^2 = \sum_{i=0}^{k-1} \frac{(n_i - np_i)^2}{p_i},$$

где p_i — вероятность попадания в интервал $(X_i, X_{i+1}]$. Если n_i определяются неоднозначно, то требуется найти максимум и минимум статистики χ^2 на области, заданной формулами (2.6) и (2.7):

$$\underline{\chi^2} = \min_{n_i} \chi^2, \ \overline{\chi^2} = \max_{n_i} \chi^2.$$

Вместо решения задачи целочисленного нелинейного программирования можно найти оценки границ статистики $\underline{\chi^2}$ и $\overline{\chi^2}$, допустив, что n_i могут принимать нецелочисленные значения:

$$\underline{\underline{\chi^2}} \leq \underline{\chi^2} \leq \chi^2 \leq \overline{\chi^2} \leq \overline{\overline{\chi^2}},$$

где $\underline{\underline{\chi^2}}$ — оценка снизу нижней границы статистики χ^2, $\overline{\overline{\chi^2}}$ — оценка сверху верхней границы статистики χ^2. Эта задача достаточно просто решается прямыми методами оптимизации с использованием штрафных функций.

Критерий отношения правдоподобия также использует группировку исходной выборки. Статистика критерия отношения правдоподобия имеет вид:

$$L = -2\sum_{i=0}^{k-1} n_i \ln\left(\frac{p_i}{n_i/n}\right),$$

где p_i — вероятность попадания в интервал $(X_i, X_{i+1}]$, n_i — количество наблюдений, попавших в интервал $(X_i, X_{i+1}]$, n — общее количество наблюдений.

Оценки верхней и нижней границы статистики отношения правдоподобия находятся аналогично.

2.2.2. Критерий согласия Колмогорова

Статистика критерия имеет вид

$$D_n = \sup_x |F_n(x) - F(x)|,$$

где $F_n(x)$ — эмпирическая функция распределения, $F(x)$ — теоретическая, согласие с которой проверяется, n — объём выборки. Преобразуем неравенство (2.4) к виду:

$$\begin{gathered}\underline{F_n}(x) - F(x) \leq F_n(x) - F(x) \leq \overline{F_n}(x) - F(x),\\ F(x) - \overline{F_n}(x) \leq F(x) - F_n(x) \leq F(x) - \underline{F_n}(x).\end{gathered}$$

Эти неравенства выполняются для всех , поэтому они сохраняются при взятии супремума:

$$\begin{aligned} &\sup_x(\underline{F_n}(x) - F(x)) \le \sup_x(F_n(x) - F(x)) \le \sup_x(\overline{F_n}(x) - F(x)),\\ &\sup_x(F(x) - \overline{F_n}(x)) \le \sup_x(F(x) - F_n(x)) \le \sup_x(F(x) - \underline{F_n}(x)). \end{aligned} \tag{2.8}$$

Объединим эти неравенства в одно и, учитывая, что статистика D_n не может быть отрицательной, получим:

$$\begin{aligned} &\underline{D_n} = \max\{\sup_x(\underline{F_n}(x) - F(x)), \sup_x(F(x) - \overline{F_n}(x)), 0\} \le\\ \le\ &D_n = \max\{\sup_x(F_n(x) - F(x)), \sup_x(F(x) - F_n(x))\} \le\\ \le\ &\overline{D_n} = \max\{\sup_x(\overline{F_n}(x) - F(x)), \sup_x(F(x) - \underline{F_n}(x))\}. \end{aligned} \tag{2.9}$$

2.2.3. Критерий согласия Смирнова

Статистика критерия имеет вид:

$$D_n^+ = \sup_x(F_n(x) - F(x)).$$

Из неравенства (2.8) следует:

$$\underline{D_n^+} = \sup_x(\underline{F_n}(x) - F(x)),\ \overline{D_n^+} = \sup_x(\overline{F_n}(x) - F(x)). \tag{2.10}$$

2.2.4. Критерий согласия ω^2 Мизеса

Построим вариационный ряд для точных значений наблюдений:

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$$

Если выборка интервальная, то каждый член вариационного ряда известен с точностью до интервала

$$\underline{x_{(i)}} \le x_{(i)} \le \overline{x_{(i)}}, \tag{2.11}$$

где $\underline{x_{(i)}}$ и $\overline{x_{(i)}}$ можно определить из неравенства (2.4), так как между вариационным рядом и эмпирической функцией распределения существует взаимно-однозначное соответствие (см. рис. 2.5) и $F_n(x_{(i)}) = i/n$:

$$\underline{x_{(i)}} = \inf\{y \mid y = \overline{F_n^{-1}}(i/n)\},\ \overline{x_{(i)}} = \sup\{y \mid y = \underline{F_n^{-1}}(i/n)\}.$$

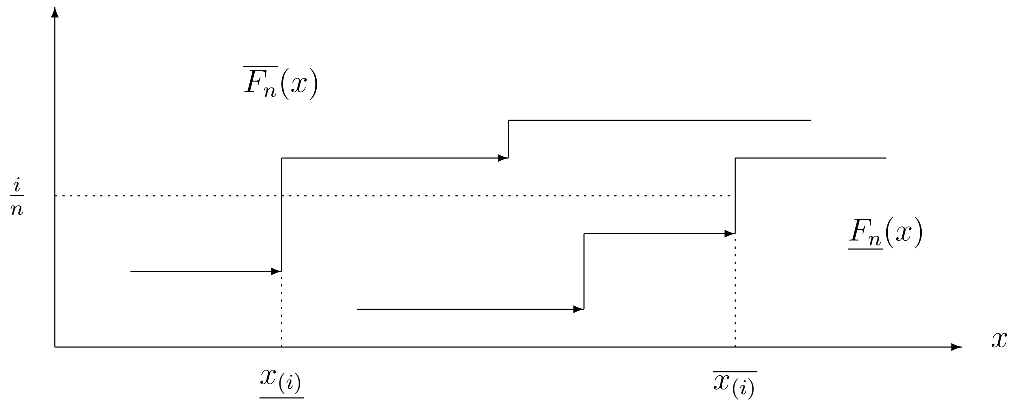


Рис. 2.5. Определение границ i-го члена вариационного ряда

Статистика критерия имеет вид:

$$n\omega_n^2 = n\int\limits_{-\infty}^{+\infty}(F_n(x)-F(x))^2dF(x) = \frac{1}{12n}+\sum_{i=1}^{n}\left[F(x_{(i)})-\frac{2i-1}{2n}\right]^2.$$

Пусть $s_i = [F(x_{(i)}) - \frac{2i-1}{2n}]^2$. Тогда из монотонности функции распределения $F(x)$ и неравенства (2.11) следует, что $\underline{s_i} \le s_i \le \overline{s_i}$, где $\underline{s_i}$ и $\overline{s_i}$ имеют вид:

$$\underline{s_i} = \begin{cases} [F(\overline{x_{(i)}}) - \frac{2i-1}{2n}]^2, & F(\underline{x_{(i)}}) \le F(\overline{x_{(i)}}) \le \frac{2i-1}{2n}, \\ 0, & F(\underline{x_{(i)}}) \le \frac{2i-1}{2n} \le F(\overline{x_{(i)}}), \\ [F(\underline{x_{(i)}}) - \frac{2i-1}{2n}]^2, & \frac{2i-1}{2n} \le F(\underline{x_{(i)}}) \le F(\overline{x_{(i)}}), \end{cases}$$

$$\overline{s_i} = \max\left\{\left[F(\underline{x_{(i)}}) - \frac{2i-1}{2n}\right]^2, \left[F(\overline{x_{(i)}}) - \frac{2i-1}{2n}\right]^2\right\}.$$

Тогда

$$\begin{array}{rcl} \underline{n\omega_n^2} & = & \frac{1}{12n} + \sum\limits_{i=1}^{n}\underline{s_i}, \\ \overline{n\omega_n^2} & = & \frac{1}{12n} + \sum\limits_{i=1}^{n}\overline{s_i}. \end{array} \qquad (2.12)$$

2.2.5. Критерий согласия Ω^2 Мизеса

Статистика критерия имеет вид:

$$n\Omega_n^2 = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_{(i)}) + \left(1-\frac{2i-1}{2n}\right)\ln(1-F(x_{(i)}))\right\}.$$

Из неравенства (2.11) и свойства монотонности функции распределения следует:

$$(F(\underline{x_{(i)}}))^{\frac{2i-1}{2n}}(1-F(\overline{x_{(i)}}))^{1-\frac{2i-1}{2n}} \le (F(\overline{x_{(i)}}))^{\frac{2i-1}{2n}}(1-F(\underline{x_{(i)}}))^{1-\frac{2i-1}{2n}};$$

$$(F(x_{(i)}))^{\frac{2i-1}{2n}}(1-F(x_{(i)}))^{1-\frac{2i-1}{2n}} \geq (F(\underline{x_{(i)}}))^{\frac{2i-1}{2n}}(1-F(\overline{x_{(i)}}))^{1-\frac{2i-1}{2n}}.$$

Отсюда

$$\begin{aligned} \underline{n\Omega_n^2} &= -n - 2\ln\prod_{i=1}^{n}\Big(F(\underline{x_{(i)}})\Big)^{\frac{2i-1}{2n}}\Big(1-F(\overline{x_{(i)}})\Big)^{1-\frac{2i-1}{2n}}; \\ \overline{n\Omega_n^2} &= -n - 2\ln\prod_{i=1}^{n}\Big(F(\overline{x_{(i)}})\Big)^{\frac{2i-1}{2n}}\Big(1-F(\underline{x_{(i)}})\Big)^{1-\frac{2i-1}{2n}}. \end{aligned} \tag{2.13}$$

2.2.6. Асимптотические свойства критериев согласия по интервальным выборкам

Очевидно, что чем меньше интервал неопределённости $[p_{min}, p_{max}]$, тем более определённые выводы можно сделать. На длину интервала неопределённости $\Delta p = p_{max} - p_{min}$ влияют

- длины интервалов $[a_i, b_i], i = 1, \ldots, n$;
- закон распределения, с которым проверяется согласие;
- критерий согласия;
- количество наблюдений.

Действительно, если задана точечная выборка, то $\Delta p = 0$. В пределе при $\min_i(b_i - a_i) \to \infty$, Δp стремится к единице. На рис. 2.6 показана зависимость Δp от величины абсолютной ошибки измерений ε для различных критериев согласия при проверке согласия выборки, смоделированной по нормальному закону объёмом 100 наблюдений, с нормальным распределением. На графике хорошо видно, что для различных критериев согласия величины Δp существенно отличаются. Однако нельзя сделать однозначного вывода о том, что тот или иной критерий лучше, так как на других модельных примерах расположение кривых $\Delta p(\varepsilon)$ может существенно отличаться от приведенного на рис. 2.6.

Влияние теоретического закона распределения, с которым проверяется согласие, на величину Δp хорошо проиллюстрировано на рис. 2.7–2.12. На этих графиках изображены верхняя и нижняя границы вероятности согласия по интервальной выборке из примера 2.2 в зависимости от значения параметра сдвига нормального распределения. Когда теоретическое распределение достаточно далеко сдвинуто относительно моделируемого, то

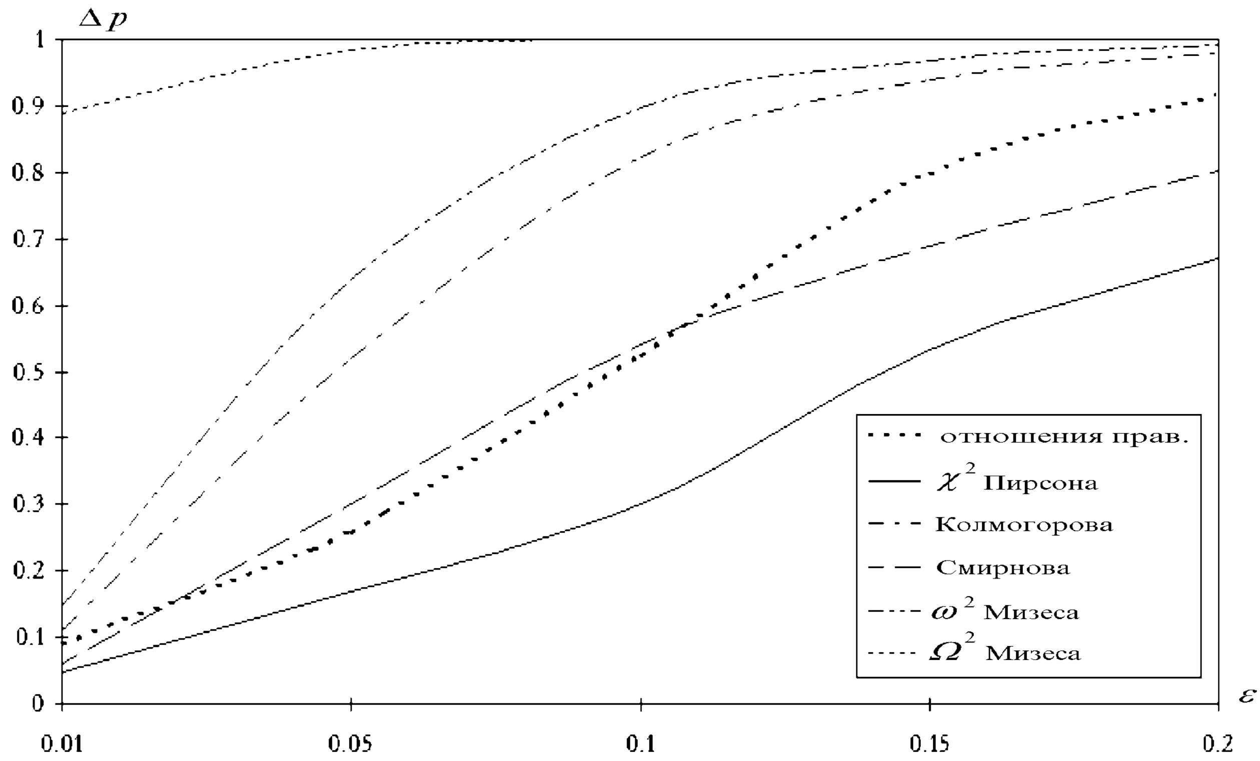


Рис. 2.6. Зависимость Δp от абсолютной погрешности измерения ε

и верхняя, и нижняя вероятности согласия равны нулю, а длина интервала неопределенности вероятности согласия равна нулю. Сравнение рисунков 2.7–2.12 позволяет сделать вывод о том, что кривые согласия по различным критериям существенно отличаются друг от друга. В частности, критерий Смирнова можно применять только для отклонения гипотезы о согласии, так как если теоретическая функция распределения $F(x)$ лежит строго выше эмпирической $F_n(x)$, то вероятность согласия равна единице (см. рис. 2.10).

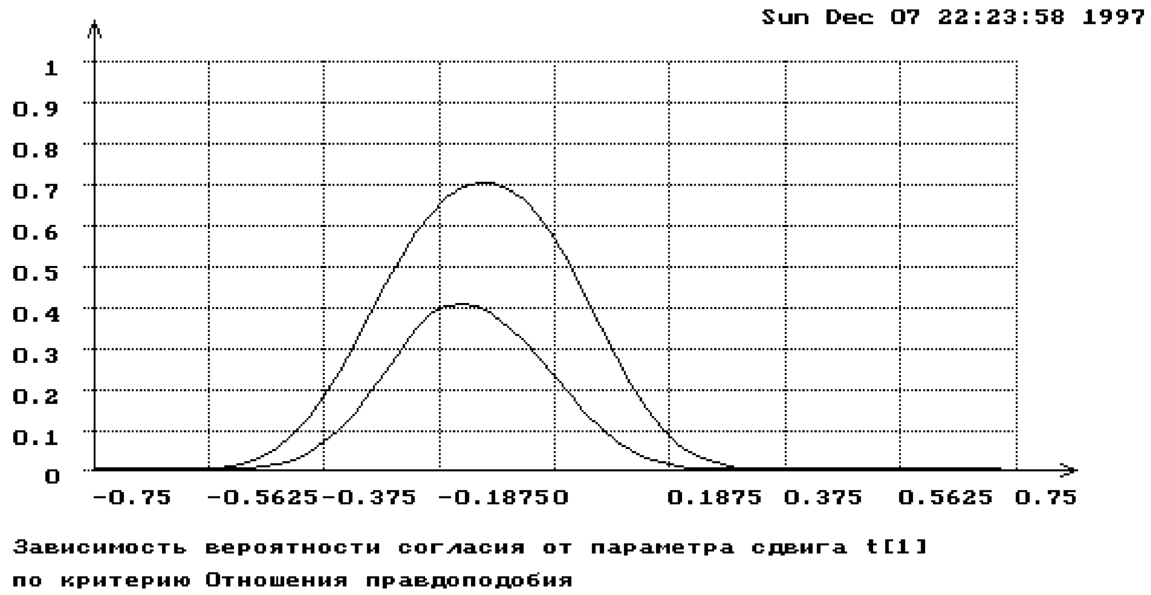


Рис. 2.7. Вероятности согласия по критерию отношения правдоподобия при изменении параметра сдвига нормального распределения
от -0.75 до 0.75

Наиболее интересный результат получается при исследовании поведения верхней и нижней границы вероятности согласия при увеличении числа наблюдений. Критерии согласия устроены таким образом, что с увеличением объема выборки максимально допустимое отклонение эмпирического распределения от теоретического (мера отклонения у каждого критерия своя) стремится к нулю (кроме критерия Смирнова). Однако, на практике из-за ошибок измерений отклонение никогда не будет равно нулю и его

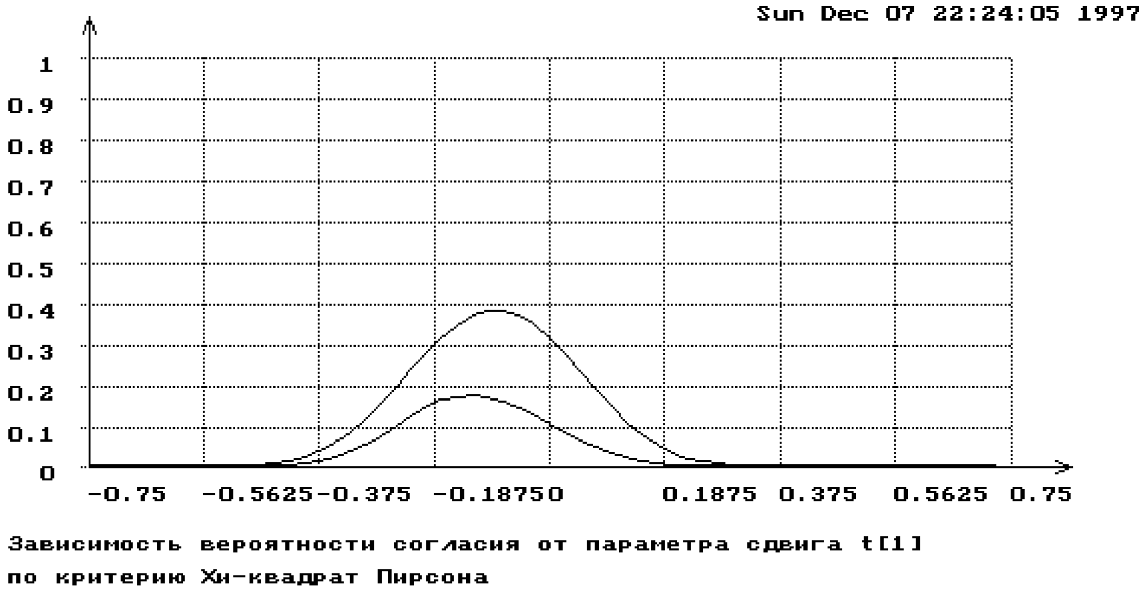


Рис. 2.8. Вероятности согласия по критерию χ^2 Пирсона при изменении параметра сдвига нормального распределения от -0.75 до 0.75

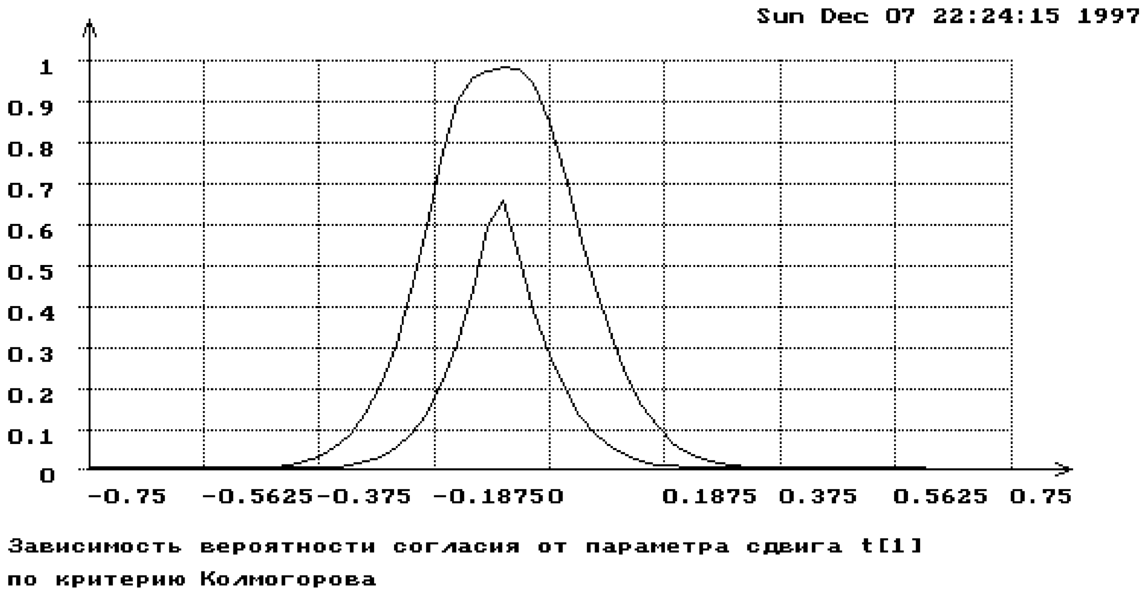


Рис. 2.9. Вероятности согласия по критерию Колмогорова при изменении параметра сдвига нормального распределения от -0.75 до 0.75

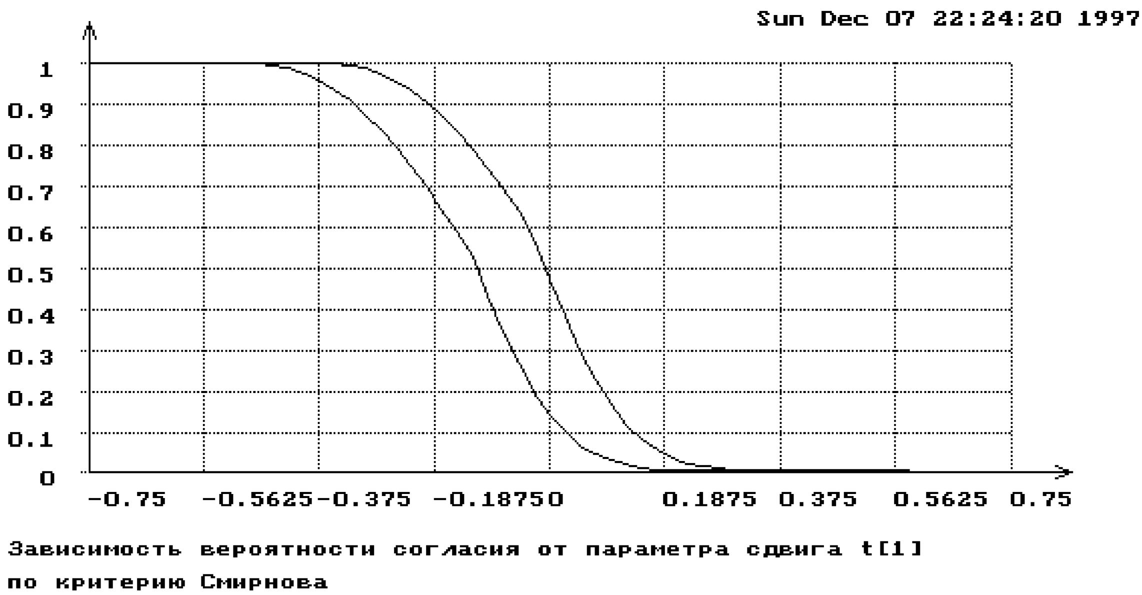


Рис. 2.10. Вероятности согласия по критерию Смирнова при изменении параметра сдвига нормального распределения от -0.75 до 0.75

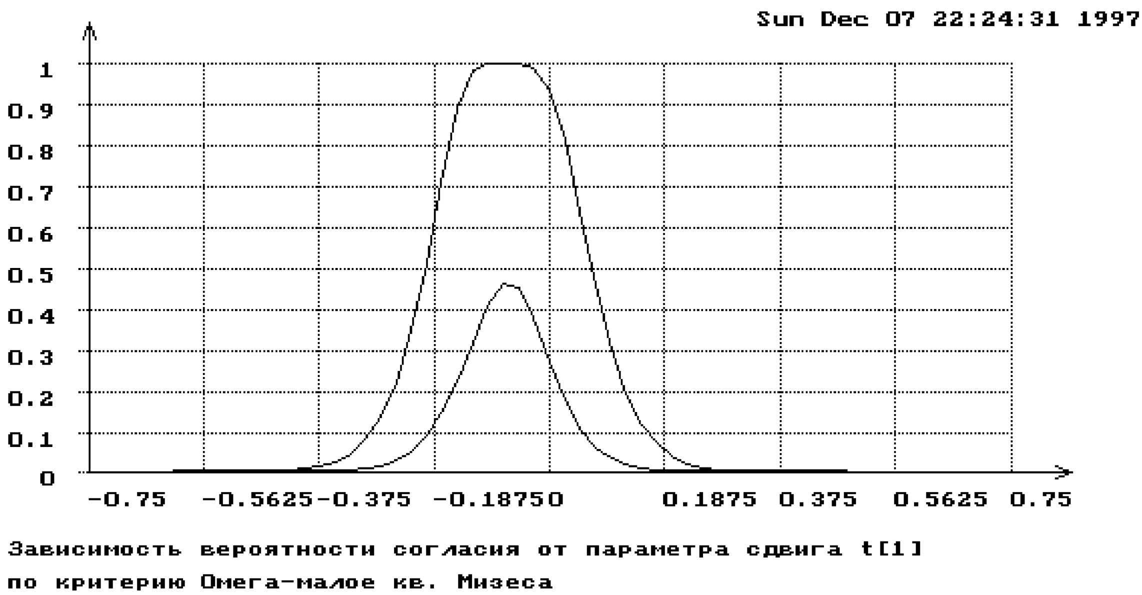


Рис. 2.11. Вероятности согласия по критерию ω^2 Мизеса при изменении параметра сдвига нормального распределения от -0.75 до 0.75

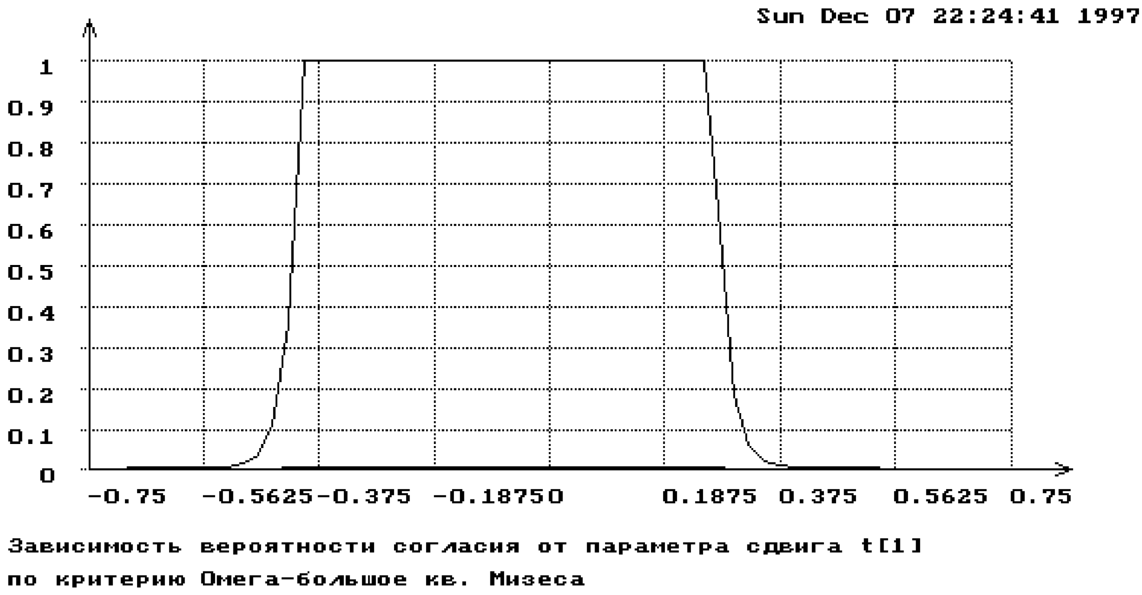


Рис. 2.12. Вероятности согласия по критерию Ω^2 Мизеса при изменении параметра сдвига нормального распределения от -0.75 до 0.75

величина будет зависеть от точности измерительного прибора.

О том, как увеличение объёма выборки влияет на Δp, говорит следующая теорема об асимптотических свойствах оценок границ статистики критерия Колмогорова по интервальной выборке.

Теорема 2.1. *Пусть задана последовательность интервальных выборок $\mathbf{X_n}$, для которых нижняя и верхняя границы эмпирической функции распределения $\underline{F_n}(x)$ и $\overline{F_n}(x)$ сходятся в равномерной метрике соответственно к $\underline{F}(x)$ и $\overline{F}(x)$ со скоростью $O(1/n)$, и $\sup\limits_x(\overline{F}(x) - \underline{F}(x)) \geq c > 0$.*

Пусть также $\mathcal{F}$ — это множество всех функций распределения, непрерывных справа, $p_{max}(F, \mathbf{X_n})$ и $p_{min}(F, \mathbf{X_n})$ — соответственно верхняя и нижняя границы вероятности согласия по критерию Колмогорова.

Тогда при $n \to \infty$:

1. $\forall F \in \mathcal{F}$, таких что $\forall x$ $(\underline{F}(x) \leq F(x) \leq \overline{F}(x))$,

$$(а)\ \ p_{max}(F, \mathbf{X_n}) \to \mathbf{1}, \qquad (б)\ \ p_{min}(F, \mathbf{X_n}) \to \mathbf{0};$$

2. $\forall F \in \mathcal{F}$, таких что $\exists x$ $((F(x) < \underline{F}(x)) \vee (F(x) > \overline{F}(x)))$,

$$(a) \;\; p_{max}(F,\mathbf{X_n}) \to \mathbf{0}, \qquad (б) \;\; p_{min}(F,\mathbf{X_n}) \to \mathbf{0}.$$

Доказательство.

Статистика $S = \frac{(6nD_n+1)^2}{18n}$ при достаточно большом n имеет распределение

$$P\{S > S^\star\} = 1 - K\left(\sqrt{\frac{S^\star}{2}}\right),$$

где $K(y) = \sum\limits_{k=-\infty}^{\infty}(-1)^k e^{-2k^2y^2}$ — функция распределения Колмогорова [88].

Для оценок границ $\underline{D_n}$ и $\overline{D_n}$ статистики D_n, определённых в (2.9), при $n \to \infty$ имеем:

$$p_{min} = 1 - K\left(\frac{6n\overline{D_n}+1}{6\sqrt{n}}\right) \to 0, \text{ если } \exists\lambda > 0 : \overline{D_n} \geq \lambda; \tag{2.14}$$

$$p_{max} = 1 - K\left(\frac{6n\underline{D_n}+1}{6\sqrt{n}}\right) \to \begin{cases} 1, \text{ если } (\underline{D_n} = 0) \vee (\underline{D_n} = O(1/n)); \\ 0, \text{ если } \exists\lambda > 0 : \underline{D_n} \geq \lambda. \end{cases} \tag{2.15}$$

Теперь для доказательства теоремы достаточно исследовать асимптотическое поведение оценок границ $\underline{D_n}$ и $\overline{D_n}$.

1. Пусть $F(x)$ — произвольная функция распределения, проходящая между $\underline{F}(x)$ и $\overline{F}(x)$.

(а) Согласно (2.9) оценка снизу для нижней границы D_n имеет вид:

$$\underline{D_n} = \max\left\{\sup_x(\underline{F_n}(x) - F(x)), \sup_x(F(x) - \overline{F_n}(x)), 0\right\}.$$

Если неравенство строгое: $\forall x \; \underline{F}(x) < F(x) < \overline{F}(x)$, то первые две величины в фигурных скобках будут отрицательными и $\underline{D_n} = 0$. Если $F(x)$ совпадает с $\underline{F}(x)$ на множестве $A \subseteq \mathrm{R}$ и с $\overline{F}(x)$ на множестве $B \subseteq \mathrm{R}$, то

$$\begin{aligned} \underline{D_n} &= \max\left\{\sup_{x\in A}(\underline{F_n}(x) - \underline{F}(x)), \sup_{x\in B}(\overline{F}(x) - \overline{F_n}(x))\right\} \leq \\ &\leq \max\left\{\sup_{x\in A}|\underline{F_n}(x) - \underline{F}(x)|, \sup_{x\in B}|\overline{F}(x) - \overline{F_n}(x)|\right\} = O(1/n). \end{aligned}$$

(б) Пусть x_0 — точка, в которой

$$\sup_x(\overline{F}(x) - \underline{F}(x)) \geq \overline{F}(x_0) - \underline{F}(x_0) \geq c > 0.$$

Обозначим $a = \overline{F}(x_0) - F(x_0) \geq 0$ и $b = F(x_0) - \underline{F}(x_0) \geq 0$. Тогда $a + b = \overline{F}(x_0) - \underline{F}(x_0) \geq c > 0$ и $\max\{a,b\} \geq c/2$. Используя оценку сверху для верхней границы D_n и введённые обозначения, получим:

$$\begin{aligned}
\overline{D_n} &= \max\left\{\sup_x(\overline{F_n}(x) - F(x)), \sup_x(F(x) - \underline{F_n}(x))\right\} = \\
&= \max\left\{\sup_x(\overline{F_n}(x) - \overline{F}(x)) + \sup_x(\overline{F}(x) - F(x))\ ,\right. \\
&\qquad \left.\sup_x(F(x) - \underline{F}(x)) + \sup_x(\underline{F}(x) - \underline{F_n}(x))\right\} \geq \\
&\geq \max\left\{O(1/n) + a\ ,\ b + O(1/n)\right\}.
\end{aligned}$$

Тогда $\exists \lambda > 0$ и $\exists n_0 : \forall n > n_0$

$$\overline{D_n} \geq \max\{a, b\} + O(1/n) \geq c/2 + O(1/n) \geq \lambda > 0.$$

2. Так как $p_{max} \geq p_{min}$, то (а) $\Rightarrow$ (б), и достаточно показать, что $p_{max} \to 0$.

Пусть x_0 — точка, в которой $F(x) > \overline{F}(x)$ (аналогично рассматривается случай, когда $F(x) < \underline{F}(x)$). Обозначим $d = F(x_0) - \overline{F}(x_0) > 0$.

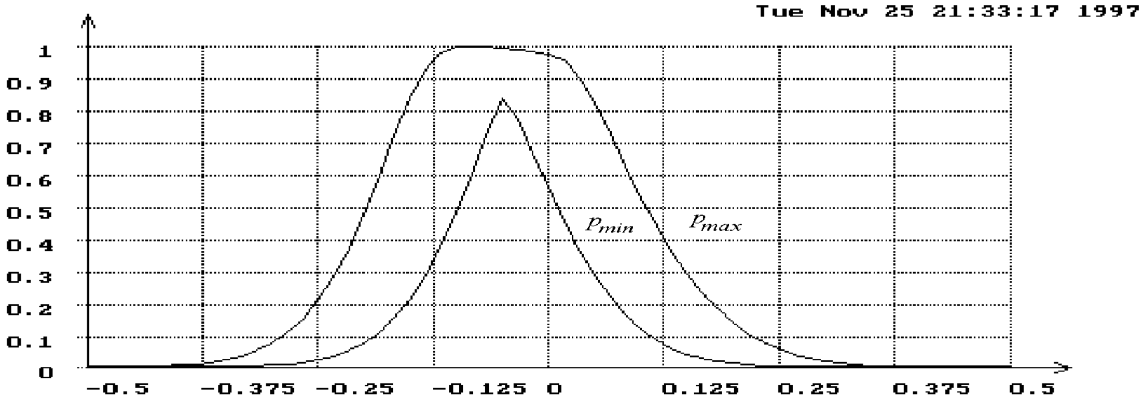
Тогда $\exists \lambda > 0$ и $\exists n_0 : \forall n > n_0$

$$\begin{aligned}
\underline{D_n} &= \max\left\{\sup_x(\underline{F_n}(x) - F(x)), \sup_x(F(x) - \overline{F_n}(x)), 0\right\} \geq \\
&\geq F(x_0) - \overline{F_n}(x_0) = F(x_0) - \overline{F}(x_0) + \overline{F}(x_0) - \overline{F_n}(x_0) \geq \\
&\geq d + O(1/n) \geq \lambda > 0.
\end{aligned}$$

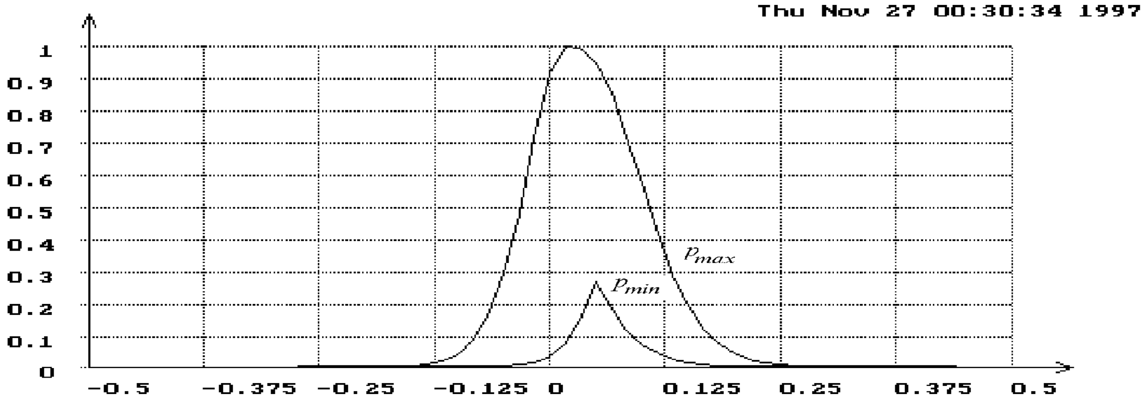
Теорема доказана.

Поведение p_{max} и p_{min} иллюстрирует следующий пример.

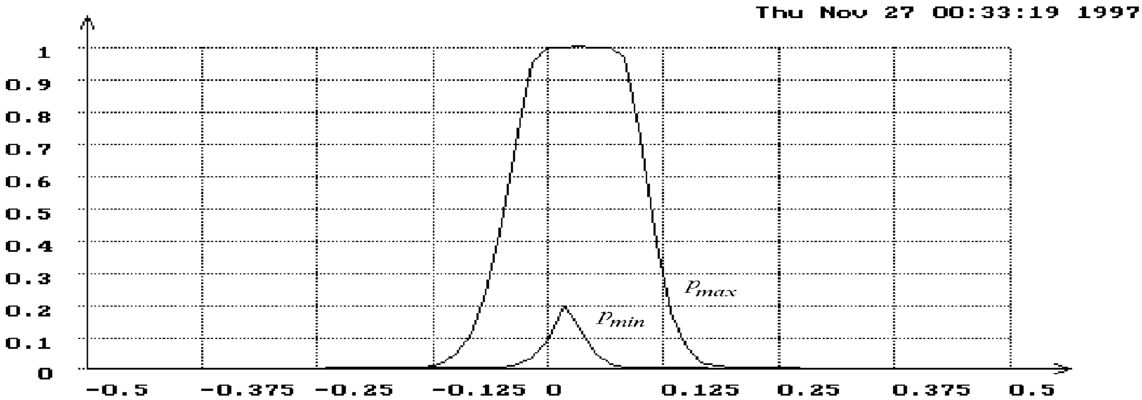
Пример 2.3. *Были сгенерированы три интервальные выборки с абсолютной погрешностью $\varepsilon = 0.05$, подчиненные одному и тому же закону распределения, объёмом* 100, 500 *и* 1000 *наблюдений соответственно. Затем исследовано поведение p_{min} и p_{max} при проверке согласия по критерию Колмогорова с нормальным распределением, у которого параметр σ зафиксирован, а параметр μ изменялся от -0.5 до 0.5* (*см. рис. 2.13*)*. Хорошо видно, что с ростом количества наблюдений верхняя кривая согласия* (p_{max}) *становится более крутой, а нижняя* (p_{min}) *становится ближе к нулю. Это означает, что множество распределений, не отвергаемых по критерию согласия, уменьшается при одном и том же уровне значимости, но неопределённость при принятии решений о согласии для этих распределений увеличивается.*



а.



б.



в.

Рис. 2.13. Согласие интервальных выборок разного объёма
с нормальным распределением по критерию Колмогорова:
(а) 100 наблюдений, (б) 500 наблюдений, (в) 1000 наблюдений

Из доказанной теоремы и рассмотренного примера вытекают два следующих практических соображения. С одной стороны, очевидно, что, опираясь на критерий Колмогорова, в случае интервальной выборки можно отсеять определённое множество законов распределения, не согласующихся с выборкой. С другой стороны, в этой же ситуации невозможно с точностью до параметров идентифицировать закон распределения, наиболее хорошо согласующийся с выборкой, если, например, для двух различных оценок параметров $p_{min} = 0$ и $p_{max} = 1$.

Таким образом, очевидно, что получение точечных оценок параметров распределений по интервальной выборке является процедурой, в значительной степени зависящей от степени оптимизма исследователя относительно соответствия выбранной модели исходным интервальным данным [21]. Действительно, нижнюю границу вероятности согласия можно рассматривать как случай наихудшего расположения точных значений наблюдений в интервалах (“крайний пессимизм”), а верхнюю — как случай наилучшего расположения точных значений наблюдений (“крайний оптимизм”).

При увеличении объёма интервальной выборки для целого множества априори допустимых для описания данной случайной величины распределений длина интервала неопределённости вероятности согласия растёт и стремится к единице. Это значит, что функцию распределения случайной величины, наблюдения которой фиксируются с неустранимой погрешностью, невозможно идентифицировать с использованием данных критериев согласия, даже при очень большом числе экспериментов. Для описания такой случайной величины лучше либо использовать интервальные оценки параметров функции распределения, либо по отдельности аппроксимировать верхнюю и нижнюю границы эмпирической функции распределения.

2.3. Оценивание параметров распределений по интервальным выборкам

Задача оценивания параметров возникает тогда, когда у исследователя есть достаточные основания считать, что все наблюдения в выборке независимы и подчинены одному и тому же закону распределения, который известен с точностью до одного или нескольких параметров.

Пусть $F(x, \theta)$ — функция распределения предполагаемого закона распределения наблюдаемой случайной величины, θ — вектор параметров. По интервальной выборке (2.2) требуется найти оценку параметра

$$\hat{\theta} = \hat{\theta}(\mathbf{X}_n). \tag{2.16}$$

По точечной выборке возможно определение как точечных, так и интервальных оценок параметров. Интервальная оценка параметра по точечной выборке учитывает статистическую неопределенность и с ростом объема выборки сходится к точечной оценке. По интервальной выборке также возможно точечное и интервальное оценивание. Однако, интервальная оценка по интервальной выборке с ростом объема выборки уже не сходится к точечной.

2.3.1. Точечное оценивание

Точечные оценки по интервальной выборке можно получать следующими способами.

1. Можно зафиксировать произвольную точечную выборку $\mathrm{X}_n \in \mathbf{X}_n$ и найти по ней точечную оценку любым известным методом оценивания. Оценка при этом может получиться смещенной, в частности, при оценивании параметров по серединам интервалов группированной выборки известны поправки Шеппарда, устраняющие смещение оценок [58].

2. Можно модифицировать *метод максимального правдоподобия*, который заключается в том, чтобы максимизировать вероятность попадания случайной выборки X_n в n-мерный параллелепипед, заданный интервальной выборкой $\mathbf{X}_n$:

$$P\{X_n \in \mathbf{X}_n\} = P\{x_1 \in [a_1, b_1] \wedge x_2 \in [a_2, b_2] \wedge \ldots \wedge x_n \in [a_n, b_n]\} =$$
$$= \prod_{i=1}^{n} P\{x_i \in [a_i, b_i]\} = \prod_{i=1}^{n} (F(b_i, \theta) - F(a_i, \theta)) \to \max_{\theta}. \tag{2.17}$$

Если в выборке присутствуют точечные наблюдения, то значение $F(b_i) - F(a_i)$ обращается в ноль. Чтобы этого не произошло модифицируем выражение (2.17) следующим образом:

$$\prod_{i=1}^{n} \frac{(F(b_i, \theta) - F(a_i, \theta))}{b_i - a_i} \to \max_{\theta}. \tag{2.18}$$

Действительно, переход к (2.18) оправдан, так как значение θ, при котором достигается максимум (2.17), совпадает со значением θ, оптимизирующим (2.18). Тогда, если i-е наблюдение точечное, т.е. $a_i = b_i = x_i$, то i-й сомножитель в произведении (2.18) будет равен $f(x_i)$.

3. Перибирая все $\mathrm{X}_n \in \mathbf{X}_n$, можно найти множество допустимых значений параметра θ и взять некоторую характерную точку этого множества. Так, например, в методе центра неопределенности [38] в качестве точечной оценки параметра берется центр масс множества допустимых значений.

Можно предложить и другие способы получения точечных оценок (например, точечные оценки “крайнего оптимиста” и “крайнего пессимиста”). Однако из теоремы 2.1 следует, что какую бы точечную оценку мы не взяли, при проверке гипотезы о согласия с интервальной выборкой $\mathbf{X}_n$ по критерию Колмогорова, с точки зрения “крайнего пессимиста” можно отвергнуть любую простую гипотезу при увеличении количества наблюдений. Кроме того, при точечном оценивании по интервальной выборке могут нарушаться статистические асимптотические свойства оценок несмещенности, эффективности, состоятельности, вследствие чего теряет смысл бесконечное увеличение объема выборки, так как статистическая погрешность оценки оказывается существенно меньше погрешности, определяемой погрешностью фиксации наблюдений.

Интервальный характер выборки обуславливает для любой точечной оценки существование некоторого интервала значений, которому она принадлежит и который определяется погрешностью наблюдений и методом оценивания. Любая точечная оценка, полученная по интервальной выборке дает лишь точку, принадлежащую этому интервалу, но не содержит информации о самом интервале.

2.3.2. Интервальное оценивание

Введем понятие нестатистической интервальной оценки.

Определение 2.3. *Пусть имеется некоторая точечная оценка параметра по точечной выборке $\theta^* = \theta^*(\mathrm{X}_n)$. Интервальной оценкой параметра θ, порожденной точечной оценкой $\theta^*(\mathrm{X}_n)$, по интервальной* *выборке будем называть интервал $[\underline{\theta^*}, \overline{\theta^*}]$, границы которого определяются из соотношений:*

$$\underline{\theta^*} = \min_{\mathrm{X}_n \in \mathbf{X}_n} \theta^*(\mathrm{X}_n), \quad \overline{\theta^*} = \max_{\mathrm{X}_n \in \mathbf{X}_n} \theta^*(\mathrm{X}_n). \tag{2.19}$$

Замечание 2.3. Вообще говоря, всякая точечная оценка, построенная по случайным наблюдениям, является случайной величиной. Следовательно, границы интервальной оценки, порожденной точечной оценкой, будут также случайными. Доверительный интервал, содержащий истинное значение параметра с заданной вероятностью, получается объединением доверительных интервалов, построенных по всем точечным выборкам из $\mathbf{X}_n$.

Приведем несколько простых примеров вычисления интервальных оценок параметров по интервальной выборке.

Пример 2.4. *Оценивание параметра сдвига нормального распределения с плотностью $f(x, \mu) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(x-\mu)^2}$. Оценка максимального правдоподобия параметра μ по точечной выборке имеет вид:*

$$\mu^* = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

В соответствии с (2.19) для интервальной оценки по интервальной выборке имеем

$$\underline{\mu^*} = \min_{x_i \in [a_i, b_i]} \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{i=1}^{n} a_i,$$

$$\overline{\mu^*} = \max_{x_i \in [a_i, b_i]} \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{i=1}^{n} b_i.$$

Пример 2.5. *Оценивание параметра масштаба нормального распределения с плотностью $f(x, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2\lambda^2}}$. Аналогично, оценка максимального правдоподобия параметра λ имеет вид:*

$$\lambda^* = \frac{1}{n} \sum_{i=1}^{n} x_i^2.$$

Тогда по (2.19)

$$\underline{\lambda^*} = \min_{x_i \in [a_i, b_i]} \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \frac{1}{n} \left\{ \sum_{i=1}^{n} b_i^2 I_{b_i \le 0} + \sum_{i=1}^{n} a_i^2 I_{a_i \ge 0} \right\},$$

$$\overline{\lambda^*} = \max_{x_i \in [a_i, b_i]} \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \frac{1}{n} \left\{ \sum_{i=1}^{n} a_i^2 I_{b_i \le 0} + \sum_{i=1}^{n} b_i^2 I_{a_i \ge 0} + \sum_{i=1}^{n} \max\{a_i^2, b_i^2\} I_{a_i < 0 < b_i} \right\},$$

где $I_A = 1$, если условие A выполнено и $I_A = 0$ в противном случае.

В этих двух примерах интервальные оценки имют аналитически простой вид, но в большинстве случаев определение интервальной оценки в явном виде невозможно. Рассмотрим общие подходы к определению интервальных оценок, порожденных классами точечных L-оценок, -оценок и D-оценок.

2.3.2.1. Интервальные L-оценки

L-оценки формируются как линейные комбинации порядковых статистик:

$$\theta^* = \sum_{i=1}^{n} c_i x_{(i)}, \quad \sum_{i=1}^{n} c_i = 1, \tag{2.20}$$

где $x_{(i)}$ — i-я порядковая статистика. Чтобы найти интервальную оценку, порожденную оценкой вида (2.20), нужно определить интервалы, в границах которых могут находиться порядковые статистики.

Воспользуемся полученными в п.2.2.4 границами для каждого члена вариационного ряда

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)}, \tag{2.21}$$

построенного по интервальной выборке $\mathbf{X}_n$:

$$\underline{x_{(i)}} = \inf\{y \mid y = \overline{F_n^{-1}}(i/n)\}, \ \overline{x_{(i)}} = \sup\{y \mid y = \underline{F_n^{-1}}(i/n)\}. \tag{2.22}$$

Подставим (2.20) в (2.19) и, учитывая (2.21-2.22), получим две задачи линейного программирования размерности n, решение которых даст нам верхнюю и нижнюю границы L-оценки.

Пример 2.6. *К классу L-оценок относится часто используемая для оценивания параметра положения выборочная медиана:*

$$m = \begin{cases} x_{((n+1)/2)}, & n - \ ; \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right), & n - \ . \end{cases}$$

Соответствующая интервальная оценка будет иметь вид:

$$\underline{m} = \begin{cases} \underline{x_{((n+1)/2)}}, & n - \ ; \\ \frac{1}{2}\left(\underline{x_{(n/2)}} + \underline{x_{(n/2+1)}}\right), & n - \ . \end{cases}$$

$$\overline{m} = \begin{cases} \overline{x_{((n+1)/2)}}, & n - \ ; \\ \frac{1}{2}\left(\overline{x_{(n/2)}} + \overline{x_{(n/2+1)}}\right), & n - \ . \end{cases}$$

2.3.2.2. Интервальные M-оценки

M-оценки получаются в результате минимизации функционала [59]:

$$\theta^*(\mathrm{X}_n) = \arg\min_\theta M(\mathrm{X}_n, \theta) = \arg\min_\theta \sum_{i=1}^n \rho(x_i, \theta), \tag{2.23}$$

где $\rho(x_i, \theta)$ — функция потерь.

Пусть

$$\underline{M}(\theta) = \min_{\mathrm{X}_n \in \mathbf{X}_n} M(\mathrm{X}_n, \theta) = \sum_{i=1}^n \min_{x_i \in [a_i, b_i]} \rho(x_i, \theta), \tag{2.24}$$

$$\overline{M}(\theta) = \max_{\mathrm{X}_n \in \mathbf{X}_n} M(\mathrm{X}_n, \theta) = \sum_{i=1}^n \max_{x_i \in [a_i, b_i]} \rho(x_i, \theta) \tag{2.25}$$

— верхняя и нижняя границы $M(X_n, \theta)$ по всем возможным точечным выборкам $\mathrm{X}_n \in \mathbf{X}_n$. Тогда

$$\underline{M}(\theta) \leq M(\mathrm{X}_n, \theta) \leq \overline{M}(\theta), \ \forall \mathrm{X}_n \in \mathbf{X}_n, \tag{2.26}$$

и, следовательно,

$$\min_\theta \underline{M}(\theta) \leq \min_\theta M(\mathrm{X}_n, \theta) \leq \min_\theta \overline{M}(\theta), \ \forall \mathrm{X}_n \in \mathbf{X}_n. \tag{2.27}$$

Обозначим

$$\overline{M} = \min_\theta \overline{M}(\theta) \tag{2.28}$$

и

$$\underline{M} = \min_\theta \underline{M}(\theta). \tag{2.29}$$

максимальное и минимальное значение $M(\theta)$.

Зафиксируем какую-нибудь точечную выборку $\mathrm{X}_n \in \mathbf{X}_n$. Пусть минимум функционала $M(\mathrm{X}_n, \theta)$ достигается при каком-то значении θ^*:

$$M(\mathrm{X}_n, \theta^*) = \min_\theta M(\mathrm{X}_n, \theta). \tag{2.30}$$

Тогда из (2.26) и (2.27) следует, что

$$\underline{M}(\theta^*) \leq M(\mathrm{X}_n, \theta^*) \leq \overline{M}. \tag{2.31}$$

Отсюда

$$\underline{M}(\theta^*) \leq \overline{M}. \tag{2.32}$$

Таким образом мы получили следующую теорему.

Теорема 2.2. *Множество*

$$T = \{\theta \mid \underline{M}(\theta) \le \overline{M}\} \tag{2.33}$$

содержит все возможные точечные M-оценки (2.23) при $\mathrm{X}_n \in \mathbf{X}_n$.

Точечную оценку, при которой достигается минимум (2.28)

$$\min_{\theta} \overline{M}(\theta) = \overline{M}(\theta_{.}), \tag{2.34}$$

можно назвать оценкой "крайнего пессимиста", а оценку, при которой достигается минимум (2.29)

$$\min_{\theta} \underline{M}(\theta) = \underline{M}(\theta_{.}), \tag{2.35}$$

можно назвать оценкой "крайнего оптимиста" (см. рис. 2.14).

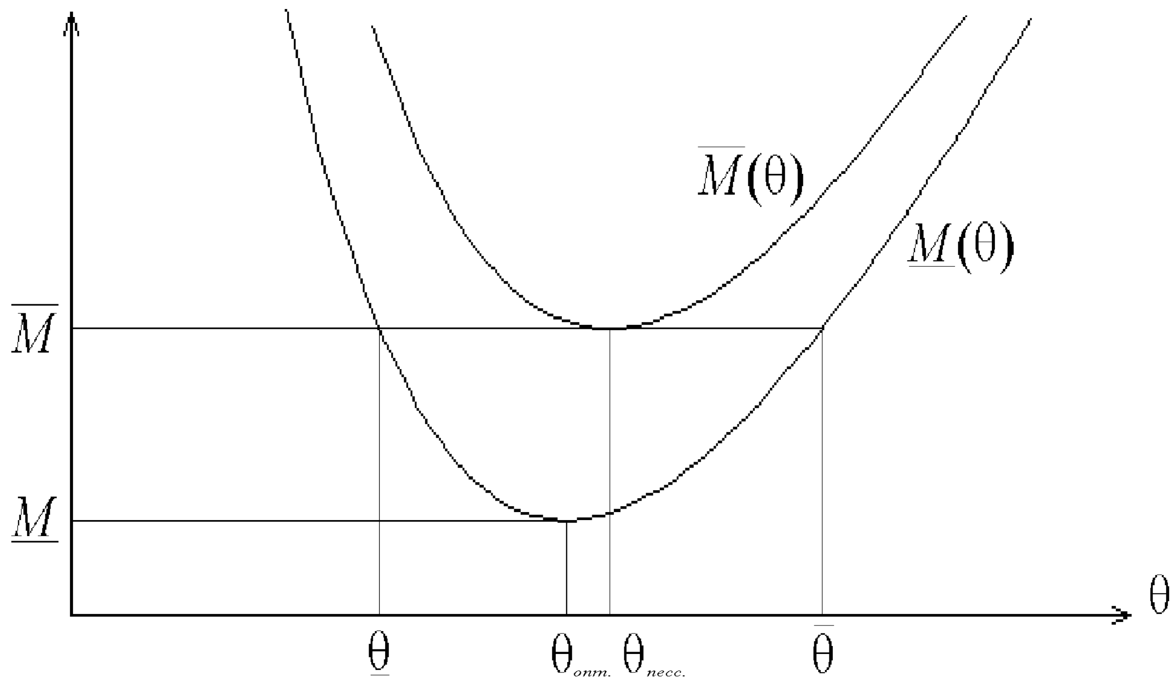


Рис. 2.14. Нахождение интервальной M-оценки

Теорема 2.2 описывает допустимое множество оценок параметров, но не дает метода его построения. Рассмотрим случай, когда θ — скалярный параметр и функционал $M(X_n, \theta)$ является выпуклым. Тогда множество T представляет собой замкнутый интервал, границы которого определяются из уравнения $\underline{M}(\theta) = \overline{M}$ (см. рис. 2.14). Это уравнение можно свести к двум задачам нелинейного программирования:

$$\underline{\theta^*} = \arg\min_{\theta<\theta_.}(\underline{M}(\theta) - \underline{M})^2, \tag{2.36}$$

$$\overline{\theta^*} = \arg\min_{\theta>\theta_.}(\underline{M}(\theta) - \underline{M})^2. \tag{2.37}$$

Частным случаем M-оценок являются оценки максимального правдоподобия, когда функция потерь имеет вид

$$\rho(x, \theta) = -\ln f(x, \theta),$$

где $f(x, \theta)$ — функция плотности распределения.

2.3.2.3. Интервальные MD-оценки

D-оценки получаются, если в выражении (2.23) в качестве минимизируемого функционала взять расстояние между эмпирической функцией распределения $F_n(x)$ и теоретической $F(x, \theta)$. Интервальные MD-оценки получаются аналогично интервальным M-оценкам, если учесть выражение (2.4), задающее границы возможного расположения эмпирической функции распределения. В качестве меры близости эмпирической и теоретической функций распределения можно использовать статистики непараметрических критериев согласия. Оценки границ $\underline{M}(\theta)$ и $\overline{M}(\theta)$ для статистик критериев Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса получены в п.2.2.2–2.2.5.

2.3.3. Свойства интервальных оценок

Различные точечные оценки параметров порождают в общем случае различные интервальные оценки. Естественным критерием для сравнения различных интервальных оценок является длина интервала $\Delta\theta = \overline{\theta} - \underline{\theta}$.

Пример 2.7. *Выборка, состоящая из 200 интервальных наблюдений, была смоделирована следующим образом: середины интервалов определялись по реализациям трех случайных величин ζ_i, ξ_i и y_i*

$$x_i = (1 + \zeta_i)y_i + \xi_i,$$

где ζ_i *— распределена равномерно на интервале* $[-\delta, \delta]$, ξ_i *— распределена равномерно на интервале* $[-\varepsilon, \varepsilon]$, y_i *— распределена по нормальному закону с параметром сдвига* $\mu = 0$ *и параметром масштаба* $\sigma = 1$. *Границы интервалов определялись по формулам*

$$a_i = (1-\delta)|x_i| - \varepsilon, \quad b_i = (1+\delta)|x_i| + \varepsilon, \tag{2.38}$$

где $\varepsilon = 0.01$ *— абсолютная погрешность наблюдения,* $\delta = 0.01$ *— относительная погрешность наблюдения. Интервальная гистограмма для полученной выборки приведена на рис. 2.15. По интервальной выборке оценивался параметр масштаба нормального распределения по трем методам: максимального правдоподобия, минимума статистики Колмогорова и минимума статистики* ω^2 *Мизеса. Полученные точечные и интервальные оценки приведены в таблице 2.1. Наихудшим из трех методов оказался метод максимального правдоподобия, так как длина интервала оценки получилась наибольшей, а наилучшим оказался метод минимума статистики Колмогорова. Отметим, что все три интервальные оценки накрыли "истинное" значение параметра, то есть то значение параметра, при котором моделировалась выборка.*

Свойства интервальных оценок во многом зависят от свойств точечных оценок, по которым они строятся. Если точечная оценка робастна, т.е. "нечувствительна" к наличию аномальных наблюдений в выборке [1], то следует ожидать, что соответствующая ей интервальная оценка будет иметь меньшую длину, чем интервальная оценка, порожденная точечной оценкой, не обладающей свойством робастности. Исследование функций влияния Хампеля показало, что оценки максимального правдоподобия по негруппированным данным для большинства используемых на практике распределений не являются робастными [60]. Этим можно объяснить тот факт, что в рассмотренном примере метод максимального правдоподобия оказался наихудшим.

Длина интервальной оценки слабо зависит от объема выборки и не стремится к нулю при увеличении числа наблюдений. Нетрудно показать, что если интервальные наблюдения получены по (2.38), а интервальная оценка максимального правдоподобия параметра масштаба нормального распределения вычисляется так же как в примере 2.5, то $\Delta\theta \geq \frac{8\varepsilon^2}{1-\delta} > 0$.

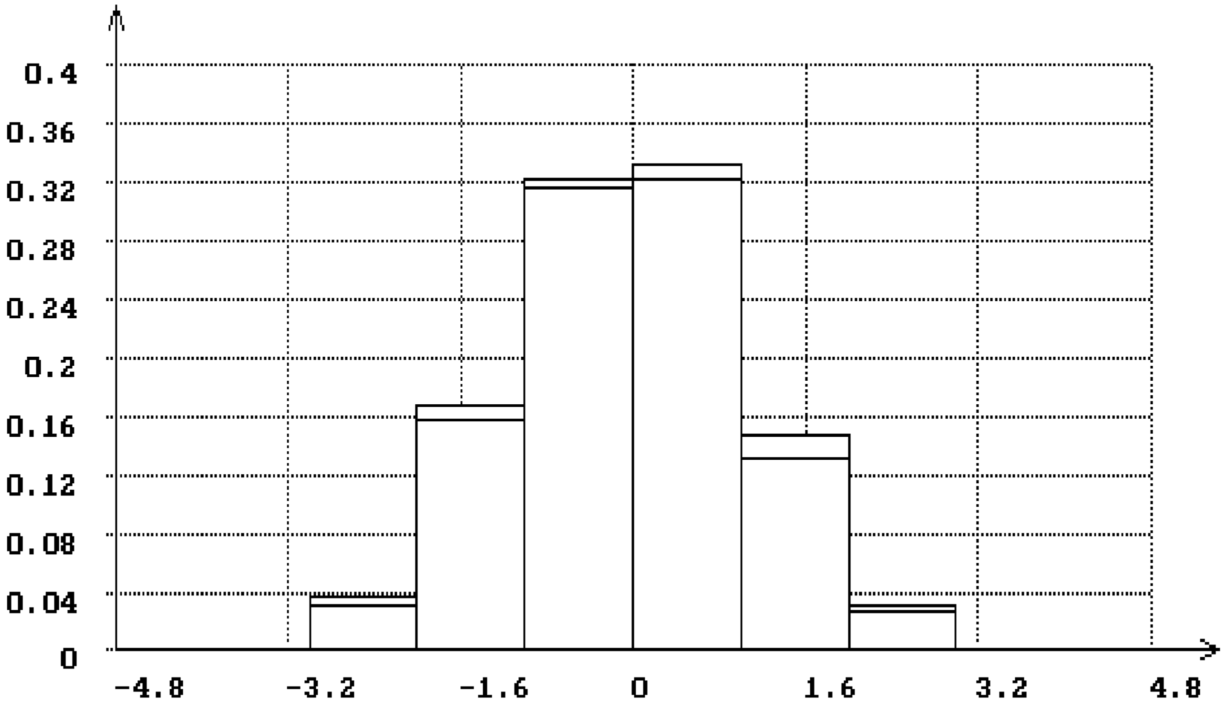


Рис. 2.15. Интервальная гистограмма для интервальной выборки объемом 200 наблюдений

Таблица 2.1

Оценивание параметра масштаба нормального распределения по интервальной выборке объемом 200 наблюдений

Метод оценивания	Класс оценки	Точечные оценки			Интервальная оценка	Длина интервала
		по серединам интервалов	“крайнего оптимиста”	“крайнего пессимиста”		
Метод максимального правдоподобия	M-оценка	1.0680	1.0493	1.0867	$[0.8798, 1.2811]$	0.4013
Метод минимума статистики Колмогорова	MD-оценка	1.0276	1.0272	1.0278	$[0.9774, 1.0946]$	0.1172
Метод минимума статистики ω^2 Мизеса	MD-оценка	1.0897	1.0809	1.0965	$[0.9544, 1.2323]$	0.2779

Таким образом, по интервальной выборке можно находить как точечные, так и интервальные оценки. Интервальные оценки параметров отражают интервальную неопределенность в задании исходных данных. В случае одного параметра интервальное оценивание сводится к решению задач линейного (при поиске L-оценок) и нелинейного (при поиске M-оценок и MD-оценок) программирования. Длина интервальной оценки существенно зависит от длин интервалов наблюдений и от метода оценивания и слабо зависит от объема выборки.

Выводы

В главе рассмотрены вопросы оценивания параметров распределений и проверки гипотез о согласии по интервальной выборке. Получены формулы для проверки гипотез о согласии по непараметрическим критериям и предложены алгоритмы для вычисления точечных и интервальных оценок параметров распределений в случае, когда исходные данные имеют интервальное представление. Все полученные формулы и алгоритмы реализованы в объектно-ориентировнной системе статистического анализа, описанной в главе 3.

На основании результатов, полученных в этой главе, можно сделать следующие выводы.

1. Статистический анализ интервальных наблюдений требует большего объема вычислений по сравнению с обработкой точечных выборок.

2. Когда выборка имеет небольшой объем, то основное значение имеет статистическая погрешность; когда выборка велика, следует учитывать метрологическую погрешность исходных данных.

3. Так как увеличение объема выборки уменьшает только статистическую погрешность, но не уменьшает погрешность, вызванную ошибками наблюдений, то существет некоторое пороговое значение объема выборки, когда дальнейшее увеличение числа наблюдений не увеличивает точности в определении параметров распределения.

3. ОБЪЕКТНО-ОРИЕНТИРОВАННАЯ СИСТЕМА СТАТИСТИЧЕСКОГО АНАЛИЗА

Работа над системой статистического анализа наблюдений одномерных непрерывных случайных величин была начата Лемешко в 1988 г. Достоинствами разработанной системы [51] были следующие возможности:

- статистический анализ группированных, негруппированных (точечных) и частично группированных наблюдений;
- оценивание параметров по методу максимального правдоподобия 26-ти наиболее часто встречающихся на практике законов распределения;
- проверка гипотез о согласии по шести критериям: отношения правдоподобия, χ^2 Пирсона, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса и принятие решения по их совокупности.
- использование *асимптотически оптимального группирования* [48] при проверке гипотез по критериям отношения правдоподобия и χ^2 Пирсона, обеспечивающее максимальную мощность этих критериев при близких альтернативных гипотезах [51];
- робастное оценивание параметров, связанное с группированием исходной выборки и оцениванием параметров по сгруппированной выборке;
- моделирование псевдослучайных выборок, распределенных по включеным в систему законам.

Следует отметить, что система [51] по своим возможностям относится к классу узкоспециализированных наукоемких программных продуктов и не имеет аналогов среди рассмотренных в [61, 62, 63, 64, 65, 66, 67, 68] статистических пакетов.

Однако практическое использование системы [51] показало, что несмотря на существенный шаг вперед, она сохранила ряд недостатков, присущих и другим программным системам статистического анализа:

- включенных в систему 26-ти законов распределения иногда оказывалось недостаточным для описания реально встречающихся данных;

- при использовании непараметрических критериев согласия Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса не учитывался факт предварительного оценивания параметров, изменяющий предельное распределение статистик критериев;

- при проверке гипотез о согласии по группированным и частично группированным данным использование непараметрических критериев было невозможным;

- не было возможности обработки интервальных наблюдений.

С 1995 года наряду с сопровождением и постоянной модификацией старой системы на основе вновь полученных результатов [55, 69, 70] велась разработка новой программной системы [71, 72, 73, 74, 75], позволяющей устранить перечисленные выше недостатки за счет:

- расширения множества параметрических моделей законов распределения смесями и усеченными законами распределения [76, 77, 78];

- моделирования предельных законов распределения статистик критериев согласия в случае оценивания одного или нескольких параметров [79, 80, 81, 82, 83];

- развития интервальной статистики, рассмотренной в главе 2, позволяющей оценивать параметры и проверять гипотезы о согласии по интервальным выборкам [56, 57].

Развитие системы потребовало коренным образом изменить структуру программного обеспечения на базе объектно-ориентированного программирования [84, 73].

3.1. Объектно-ориентированное программирование и его приложение к статистике

В связи с усложнением задач, решаемых с помощью вычислительной техники, классические процедурные языки программирования (Фортран, Си, Паскаль) оказались не в состоянии обеспечить необходимый уровень надежности вычислений и удобства разработки крупных программных систем. На смену им пришли языки объектно-ориентированного программирования (ООП), позволяющие моделировать объекты с помощью иерархической системы классов. Каждый класс может содержать набор данных и методов их обработки. Такая взаимосвязь данных и операций позволяет повысить надежность программного обеспечения за счет того, что данные можно сделать *защищенными*, а доступ к ним осуществлять только с помощью функций, и, таким образом, исключить случайное использование одних и тех же переменных в программе. Далее, помимо надежности, повышается и наглядность программирования — в отличие от обычной переменной переменная-объект класса содержит и методы работы с ней. Обычные операции ("=", "+", "−" и др.) допускают *переопределение*, что позволяет, например, естественным образом записывать матричные операции. Наконец, одни классы могут быть *производными* от других классов. В базовых классах реализуются (или декларируются) самые общие свойства целой группы объектов, а в производных классах происходит конкретизация или переопределение этих свойств.

ООП нашло широкое применение на практике. Созданы специальные библиотеки классов, так, например, в реализованной программной системе статистического анализа используется две библиотеки классов: "Turbo Vision— для создания пользовательского интерфейса и "Matrix— для работы с матрицами и векторами.

Статистическая обработка данных представляет собой достаточно сложный набор взаимосвязанных задач, и поэтому использование ООП в этом случае является очень эффективным, а иногда и просто незаменимым. ООП упрощает расширение системы и добавление новых функций в программу, и, таким образом, позволяет задействовать большие коллективы разработчиков программного обеспечения.

3.2. Иерархия классов системы

При описании *случайных явлений* существует два аксиоматических подхода: система аксиом Колмогорова и система аксиом Мизеса. Аксиоматика Колмогорова лежит в основе вероятностных моделей и ее базовыми понятиями являются случайное событие и вероятность. Она может рассматриваться как специальное применение общей теории аддитивных функций множеств.

В основе аксиоматики Мизеса лежит частотный принцип. Статистические модели строятся предельным переходом от эмпирических характеристик при неограниченном увеличении объема экспериментальных данных, числа испытаний. В этом случае понятие вероятности не является основным и выражается через другие понятия. Теория вероятностей не может сказать как найти вероятности того или иного события, она лишь позволяет вычислить вероятность одного события через вероятности других событий. Математическая статистика дает возможность найти вероятности события, исходя из ряда экспериментальных наблюдений.

Обобщая эти два подхода к описанию случайных явлений, можно выделить два основных объекта с которыми должна оперировать система статистического анализа. С точки зрения теории вероятностей базовым является объект “Распределение” (класс “Distribution”), с точки зрения статистики базовым является объект “Наблюдение” (класс “Observation”). Совокупность наблюдений есть объект “Выборка” (класс “Sample”). В то же время, выборку можно рассматривать как эмпирическое распределение, и, таким образом, объект "Выборка"должен наследовать все свойства распределений. Совокупность из двух объектов — выборки и распределения — образует объект “Статистика” (класс “Stat”). Совокупность из n выборок и m распределений образует объект “Идентификация” (класс “Ident”).

Базовый класс “Distribution” является *абстрактным*, т.е. основные функции, характеризующие распределение, являются *чисто виртуальными* и определяются в производных классах.

На рис. 3.1 приведена иерархия базовых классов системы.

3.3. Представление исходных данных

В главе 1 было показано, что наиболее общей формой представления одномерного наблюдения являются нечеткие и размытые числа. Поскольку в данной работе задача статистического анализа нечетких и размытых данных не ставилась, то было реализовано интервальное представление наблюдений.

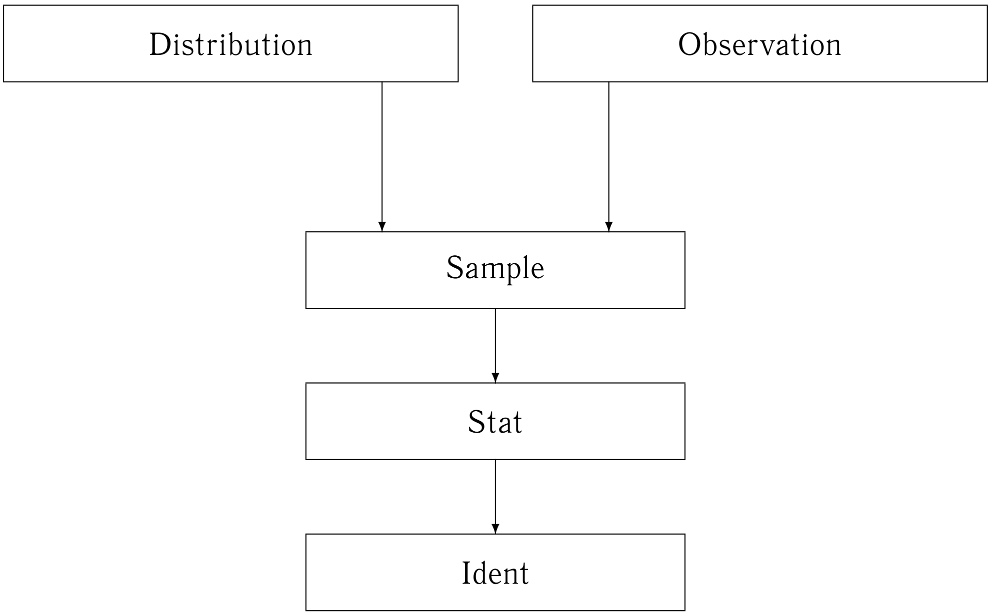


Рис. 3.1. Иерархия базовых классов

3.3.1. Интервальное наблюдение

Интервальное наблюдение (класс "Observation") характеризуется номером наблюдения, левой границей, правой границей и числом точечных наблюдений, сконцентрированных в данном интервале. Доступ к данным осуществляется через функции, поэтому в дальнейших версиях системы данный класс можно переопределить так, что получаемый интервал будет зависеть от уровня значимости (в случае представления наблюдения размытым числом) или от α-уровня (в случае представления наблюдения нечетким числом). Над наблюдениями определены следующие функции: вычисление середины интервала, вычисление длины интервала, определение типа наблюдения (точечное, интервальное, бесконечное слева, бесконечное справа).

3.3.2. Интервальная выборка

Интервальная выборка (класс “Sample”) представляет собой массив объектов типа “Observation”. Так как класс “Sample” является производным от класса “Distribution”, то в нем должны быть переопределены основные функции, характеризующие распределение. Вообще говоря, выборке соответствует дискретное распределение, но работать с ним трудно, так как эмпирическая функция распределения не является дифференцируемой. Поэтому эмпирическое распределение аппроксимируется с помощью ядерных функций (см. п. 3.4.5).

Над выборкой определены следующие функции: создание выборки, подчиненной заданному закону распределения, считывание выборки из текстового файла, сортировка выборки (по левой границе, по правой границе, по серединам интервалов), вычисление выборочных характеристик: порядковых статистик, медианы, среднего, среднеквадратического отклонения (в случае, если выборка точечная).

3.3.3. Преобразование выборки

Класс "Sample"является базовым для целого ряда производных классов, выполняющих различные преобразования над выборкой. Эти преобразования аналогичны операциям над распределениями (см. п. 3.4.1). На рис. 3.2. приведена иерархическая схема их взаимосвязи.

3.3.4. Моделирование псевдослучайной выборки

Моделирование псевдослучайной выборки, подчиненной заданному распределению с функцией распределения $F(x)$, осуществляется по методу обратной функции [85]:

$$x_i = F^{-1}(\alpha_i),\ i = 1, \dots, n;$$

где α_i — случайная величина, подчиненная равномерному распределению на интервале $[0, 1]$.

Обратная функция $F^{-1}(y)$ достаточно быстро находится по методу Ньютона.

3.4. Представление вероятностной модели

Для описания вероятностной модели служит абстрактный класс "Distribution". Основные функции, выполняемые над абстрактным распределением: вычисление информационной матрицы Фишера, вычисление обратной функции распределения, генерирование случайного числа, нахождение точек равновероятного и асимптотически оптимального группирования. Распределение задается функцией распределения $F(x, \Theta)$, плотностью $f(x, \Theta)$ и их производными по параметрам $\frac{\partial F(x,\Theta)}{\partial \theta_i}, \frac{\partial f(x,\Theta)}{\partial \theta_i}, i = 1, \dots, n$, которые необходимы для вычисления информационной матрицы Фишера по группиро-

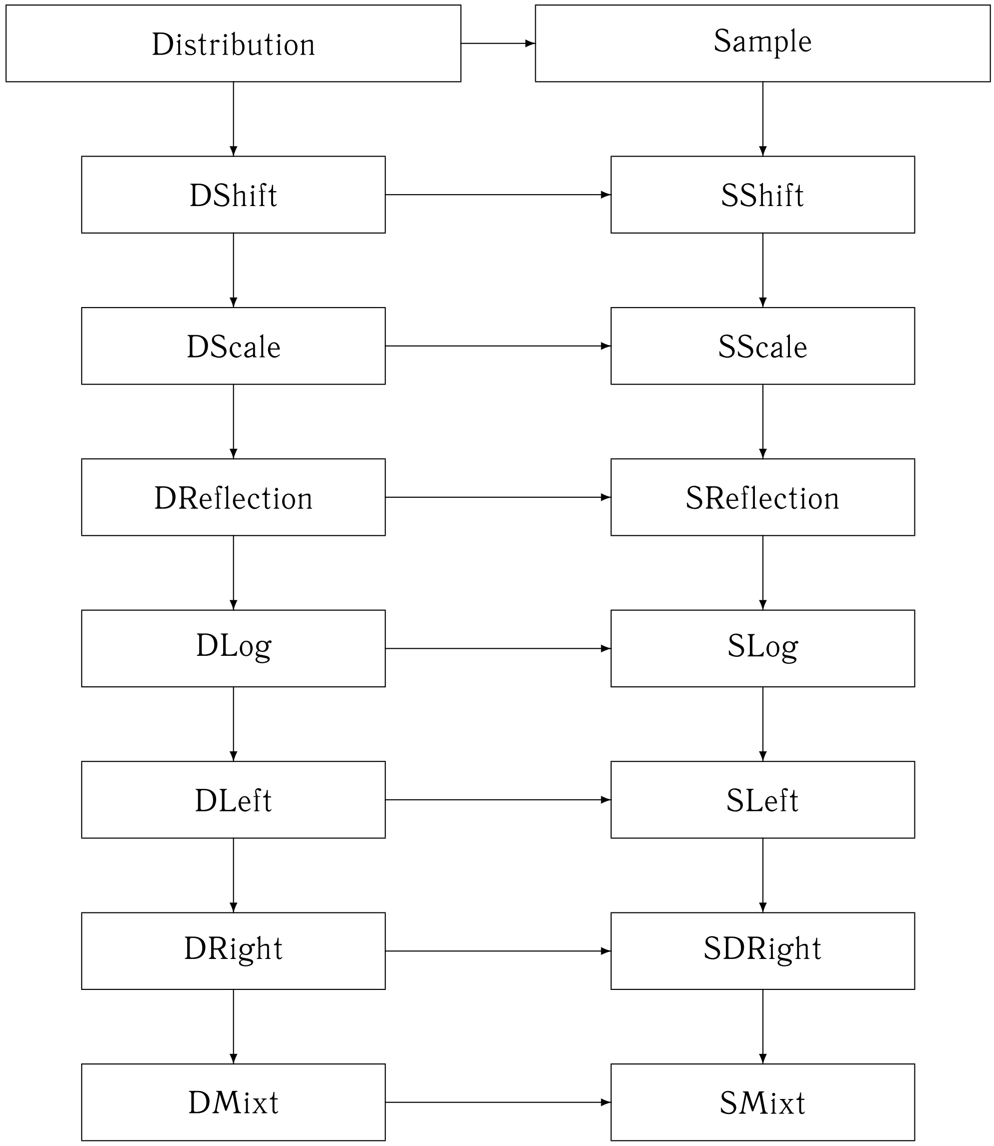


Рис. 3.2. Взаимосвязь операций над распределениями и преобразований выборки

ванным и по негруппированным данным, а также при поиске оценок параметров по методу сопряженных градиентов. Могут понадобиться также вторые производные по параметрам, если при поиске использовать более быстро сходящиеся методы второго порядка. Учитывая, что над распределениями могут быть применены операции сдвига и масштаба (см. п. 3.4.1.1 и 3.4.1.2), для вычисления первых и вторых производных по параметрам сдвига и масштаба (см. таблицы 3.2 и 3.3) нужно знать $\frac{\partial f(x,\Theta)}{\partial x}$, $\frac{\partial^2 f(x,\Theta)}{\partial x \partial \theta_i}$ и $\frac{\partial^2 f(x,\Theta)}{\partial x^2}$.

От абстрактного класса "Distribution" образовано несколько производных классов, перечисленных в таблице 3.1. Иерархия производных классов от класса "Distribution" приведена на рис. 3.2 и 3.3.

3.4.1. Операции над распределениями

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 $f(x,\Theta)$, и известны их производные по x и по параметрам θ_i, θ_j, где $i,j=1,\ldots,n$:

$$\begin{array}{lll} F(x,\Theta), & \frac{\partial F(x,\Theta)}{\partial \theta_i}, & \frac{\partial^2 F(x,\Theta)}{\partial \theta_i \partial \theta_j}; \\ f(x,\Theta), & \frac{\partial f(x,\Theta)}{\partial \theta_i}, & \frac{\partial^2 f(x,\Theta)}{\partial \theta_i \partial \theta_j}; \\ \frac{f(x,\Theta)}{\partial x}, & \frac{\partial^2 f(x,\Theta)}{\partial x \partial \theta_i}; & \\ \frac{\partial^2 f(x,\Theta)}{\partial x^2}. & & \end{array} \tag{3.1}$$

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

$$F_\eta(y) = F_\xi(g^{-1}(y)), \tag{3.2}$$

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

$$F_\eta(y) = 1 - F_\xi(g^{-1}(y)). \tag{3.3}$$

Преобразование $g(x)$ случайной величины ξ в случайную величину η может иметь свои параметры. Для всех операций над распределениями требуется найти аналитический вид для производных функции распределения (3.1) и область определения случайной величины η.

Таблица 3.1

Производные классы от “Distribution”

Наименование	Базовый класс	Описание
Sample	Distribution	Эмпирическое распределение
DShift	Distribution	Сдвиг
DScale	Distribution	Масштаб
DReflection	Distribution	Зеркальное отражение
DLeft	Distribution	Усечение слево
DRight	Distribution	Усечение справа
DLog	Distribution	Логарифмирование
DMixt	Distribution	Смесь
DMult	Distribution	Произведение
DGamma	Distribution	Семейство гамма-распределений
DBeta	Distribution	Семейство бета-распределений
DJ	Distribution	Семейство распределений Джонсона
D0	Distribution	Равномерное распределение
D1	Distribution	Экспоненциальное распределение
D2	Distribution	Полунормальное распределение
D3	Distribution	Распределение Релея
D4	Distribution	Распределение Максвелла
D5	DGamma	Распределение модуля n-мерного нормального распределения
D6	Distribution	Распределение Парето
D7	Distribution	Распределение Эрланга
D8	Distribution	Распределение Лапласа
D9	Distribution	Нормальное распределение
D10	Distribution	Логарифмически (ln) нормальное распределение
D11	Distribution	Логарифмически (lg) нормальное распределение

Продолжение табл. 3.1

Наименование	Базовый класс	Описание
D12	Distribution	Распределение Коши
D13	Distribution	Логистическое распределение
D14	Distribution	Распределение Вейбулла
D15	Distribution	Распределение минимального значения
D16	Distribution	Распределение максимального значения
D17	DGamma	Обобщенное распределение минимального значения
D18	Distribution	Распределение Накагами
D19	Distribution	Гамма-распределение
D20	DBeta	Бета-распределение 1-го рода
D21	DBeta	Бета-распределение 2-го рода
D22	DBeta	Бета-распределение 3-го рода
D23	DJ	Распределение Sb-Джонсона
D24	DJ	Распределение Sl-Джонсона
D25	DJ	Распределение Su-Джонсона
D26	Distribution	Двустороннее экспоненциальное распределение
D27	Distribution	H-распределение
D28	DGamma	Г-распределение
D29	DBeta	L-распределение

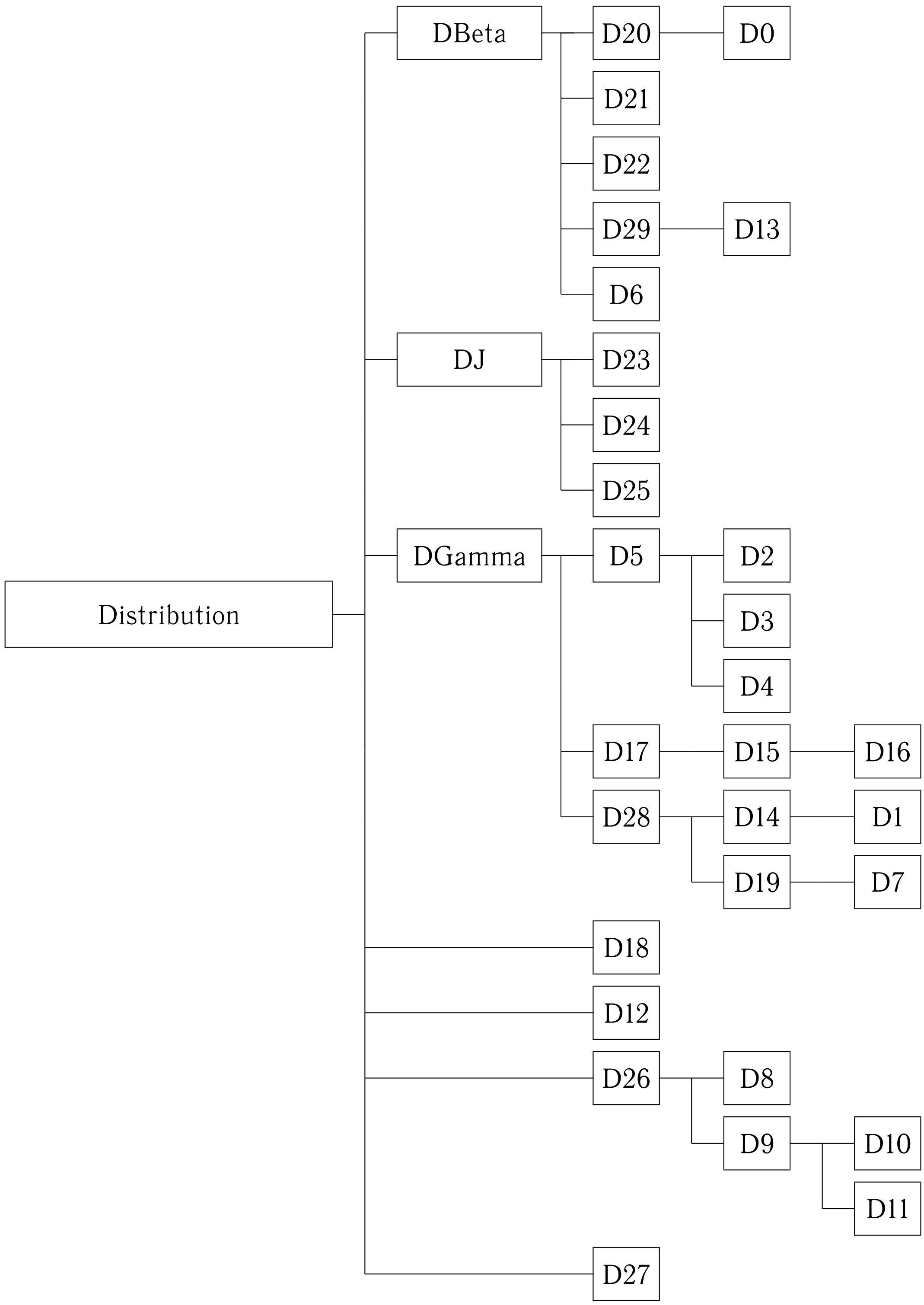


Рис. 3.3. Взаимосвязь классов распределений

Замечание 3.1. 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 $F\left(\frac{x}{\lambda}-\mu\right)$.

3.4.1.1. 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 согласно (3.2) имеет вид:

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

Производные функции распределения $G(x, \theta_1, \theta_2, \ldots, \theta_n, \mu)$ приведены в таблице 3.2, где $u = x - \mu$.

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

Пример сдвига распределения показан на рис. 3.4.

3.4.1.2. 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 согласно (3.2) имеет вид:

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

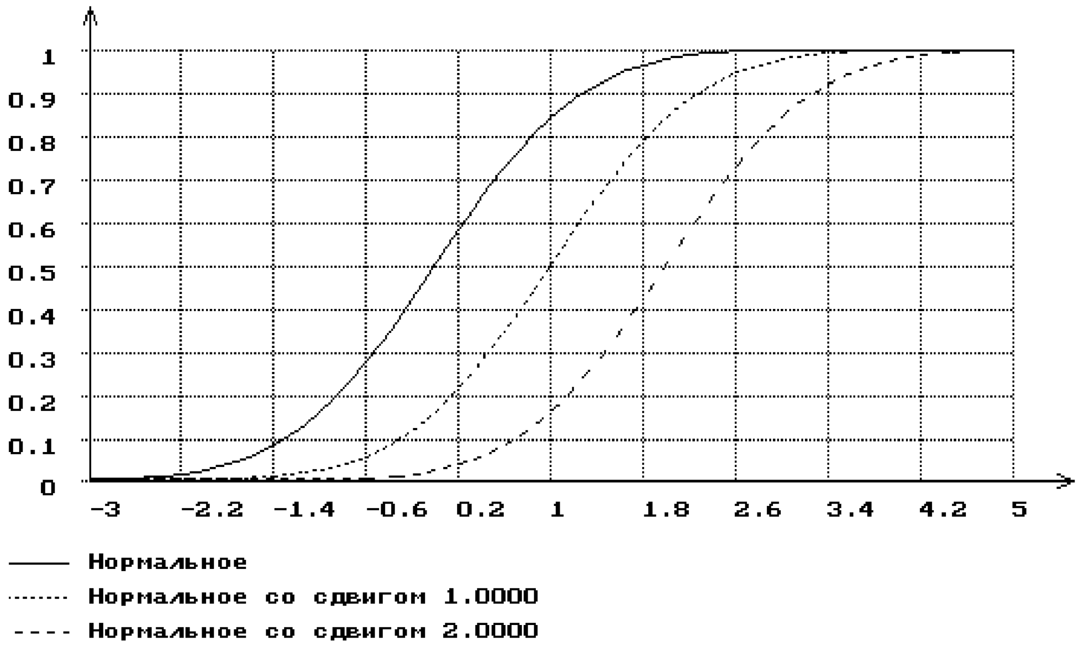


Рис. 3.4. Функции распределения нормального распределения с параметрами сдвига 0, 1 и 2

Таблица 3.2

Частные производные функции распределения $G(x, \theta_1, \theta_2, \ldots, \theta_n, \mu) = F(u, \Theta)$, получаемой при сдвиге распределения $F(x, \Theta)$ на величину μ

	1	$\frac{\partial}{\partial\mu}$	$\frac{\partial}{\partial\theta_i}$	$\frac{\partial^2}{\partial\mu^2}$	$\frac{\partial^2}{\partial\theta_i\partial\theta_j}$	$\frac{\partial^2}{\partial\theta_i\partial\mu}$
1	$F(u)$	$-f(u)$	$F'_{\theta_i}(u)$	$f'_x(u)$	$F''_{\theta_i\theta_j}(u)$	$-f'_{\theta_i}(u)$
$\frac{\partial}{\partial x}$	$f(u)$	$-f'_x(u)$	$f'_{\theta_i}(u)$	$f''_{xx}(u)$	$f''_{\theta_i\theta_j}(u)$	$-f''_{x\theta_i}(u)$
$\frac{\partial^2}{\partial x^2}$	$f'_x(u)$	$-f''_{xx}(u)$	$f''_{x\theta_i}(u)$			
$\frac{\partial^3}{\partial x^3}$	$f''_{xx}(u)$					

Производные функции распределения $G(x, \theta_1, \theta_2, \ldots, \theta_n, \lambda)$ приведены в таблице 3.3, где $u = x/\lambda$.

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

Пример масштабирования распределения показан на рис. 3.5. и 3.6.

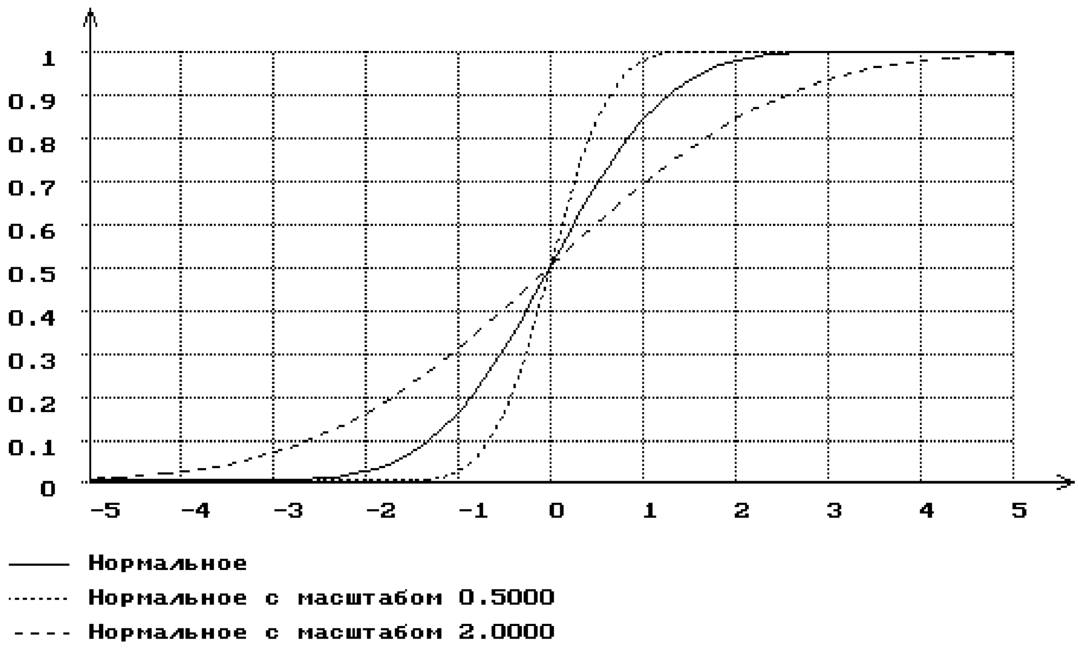


Рис. 3.5. Нормальное распределение с параметрами масштаба 1, 0.5 и 2

3.4.1.3. 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 согласно (3.3) имеет вид:

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

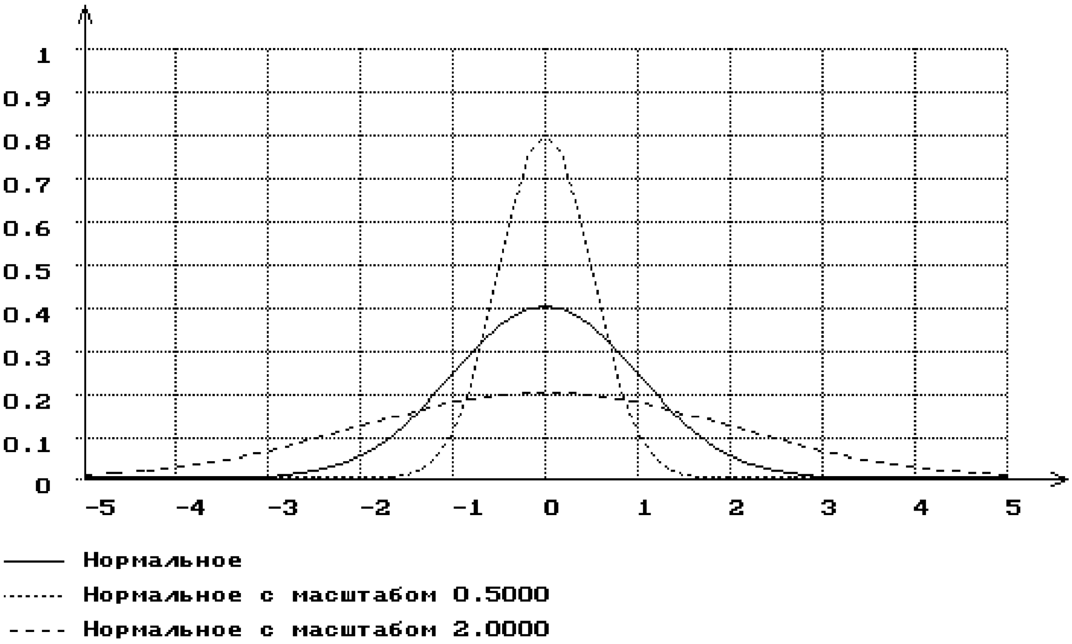


Рис. 3.6. Функции плотности нормального распределения с параметрами масштаба 1, 0.5 и 2

Таблица 3.3

Частные производные функции распределения $G(x, \theta_1, \theta_2, \ldots, \theta_n, \lambda) = F(u, \Theta)$, получаемой при масштабировании распределения $F(x, \Theta)$ на величину λ

	1	$\frac{\partial}{\partial\lambda}$	$\frac{\partial}{\partial\theta_i}$
1	$F(u)$	$-\frac{uf(u)}{\lambda}$	$F'_{\theta_i}(u)$
$\frac{\partial}{\partial x}$	$\frac{f(u)}{\lambda}$	$-\frac{f(u)+uf'_x(u)}{\lambda^2}$	$\frac{f'_{\theta_i}(u)}{\lambda}$
$\frac{\partial^2}{\partial x^2}$	$\frac{f'_x(u)}{\lambda^2}$	$-\frac{2f'_x(u)+uf''_{xx}(u)}{\lambda^3}$	
$\frac{\partial^3}{\partial x^3}$	$\frac{f''_{xx}(u)}{\lambda^3}$		

	$\frac{\partial^2}{\partial\lambda^2}$	$\frac{\partial^2}{\partial\theta_i\partial\theta_j}$	$\frac{\partial^2}{\partial\theta_i\partial\lambda}$
1	$\frac{2uf(u)+u^2f'_x(u)}{\lambda^2}$	$F''_{\theta_i\theta_j}(u)$	$-\frac{uf'_{\theta_i}(u)}{\lambda^2}$
$\frac{\partial}{\partial x}$	$\frac{2f(u)+4uf'_x(u)+u^2f''_{xx}(u)}{\lambda^3}$	$\frac{f''_{\theta_i\theta_j}(u)}{\lambda}$	$-\frac{f'_{\theta_i}(u)+uf''_{x\theta_i}(u)}{\lambda^2}$

Производные функции распределения $G(x, \theta_1, \theta_2, \ldots, \theta_n)$ приведены в таблице 3.4, где $u = -x$.

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

Замечание 3.2. 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 не меняется).

Пример зеркального отражения распределения показан на рис. 3.7. и 3.8.

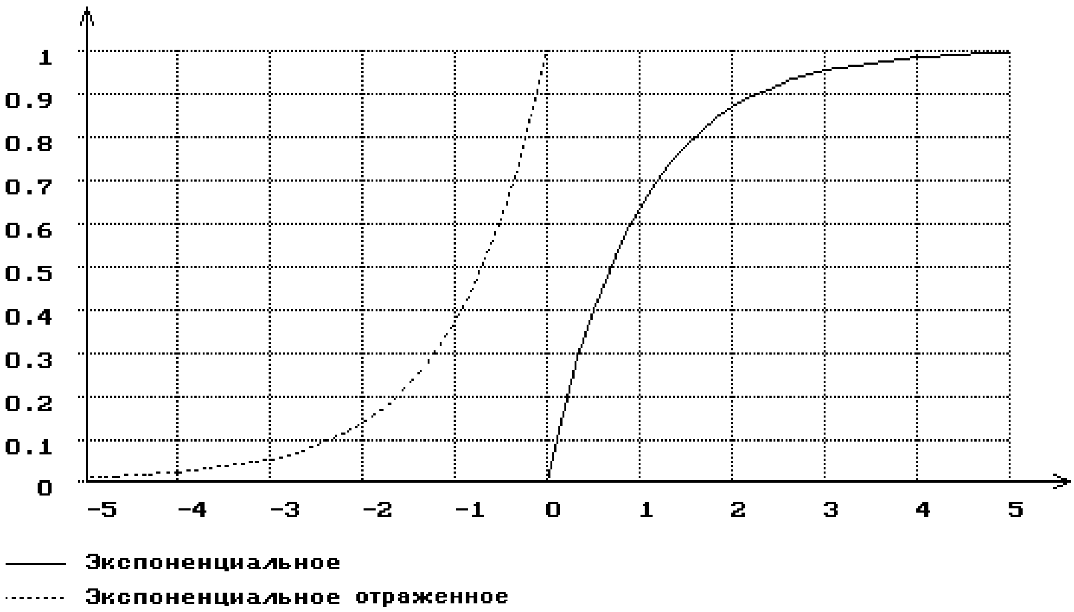


Рис. 3.7. Функции распределения экспоненциального и зеркального экспоненциального распределений

3.4.1.4. Усечение слева

Операция усечения слева, преобразующая случайную величину ξ в случайную величину $\eta = (\xi | \xi > a)$, где a — параметр усечения, задается с

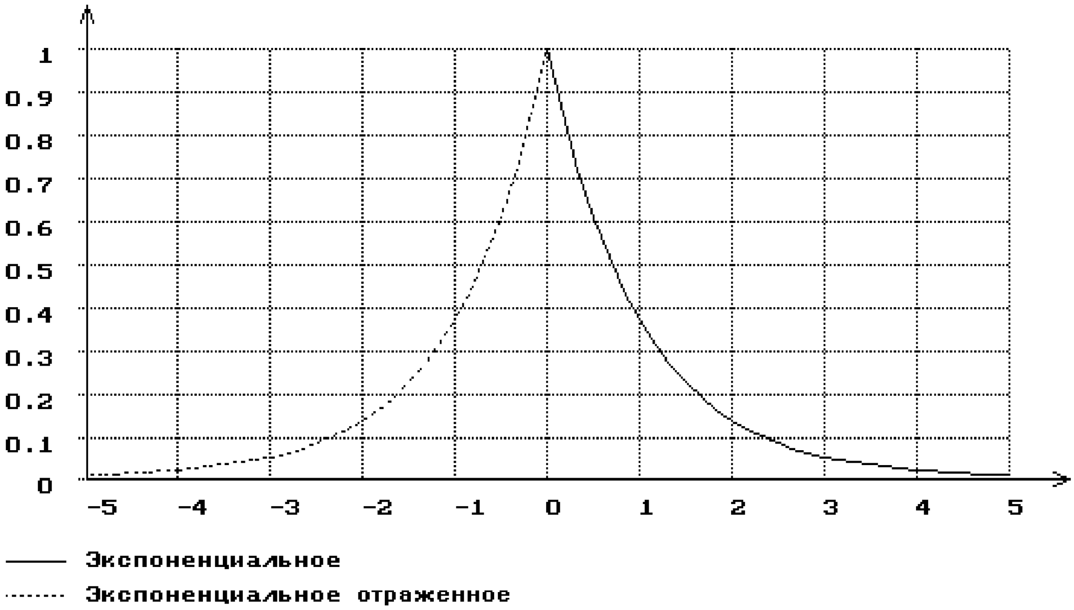


Рис. 3.8. Функции плотности экспоненциального и зеркального экспоненциального распределений

Таблица 3.4

Частные производные функции распределения $G(x, \theta_1, \theta_2, \ldots, \theta_n) = F(u, \Theta)$, получаемой при зеркальном отражении распределения $F(x, \Theta)$

	1	$\frac{\partial}{\partial\theta_i}$	$\frac{\partial^2}{\partial\theta_i\partial\theta_j}$
1	$1 - F(u)$	$-F'_{\theta_i}(u)$	$-F''_{\theta_i\theta_j}(u)$
$\frac{\partial}{\partial x}$	$f(u)$	$f'_{\theta_i}(u)$	$f''_{\theta_i\theta_j}(u)$
$\frac{\partial^2}{\partial x^2}$	$-f'_x(u)$	$-f''_{x\theta_i}(u)$	
$\frac{\partial^3}{\partial x^3}$	$f''_{xx}(u)$		

помощью условной функции распределения:

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

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

Производные функции распределения $G(x,\theta_1,\theta_2,\ldots,\theta_n,a)$ приведены в таблице 3.5.

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

Пример усечения распределения слева показан на рис. 3.9. и 3.10.

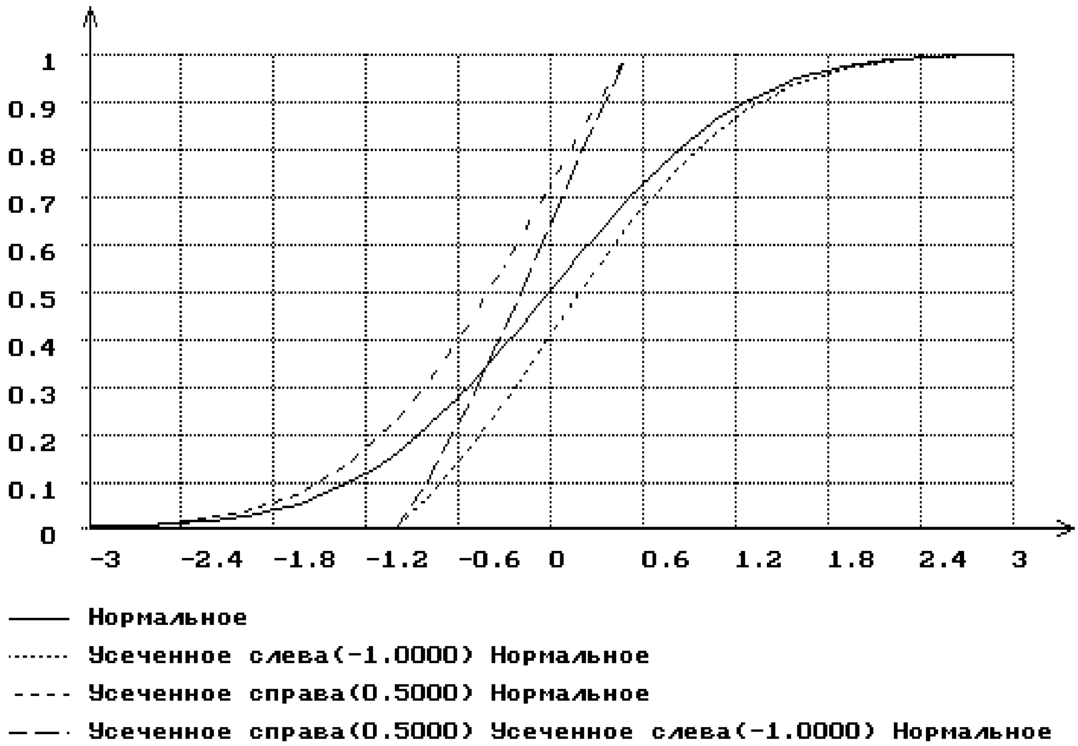


Рис. 3.9. Функции распределения усеченного слева в точке -1, усеченного справа в точке 0.5 и усеченного с двух сторон нормального распределения

3.4.1.5. Усечение справа

Операция усечения справа, преобразующая случайную величину ξ в случайную величину $\eta = (\xi|\xi < b)$, где b — параметр усечения, задается с

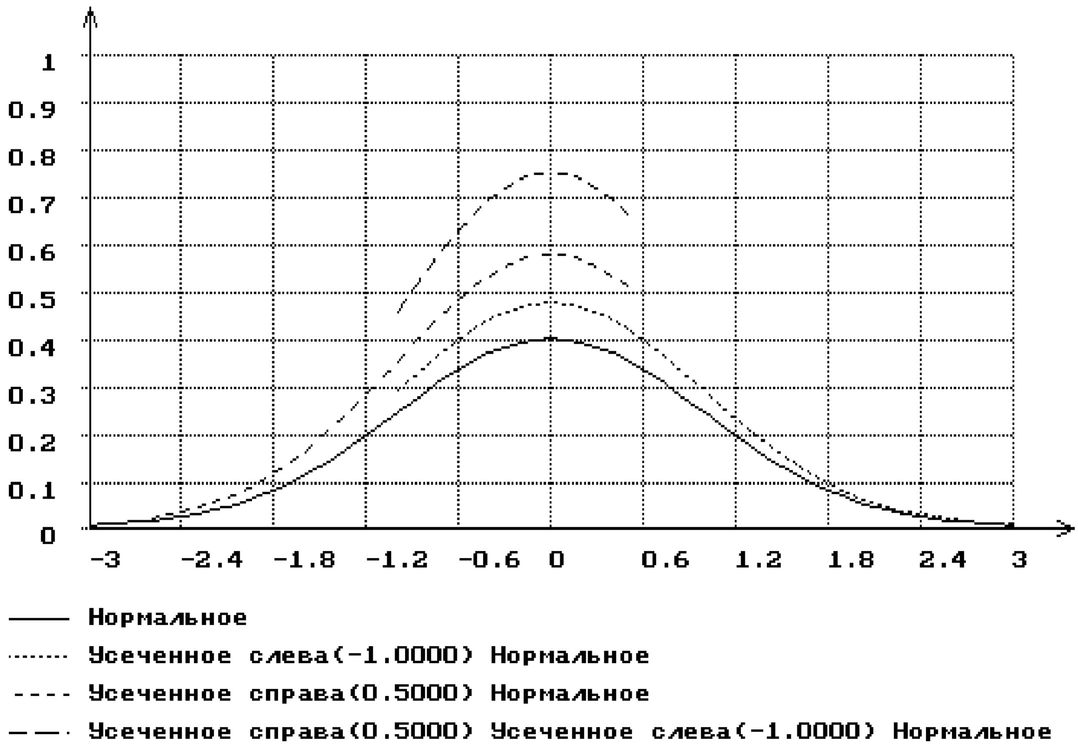


Рис. 3.10. Функции плотности распределения усеченного слева в точке -1, усеченного справа в точке 0.5 и усеченного с двух сторон нормального распределения

Таблица 3.5

Частные производные функции распределения $G(x, \theta_1, \theta_2, \ldots, \theta_n, a)$, получаемой при усечении слева распределения $F(x, \Theta)$ в точке a

	1	$\frac{\partial}{\partial a}$	$\frac{\partial}{\partial \theta_i}$	$\frac{\partial^2}{\partial a^2}$
1	$\frac{F(x)-F(a)}{1-F(a)}$	$\frac{(F(x)-1)f(a)}{(1-F(a))^2}$	$\frac{F'_{\theta_i}(x)}{1-F(a)} + \frac{F'_{\theta_i}(a)(F(x)-1)}{(1-F(a))^2}$	$\frac{(F(x)-1)(f'_x(a)(1-F(a))+2f^2(a))}{(1-F(a))^3}$
$\frac{\partial}{\partial x}$	$\frac{f(x)}{1-F(a)}$	$\frac{f(x)f(a)}{(1-F(a))^2}$	$\frac{f'_{\theta_i}(x)}{1-F(a)} + \frac{F'_{\theta_i}(a)f(x)}{(1-F(a))^2}$	$\frac{f(x)(f'_x(a)(1-F(a))+2f^2(a))}{(1-F(a))^3}$
$\frac{\partial^2}{\partial x^2}$	$\frac{f'_x(x)}{1-F(a)}$	$\frac{f'_x(x)f(a)}{(1-F(a))^2}$	$\frac{f''_{x\theta_i}(x)}{1-F(a)} + \frac{F'_{\theta_i}(a)f'_x(x)}{(1-F(a))^2}$	
$\frac{\partial^3}{\partial x^3}$	$\frac{f''_{xx}(x)}{1-F(a)}$			

	$\frac{\partial^2}{\partial \theta_i \partial \theta_j}$
1	$\frac{F''_{\theta_i\theta_j}(x)(1-F(a))+F'_{\theta_i}(x)F'_{\theta_j}(a)+F'_{\theta_i}(a)F'_{\theta_j}(x)+F''_{\theta_i\theta_j}(a)(F(x)-1)}{(1-F(a))^2} + \frac{2(F(x)-1)F'_{\theta_i}(a)F'_{\theta_j}(a)}{(1-F(a))^3}$
$\frac{\partial}{\partial x}$	$\frac{f''_{\theta_i\theta_j}(x)(1-F(a))+f'_{\theta_i}(x)F'_{\theta_j}(a)+F'_{\theta_i}(a)f'_{\theta_j}(x)+F''_{\theta_i\theta_j}(a)f(x)}{(1-F(a))^2} + \frac{2f(x)F'_{\theta_i}(a)F'_{\theta_j}(a)}{(1-F(a))^3}$

	$\frac{\partial^2}{\partial \theta_i \partial a}$
1	$\frac{F'_{\theta_i}(x)f(a)+(F(x)-1)f'_{\theta_i}(a)}{(1-F(a))^2} + \frac{2f(a)F'_{\theta_i}(a)(F(x)-1)}{(1-F(a))^3}$
$\frac{\partial}{\partial x}$	$\frac{f'_{\theta_i}(x)f(a)+f(x)f'_{\theta_i}(a)}{(1-F(a))^2} + \frac{2f(a)F'_{\theta_i}(a)f(x)}{(1-F(a))^3}$

помощью условной функции распределения:

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

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

Производные функции распределения $G(x, \theta_1, \theta_2, \ldots, \theta_n, b)$ приведены в таблице 3.6.

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

Пример усечения распределения справа показан на рис. 3.9. и 3.10.

3.4.1.6. Двустороннее усечение

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

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

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

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

Пример двустороннего усечения распределения показан на рис. 3.9. и 3.10.

3.4.1.7. Логарифмирование

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

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

Тогда

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

Таблица 3.6

Частные производные функции распределения $G(x, \theta_1, \theta_2, \ldots, \theta_n, b)$, получаемой при усечении справа распределения $F(x, \Theta)$ в точке b

	1	$\frac{\partial}{\partial b}$	$\frac{\partial}{\partial \theta_i}$	$\frac{\partial^2}{\partial b^2}$
1	$\frac{F(x)}{F(b)}$	$-\frac{F(x)f(b)}{(F(b))^2}$	$\frac{F'_{\theta_i}(x)}{F(b)} - \frac{F'_{\theta_i}(b)F(x)}{(F(b))^2}$	$-\frac{F(x)(f'_x(b)F(b)-2f^2(b))}{(F(b))^3}$
$\frac{\partial}{\partial x}$	$\frac{f(x)}{F(b)}$	$-\frac{f(x)f(b)}{(F(b))^2}$	$\frac{f'_{\theta_i}(x)}{F(b)} - \frac{F'_{\theta_i}(b)f(x)}{(F(b))^2}$	$-\frac{f(x)(f'_x(b)F(b)+2f^2(b))}{(F(b))^3}$
$\frac{\partial^2}{\partial x^2}$	$\frac{f'_x(x)}{F(b)}$	$-\frac{f'_x(x)f(b)}{(F(b))^2}$	$\frac{f''_{x\theta_i}(x)}{F(b)} - \frac{F'_{\theta_i}(b)f'_x(x)}{(F(b))^2}$	
$\frac{\partial^3}{\partial x^3}$	$\frac{f''_{xx}(x)}{F(b)}$			

	$\frac{\partial^2}{\partial \theta_i \partial \theta_j}$
1	$\frac{F''_{\theta_i\theta_j}(x)F(b)-F'_{\theta_i}(x)F'_{\theta_j}(b)-F'_{\theta_i}(b)F'_{\theta_j}(x)-F''_{\theta_i\theta_j}(b)F(x))}{(F(b))^2} + \frac{2F(x)F'_{\theta_i}(b)F'_{\theta_j}(b)}{(F(b))^3}$
$\frac{\partial}{\partial x}$	$\frac{f''_{\theta_i\theta_j}(x)F(b)-f'_{\theta_i}(x)F'_{\theta_j}(b)-F'_{\theta_i}(b)f'_{\theta_j}(x)-F''_{\theta_i\theta_j}(b)f(x)}{(F(b))^2} + \frac{2f(x)F'_{\theta_i}(b)F'_{\theta_j}(b)}{(F(b))^3}$

	$\frac{\partial^2}{\partial \theta_i \partial b}$
1	$-\frac{F'_{\theta_i}(x)f(b)+F(x)f'_{\theta_i}(b)}{(F(b))^2} + \frac{2f(b)F'_{\theta_i}(b)F(x)}{(F(b))^3}$
$\frac{\partial}{\partial x}$	$-\frac{f'_{\theta_i}(x)f(b)+f(x)f'_{\theta_i}(b)}{(F(b))^2} + \frac{2f(b)F'_{\theta_i}(b)f(x)}{(F(b))^3}$

и функция распределения согласно (3.2) имеет вид:

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

Производные функции распределения $G(x,\theta_1,\theta_2,\ldots,\theta_n)$ приведены в таблице 3.7, где $u=\ln x$.

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

Пример применения операции логарифмирования показан на рис. 3.11.

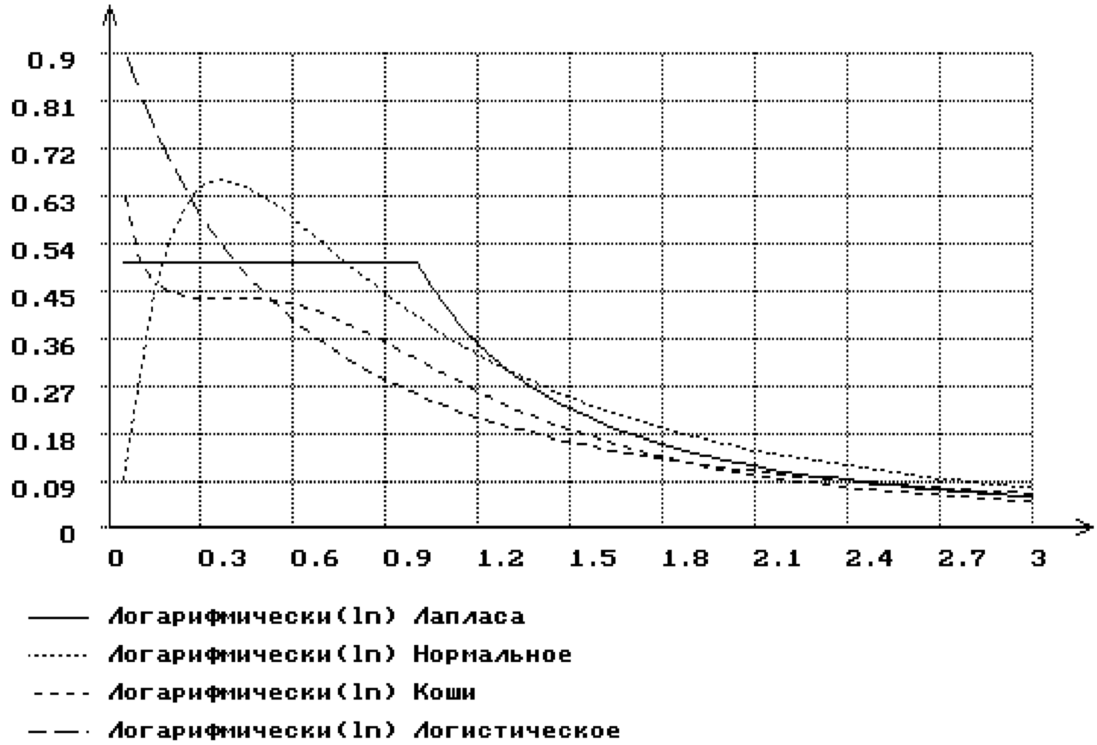


Рис. 3.11. Функции плотности распределений, полученных при выполнении операции логарифмирования

Таблица 3.7

Частные производные функции распределения $G(x, \theta_1, \theta_2, \ldots, \theta_n) = F(u, \Theta)$, получаемые логарифмированием распределения $F(x, \Theta)$

	1	$\frac{\partial}{\partial \theta_i}$	$\frac{\partial^2}{\partial \theta_i \partial \theta_j}$
1	$F(u)$	$F'_{\theta_i}(u)$	$F''_{\theta_i \theta_j}(u)$
$\frac{\partial}{\partial x}$	$\frac{f(u)}{x}$	$\frac{f'_{\theta_i}(u)}{x}$	$\frac{f''_{\theta_i \theta_j}(u)}{x}$
$\frac{\partial^2}{\partial x^2}$	$\frac{f'_x(u) - f(u)}{x^2}$	$\frac{f''_{x\theta_i}(u) - f'_{\theta_i}(u)}{x^2}$	
$\frac{\partial^3}{\partial x^3}$	$\frac{f''_{xx}(u) - 3f'_x(u) + 2f(u)}{x^3}$		

3.4.1.8. Смесь

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

$$G_\eta(x, \Theta_1, \Theta_2, \mathrm{w}) = \mathrm{w}\mathrm{F}_{\xi_1}(\mathrm{x}, \Theta_1) + (1 - \mathrm{w})\mathrm{F}_{\xi_2}(\mathrm{x}, \Theta_2),$$

где $F_{\xi_1}(x, \Theta_1)$ и $F_{\xi_2}(x, \Theta_2)$ — функции распределения случайных величин ξ_1 и ξ_2 соответственно, w — это параметр смеси, принадлежащий интервалу $[a, b]$, a и b вычисляются по формулам [76]:

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

и

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

Смеси распределений очень часто встречаются на практике [86]. 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.

Производные функции распределения $G_\eta(x, \Theta_1, \Theta_2, \mathrm{w})$ приведены в таблице 3.8.

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 $[l_1, r_1] \cup [l_2, r_2]$.

Пример применения операции смеси показан на рис. 3.12. и 3.13

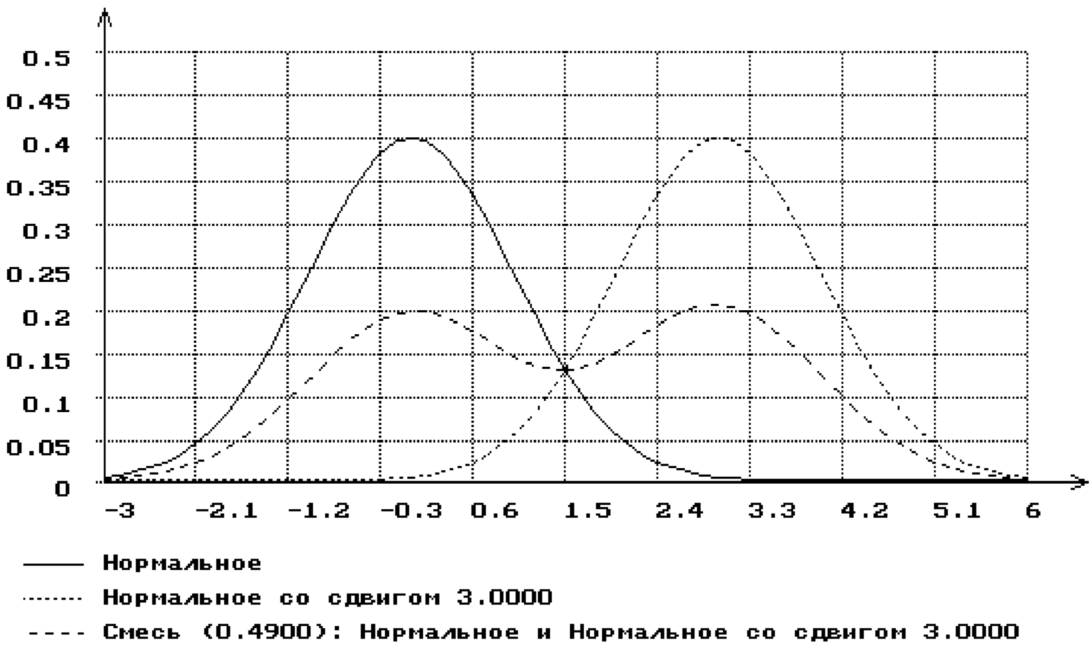


Рис. 3.12. Смесь двух нормальных распределений с параметром смеси $\mathrm{w} \in [0, 1]$

3.4.1.9. 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.

Производные функции распределения $G_\eta(x, \Theta_1, \Theta_2)$ приведены в таблице 3.9.

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 $[l_1, r_1] \cap [l_2, r_2]$.

Таблица 3.8

Частные производные функции распределения $G(x, \Theta_1, \Theta_2, \mathrm{w})$, получаемой при смешивании распределений $F_1(x, \Theta_1)$ и $F_2(x, \Theta_2)$ с параметром w

	1	$\frac{\partial}{\partial \mathrm{w}}$	$\frac{\partial}{\partial \theta_i}, \theta_i \in \Theta_1$	$\frac{\partial}{\partial \theta_i}, \theta_i \in \Theta_2$
1	$\mathrm{wF_1(x) + (1-w)F_2(x)}$	$F_1(x) - F_2(x)$	$\mathrm{w}\frac{\partial \mathrm{F_1(x)}}{\partial \theta_\mathrm{i}}$	$(1-\mathrm{w})\frac{\partial \mathrm{F_2(x)}}{\partial \theta_\mathrm{i}}$
$\frac{\partial}{\partial x}$	$\mathrm{wf_1(x) + (1-w)f_2(x)}$	$f_1(x) - f_2(x)$	$\mathrm{w}\frac{\partial \mathrm{f_1(x)}}{\partial \theta_\mathrm{i}}$	$(1-\mathrm{w})\frac{\partial \mathrm{f_2(x)}}{\partial \theta_\mathrm{i}}$
$\frac{\partial^2}{\partial x^2}$	$\mathrm{w}\frac{\partial \mathrm{f_1(x)}}{\partial \mathrm{x}} + (1-\mathrm{w})\frac{\partial \mathrm{f_2(x)}}{\partial \mathrm{x}}$	$\frac{\partial f_1(x)}{\partial x} - \frac{\partial f_2(x)}{\partial x}$	$\mathrm{w}\frac{\partial^2 \mathrm{f_1(x)}}{\partial \mathrm{x} \partial \theta_\mathrm{i}}$	$(1-\mathrm{w})\frac{\partial^2 \mathrm{f_2(x)}}{\partial \mathrm{x} \partial \theta_\mathrm{i}}$
$\frac{\partial^3}{\partial x^3}$	$\mathrm{w}\frac{\partial^2 \mathrm{f_1(x)}}{\partial \mathrm{x}^2} + (1-\mathrm{w})\frac{\partial^2 \mathrm{f_2(x)}}{\partial \mathrm{x}^2}$			

	$\frac{\partial^2}{\partial \mathrm{w}^2}$	$\frac{\partial^2}{\partial \theta_i \partial \theta_j}, \theta_i, \theta_j \in \Theta_1$	$\frac{\partial^2}{\partial \theta_i \partial \theta_j}, \theta_i, \theta_j \in \Theta_2$
1	0	$\mathrm{w}\frac{\partial^2 \mathrm{F_1(x)}}{\partial \theta_\mathrm{i} \partial \theta_\mathrm{j}}$	$(1-\mathrm{w})\frac{\partial^2 \mathrm{F_2(x)}}{\partial \theta_\mathrm{i} \partial \theta_\mathrm{j}}$
$\frac{\partial}{\partial x}$	0	$\mathrm{w}\frac{\partial^2 \mathrm{f_1(x)}}{\partial \theta_\mathrm{i} \partial \theta_\mathrm{j}}$	$(1-\mathrm{w})\frac{\partial^2 \mathrm{f_2(x)}}{\partial \theta_\mathrm{i} \partial \theta_\mathrm{j}}$

	$\frac{\partial^2}{\partial \theta_i \partial \theta_j}, \theta_i \in \Theta_1, \theta_j \in \Theta_2$	$\frac{\partial^2}{\partial \mathrm{w} \partial \theta_\mathrm{i}}, \theta_i \in \Theta_1$	$\frac{\partial^2}{\partial \mathrm{w} \partial \theta_\mathrm{i}}, \theta_i \in \Theta_2$
1	0	$\frac{\partial F_1(x)}{\partial \theta_i}$	$-\frac{\partial F_2(x)}{\partial \theta_i}$
$\frac{\partial}{\partial x}$	0	$\frac{\partial f_1(x)}{\partial \theta_i}$	$-\frac{\partial f_2(x)}{\partial \theta_i}$

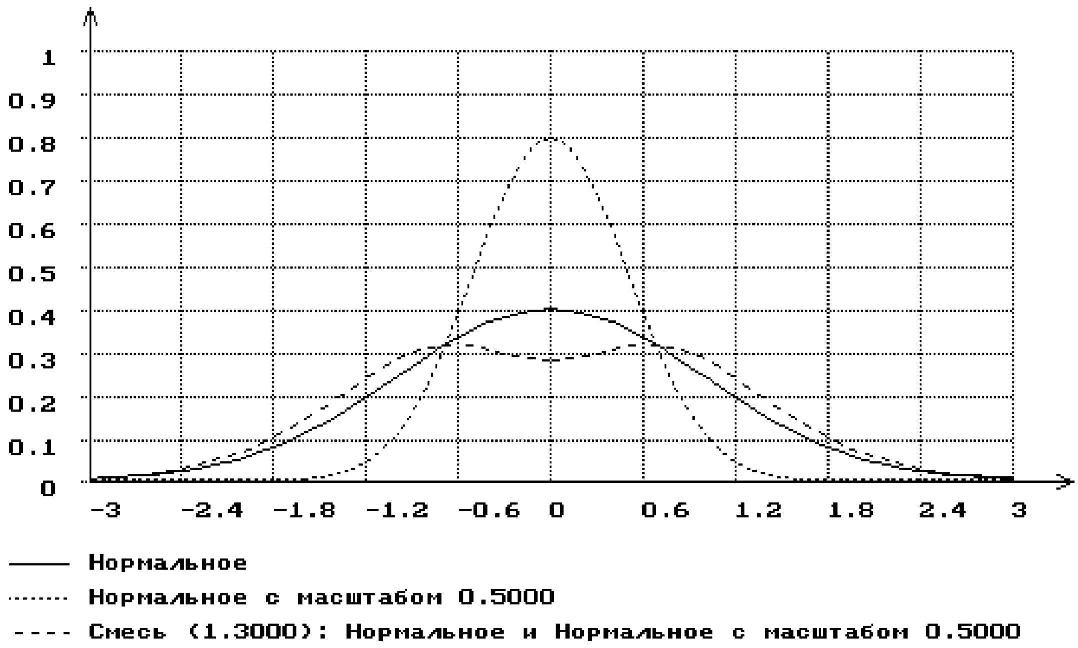


Рис. 3.13. Смесь двух нормальных распределений с параметром смеси w > 1

Пример применения операции произведения показан на рис. 3.14.

3.4.2. Семейства распределений

Многие распределения, широко применяемые на практике, можно свести к нескольким семействам распределений. Семейства распределений можно получать двумя способами [59]. *Групповые семейства* получаются применением операций над распределениями (см. п. 3.4.1), 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

$$F(x, \Theta) = H(g(x), \Theta),\ f(x, \Theta) = h(g(x), \Theta)g'(x), \tag{3.4}$$

где Θ — вектор параметров, g(x) — генерирующая функция, а конкретное распределение из семейства получается при подстановке в (3.2) конкретной

Таблица 3.9

Частные производные функции $G(x,\Theta_1,\Theta_2) = F_1(x,\Theta_1)F_2(x,\Theta_2)$

	1
1	$F_1(x)F_2(x)$
$\frac{\partial}{\partial x}$	$f_1(x)F_2(x) + F_1(x)f_2(x)$
$\frac{\partial^2}{\partial x^2}$	$\frac{\partial f_1(x)}{\partial x}F_2(x) + 2f_1(x)f_2(x) + F_1(x)\frac{\partial f_2(x)}{\partial x}$
$\frac{\partial^3}{\partial x^3}$	$\frac{\partial^2 f_1(x)}{\partial x^2}F_2(x) + 3\frac{\partial f_1(x)}{\partial x}f_2(x) + 3f_1(x)\frac{\partial f_2(x)}{\partial x} + F_1(x)\frac{\partial^2 f_2(x)}{\partial x^2}$

	$\frac{\partial}{\partial \theta_i}, \theta_i \in \Theta_1$
1	$\frac{\partial F_1(x)}{\partial \theta_i}F_2(x)$
$\frac{\partial}{\partial x}$	$\frac{\partial f_1(x)}{\partial \theta_i}F_2(x) + \frac{\partial F_1(x)}{\partial \theta_i}f_2(x)$
$\frac{\partial^2}{\partial x^2}$	$\frac{\partial^2 f_1(x)}{\partial x\partial \theta_i}F_2(x) + 2\frac{\partial f_1(x)}{\partial \theta_i}f_2(x) + \frac{\partial F_1(x)}{\partial \theta_i}\frac{\partial f_2(x)}{\partial x}$

	$\frac{\partial}{\partial \theta_i}, \theta_i \in \Theta_2$
1	$F_1(x)\frac{\partial F_2(x)}{\partial \theta_i}$
$\frac{\partial}{\partial x}$	$f_1(x)\frac{\partial F_2(x)}{\partial \theta_i} + F_1(x)\frac{\partial f_2(x)}{\partial \theta_i}$
$\frac{\partial^2}{\partial x^2}$	$\frac{\partial f_1(x)}{\partial x}\frac{\partial F_2(x)}{\partial \theta_i} + 2f_1(x)\frac{\partial f_2(x)}{\partial \theta_i} + F_1(x)\frac{\partial^2 f_2(x)}{\partial x\partial \theta_i}$

	$\frac{\partial^2}{\partial \theta_i\partial \theta_j}, \theta_i \in \Theta_1, \theta_j \in \Theta_1$	$\frac{\partial^2}{\partial \theta_i\partial \theta_j}, \theta_i \in \Theta_2, \theta_j \in \Theta_2$
1	$\frac{\partial^2 F_1(x)}{\partial \theta_i\partial \theta_j}F_2(x)$	$F_1(x)\frac{\partial^2 F_2(x)}{\partial \theta_i\partial \theta_j}$
$\frac{\partial}{\partial x}$	$\frac{\partial^2 f_1(x)}{\partial \theta_i\partial \theta_j}F_2(x) + \frac{\partial^2 F_1(x)}{\partial \theta_i\partial \theta_j}f_2(x)$	$f_1(x)\frac{\partial^2 F_2(x)}{\partial \theta_i\partial \theta_j} + F_1(x)\frac{\partial^2 f_2(x)}{\partial \theta_i\partial \theta_j}$

	$\frac{\partial^2}{\partial \theta_i\partial \theta_j}, \theta_i \in \Theta_1, \theta_j \in \Theta_2$
1	$\frac{\partial F_1(x)}{\partial \theta_i}\frac{\partial F_2(x)}{\partial \theta_j}$
$\frac{\partial}{\partial x}$	$\frac{\partial f_1(x)}{\partial \theta_i}\frac{\partial F_2(x)}{\partial \theta_j} + \frac{\partial F_1(x)}{\partial \theta_i}\frac{\partial f_2(x)}{\partial \theta_j}$

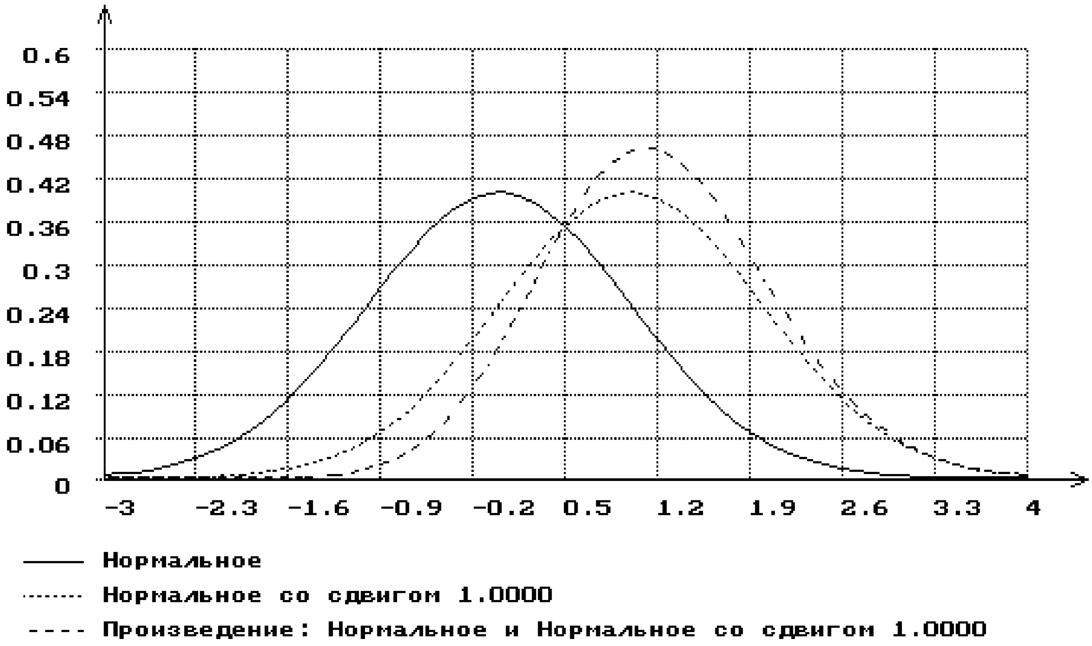


Рис. 3.14. Произведение двух нормальных распределений

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:

$$f(x,\Theta) = \mathrm{e}^{\sum\limits_{i=1}^{s} \eta_i(\Theta) T_i(x) - B(\Theta)} h(x),$$

где η_i и B — вещественнозначные функции от параметров, T_i — вещественнозначные статистики [59].

3.4.2.1. 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.

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, $g(x) = \ln(x + \sqrt{x^2+1}),\ -\infty < x < +\infty$.

Пусть известны производные функции $g(x)$ до третьего порядка включительно. Тогда можно вычислить производные от функции распределения и плотности функции распределения по x и по параметрам α и β.

Производные функции распределения по параметрам имеют вид:

$$\begin{aligned}
F'_\alpha(x) &= \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2},\\
F'_\beta(x) &= \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2} g(x),\\
F''_{\alpha\alpha}(x) &= -\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}(\alpha+\beta g(x)),\\
F''_{\beta\beta}(x) &= -\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}(\alpha+\beta g(x))g^2(x),\\
F''_{\alpha\beta}(x) &= -\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}(\alpha+\beta g(x))g(x).
\end{aligned}$$

Функция плотности распределения и ее производные по параметрам имеют вид:

$$\begin{aligned}
f(x) &= \frac{\beta}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2} g'_x(x);\\
f'_\alpha(x) &= -\frac{\beta}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}(\alpha+\beta g(x))g'_x(x),\\
f'_\beta(x) &= -\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}(\beta(\alpha+\beta g(x))g(x)-1)g'_x(x);\\
f''_{\alpha\alpha}(x) &= \frac{\beta}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}((\alpha+\beta g(x))^2-1)g'_x(x),\\
f''_{\beta\beta}(x) &= \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}(\beta((\alpha+\beta g(x))^2-1)g^2(x) -
\end{aligned}$$

$$
\begin{aligned}
& -2(\alpha+\beta g(x))g(x))g'_x(x),\\
f''_{\alpha\beta}(x) &= -\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}(\alpha+2\beta g(x)-\beta(\alpha+\beta g(x))^2 g(x))g'_x(x).
\end{aligned}
$$

Производная функции плотности по x и ее производные по параметрам имеют вид:

$$
\begin{aligned}
f'_x(x) &= \frac{\beta}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}\{-\beta(\alpha+\beta g(x))(g'_x(x))^2+g''_{xx}(x)\};\\
f''_{x\alpha}(x) &= \frac{\beta}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}\{\beta[(\alpha+\beta g(x))^3-1](g'_x(x))^2+\\
&\quad +(\alpha+\beta g(x))^2 g''_{xx}(x)\},\\
f''_{x\beta}(x) &= \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}\{[-\beta(\alpha+\beta g(x))(g'_x(x))^2+g''_{xx}(x)]\times\\
&\quad \times[1-\beta(\alpha+\beta g(x))g(x)]+\beta(\alpha-2\beta g(x))(g'_x(x))^2\}.
\end{aligned}
$$

Вторая производная функции плотности по x имеет вид:

$$
f''_{xx}(x)=\frac{\beta}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta g(x))^2}\{\beta^2[(\alpha+\beta g(x))^2-1](g'_x(x))^3-
$$

$$
-3\beta(\alpha+\beta g(x))g'_x(x)g''_{xx}(x)+g'''_{xxx}(x)\}.
$$

3.4.2.2. Семейство гамма-распределений

Семейство гамма-распределений задается формулой

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

где $\Gamma(\delta)=\int\limits_0^\infty y^{\delta-1}\mathrm{e}^{-y}dy$ и $\gamma(x,\delta)=\int\limits_0^x y^{\delta-1}\mathrm{e}^{-y}dy$ — 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, $g(x) = \mathrm{e}^x,\ -\infty < x < +\infty$.

Пусть известны производные генерирующей функции $g(x,\alpha)$ до третьего порядка включительно. Тогда можно вычислить производные от функции распределения и плотности функции распределения по x и по параметрам α и δ.

Производные функции распределения по параметрам имеют вид:

$$\begin{aligned}
F'_\alpha(x) &= \frac{1}{\Gamma(\delta)}(g(x))^{\delta-1}\mathrm{e}^{-g(x)}g'_\alpha(x),\\
F'_\delta(x) &= \frac{1}{\Gamma(\delta)}\{\gamma'_\delta(g(x),\delta) - \psi(\delta)\gamma(g(x),\delta)\},\\
F''_{\alpha\alpha}(x) &= \frac{1}{\Gamma(\delta)}(g(x))^{\delta-2}\mathrm{e}^{-g(x)}\{(\delta-1-g(x))(g'_\alpha(x))^2 + g(x)g''_{\alpha\alpha}(x)\},\\
F''_{\delta\delta}(x) &= \frac{1}{\Gamma(\delta)}\{\gamma''_\delta(g(x),\delta) - 2\psi(\delta)\gamma'_\delta(g(x),\delta) + [\psi^2(\delta) - \psi'_\delta(\delta)]\gamma(g(x),\delta)\},\\
F''_{\alpha\delta}(x) &= \frac{1}{\Gamma(\delta)}g^{\delta-1}(x)\mathrm{e}^{-g(x)}g'_\alpha(x)\{\ln g(x) - \psi(\delta)\}.
\end{aligned}$$

Функция плотности распределения и ее производные по параметрам имеют вид:

$$\begin{aligned}
f(x) &= \frac{1}{\Gamma(\delta)}(g(x))^{\delta-1}\mathrm{e}^{-g(x)}g'_x(x),\\
f'_\alpha(x) &= \frac{1}{\Gamma(\delta)}(g(x))^{\delta-2}\mathrm{e}^{-g(x)}\{[\delta-1-g(x)]g'_\alpha(x)g'_x(x) + g(x)g''_{x\alpha}(x)\},\\
f'_\delta(x) &= \frac{1}{\Gamma(\delta)}(g(x))^{\delta-1}\mathrm{e}^{-g(x)}\{\ln g(x) - \psi(\delta)\}g'_x(x),\\
f''_{\alpha\alpha}(x) &= \frac{1}{\Gamma(\delta)}(g(x))^{\delta-3}\mathrm{e}^{-g(x)}\{[(\delta-1-g(x))^2 - \delta + 1](g'_\alpha(x))^2 g'_x(x) +\\
&\quad +(\delta-1-g(x))g(x)[2g'_\alpha(x)g''_{x\alpha}(x) + g''_{\alpha\alpha}(x)g'_x(x)] +\\
&\quad +(g(x))^2 g'''_{x\alpha\alpha}(x)\},\\
f''_{\delta\delta}(x) &= \frac{1}{\Gamma(\delta)}(g(x))^{\delta-1}\mathrm{e}^{-g(x)}\{\ln^2 g(x) - 2\psi(\delta)\ln g(x) + \psi^2(\delta) - \psi'_\delta(\delta)\}g'_x(x),\\
f''_{\alpha\delta}(x) &= \frac{1}{\Gamma(\delta)}g^{\delta-2}(x)\mathrm{e}^{-g(x)}\{[\ln g(x) - \psi(\delta)][(\delta-1-g(x)]g'_\alpha(x)g'_x(x) +\\
&\quad +g(x)g''_{x\alpha}(x) + g'_\alpha(x)g'_x(x)\}.
\end{aligned}$$

Производная функции плотности по x и ее производные по параметрам имеют вид:

$$\begin{aligned}
f'_x(x) &= \frac{1}{\Gamma(\delta)}(g(x))^{\delta-2}\mathrm{e}^{-g(x)}\{[\delta-1-g(x)](g'_x(x))^2+g(x)g''_{xx}(x)\},\\
f''_{x\alpha}(x) &= \frac{1}{\Gamma(\delta)}(g(x))^{\delta-3}\mathrm{e}^{-g(x)}\{(\delta-1-g(x))g'_\alpha(x)\times\\
&\times[(\delta-1-g(x))(g'_x(x))^2+g(x)g''_{xx}(x)]+(1-\delta)g'_\alpha(x)(g'_x(x))^2+\\
&+2(\delta-1-g(x))g(x)g'_x(x)g''_{x\alpha}(x)+(g(x))^2g'''_{x\alpha\alpha}(x)\},\\
f''_{x\delta}(x) &= \frac{1}{\Gamma(\delta)}(g(x))^{\delta-2}\mathrm{e}^{-g(x)}\{(\ln g(x)-\psi(\delta))\times\\
&\times[(\delta-1-g(x))(g'_x(x))^2+g(x)g''_{xx}(x)]+(g'_x(x))^2\},
\end{aligned}$$

Вторая производная функции плотности по x имеет вид:

$$f'_x(x)=\frac{1}{\Gamma(\delta)}(g(x))^{\delta-3}\mathrm{e}^{-g(x)}\{[(\delta-1-g(x))^2-\delta+1](g'_x(x))^3+$$

$$+3(\delta-1-g(x))g(x)g'_x(x)g''_{xx}(x)+(g(x))^2g'''_{xxx}(x)\}.$$

3.4.2.3. Семейство бета-распределений

Семейство бета-распределений задается формулой

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

где $\mathrm{B}(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}=\int\limits_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy$ и $\mathcal{B}(x,\alpha,\beta)=\int\limits_0^x y^{\alpha-1}(1-y)^{\beta-1}dy$ — полная и неполная бета-функции, α,β и δ — параметры, $g(x,\delta)$ — непрерывная, монотонно возрастающая от 0 до 1 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, $g(x)=\frac{\delta x}{1+(\delta-1)x},\ 0\le x<1;$
- распределение Парето, $g(x)=\frac{x-1}{x},\ 1\le x<\infty,\ \alpha=1;$

- L-распределение (обобщенное логистическое), $g(x) = \frac{\mathrm{e}^x}{1+\mathrm{e}^x}$, $-\infty < x < +\infty$;

Пусть известны производные генерирующей функции $g(x,\delta)$ до третьего порядка включительно. Тогда можно вычислить производные от функции распределения и плотности функции распределения по x и по параметрам α, β и δ.

Введем обозначение:

$$K = (\alpha - 1)(1 - g(x,\delta)) - (\beta - 1)g(x,\delta),$$

и обозначим неполную гамма-функцию и ее первые и вторые производные по параметрам α и β символами $\mathcal{B}, \mathcal{B}'_\alpha, \mathcal{B}'_\beta, \mathcal{B}''_{\alpha\alpha}, \mathcal{B}''_{\beta\beta}, \mathcal{B}''_{\alpha\beta}$ соответственно.

Производные функции распределения по параметрам имеют вид:

$$\begin{aligned}
F'_\alpha(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}\{\mathcal{B}'_\alpha - (\psi(\alpha) - \psi(\alpha+\beta))\mathcal{B}\},\\
F'_\beta(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}\{\mathcal{B}'_\beta - (\psi(\beta) - \psi(\alpha+\beta))\mathcal{B}\},\\
F'_\delta(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-1}(1-g(x,\delta))^{\beta-1}g'_\delta;\\
F''_{\alpha\alpha}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}\{\mathcal{B}''_{\alpha\alpha} - 2(\psi(\alpha) - \psi(\alpha+\beta))\mathcal{B}'_\alpha + (\psi(\alpha) - \psi(\alpha+\beta))^2\mathcal{B}\},\\
F''_{\beta\beta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}\{\mathcal{B}''_{\beta\beta} - 2(\psi(\beta) - \psi(\alpha+\beta))\mathcal{B}'_\beta + (\psi(\beta) - \psi(\alpha+\beta))^2\mathcal{B}\},\\
F''_{\delta\delta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-2}(1-g(x,\delta))^{\beta-2}\times\\
&\quad \times[K(g'_x)^2 + g(x,\delta)(1-g(x,\delta))g''_{xx}];\\
F''_{\alpha\beta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}\{\mathcal{B}''_{\alpha\beta} - (\psi(\beta) - \psi(\alpha+\beta))\mathcal{B}'_\alpha - (\psi(\alpha) - \psi(\alpha+\beta))\mathcal{B}'_\beta +\\
&\quad +[(\psi(\beta) - \psi(\alpha+\beta))(\psi(\beta) - \psi(\alpha+\beta)) + \psi'(\alpha+\beta)]\mathcal{B}\},\\
F''_{\alpha\delta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-1}(1-g(x,\delta))^{\beta-1}g'_\delta\times\\
&\quad \times[\ln g(x,\delta) - \psi(\alpha) + \psi(\alpha+\beta)],\\
F''_{\beta\delta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-1}(1-g(x,\delta))^{\beta-1}g'_\delta\times\\
&\quad \times[\ln(1-g(x,\delta)) - \psi(\beta) + \psi(\alpha+\beta)].
\end{aligned}$$

Функция плотности распределения и ее производные по параметрам имеют вид:

$$
\begin{aligned}
f(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-1}(1-g(x,\delta))^{\beta-1}g'_x,\\
f'_\alpha(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-1}(1-g(x,\delta))^{\beta-1}g'_x \times\\
&\times[\ln g(x,\delta)-\psi(\alpha)+\psi(\alpha+\beta)],\\
f'_\beta(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-1}(1-g(x,\delta))^{\beta-1}g'_x \times\\
&\times[\ln(1-g(x,\delta))-\psi(\beta)+\psi(\alpha+\beta)],\\
f'_\delta(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-2}(1-g(x,\delta))^{\beta-2}\{Kg'_\delta g'_x+\\
&+g(x,\delta)(1-g(x,\delta))g''_{x\delta}\};\\
f''_{\alpha\alpha}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-1}(1-g(x,\delta))^{\beta-1}g'_x \times\\
&\times\{[\ln g(x,\delta)-\psi(\alpha)+\psi(\alpha+\beta)]^2-\psi'(\alpha)+\psi'(\alpha+\beta)\},\\
f''_{\beta\beta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-1}(1-g(x,\delta))^{\beta-1}g'_x \times\\
&\times\{[\ln(1-g(x,\delta))-\psi(\beta)+\psi(\alpha+\beta)]^2-\psi'(\beta)+\psi'(\alpha+\beta)\},\\
f''_{\delta\delta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-3}(1-g(x,\delta))^{\beta-3}\{[K^2-(\alpha-1)(1-g(x,\delta))^2-\\
&-(\beta-1)g^2](g'_\delta)^2g'_x+Kg(x,\delta)(1-g(x,\delta))(g''_{\delta\delta}g'_x+2g'_\delta g''_{x\delta})+\\
&+(g(x,\delta))^2(1-g(x,\delta))^2g'''_{x\delta\delta}\};\\
f''_{\alpha\beta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-1}(1-g(x,\delta))^{\beta-1}g'_x \times\\
&\times\{[\ln g(x,\delta)-\psi(\alpha)+\psi(\alpha+\beta)]\times\\
&\times[\ln(1-g(x,\delta))-\psi(\beta)+\psi(\alpha+\beta)]+\psi'(\alpha+\beta)\},\\
f''_{\alpha\delta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-2}(1-g(x,\delta))^{\beta-2}\times\\
&\times\{[K[\ln g(x,\delta)-\psi(\alpha)+\psi(\alpha+\beta)]+1-g(x,\delta)]g'_xg'_\delta+\\
&+[\ln g(x,\delta)-\psi(\alpha)+\psi(\alpha+\beta)]g(x,\delta)(1-g(x,\delta))g''_{x\delta}\},\\
f''_{\beta\delta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-2}(1-g(x,\delta))^{\beta-2}\times\\
&\times\{[K[\ln(1-g(x,\delta))-\psi(\beta)+\psi(\alpha+\beta)]+1-g(x,\delta)]g'_xg'_\delta+\\
&+[\ln(1-g(x,\delta))-\psi(\beta)+\psi(\alpha+\beta)]g(x,\delta)(1-g(x,\delta))g''_{x\delta}\}.
\end{aligned}
$$

Производная функции плотности по x и ее производные по параметрам имеют вид:

$$\begin{aligned}
f'_x(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-2}(1-g(x,\delta))^{\beta-2}\times\\
&\times[K(g'_x)^2+g(x,\delta)(1-g(x,\delta))g''_{xx}];\\
f''_{x\alpha}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-2}(1-g(x,\delta))^{\beta-2}\times\\
&\times\{[\ln g(x,\delta)-\psi(\alpha)+\psi(\alpha+\beta)]\times\\
&\times(K(g'_x)^2+g(x,\delta)(1-g(x,\delta))g''_{xx})+(1-g(x,\delta))(g'_x)^2\},\\
f''_{x\beta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-2}(1-g(x,\delta))^{\beta-2}\times\\
&\times\{[\ln(1-g(x,\delta))-\psi(\beta)+\psi(\alpha+\beta)]\times\\
&\times(K(g'_x)^2+g(x,\delta)(1-g(x,\delta))g''_{xx})-g(x,\delta)(g'_x)^2\},\\
f''_{x\delta}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-3}(1-g(x,\delta))^{\beta-3}\times\\
&\times\{[K\cdot((\alpha-2)(1-g(x,\delta))-(\beta-2)g(x,\delta)))+\\
&+g(x,\delta)(1-g(x,\delta))(2-\alpha-\beta)]g'_\delta(g'_x)^2+\\
&+[\alpha-1+(2-\alpha-\beta)g(x,\delta)]g(x,\delta)(1-g(x,\delta))g'_\delta g''_{xx}+\\
&+2Kg(x,\delta)(1-g(x,\delta))g'_x g''_{x\delta}+(g(x,\delta))^2(1-g(x,\delta))^2 g'''_{xx\delta}\}.
\end{aligned}$$

Вторая производная функции плотности по x имеет вид:

$$\begin{aligned}
f''_{xx}(x) &= \frac{1}{\mathrm{B}(\alpha,\beta)}(g(x,\delta))^{\alpha-3}(1-g(x,\delta))^{\beta-3}\{[K^2-(\alpha-1)(1-g(x,\delta))^2-\\
&-(\beta-1)g^2](g'_x)^3+3Kg(x,\delta)(1-g(x,\delta))g'_x g''_{xx}+\\
&+(g(x,\delta))^2(1-g(x,\delta))^2 g'''_{xxx}\}.
\end{aligned}$$

3.4.3. Стандартные распределения

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

3.4.3.1. Равномерное распределение

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

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x < 0; \\ x, & 0 \le x \le 1; \\ 1, & x > 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x < 0 \vee x > 1; \\ 1, & 0 \le x \le 1. \end{cases}$$

Производные функция плотности распределения равны нулю:

$$f'_x(x) = f''_{xx}(x) = 0.$$

Равномерное распределение является частным случаем Бета-распределения I-го рода при $\alpha = \beta = 1$ (см. п. 3.4.3.21) и предельным для двустороннего экспоненциального при $\alpha \to +\infty$ (см. п. 3.4.3.27).

График функции плотности приведен на рис. 3.15.

3.4.3.2. Экспоненциальное распределение

Случайная величина, распределенная по экспоненциальному закону, имеет область определения $[0, +\infty)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x < 0; \\ 1 - \mathrm{e}^{-x}, & x \ge 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x < 0; \\ \mathrm{e}^{-x}, & x \ge 0. \end{cases}$$

Экспоненциальное распределение является частным случаем распределения Вейбулла при $\alpha = 1$ (см. п. 3.4.3.15).

График функции плотности приведен на рис. 3.16.

3.4.3.3. Полунормальное распределение

Случайная величина, распределенная по полунормальному закону, имеет область определения $[0, +\infty)$.

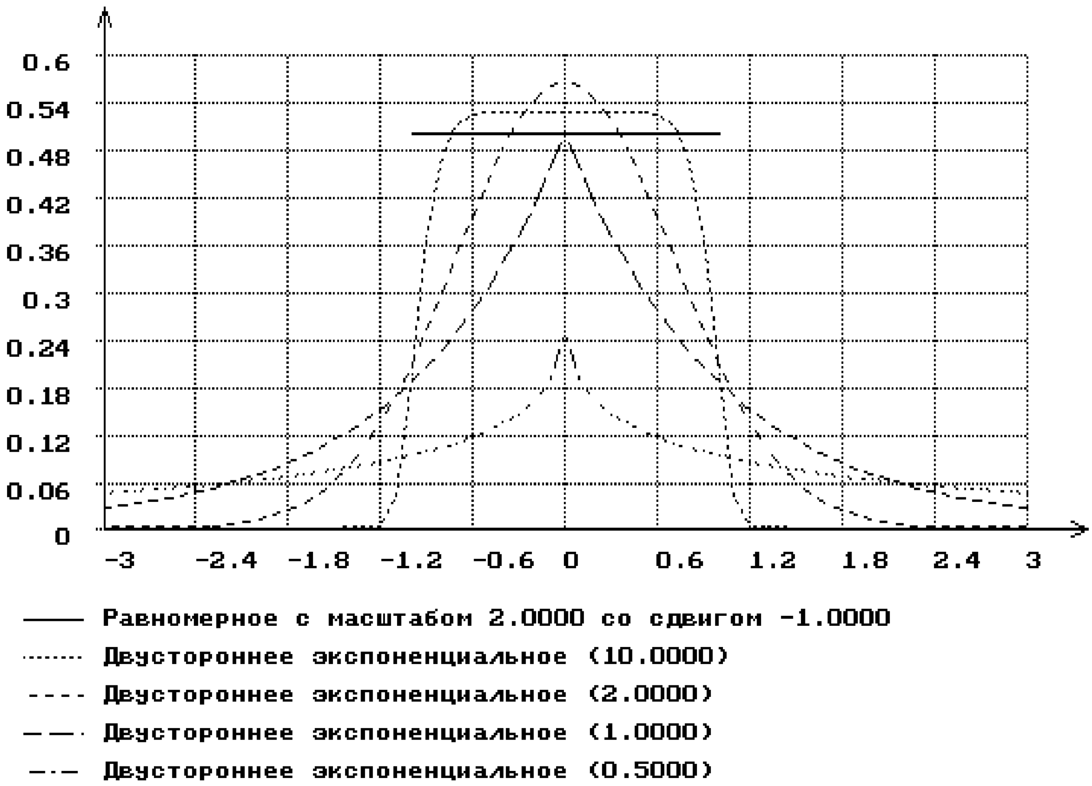


Рис. 3.15. Функции плотности равномерного и двустороннего экспоненциального распределений

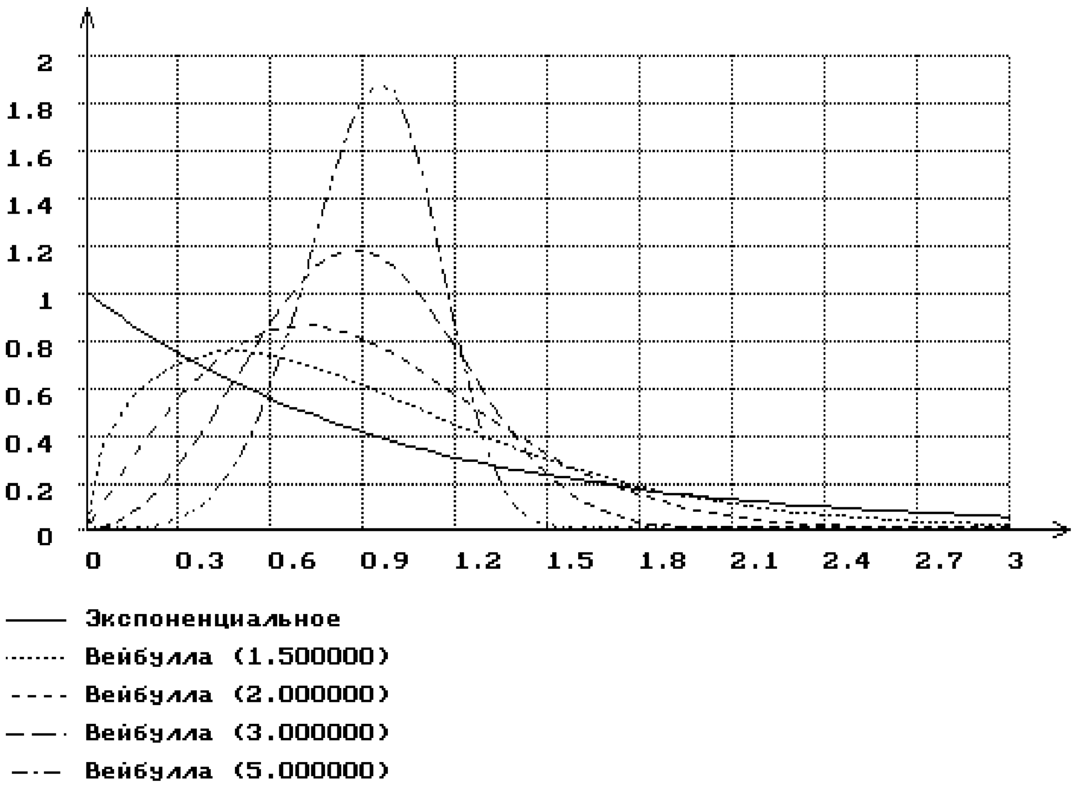


Рис. 3.16. Функции плотности распределения Вейбулла с параметром формы $\alpha = 1, 1.5, 2, 3, 5$

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x < 0; \\ 2\Phi_0(x) = \frac{1}{\Gamma(1/2)}\gamma(\frac{x^2}{2}, \frac{1}{2}), & x \geq 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x < 0; \\ \frac{1}{\sqrt{\pi/2}} \mathrm{e}^{-\frac{x^2}{2}}, & x \geq 0. \end{cases}$$

Полунормальное распределение является частным случаем распределения модуля многомерного нормального вектора при $n = 1$ (см. п. 3.4.3.6).

График функции плотности приведен на рис. 3.17.

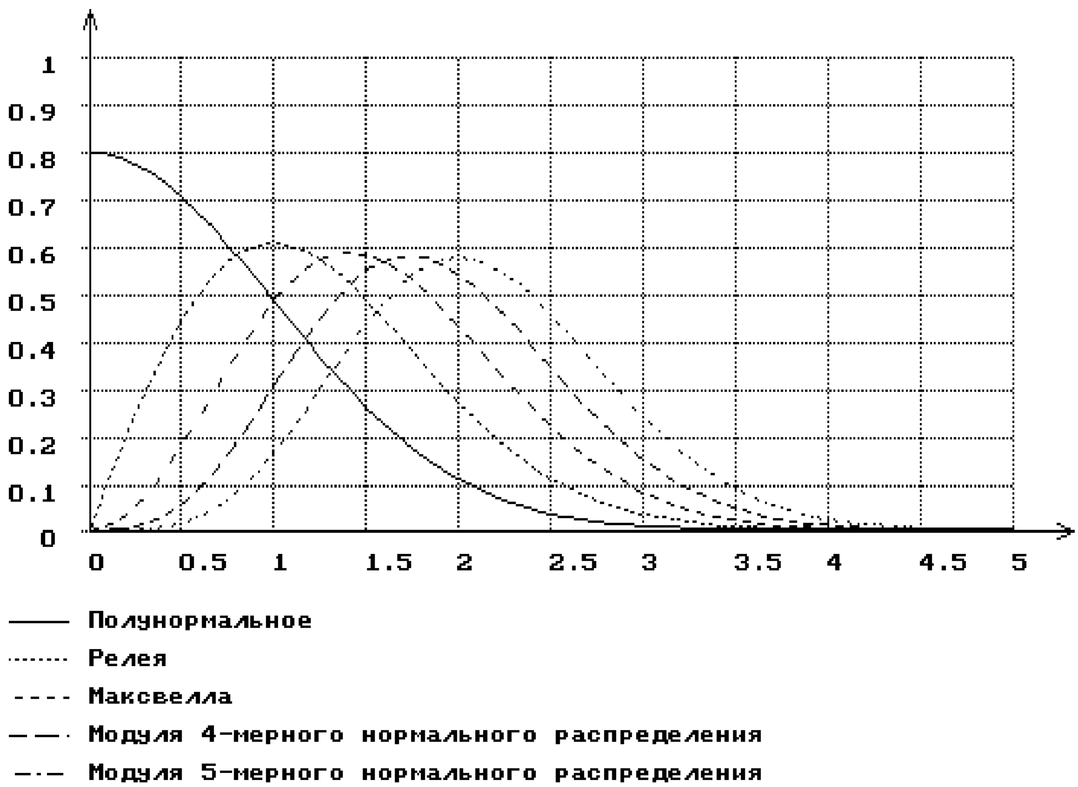


Рис. 3.17. Функции плотности модуля n-мерного нормального распределения с параметром $n = 1, 2, 3, 4, 5$

3.4.3.4. Распределение Рэлея

Случайная величина, распределенная по закону Рэлея, имеет область определения $[0, +\infty)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x < 0; \\ 1 - \mathrm{e}^{-\frac{x^2}{2}}, & x \geq 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x < 0; \\ x\mathrm{e}^{-\frac{x^2}{2}}, & x \geq 0. \end{cases}$$

Распределение Рэлея является частным случаем распределения модуля многомерного нормального вектора при $n = 2$ (см. п. 3.4.3.6).

График функции плотности приведен на рис. 3.17.

3.4.3.5. Распределение Максвелла

Случайная величина, распределенная по закону Максвелла, имеет область определения $[0, +\infty)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x < 0; \\ \frac{1}{\sqrt{\pi}/2}\gamma(\frac{x^2}{2}, \frac{3}{2}), & x \geq 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x < 0; \\ \frac{2x^2}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}, & x \geq 0. \end{cases}$$

Распределение Максвелла является частным случаем распределения модуля многомерного нормального вектора при $n = 3$ (см. п. 3.4.3.6).

График функции плотности приведен на рис. 3.17.

3.4.3.5. Распределение модуля многомерного нормального вектора

Случайная величина, распределенная по закону модуля многомерного нормального вектора, имеет область определения $[0, +\infty)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x < 0; \\ \frac{1}{\Gamma(n/2)}\gamma(\frac{x^2}{2}, \frac{n}{2}), & x \geq 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x < 0; \\ \frac{2x^{n-1}}{2^{n/2}\Gamma(n/2)} \mathrm{e}^{-\frac{x^2}{2}}, & x \geq 0. \end{cases}$$

График функции плотности приведен на рис. 3.17.

Распределение модуля n-мерного нормального вектора относится к семейству гамма-распределений (см. п. 3.4.2.2) с генерирующей функцией $g(x) = \frac{x^2}{2}$ и целым параметром $n = 2\delta$.

Производные функции $g(x)$:

$$g'_x(x) = x; \;\; g''_{xx}(x) = 1; \;\; g'''_{xxx}(x) = 0.$$

3.4.3.7. Распределение Парето

Случайная величина, распределенная по закону Парето, имеет область определения $[1, +\infty)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x < 1; \\ 1 - \frac{1}{x^\beta}, & x \geq 1, \end{cases}$$

где $\beta > 0$ — параметр.

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x < 1; \\ \frac{\beta}{x^{\beta+1}}, & x \geq 1. \end{cases}$$

Распределение Парето относится к семейству бета-распределений (см. п. 3.4.2.3) с генерирующей функцией $g(x) = \frac{x-1}{x}, x > 1$ и параметром $\alpha = 1$.

Производные функции $g(x)$:

$$g'_x(x) = \frac{1}{x^2}; \;\; g''_{xx}(x) = -\frac{2}{x^3}; \;\; g'''_{xxx}(x) = \frac{6}{x^4}.$$

3.4.3.8. Распределение Эрланга

Случайная величина, распределенная по закону Эрланга, имеет область определения $[0, +\infty)$. Распределение Эрланга (или показательно-степенное распределение [87]) является частным случаем гамма-распределения с целым параметром $n = \alpha$ (см. п. 3.4.3.20).

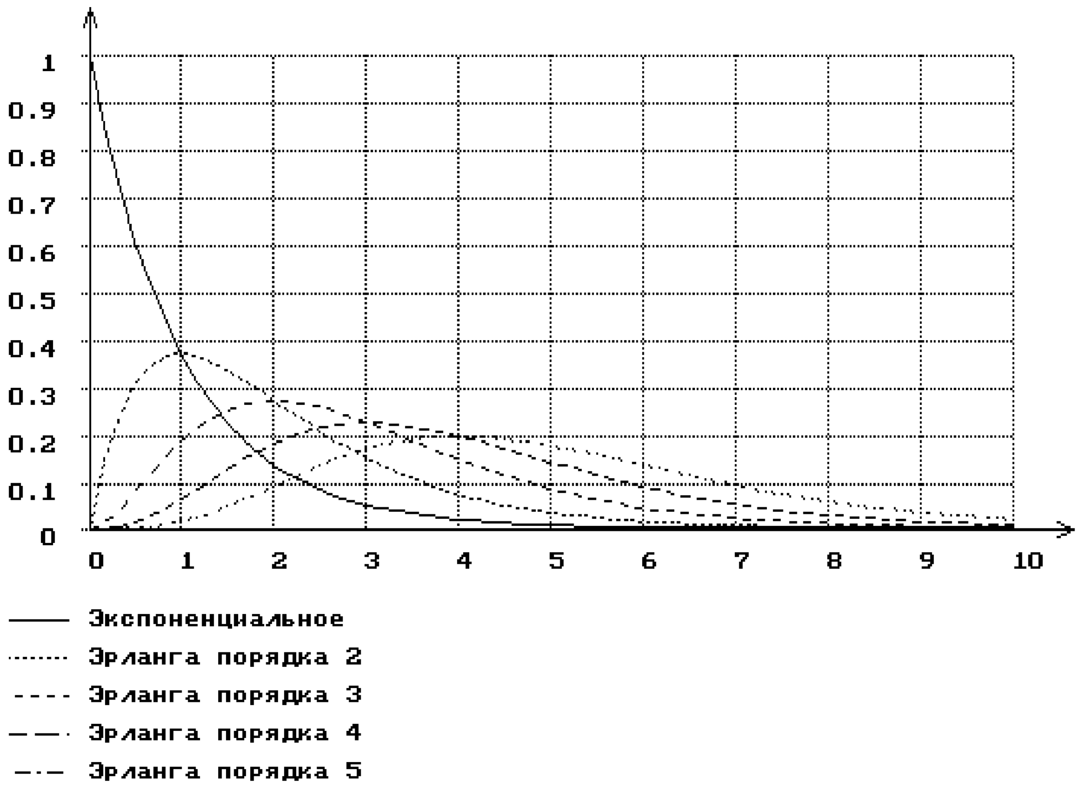


Рис. 3.18. Функции плотности распределения Эрланга с параметром формы $\alpha = 1, 2, 3, 4, 5$

График функции плотности приведен на рис. 3.18.

3.4.3.9. Распределение Лапласа

Случайная величина, распределенная по закону Лапласа, имеет область определения $(-\infty, +\infty)$.

Функция распределения [87]:

$$F(x) = \frac{1}{2} + \frac{\mathrm{sign(x)}}{2}\mathrm{e}^{-|x|}.$$

Функция плотности распределения:

$$f(x) = \frac{1}{2}\mathrm{e}^{-|x|}.$$

Распределение Лапласа является частным случаем двустороннего экспоненциального распределения с параметром $\alpha = 1$ (см. п. 3.4.3.27).

График функции плотности приведен на рис. 3.19.

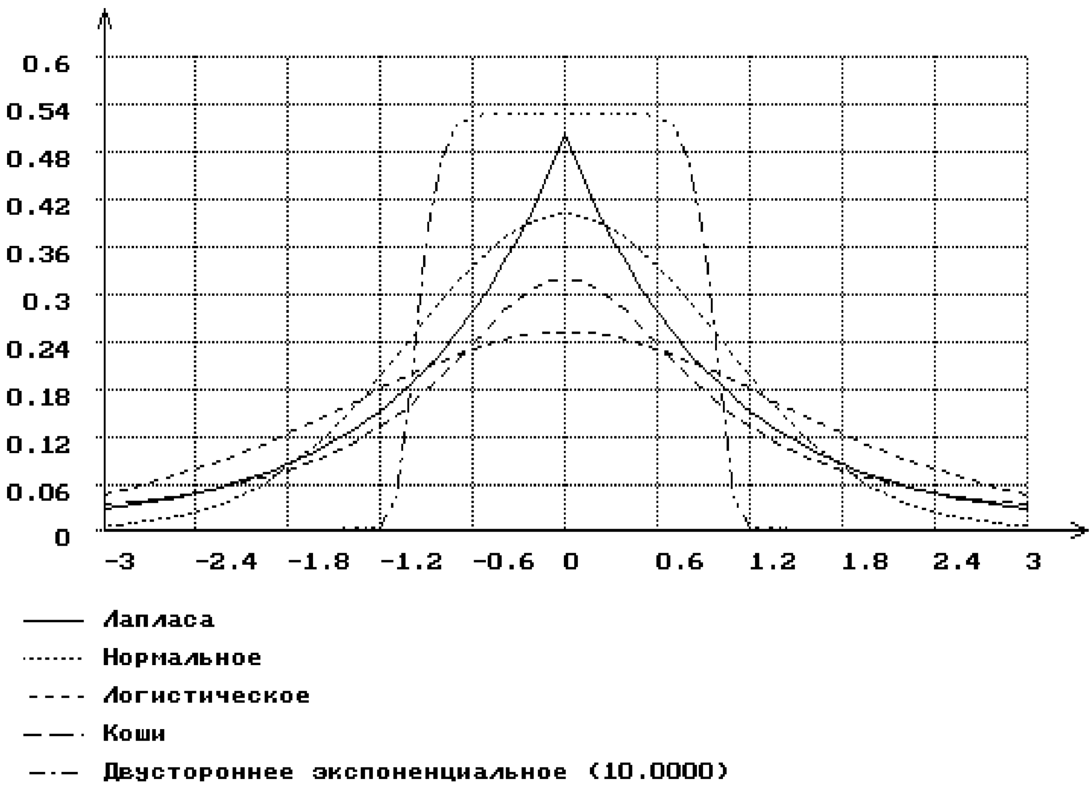


Рис. 3.19. Множество симметричных распределений: двустороннее экспоненциальное с параметром $\alpha = 10$, Лапласа, нормальное, логистическое, Коши

3.4.3.10. Нормальное распределение

Случайная величина, распределенная по нормальному закону, имеет область определения $(-\infty, +\infty)$.

Функция распределения [87]:

$$F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} \mathrm{e}^{-\frac{x^2}{2}} dx = \frac{1}{2} + \frac{\mathrm{sign(x)}}{2\sqrt{\pi}} \gamma(\frac{x^2}{2}, \frac{1}{2}).$$

Функция плотности распределения имеет вид:

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

Нормальное распределение является частным случаем двустороннего экспоненциального распределения с параметром $\alpha = 2$ и масштабом $\sqrt{2}$ (см. п. 3.4.3.27).

График функции плотности приведен на рис. 3.19.

3.4.3.11. Логарифмически (ln) нормальное распределение

Случайная величина, распределенная по логарифмически (ln) нормальному закону, имеет область определения $(0, +\infty)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x \le 0; \\ \Phi(\ln x) = \frac{1}{2} + \frac{\mathrm{sign(\ln x)}}{2\sqrt{\pi}} \gamma(\frac{(\ln x)^2}{2}, \frac{1}{2}), & x > 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \le 0; \\ \frac{1}{x\sqrt{2\pi}} \mathrm{e}^{-\frac{(\ln x)^2}{2}} & x > 0. \end{cases}$$

Логарифмически нормальное распределение получается из нормального применением операции логарифмирования (см. п. 3.4.1.7) и является частным случаем распределения S_L-Джонсона при $\alpha = \beta = 1$ (см. п. 3.4.3.25).

График функции плотности приведен на рис. 3.20.

3.4.3.12. Логарифмически (lg) нормальное распределение

Случайная величина, распределенная по логарифмически (lg) нормальному закону, имеет область определения $(0, +\infty)$.

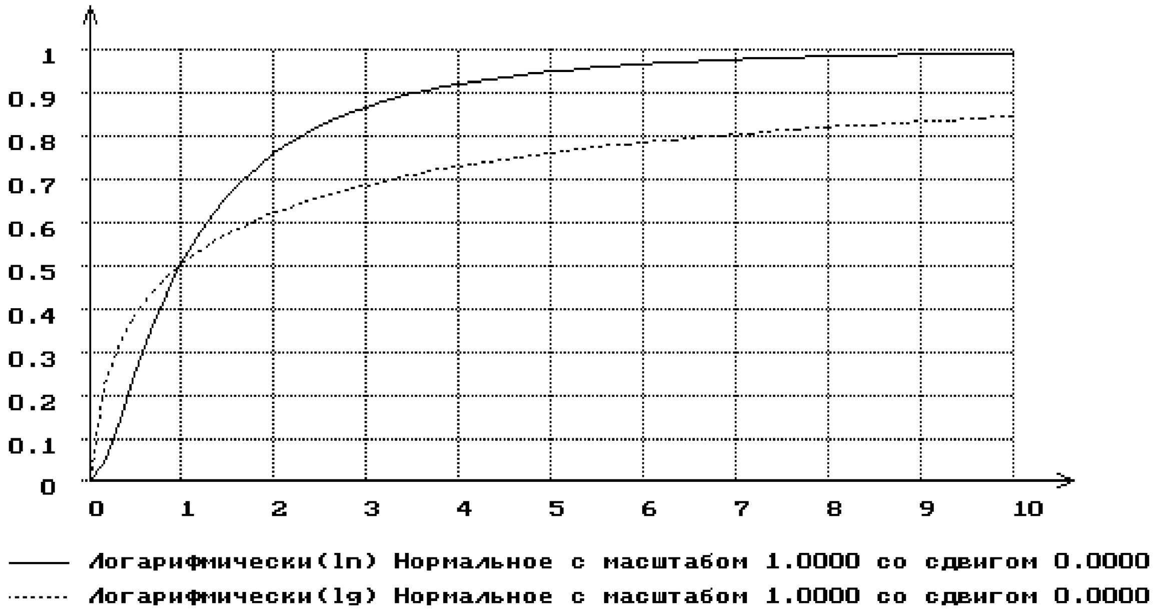


Рис. 3.20. Функции распределения логарифмически (ln) и логарифмически (lg) нормальных распределений

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x \le 0; \\ \Phi(\lg x) = \frac{1}{2} + \frac{\mathrm{sign(lg\,x)}}{2\sqrt{\pi}}\gamma(\frac{(\lg x)^2}{2}, \frac{1}{2}), & x > 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \le 0; \\ \frac{1}{x \ln 10 \sqrt{2\pi}} \mathrm{e}^{-\frac{(\lg x)^2}{2}} & x > 0. \end{cases}$$

Логарифмически (lg) нормальное распределение получается из логарифмически (ln) нормального применением операции масштабирования с параметром $\lambda = \ln 10$ (см. п. 3.4.1.2).

График функции плотности приведен на рис. 3.20.

3.4.3.13. Распределение Коши

Случайная величина, распределенная по закону Коши (Брейта-Вигнера, арктангенса), имеет область определения $(-\infty, +\infty)$.

Функция распределения [87]:

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\mathrm{arctg}x.$$

Функция плотности распределения:

$$f(x) = \frac{1}{\pi(1 + x^2)}.$$

Производные функция плотности распределения:

$$f'_x(x) = \frac{2x}{\pi(1 + x^2)^2},$$

$$f''_{xx}(x) = \frac{8x^2(1 + x^2)}{\pi(1 + x^2)^3} - \frac{2}{\pi(1 + x^2)^2}.$$

График функции плотности приведен на рис. 3.19.

3.4.3.14. Логистическое распределение

Случайная величина, распределенная по логистическому закону, имеет область определения $(-\infty, +\infty)$.

Функция распределения [87]:

$$F(x) = \frac{1}{1 + \mathrm{e}^{-x}}.$$

Функция плотности распределения:

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^x}{(1+e^x)^2}.$$

Логистическое распределение является частным случаем L-распределения при $\alpha = \beta = 1$ (см. п.3.4.3.30).

График функции плотности приведен на рис. 3.19.

3.4.3.15. Распределение Вейбулла

Случайная величина, распределенная по закону Вейбулла (Вейбулла-Гнеденко, экстремальных значений III-го типа), имеет область определения $(0, +\infty)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x \leq 0; \\ 1 - e^{-x^\alpha}, & x > 0. \end{cases}$$

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \leq 0; \\ \alpha x^{\alpha-1} e^{-x^\alpha} & x > 0. \end{cases}$$

Распределение Вейбулла является частным случаем Г-распределения при $\delta = 1$ (см. п. 3.4.3.29).

График функции плотности приведен на рис. 3.16.

3.4.3.16. Распределение минимального значения

Случайная величина, распределенная по закону минимального значения (Гумбеля, экстремальных значений II-го типа), имеет область определения $(-\infty, +\infty)$.

Функция распределения [87]:

$$F(x) = 1 - e^{-e^x}.$$

Функция плотности распределения:

$$f(x) = e^{x - e^x}.$$

Распределение минимального значения является частным случаем обобщенного распределения минимального значения при $\delta = 1$ (см. п.3.4.3.18).

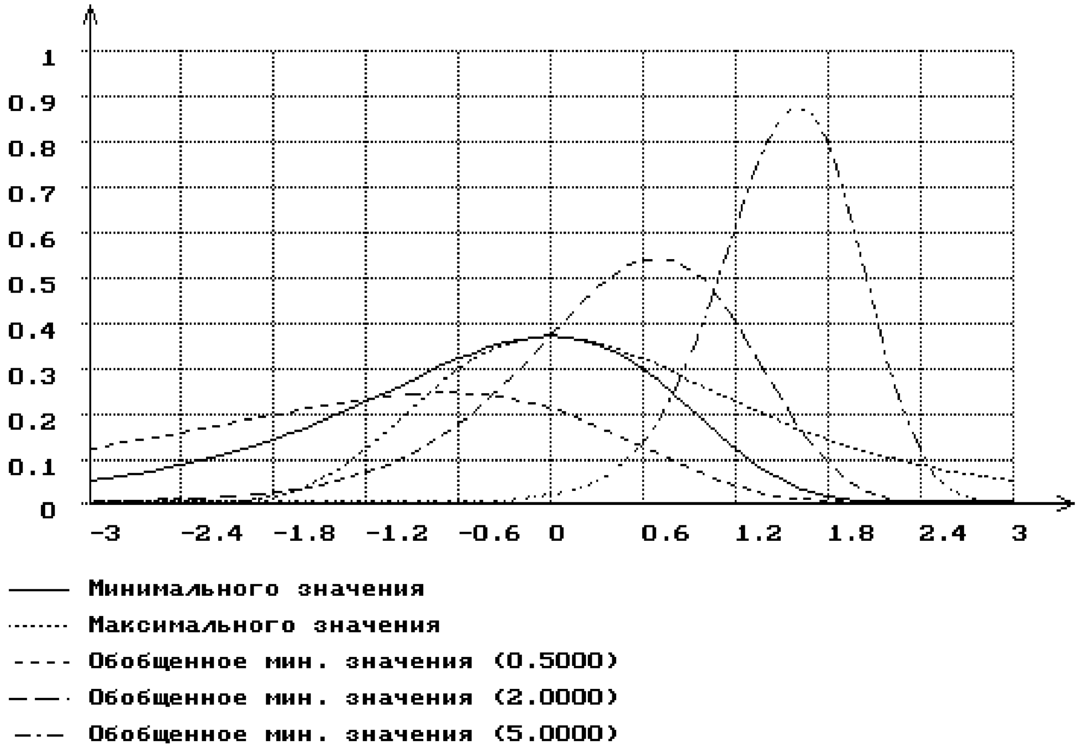


Рис. 3.21. Функции плотности распределений минимального значения, максимального значения и обобщенного минимального значения с параметром $\delta = 0.5, 1, 2, 5$

График функции плотности приведен на рис. 3.21.

3.4.3.17. 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 [87]:

$$F(x) = \mathrm{e}^{-\mathrm{e}^{-x}}.$$

Функция плотности распределения:

$$f(x) = \mathrm{e}^{-x-\mathrm{e}^{-x}}.$$

Распределение максимального значения получается из распределения минимального значения применением операции поворота (см. п.3.4.1.3).

График функции плотности приведен на рис. 3.21.

3.4.3.18. Обобщенное распределение минимального значения

Случайная величина, распределенная по закону обобщенного минимального значения, имеет область определения $(-\infty, +\infty)$.

Функция распределения:

$$F(x) = \frac{1}{\Gamma(\delta)}\gamma(\mathrm{e}^x, \delta),$$

где параметр $\delta > 0$.

Функция плотности распределения:

$$f(x) = \frac{1}{\Gamma(\delta)}\mathrm{e}^{\delta x-\mathrm{e}^x}.$$

Обобщенное распределение минимального значения принадлежит к семейству гамма-распределений (см. п.3.4.2.2) с генерирующей функцией $g(x) = \mathrm{e}^x$.

Производные функции $g(x)$:

$$g'_x(x) = g''_{xx}(x) = g'''_{xxx}(x) = \mathrm{e}^x.$$

График функции плотности приведен на рис. 3.21.

3.4.3.19. Распределение Накагами

Случайная величина, распределенная по закону Накагами, имеет область определения $(0, +\infty)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x \le 0; \\ \frac{1}{\Gamma(\alpha)}\gamma(\alpha x^2, \alpha), & x > 0. \end{cases},$$

где параметр $\alpha \ge 1/2$.

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \le 0; \\ \frac{2\alpha^\alpha x^{2\alpha-1}}{\Gamma(\alpha)}\mathrm{e}^{-\alpha x^2}, & x > 0. \end{cases}$$

Распределение Накагами также принадлежит к семейству гамма-распределений (см. п.3.4.2.2), но имеет особенность в том, что генерирующая функция $g(x) = \alpha x^2$ зависит от основного параметра, и поэтому все производные принимают другой вид.

Производные функции распределения по параметру имеют вид:

$$\begin{aligned} F'_\alpha(x) &= \frac{1}{\Gamma(\alpha)}\{\gamma'_\alpha(\alpha x^2, \alpha) + \alpha^{\alpha-1}x^{2\alpha}\mathrm{e}^{-\alpha x^2} - \psi(\alpha)\gamma(\alpha x^2, \alpha)\}, \\ F''_{\alpha\alpha}(x) &= \frac{1}{\Gamma(\alpha)}\{\gamma''_{\alpha\alpha}(\alpha x^2, \alpha) + \alpha^{\alpha-1}x^{2\alpha}\mathrm{e}^{-\alpha x^2}[2\ln(\alpha x^2) + \frac{\alpha-1}{\alpha} - x^2 - \\ &\quad -2\psi(\alpha)] - 2\psi(\alpha)\gamma'_\alpha(\alpha x^2, \alpha) + [\psi^2(\alpha) - \psi'_\alpha(\alpha)]\gamma(\alpha x^2, \alpha)\}, \end{aligned}$$

Производные функции плотности распределения по параметру имеют вид:

$$\begin{aligned} f'_\alpha(x) &= \frac{2}{\Gamma(\alpha)}\alpha^\alpha x^{2\alpha-1}\mathrm{e}^{-\alpha x^2}\{1 + \ln(\alpha x^2) - x^2 - \psi(\alpha)\}, \\ f''_{\alpha\alpha}(x) &= \frac{2}{\Gamma(\alpha)}\alpha^\alpha x^{2\alpha-1}\mathrm{e}^{-\alpha x^2}\{[1 + \ln(\alpha x^2) - x^2 - \psi(\alpha)]^2 + 1/\alpha - \psi'_\alpha(\alpha)\}. \end{aligned}$$

Производная функции плотности по x и ее производные по параметру имеют вид:

$$\begin{aligned} f'_x(x) &= \frac{2}{\Gamma(\alpha)}\alpha^\alpha x^{2\alpha-2}\mathrm{e}^{-\alpha x^2}\{2\alpha(1 - x^2) - 1\}, \\ f''_{x\alpha}(x) &= \frac{2}{\Gamma(\alpha)}\alpha^\alpha x^{2\alpha-2}\mathrm{e}^{-\alpha x^2}\{(2\alpha(1 - x^2) - 1)(1 + \ln(\alpha x^2) - x^2 - \psi(\alpha)) + \\ &\quad +2(1 - x^2)\} \end{aligned}$$

Вторая производная функции плотности по x имеет вид:

$$f''_{xx}(x) = \frac{2}{\Gamma(\alpha)}\alpha^{\alpha}x^{2\alpha-3}\mathrm{e}^{-\alpha x^2}\{2(2\alpha(1-x^2)-1)(\alpha-1-\alpha x^2)-4\alpha x^2\}.$$

3.4.3.20. Гамма-распределение

Случайная величина, имеющая Гамма-распределение (распределение Пирсона III), определена на области $(0,+\infty)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x \le 0; \\ \frac{1}{\Gamma(\delta)}\gamma(x,\delta), & x > 0. \end{cases},$$

где параметр $\delta > 0$.

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \le 0; \\ \frac{1}{\Gamma(\delta)}x^{\delta-1}\mathrm{e}^{-x}, & x > 0. \end{cases}$$

Гамма-распределение является частным случаем Г-распределения при $\alpha = 1$ (см. п. 3.4.3.29).

3.4.3.21. Бета-распределение I-го рода

Случайная величина, имеющая бета-распределение I-го рода, определена на области $[0,1]$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x < 0; \\ \frac{1}{\mathrm{B}(\alpha,\beta)}\mathcal{B}(x,\alpha,\beta), & 0 \le x \le 1; \\ 1, & x > 1; \end{cases}$$

где параметры $\alpha > 0$ и $\beta > 0$.

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \le 0 \vee x \ge 1; \\ \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, & 0 < x < 1. \end{cases}$$

Бета-распределение I-го рода принадлежит к семейству бета-распределений с генерирующей функцией $g(x) = x$ (см. п. 3.4.2.3).

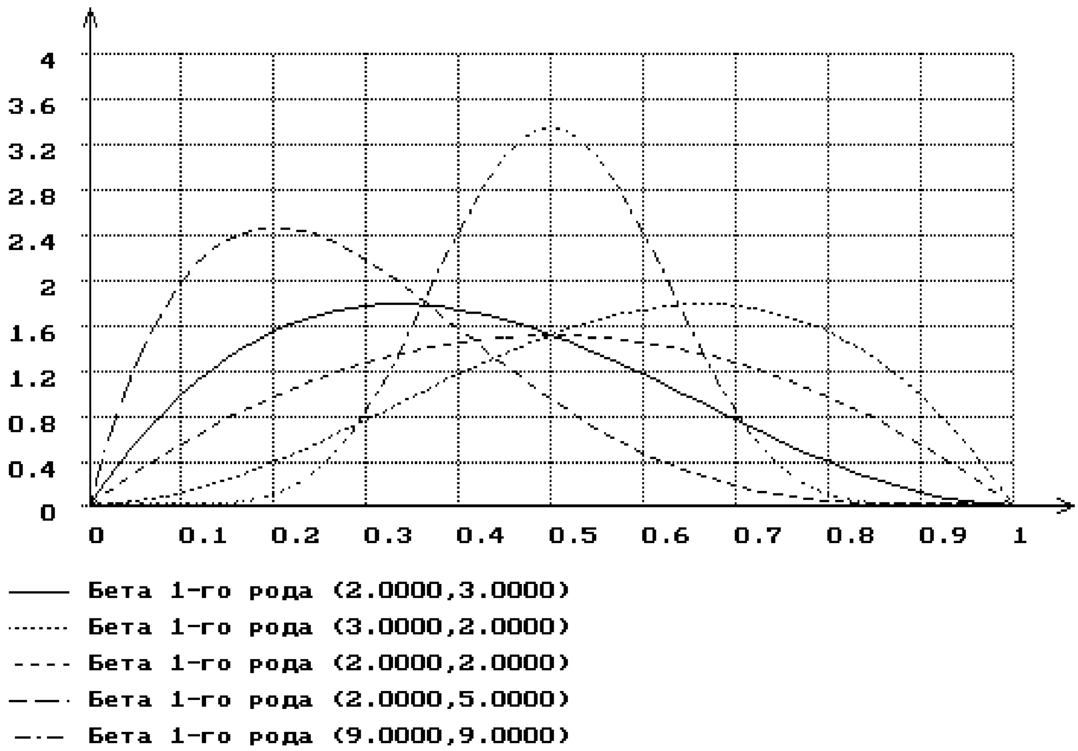


Рис. 3.22. Функции плотности Бета-распределения I-го рода

Производные функции $g(x)$:

$$g'_x(x) = 1,\ g''_{xx}(x) = g'''_{xxx}(x) = 0.$$

График функции плотности приведен на рис. 3.22.

3.4.3.22. Бета-распределение II-го рода

Случайная величина, имеющая бета-распределение II-го рода, определена на области $[0, +\infty)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x < 0; \\ \frac{1}{\mathrm{B}(\alpha,\beta)}\mathcal{B}(\frac{x}{1+x}, \alpha, \beta), & x \leq 0; \end{cases}$$

где параметры $\alpha > 0$ и $\beta > 0$.

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \leq 0; \\ \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1+x)^{-\alpha-\beta}, & x > 0. \end{cases}$$

Бета-распределение II-го рода принадлежит к семейству бета-распределений с генерирующей функцией $g(x) = \frac{x}{1+x}$ (см. п. 3.4.2.3).

Производные функции $g(x)$:

$$g'_x(x) = \frac{1}{(1+x)^2},\ g''_{xx}(x) = -\frac{2}{(1+x)^3},\ g'''_{xxx}(x) = \frac{6}{(1+x)^4}.$$

График функции плотности приведен на рис. 3.23.

3.4.3.23. Бета-распределение III-го рода

Случайная величина, имеющая бета-распределение III-го рода, определена на области $[0, 1]$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x < 0; \\ \frac{1}{\mathrm{B}(\alpha,\beta)}\mathcal{B}(\frac{\delta x}{1+(\delta-1)x}, \alpha, \beta), & 0 \leq x \leq 1; \\ 1, & x > 1; \end{cases}$$

где параметры $\alpha > 0$, $\beta > 0$ и $\delta > 0$.

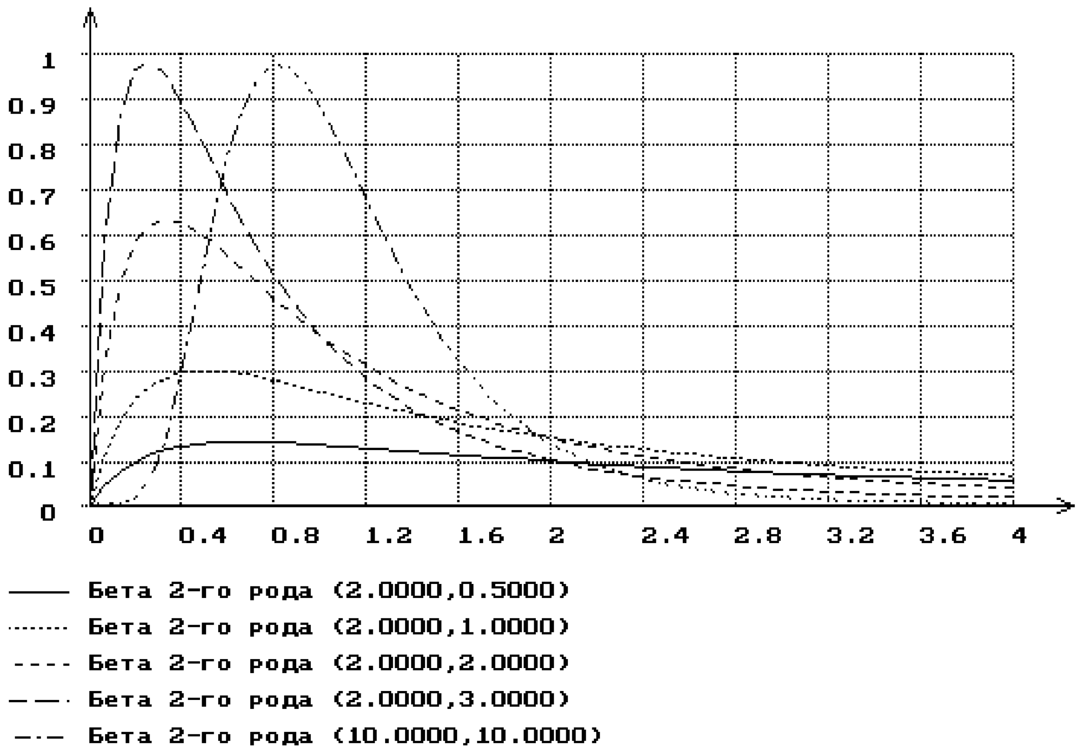


Рис. 3.23. Функции плотности Бета-распределения II-го рода

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \le 0 \vee x \ge 1; \\ \frac{\delta^{\alpha}}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}(1+(\delta-1)x)^{-\alpha-\beta}, & 0 < x < 1. \end{cases}$$

Бета-распределение III-го рода принадлежит к семейству бета-распределений с генерирующей функцией $g(x,\delta) = \frac{\delta x}{1+(\delta-1)x}$ (см. п. 3.4.2.3).

Производные функции $g(x,\delta)$:

$$\begin{aligned} g'_x(x) &= \frac{\delta}{[1+(\delta-1)x]^2}, \\ g'_\delta(x) &= \frac{x(1-x)}{[1+(\delta-1)x]^2}; \\ g''_{xx}(x) &= -\frac{2\delta(\delta-1)}{[1+(\delta-1)x]^3}, \\ g''_{\delta\delta}(x) &= -\frac{2x^2(1-x)}{[1+(\delta-1)x]^3}, \\ g''_{x\delta}(x) &= \frac{1-x(1+\delta)}{[1+(\delta-1)x]^3}; \\ g'''_{xxx}(x) &= \frac{6\delta(\delta-1)^2}{[1+(\delta-1)x]^4}, \\ g'''_{xx\delta}(x) &= \frac{2[(\delta^2-1)x+1-2\delta]}{[1+(\delta-1)x]^4}, \\ g'''_{x\delta\delta}(x) &= \frac{2x(1-x(\delta+2))}{[1+(\delta-1)x]^4}. \end{aligned}$$

3.4.3.24. Распределение S_B-Джонсона

Случайная величина, распределенная по закону S_B-Джонсона, определена на области $(0,1)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x \le 0; \\ \Phi(\alpha + \beta \ln \frac{x}{1-x}), & 0 < x < 1; \\ 1, & x > 1, \end{cases}$$

где параметры $|\alpha| < \infty$ и $\beta > 0$.

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \le 0 \vee x \ge 1; \\ \frac{\beta}{x(1-x)\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(\alpha+\beta \ln \frac{x}{1-x})^2}, & 0 < x < 1. \end{cases}$$

Распределение S_B-Джонсона принадлежит к семейству распределений Джонсона с генерирующей функцией $g(x) = \ln\frac{x}{1-x}$ (см. п. 3.4.2.1).

Производные функции $g(x)$:

$$g'_x(x) = \frac{1}{x(1-x)},\ g''_{xx}(x) = \frac{1-2x}{x^2(1-x)^2},\ g'''_{xxx}(x) = \frac{2(x^2-3x+1)}{x^3(1-x)^3}.$$

3.4.3.25. Распределение S_L-Джонсона

Случайная величина, распределенная по закону S_L-Джонсона, определена на области $(0, +\infty)$.

Функция распределения [87]:

$$F(x) = \begin{cases} 0, & x \le 0; \\ \Phi(\alpha + \beta\ln x), & x > 0, \end{cases}$$

где параметры $|\alpha| < \infty$ и $\beta > 0$.

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \le 0; \\ \frac{\beta}{x\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta\ln x)^2}, & x > 0. \end{cases}$$

Распределение S_L-Джонсона принадлежит к семейству распределений Джонсона с генерирующей функцией $g(x) = \ln x$ (см. п. 3.4.2.1).

Производные функции $g(x)$:

$$g'_x(x) = \frac{1}{x},\ g''_{xx}(x) = -\frac{1}{x^2},\ g'''_{xxx}(x) = \frac{2}{x^3}.$$

3.4.3.26. Распределение S_U-Джонсона

Случайная величина, распределенная по закону S_U-Джонсона, определена на области $(-\infty, +\infty)$.

Функция распределения [87]:

$$F(x) = \Phi(\alpha + \beta\ln(x + \sqrt{x^2+1})),$$

где параметры $|\alpha| < \infty$ и $\beta > 0$.

Функция плотности распределения:

$$f(x) = \frac{\beta}{\sqrt{2\pi}\sqrt{x^2+1}}\mathrm{e}^{-\frac{1}{2}(\alpha+\beta\ln(x+\sqrt{x^2+1}))^2}.$$

Распределение S_U-Джонсона принадлежит к семейству распределений Джонсона с генерирующей функцией $g(x)=\ln(x+\sqrt{x^2+1})$ (см. п. 3.4.2.1).

Производные функции $g(x)$:

$$g'_x(x)=\frac{1}{\sqrt{x^2+1}},\ g''_{xx}(x)=-\frac{x}{(x^2+1)^{3/2}},\ g'''_{xxx}(x)=\frac{-2x^2+3x-1}{2(x^2+1)^{5/2}}.$$

3.4.3.27. Двустороннее экспоненциальное распределение

Случайная величина, распределенная по двусторонему экспоненциальному закону (класс экспоненциальных распределений), определена на области $(-\infty,+\infty)$.

Функция распределения [87]:

$$F(x)=\frac{1}{2}\left[1+\frac{\mathrm{sign(x)}}{\Gamma(1/\alpha)}\gamma(|x|^\alpha,1/\alpha)\right],$$

где параметр $\alpha>0$.

Функция плотности распределения:

$$f(x)=\frac{\alpha}{2\Gamma(1/\alpha)}\mathrm{e}^{-|x|^\alpha}.$$

График функции плотности приведен на рис. 3.15.

Двустороннее экспоненциальное распределение является частным случаем H-распределения при $\delta=1/\alpha$ (см. п. 3.4.3.28). Однако, как и в случае распределения Накагами, из-за функциональной связи между параметрами, производные будут иметь другой вид.

Производные функции распределения по параметру имеют вид:

$$\begin{aligned}
F'_\alpha(x) &= \frac{\mathrm{sign(x)}}{2\alpha^2\Gamma(1/\alpha)}\left\{-\gamma'(|x|^\alpha,1/\alpha)+\gamma(|x|^\alpha,1/\alpha)\psi(1/\alpha)+\alpha^2\mathrm{e}^{-|x|^\alpha}\ln|x|\right\},\\
F''_{\alpha\alpha}(x) &= \frac{\mathrm{sign(x)}}{2\alpha^4\Gamma(1/\alpha)}\left\{\gamma''(|x|^\alpha,1/\alpha)+(\alpha-\psi(1/\alpha))\gamma'(|x|^\alpha,1/\alpha)-\right.\\
&\quad -(2\alpha+\psi'(1/\alpha)-\psi^2(1/\alpha))\gamma(|x|^\alpha,1/\alpha)+\alpha^2\mathrm{e}^{-|x|^\alpha}\times\\
&\quad \left.\times(2\alpha\psi(1/\alpha)-\alpha\ln|x|)\ln|x|\right\}.
\end{aligned}$$

Производные функции плотности распределения по параметру имеют вид:

$$f'_\alpha(x) = \frac{1}{2\Gamma(1/\alpha)}\mathrm{e}^{-|x|^\alpha}\left\{1-\alpha|x|^\alpha\ln|x|+\frac{1}{\alpha}\psi(1/\alpha)\right\},$$

$$f''_{\alpha\alpha}(x) = \frac{1}{2\Gamma(1/\alpha)}\mathrm{e}^{-|x|^\alpha}\left\{[-2+\alpha|x|^\alpha\ln|x| - \frac{\alpha+1}{\alpha}\psi(1/\alpha) - \alpha\ln|x|]\times\right.$$
$$\left.\times|x|^\alpha\ln|x| - \frac{1}{\alpha^3}[\psi'(1/\alpha) - \psi^2(1/\alpha)]\right\}.$$

Производная функции плотности по x и ее производные по параметру имеют вид:

$$f'_x(x) = -\frac{\alpha^2|x|^{\alpha-1}\mathrm{sign(x)}}{2\Gamma(1/\alpha)}\mathrm{e}^{-|x|^\alpha},$$
$$f''_{x\alpha}(x) = -\frac{\alpha\,\mathrm{sign(x)|x|}^{\alpha-1}}{2\Gamma(1/\alpha)}\mathrm{e}^{-|x|^\alpha}\left\{2+\alpha(1-|x|^\alpha)\ln|x| + \frac{1}{\alpha}\psi(1/\alpha)\right\}.$$

Вторая производная функции плотности по x имеет вид:

$$f''_{xx}(x) = \frac{\alpha^3|x|^{\alpha-2}(1/\alpha - 1 + |x|^{\alpha-1})}{2\Gamma(1/\alpha)}\mathrm{e}^{-|x|^\alpha}.$$

3.4.3.28. Н-распределение

Случайная величина, имеющая Н-распределение, определена на области $(-\infty, +\infty)$.

Функция распределения [87]:

$$F(x) = \frac{1}{2}\left[1 + \frac{\mathrm{sign(x)}}{\Gamma(\delta)}\gamma(|x|^\alpha, \delta),\right]$$

где параметры $\alpha > 0$ и $\delta > 0$.

Функция плотности распределения:

$$f(x) = \frac{\alpha}{\Gamma(\delta)}|x|^{\alpha\delta-1}\mathrm{e}^{-|x|^\alpha},$$

Н-распределение в общем случае (при $\delta \neq 1/\alpha$) является двухмодальным и симметричным относительно нуля.

График функции плотности приведен на рис. 3.24.

Производные функции распределения по параметрам имеют вид:

$$F'_\alpha(x) = \frac{\mathrm{sign(x)}}{2\Gamma(\delta)}|x|^{\alpha\delta}\mathrm{e}^{-|x|^\alpha}\ln|x|,$$
$$F'_\delta(x) = \frac{\mathrm{sign(x)}}{2\Gamma(\delta)}\left\{\gamma'(|x|^\alpha, \delta) - \psi(\delta)\gamma(|x|^\alpha, \delta)\right\},$$
$$F''_{\alpha\alpha}(x) = \frac{\mathrm{sign(x)}}{2\Gamma(\delta)}|x|^{\alpha\delta}\mathrm{e}^{-|x|^\alpha}(\delta - |x|^\alpha)\ln^2|x|,$$

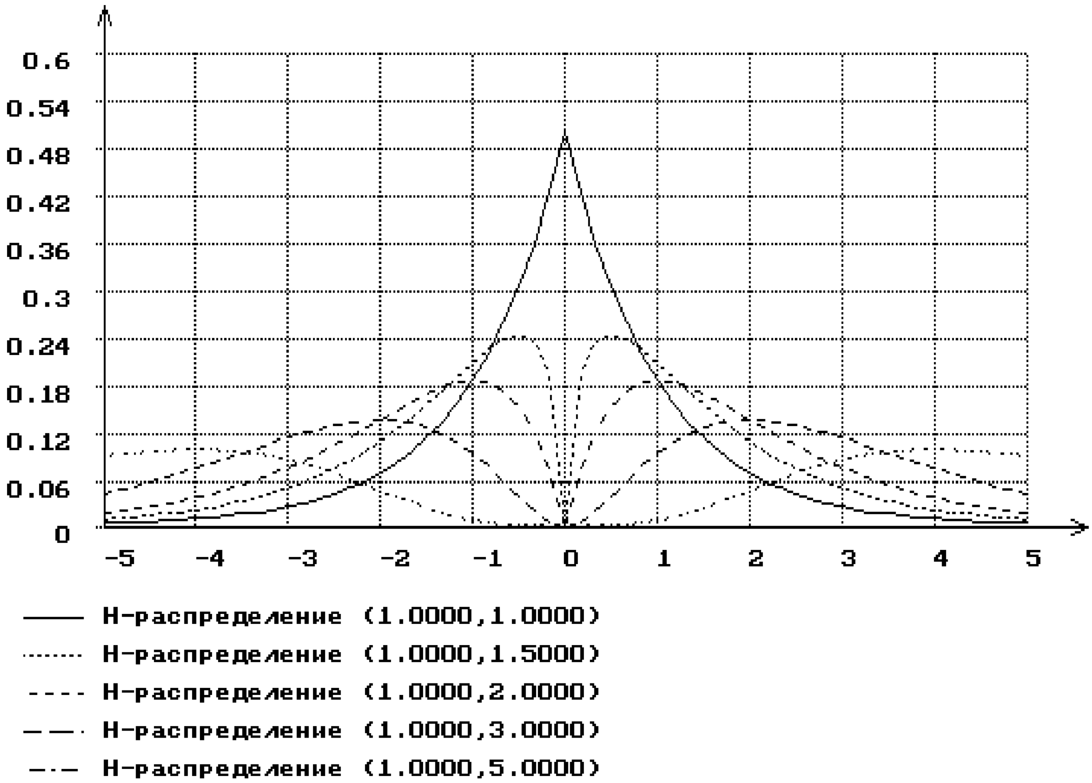


Рис. 3.24. Функции плотности Н-распределения

$$
\begin{aligned}
F''_{\delta\delta}(x) &= \frac{\mathrm{sign(x)}}{2\Gamma(\delta)}\left\{\gamma''(|x|^{\alpha},\delta) - 2\psi(\delta)\gamma'(|x|^{\alpha},\delta)[psi^2(\delta) - psi'(\delta)]\gamma(|x|^{\alpha},\delta)\right\}, \\
F''_{\alpha\delta}(x) &= \frac{\mathrm{sign(x)}}{2\Gamma(\delta)}|x|^{\alpha\delta}\mathrm{e}^{-|x|^{\alpha}}\ln|x|\left\{\alpha\ln|x| - \psi(\delta)\right\}.
\end{aligned}
$$

Производные функции плотности распределения по параметрам имеют вид:

$$
\begin{aligned}
f'_{\alpha}(x) &= \frac{1}{2\Gamma(\delta)}|x|^{\alpha\delta-1}\mathrm{e}^{-|x|^{\alpha}}\left\{1 + \alpha(\delta - |x|^{\alpha})\ln|x|\right\}, \\
f'_{\delta}(x) &= \frac{\alpha}{2\Gamma(\delta)}|x|^{\alpha\delta-1}\mathrm{e}^{-|x|^{\alpha}}\left\{\alpha\ln|x| - \psi(\delta)\right\}, \\
f''_{\alpha\alpha}(x) &= \frac{1}{2\Gamma(\delta)}|x|^{\alpha\delta-1}\mathrm{e}^{-|x|^{\alpha}}\left\{2(\delta - |x|^{\alpha}) + \alpha(\delta - |x|^{\alpha})^2\ln|x| - \alpha|x|^{\alpha}\ln^2|x|\right\}, \\
f''_{\alpha\delta}(x) &= \frac{1}{2\Gamma(\delta)}|x|^{\alpha\delta-1}\mathrm{e}^{-|x|^{\alpha}}\left\{[1 + \alpha(\delta - |x|^{\alpha})\ln|x|][\alpha\ln|x| - \psi(\delta)] + \right. \\
&\quad \left. +\alpha\ln|x|\right\}.
\end{aligned}
$$

Производная функции плотности по x и ее производные по параметрам имеют вид:

$$
\begin{aligned}
f'_{x}(x) &= \frac{\alpha\,\mathrm{sign(x)}}{2\Gamma(\delta)}|x|^{\alpha\delta-2}\mathrm{e}^{-|x|^{\alpha}}\left\{\alpha(\delta - |x|^{\alpha}) - 1\right\}, \\
f''_{x\alpha}(x) &= \frac{\mathrm{sign(x)}}{2\Gamma(\delta)}|x|^{\alpha\delta-2}\mathrm{e}^{-|x|^{\alpha}}\left\{[\alpha(\delta - |x|^{\alpha}) - 1][2 + \alpha\delta\ln|x| - \alpha|x|^{\alpha}\ln|x|] + \right. \\
&\quad \left. +1 - \alpha^2|x|^{\alpha}\ln|x|\right\}, \\
f''_{x\delta}(x) &= \frac{\alpha\,\mathrm{sign(x)}}{2\Gamma(\delta)}|x|^{\alpha\delta-2}\mathrm{e}^{-|x|^{\alpha}}\left\{[\alpha(\delta - |x|^{\alpha}) - 1][\alpha\ln|x| - \psi(\delta)] + \alpha\right\}.
\end{aligned}
$$

Вторая производная функции плотности по x имеет вид:

$$
f''_{xx}(x) = \frac{\alpha}{2\Gamma(\delta)}|x|^{\alpha\delta-3}\mathrm{e}^{-|x|^{\alpha}}\left\{[\alpha(\delta - |x|^{\alpha}) - 1][\alpha(\delta - |x|^{\alpha}) - 2] - \alpha^2|x|^{\alpha}\right\}.
$$

3.4.3.29. Г-распределение

Случайная величина, имеющая Г-распределение, определена на области $(0, +\infty)$.

Функция распределения [87]:

$$
F(x) = \begin{cases} 0, & x \le 0; \\ \frac{1}{\Gamma(\delta)}\gamma(x^{\alpha},\delta), & x > 0, \end{cases}
$$

где параметры $\alpha > 0$ и $\delta > 0$.

Функция плотности распределения:

$$f(x) = \begin{cases} 0, & x \le 0; \\ \frac{\alpha}{\Gamma(\delta)} x^{\alpha\delta - 1} \mathrm{e}^{-x^{\alpha}}, & x > 0. \end{cases}$$

Г-распределение принадлежит к семейству гамма-распределений с генерирующей функцией $g(x, \alpha) = x^{\alpha}$ (см. п. 3.4.2.2).

Производные функции $g(x, \alpha)$:

$$\begin{aligned}
g'_x(x) &= \alpha x^{\alpha-1}, \\
g'_\alpha(x) &= x^{\alpha} \ln x, \\
g''_{xx}(x) &= \alpha(\alpha-1)x^{\alpha-2}, \\
g''_{x\alpha}(x) &= x^{\alpha-1}(1 + \alpha \ln x), \\
g''_{\alpha\alpha}(x) &= x^{\alpha} \ln^2 x, \\
g'''_{xxx}(x) &= \alpha(\alpha-1)(\alpha-2)x^{\alpha-3}, \\
g'''_{xx\alpha}(x) &= x^{\alpha-2}(2\alpha - 1 + \alpha(\alpha-1)\ln x), \\
g'''_{x\alpha\alpha}(x) &= x^{\alpha-1} \ln x (2 + \alpha \ln x), \\
g'''_{\alpha\alpha\alpha}(x) &= x^{\alpha} \ln^3 x.
\end{aligned}$$

3.4.3.30. Обобщенное логистическое распределение

Случайная величина, имеющая обобщенное логистическое распределение, определена на области $(-\infty, +\infty)$.

Функция распределения [87]:

$$F(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} \mathcal{B}\left(\frac{\mathrm{e}^x}{1 + \mathrm{e}^x}, \alpha, \beta\right),$$

где параметры $\alpha > 0$ и $\beta > 0$.

Функция плотности распределения:

$$f(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} \frac{\mathrm{e}^{\alpha x}}{(1 + \mathrm{e}^x)^{\alpha + \beta}}.$$

Обобщенное логистическое распределение принадлежит к семейству бета-распределений с генерирующей функцией $g(x) = \frac{\mathrm{e}^x}{1 + \mathrm{e}^x}$ (см. п. 3.4.2.3).

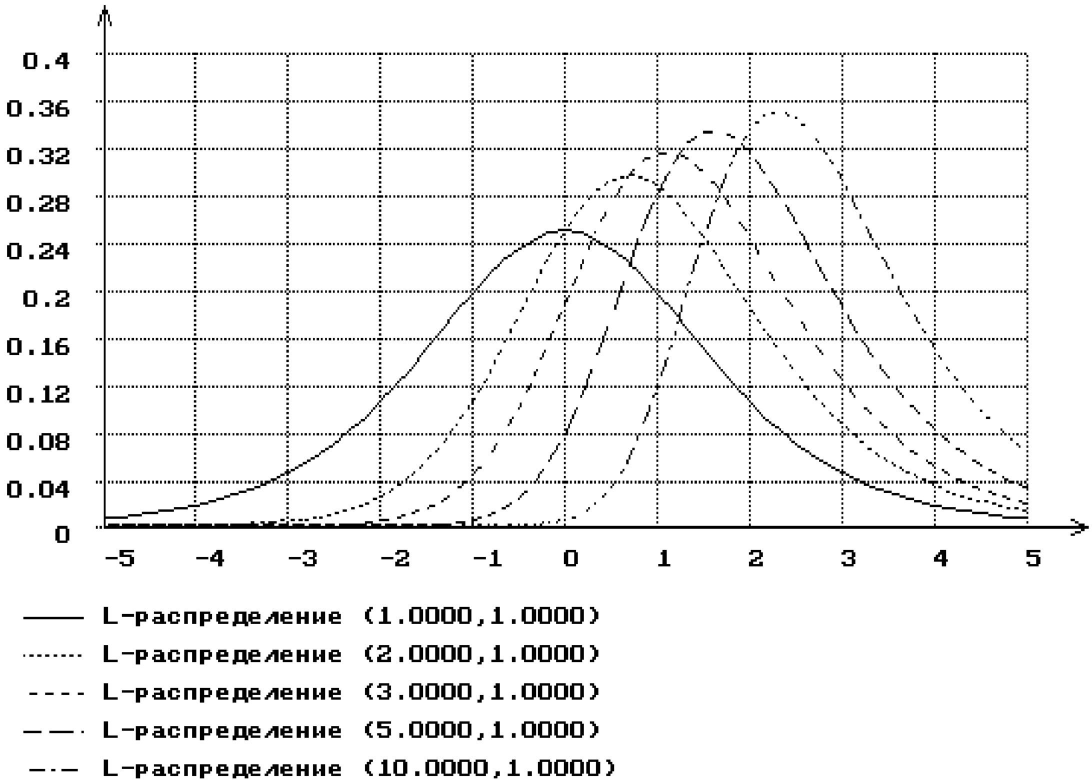


Рис. 3.25. Функции плотности обобщенного логистического распределения

Производные функции $g(x)$:

$$g'_x(x) = \frac{e^x}{(1+e^x)^2},\ g''_{xx}(x) = \frac{e^x(1-e^x)}{(1+e^x)^3},\ g'''_{xxx}(x) = \frac{e^x(1+2e^x-5e^{2x})}{(1+e^x)^4}.$$

График функции плотности приведен на рис. 3.25.

3.4.4. Распределения статистик критериев согласия

Распределения, приведенные в этом пункте, являются предельными для распределений статистик соответствующих критериев согласия. При вычислении вероятности согласия достаточно знать только функцию распределения.

3.4.4.1. Распределение χ^2 Пирсона

Случайная величина, распределенная по закону χ^2 Пирсона, определена на интервале $(0, +\infty)$.

Функция распределения [88]:

$$F(x) = \begin{cases} 0, & x \le 0; \\ \frac{1}{\Gamma(k/2)}\gamma(x, k/2), & x > 0, \end{cases}$$

где $k = 1, 2, \ldots$ — число степеней свободы.

Распределение χ^2 является частным случаем гамма-распределения при $\delta = k/2$ с параметром масштаба 2 (см. п. 3.4.3.20).

Распределение χ^2 является предельным для распределения статистики χ^2 Пирсона, статистики отношения правдоподобия и статистики Смирнова. На рис. 3.26 приведены результаты проверки согласия эмпирического распределения статистики Смирнова, полученного по 5000 выборок объемом 100 наблюдений, и распределения χ^2 с двумя степенями свободы.

3.4.4.2. Распределение Колмогорова

Случайная величина, распределенная по закону Колмогорова, определена на интервале $(0, +\infty)$.

Функция распределения [88]:

$$F(x) = \begin{cases} 0, & x \le 0; \\ \sum\limits_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2x^2}, & x > 0. \end{cases}$$

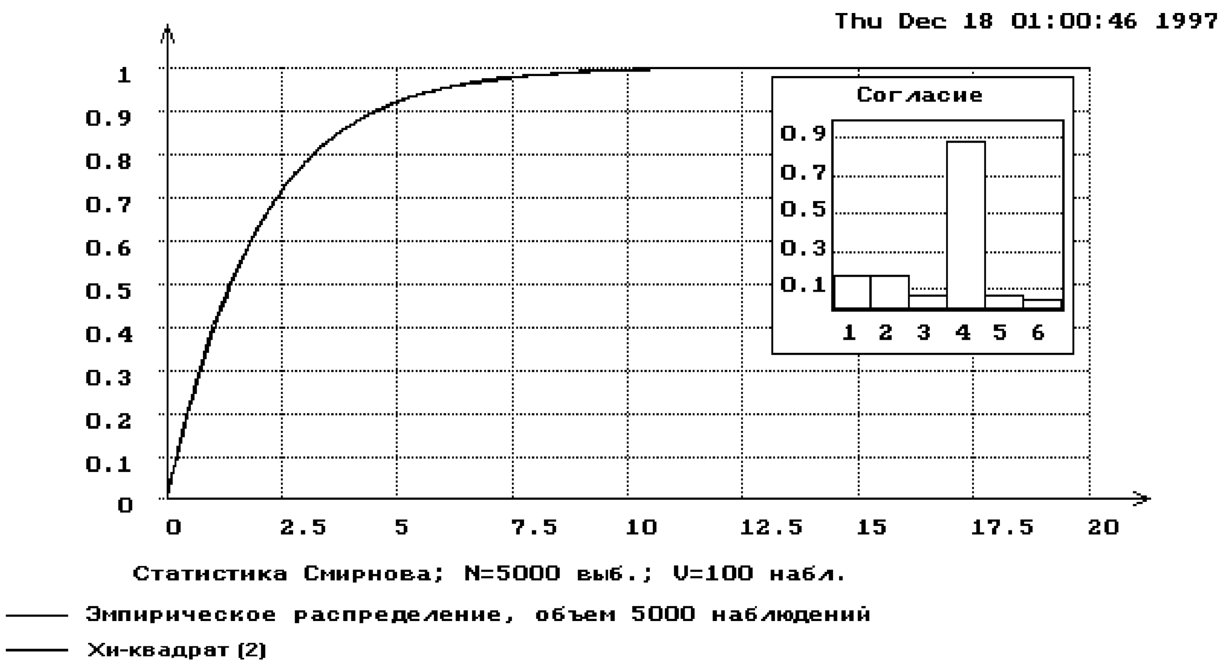


Рис. 3.26. Согласие эмпирического распределения статистики Смирнова с распределением χ^2 с двумя степенями свободы

Распределение Колмогорова является предельным для распределения статистики Колмогорова. На рис. 3.27 приведены результаты проверки согласия эмпирического распределения статистики Колмогорова, полученного по 5000 выборок объемом 100 наблюдений, и распределения Колмогорова.

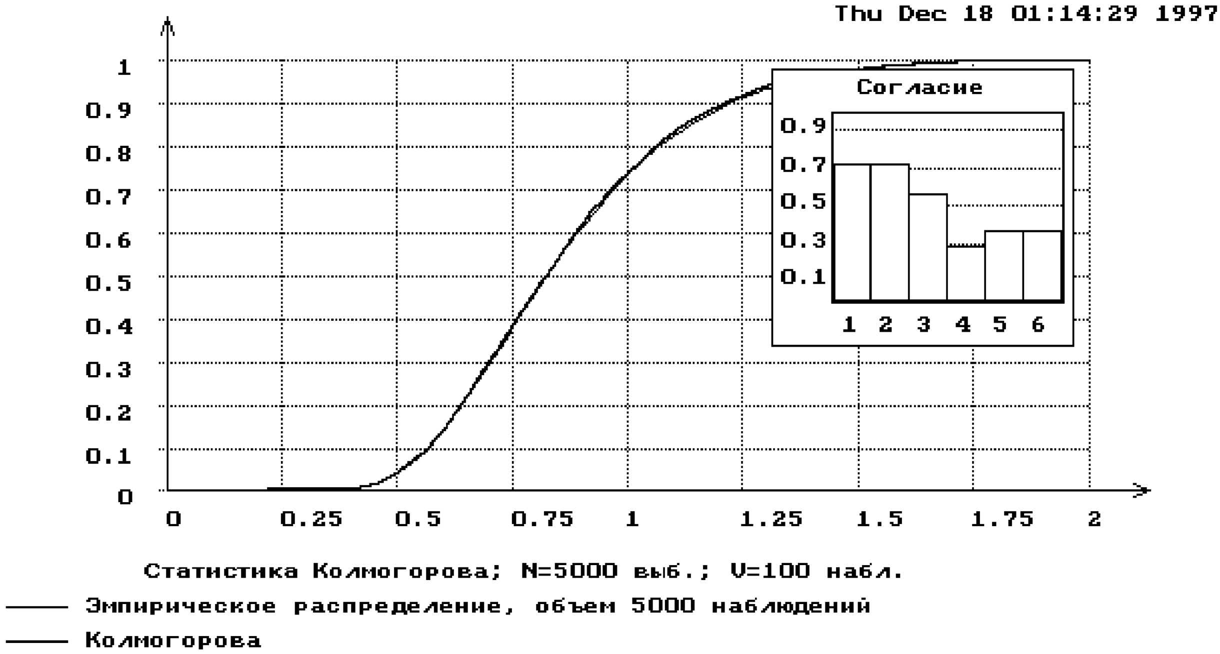


Рис. 3.27. Согласие эмпирического распределения статистики Колмогорова с распределением Колмогорова

3.4.4.3. Распределение статистики ω^2 Мизеса

Статистика ω^2 Мизеса распределена по закону $a1$ с функцией распределения [88]:

$$F(x) = \begin{cases} 0, & x \le 0; \\ a1(x), & x > 0. \end{cases}$$

где

$$\begin{aligned} a1(x) \;&=\; \frac{1}{\sqrt{2x}} \sum_{j=0}^{\infty} \frac{\Gamma(j+1/2)\sqrt{4j+1}}{\Gamma(1/2)\Gamma(j+1)} \mathrm{e}^{-\frac{(4j+1)^2}{16x}} \times \\ &\times \left\{ I_{-\frac{1}{4}}\left[\frac{(4j+1)^2}{16x}\right] - I_{\frac{1}{4}}\left[\frac{(4j+1)^2}{16x}\right] \right\}, \end{aligned}$$

и $I_{-\frac{1}{4}}(\cdot)$ и $I_{\frac{1}{4}}(\cdot)$ — модифицированные функции Бесселя.

Распределение $a1$ является предельным для распределения статистики ω^2 Мизеса. На рис. 3.28 приведены результаты проверки согласия эмпирического распределения статистики ω^2 Мизеса, полученного по 5000 выборок объемом 100 наблюдений, и распределения $a1$.

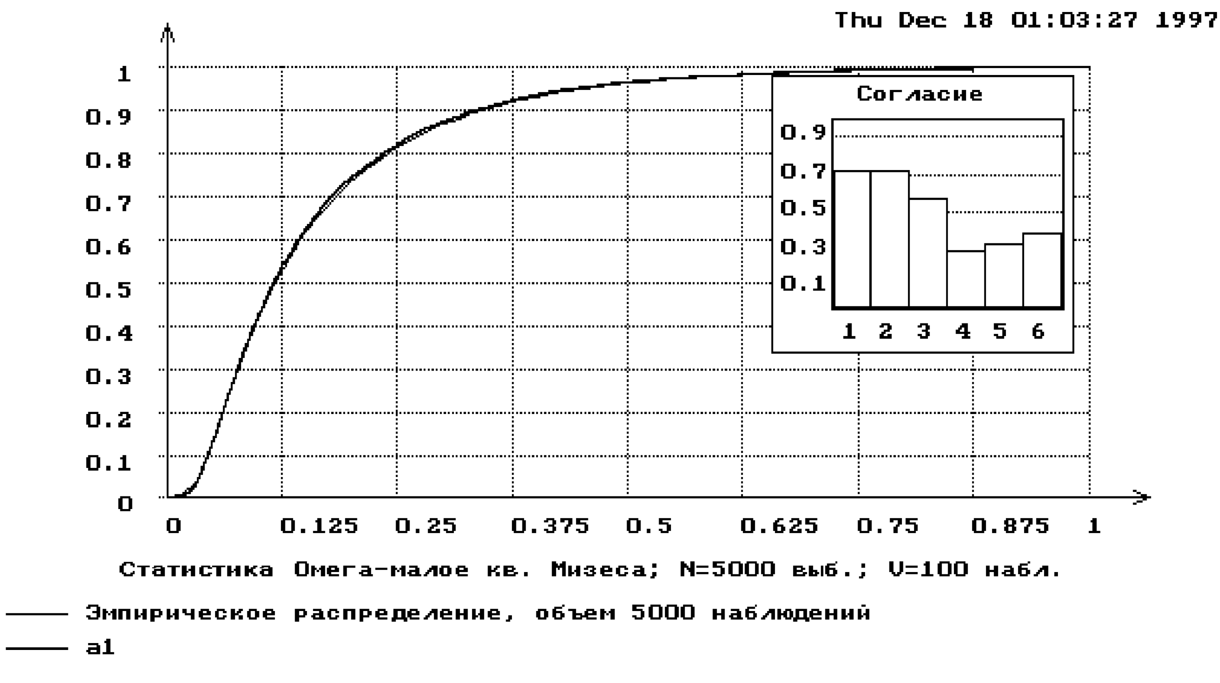


Рис. 3.28. Согласие эмпирического распределения статистики ω^2 Мизеса с распределением $a1$

3.4.4.4. Распределение статистики Ω^2 Мизеса

Статистика Ω^2 Мизеса распределена по закону $a2$ с функцией распределения [88]:

$$F(x) = \begin{cases} 0, & x \le 0; \\ a2(x), & x > 0. \end{cases}$$

где

$$\begin{aligned} a2(x) \; = \; & \frac{\sqrt{2\pi}}{x} \sum_{j=0}^{\infty} (-1)^j \frac{\Gamma(j+1/2)\sqrt{4j+1}}{\Gamma(1/2)\Gamma(j+1)} \mathrm{e}^{-\frac{(4j+1)^2\pi^2}{8x}} \times \\ & \times \int\limits_0^{\infty} \exp\left\{ \frac{x}{8(y^2+1)} - \frac{(4j+1)^2\pi^2 y^2}{8x} \right\} dy. \end{aligned}$$

Распределение $a2$ является предельным для распределения статистики Ω^2 Мизеса. На рис. 3.29 приведены результаты проверки согласия эмпирического распределения статистики Ω^2 Мизеса, полученного по 5000 выборок объемом 100 наблюдений, и распределения $a2$.

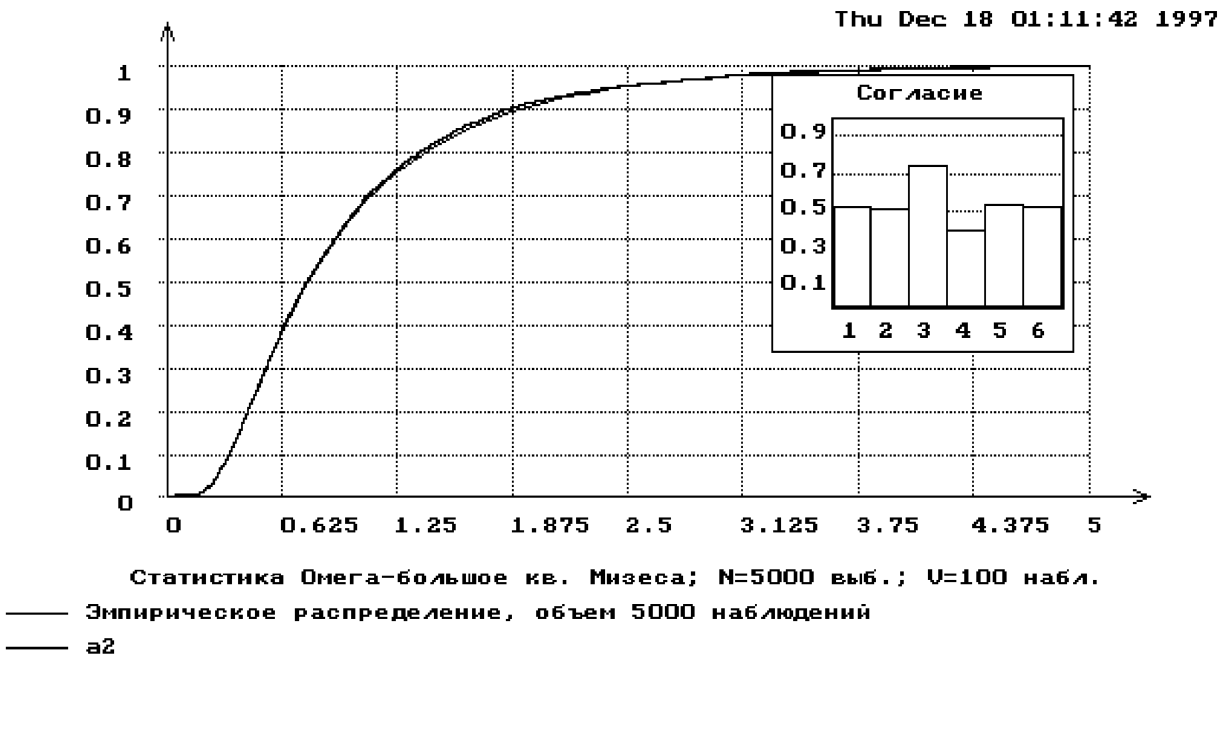


Рис. 3.29. Согласие эмпирического распределения статистики Ω^2 Мизеса с распределением $a2$

3.4.5. Эмпирическое распределение

Эмпирическая функция распределения $F_n(x)$, построенная по точечной выборке $\{x_1, x_2, \ldots, x_n\}$, является непараметрической оценкой функции распределения $F(x)$ наблюдаемой случайной величины. Эта оценка является несмещенной и состоятельной [89]. Состоятельность следует из теоремы Гливенко: $P\{\lim\limits_{n\to\infty} \sup\limits_{x} |F_n(x) - F(x)| = 0\} = 1$.

Непараметрической оценкой функции плотности распределения $f(x)$ является функция $f_n(x)$

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i),$$

где

$$\delta(x - x_i) = \begin{cases} 0, & x \neq x_i, \\ \infty & x = x_i \end{cases}$$

и

$$\int\limits_{-\infty}^{\infty} \delta(x - x_i)dx = 1;$$

которая получается дифференцированием эмпирической функции распределения. Такая оценка функции плотности "запрещает" появление наблюдений, отличных от $x_i, i = 1, 2, \ldots, n$, и поэтому является нереалистичной. Для получения более гладкой оценки плотности распределения можно заменить дельта-функцию на некоторую колоколообразную функцию $K(x, x_i)$, которая будет достигать максимума в точке x_i. Тогда

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i). \tag{3.5}$$

Оценка (3.5) является асимптотически несмещенной в каждой точке непрерывности $f_n(x)$, если выполняются условия (1.14). В этом случае функцию $K(x, x_i)$ называют "ядром". В частности в качестве ядра можно взять функцию плотности нормального распределения со сдвигом $\mu = x_i$, и параметром масштаба λ. Тогда параметр масштаба λ имеет смысл параметра "размытости": если $\lambda \to 0$, то мы получим дельта-функцию, а если $\lambda \to \infty$, то функция плотности $f_n(x) \to 0, \forall x$. Естественно, что эти два крайних случая не являются приемлемыми, поэтому возникает задача об оптимальном выборе параметра размытости. В [90] дается рекомендация о выборе $\lambda \approx n^{-1/4}$. На рис. 3.30, 3.31 и 3.32 приведены ядерные оценки функции плотности, построенные по выборкам из стандартного нормального распределения разного объема. На рисунках хорошо видно, что с ростом объема выборки ядерная оценка плотности приближается к плотности нормального распределения.

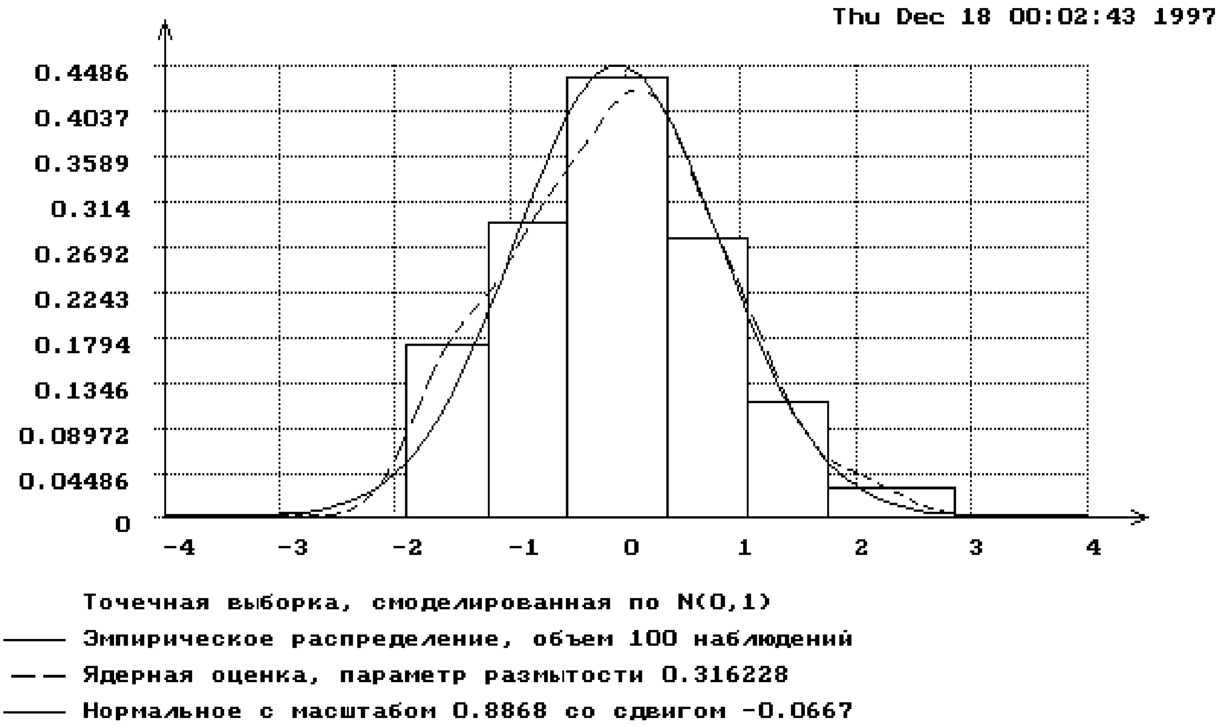


Рис. 3.30. Ядерная оценка функции плотности по точечной выборке

3.5. Статистический анализ выборки

Задачи статистического анализа решаются в классе “Stat”. Основными данными этого класса являются выборка и распределение. Над объектом типа “Stat” определены следующие методы: оценивание параметров распределения по выборке, проверка простых и сложных гипотез о согласии выборки с распределением, выделение аномальных наблюдений в выборке, группирование выборки одним из четырех методов.

3.5.1. Оценивание параметров распределений

Оценивание параметров распределений в системе производится по трем методам: метод максимального правдоподобия, метод минимума статисти-

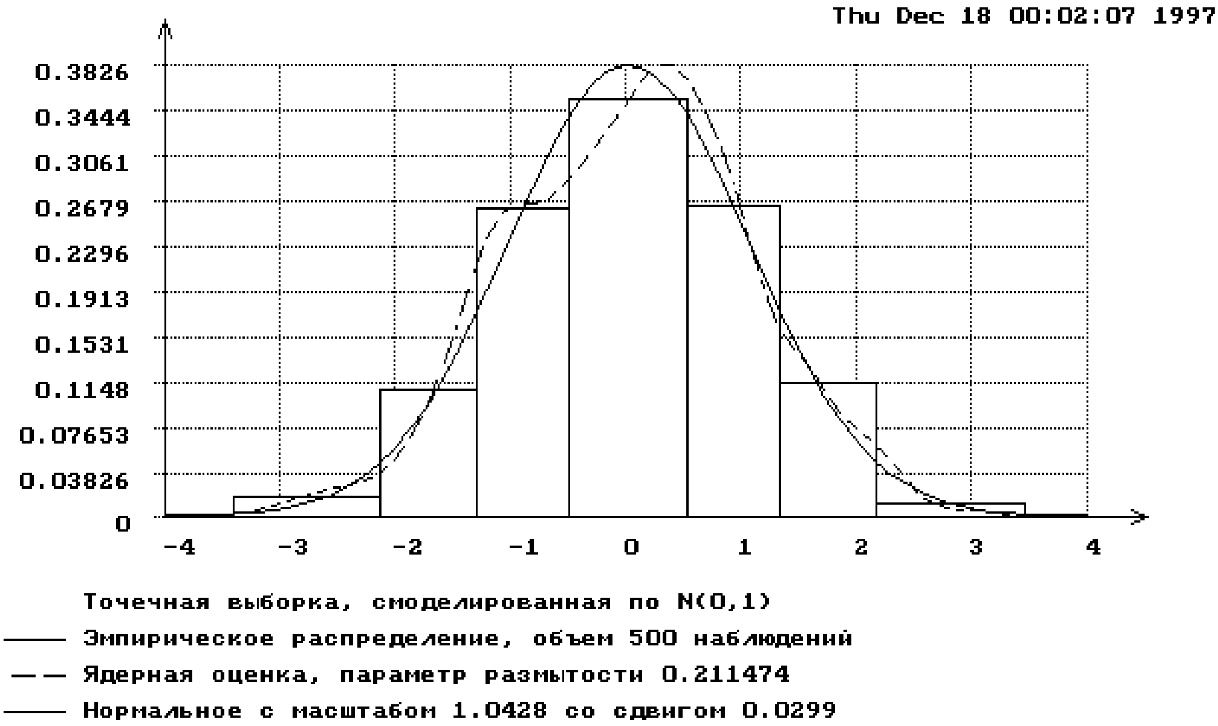


Рис. 3.31. Ядерная оценка функции плотности по точечной выборке объемом 500 наблюдений

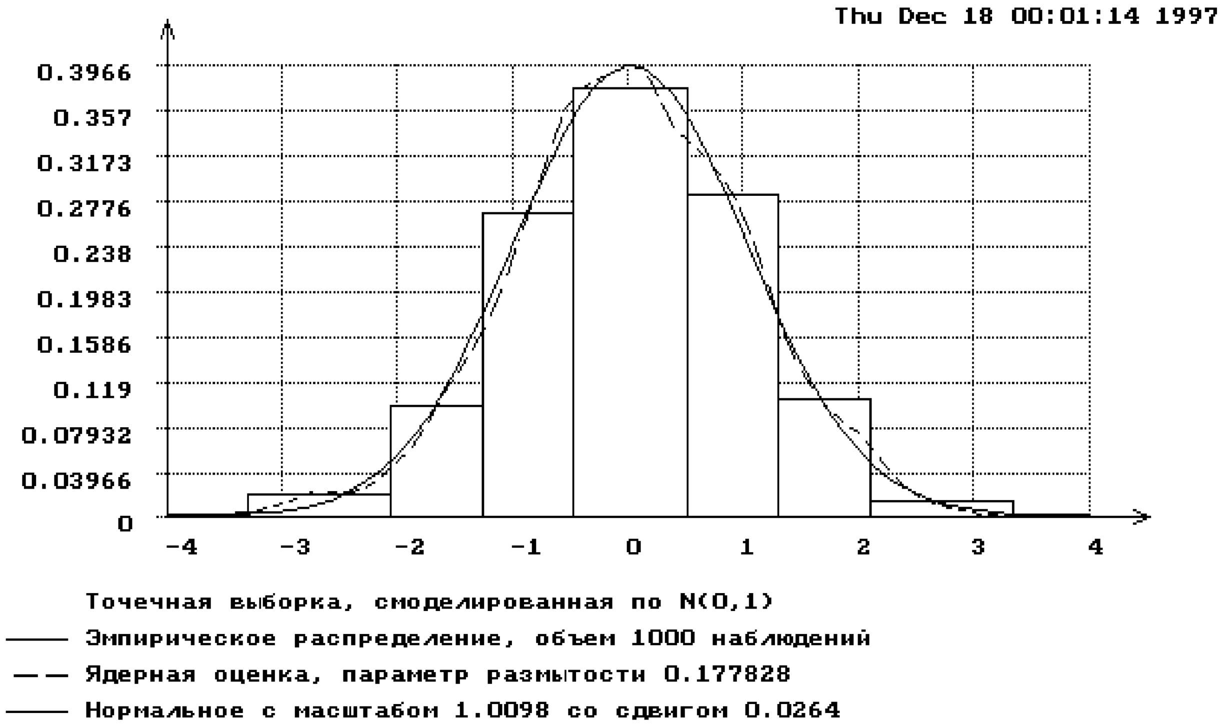


Рис. 3.32. Ядерная оценка функции плотности по точечной выборке объемом 1000 наблюдений

ки Колмогорова и метод минимума статистики ω^2 Мизеса. В случае интервальной выборки допускается вычисление как точечных, так и интервальных оценок. При вычислении точечных оценок по интервальной выборке используются различные стратегии: оценка по серединам интервалов, оценка “крайнего оптимиста”, оценка “крайнего пессимиста”, оценка максимального правдоподобия.

3.5.2. Проверка гипотез о согласии

Проверка гипотез о согласии выборки с теоретическим распределением производится по шести критериям (см. таблицу 3.10). Первые два критерия предусматривают группировку данных, остальные вычисляются по негруппированным данным. Теоретические распределения известны для всех статистик при проверке простой гипотезы (т.е. при проверке согласия с фиксированным распределением) и приведены в п. 3.4.4. Если проверяется сложная гипотеза, т.е. гипотеза о согласии выборки с параметрическим семейством распределений, то известны предельные распределения для статистик критериев отношения правдоподобия и χ^2 Пирсона. Сложная гипотеза проверяется следующим образом: сначала по выборке оцениваются параметры, а затем проверяется согласие. Параметрические критерии (отношения правдоподобия и χ^2 Пирсона) учитывают факт оценивания уменьшением числа степеней свободы предельного χ^2 распределения на число оцениваемых параметров. Исследование зависимости распределения статистик χ^2 и отношения правдоподобия от способа группирования, при оценивании параметров по методу максимального правдоподобия проведено нами в [81, 83].

Непараметрические критерии (Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса) при оценивании теряют свойство "свободы от распределения", т.е. предельные распределения этих статистик будут зависеть от распределения, которому подчинена выборка.

В литературе известны следующие подходы к использованию критериев согласия в этом случае.

1. Если выборка достаточно большая, то ее можно разбить на две равные части: по одной оценивать параметры, а по другой проверять согласие

[91]. Если объем выборки невелик, то оценки параметров будут зависеть от способа разбиения выборки на две части, и, следовательно, результаты проверки согласия также будут неоднозначны.

2. Для случая нормального распределения предельные распределения статистики критерия ω^2 Мизеса при оценивании одного или обоих параметров по методу максимального правдоподобия получены аналитически и табулированы в [92].

3. Методом статистического моделирования получены таблицы процентных точек для предельных распределений непараметрических статистик при оценивании параметров экспоненциального, нормального, экстремальных значений, Вейбулла и некоторых других законов распределений [93, 94, 95, 96].

4. В работах [97, 98, 99, 100] для статистик типа Колмогорова-Смирнова получены формулы для *приближенного* вычисления вероятностей согласия. С помощью этих формул вычисляется вероятность согласия в пакете STADIA [63].

5. Нами в результате моделирования предельных законов распределения статистик непараметрических критериев найдены такие законы распределения вероятностей, которые хорошо аппроксимируют предельные распределения статистик непараметрических критериев согласия в тех случаях, когда по выборке оцениваются параметры по методу максимального правдоподобия [80, 83].

Тем не менее, полученные за 40 лет исследований таблицы процентных точек и предельные распределения пригодны лишь в относительно небольшом числе случаев. В самом деле, распределения статистик (или их процентные точки) получены для 10-15 законов, в то время как число законов в разрабатываемой системе за счет применения операций смеси, усечения и т.д. существенно больше. Далее, во всех работах при нахождении распределений статистик были зафиксированы методы оценивания, но *способ получения оценок также сильно влияет на распределение статистики критерия согласия*.

Таблица 3.10

Статистики критериев согласия

ь	Критерий	Статистика
1	Отношения правдоподобия	$-2\sum\limits_{i=1}^{k} n_i \ln\left(\frac{P_i(\theta)}{n_i/n}\right),$
2	χ^2 Пирсона	$n\sum\limits_{i=1}^{k} \frac{(n_i/n-P_i(\theta))^2}{P_i(\theta)}$
3	Колмогорова	$(6n\sup\limits_{x}\lvert F_n(x)-F(x)\rvert+1)/6\sqrt{n}$
4	Смирнова	$(6n\sup\limits_{x}(F_n(x)-F(x))+1)^2)/9n$
5	ω^2 Мизеса	$\frac{1}{12n}+\sum\limits_{i=1}^{n}\left[F(x_{(i)})-\frac{2i-1}{2n}\right]^2$
6	Ω^2 Мизеса	$-n-2\sum\limits_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_{(i)})+\left(1-\frac{2i-1}{2n}\right)\ln(1-F(x_{(i)}))\right\}$

Вообще, распределение (не предельное!) статистики критерия согласия зависит от множества факторов:

- от объема выборки (см. рис. 3.33);
- от распределения (см. рис. 3.34);
- от числа оцениваемых параметров и от того, какие параметры оцениваются (см. рис. 3.35);
- от метода оценивания (см. рис. 3.36);
- от способа группирования, если критерий предусматривает группировку данных (см. рис. 3.37).

Распределения на рис. 3.33–3.37 были получены методом статистического моделирования и позволяют сделать следующие выводы:

- Объем выборки существенно влияет на распределение статистки Колмогорова при $n < 10$.
- Хотя критерий согласия Колмогорова теряет свойство "свободы от распределения"при оценивании, тем не менее для целого класса законов

(Лапласа, логистического, Коши, нормального) распределения статистики Колмогорова совпали, *когда оценки находились с помощью минимизации статистики Колмогорова*. При оценивании по методу максимального правдоподобия распределения статистики Колмогорова для этих же законов заметно отличаются.

Таким образом, когда закон распределения статистики критерия неизвестен, то для вычисления вероятности согласия необходимо применять метод статистического моделирования с учетом всех факторов, влияющих на распределение статистики.

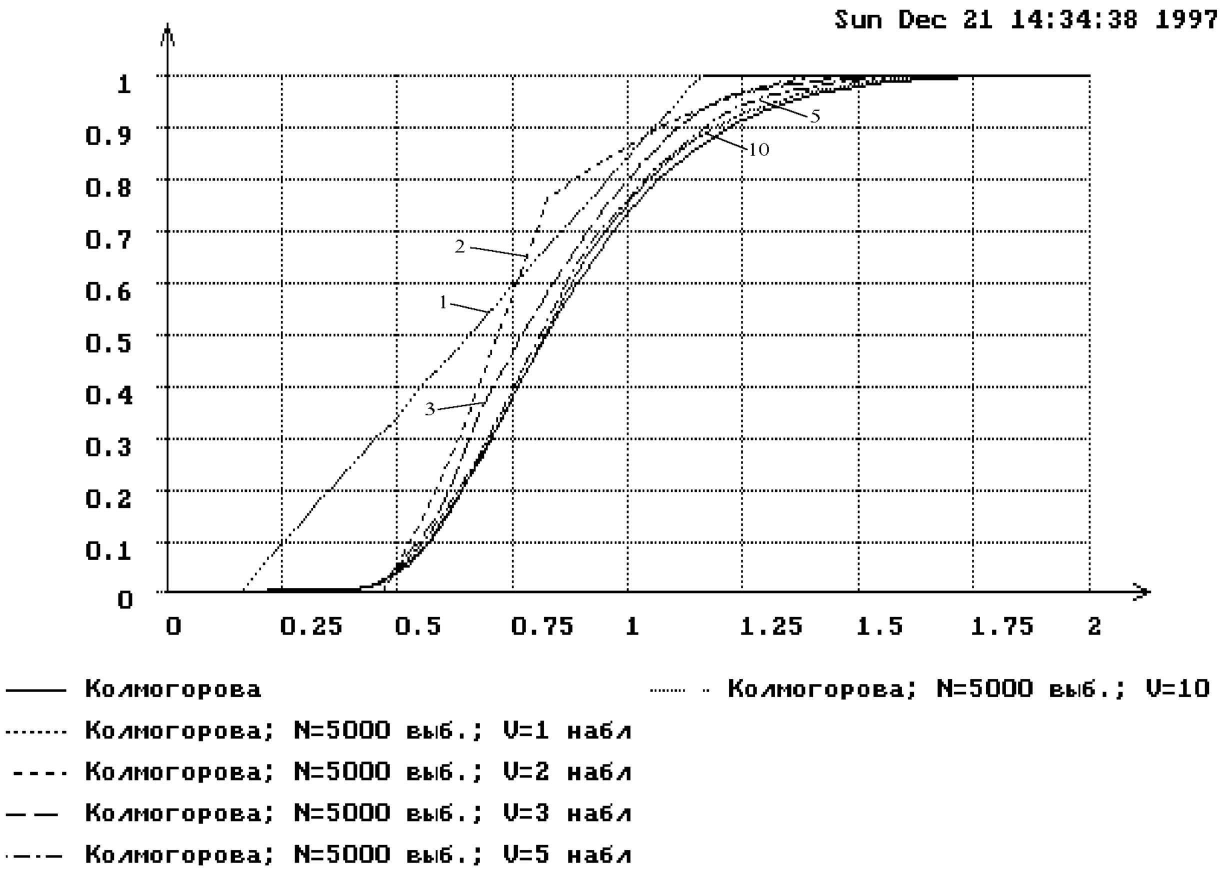


Рис. 3.33. Распределение статистики Колмогорова при различных объемах выборки.

В системе задача вычисления вероятности согласия выборки с заданным распределением решается следующим образом. Если гипотеза о согласии простая, то используются предельные теоретические распределения статистик критериев. Если гипотеза сложная, то вероятности согласия нахо-

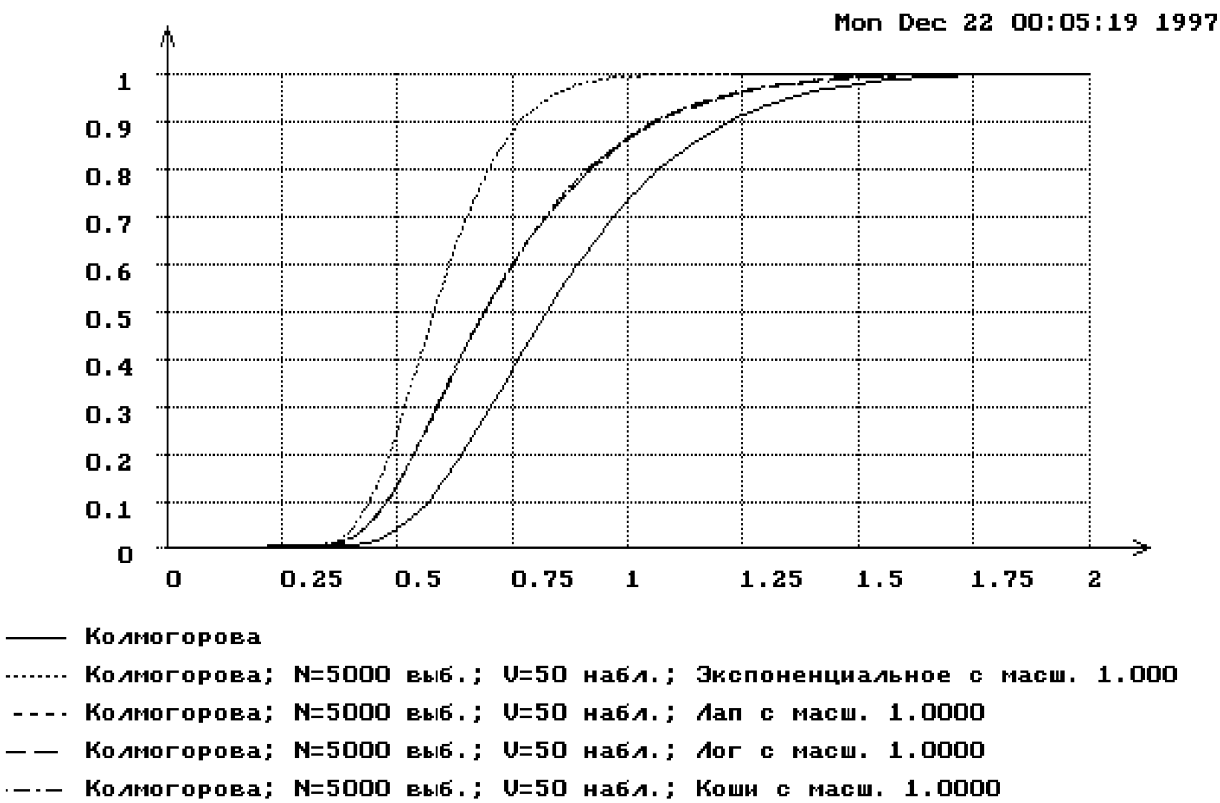


Рис. 3.34. Распределение статистики Колмогорова при оценивании по методу минимума статистики Колмогорова параметров масштаба экспоненциального распределения, распределения Лапласа, логистического распределения и распределения Коши

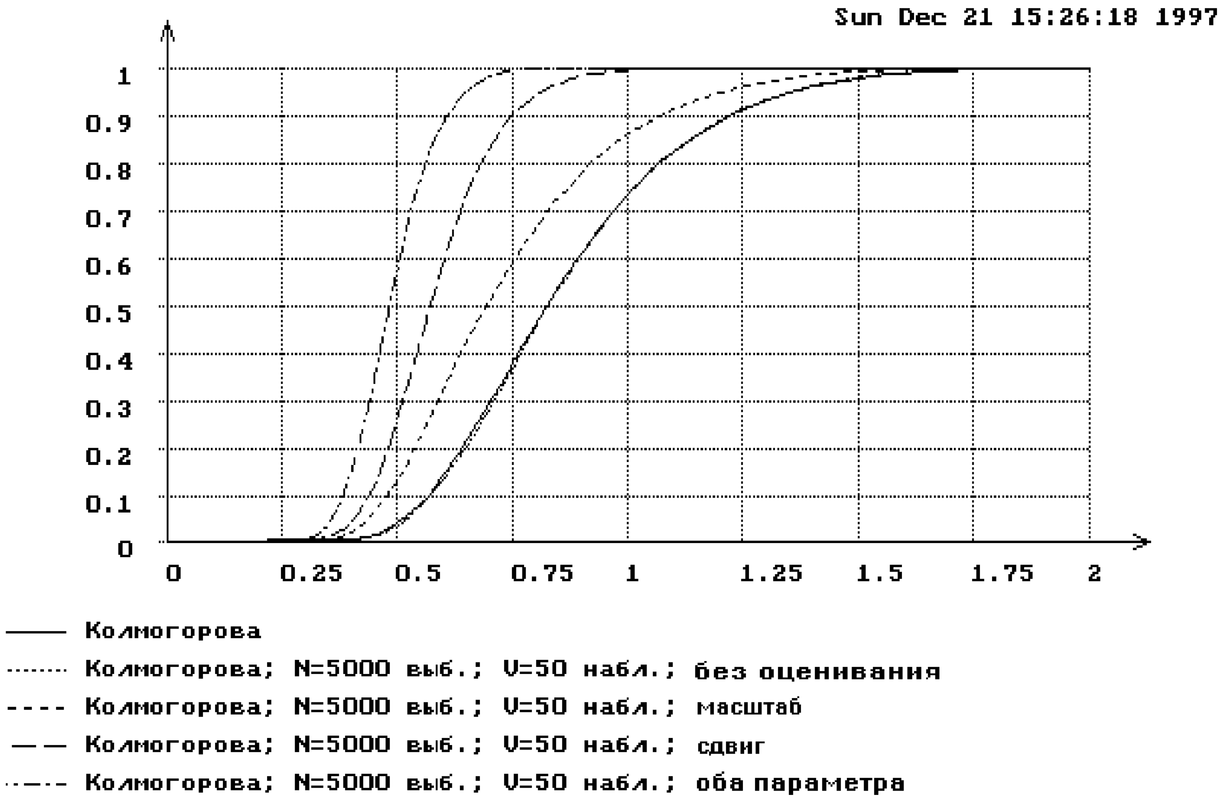


Рис. 3.35. Распределение статистики Колмогорова при оценивании по методу минимума статистики Колмогорова параметров нормального распределения

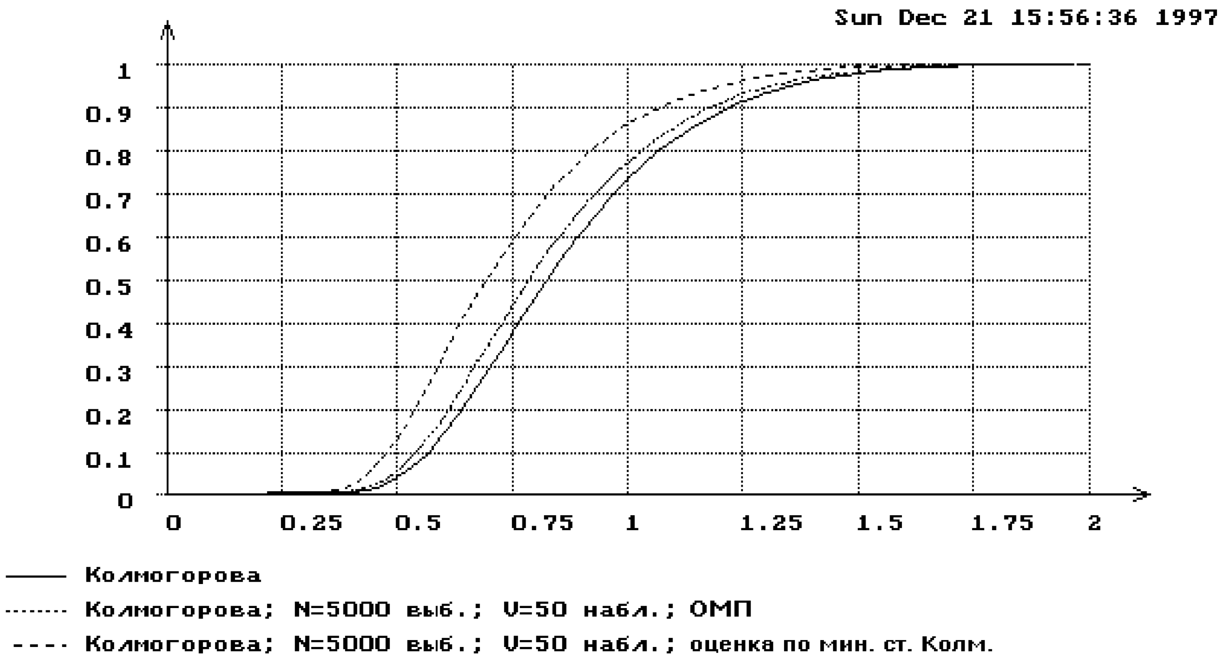


Рис. 3.36. Распределение статистики Колмогорова при оценивании параметра масштаба нормального распределения по методу максимального правдоподобия и по методу минимума статистики Колмогорова

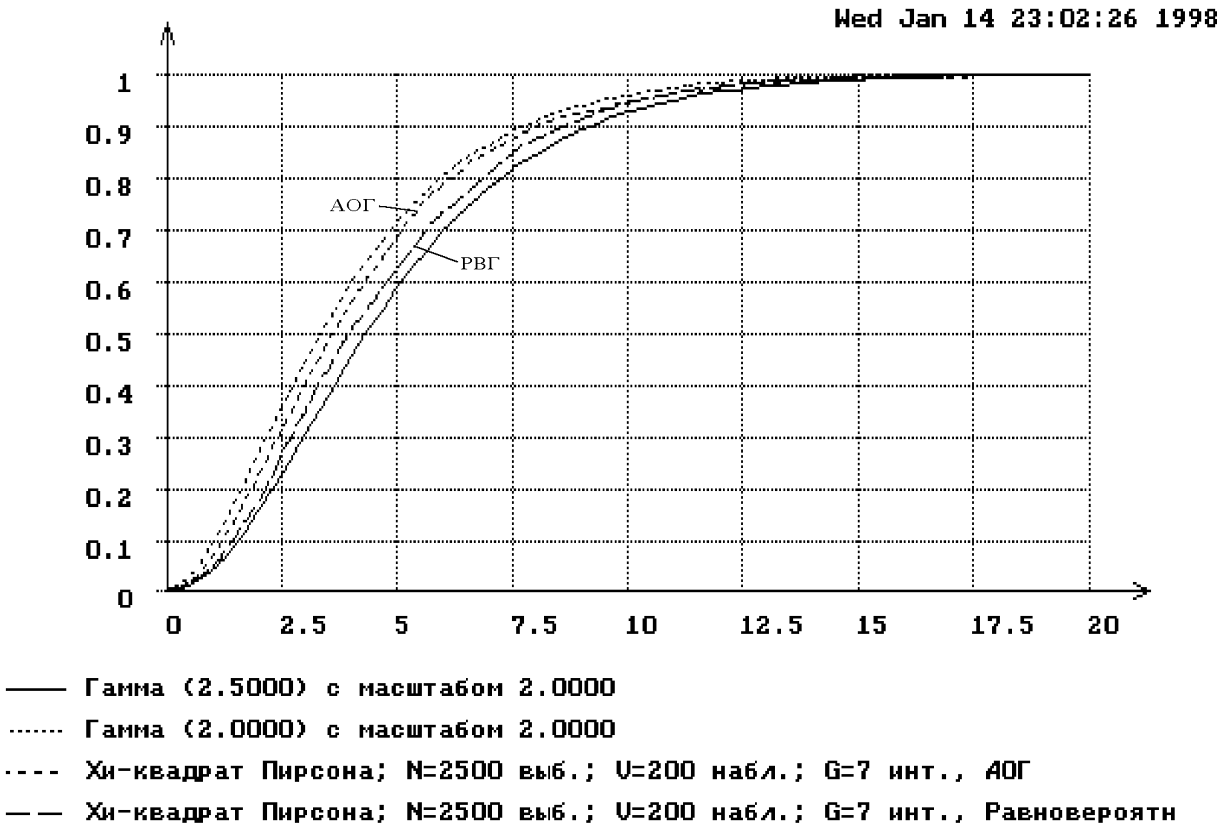


Рис. 3.37. Распределение статистики χ^2 Пирсона при оценивании параметра масштаба нормального распределения по методу максимального правдоподобия, при асимптотически оптимальном (АОГ) и равновероятном (РВГ) группировании

дятся с помощью метода Монте-Карло [85]. Алгоритм вычисления вероятности согласия по методу Монте-Карло приведен на рис. 3.38. Точность вычисления зависит от числа смоделированных выборок N. Чем больше N, тем выше точность вычисления вероясности согласия, и тем дольше производятся вычисления. Поэтому практически достаточно выбрать такое N, чтобы ошибка моделирования распределения была существенно меньше ошибки при использовании теоретических законов, не учитывающих оценивание параметров. Рассмотрим следующий пример.

Пример 3.10. *Была смоделирована точечная выборка по стандартному нормальному закону объемом 100 наблюдений. Далее, была проверена гипотеза о согласии полученной выборки с нормальным распределением с параметром сдвига $\mu = 0$. Так как параметры не оценивались, то допустимо использование предельных теоретических законов распределения. Результаты проверки согласия приведены на рис. 3.39.*

Потом по выборке был оценен параметр сдвига по методу максимального правдоподобия и проверено согласие с нормальным распределением с параметром $\mu = 0.0868$ (см. рис. 3.40). Вероятности согласия при этом были получены по методу Монте-Карло с $N = 5000$. В таблице 3.11 приведены ошибки вычисления вероятности согласия без учета факта оценивания параметра и при меньших значениях N (в процентах от максимально возможных).

Из рассмотренного примера хорошо видно, что уже при небольшом количестве моделируемых выборок ($N = 10$) ошибка получается меньше, чем при использовании предельных теоретических законов, не учитывающих факт оценивания параметров. Но в то же время нельзя однозначно сказать какого количества выборок достаточно для получения приемлемой точности вероятности согласия, так как для разных критериев требуется разный объем моделирования.

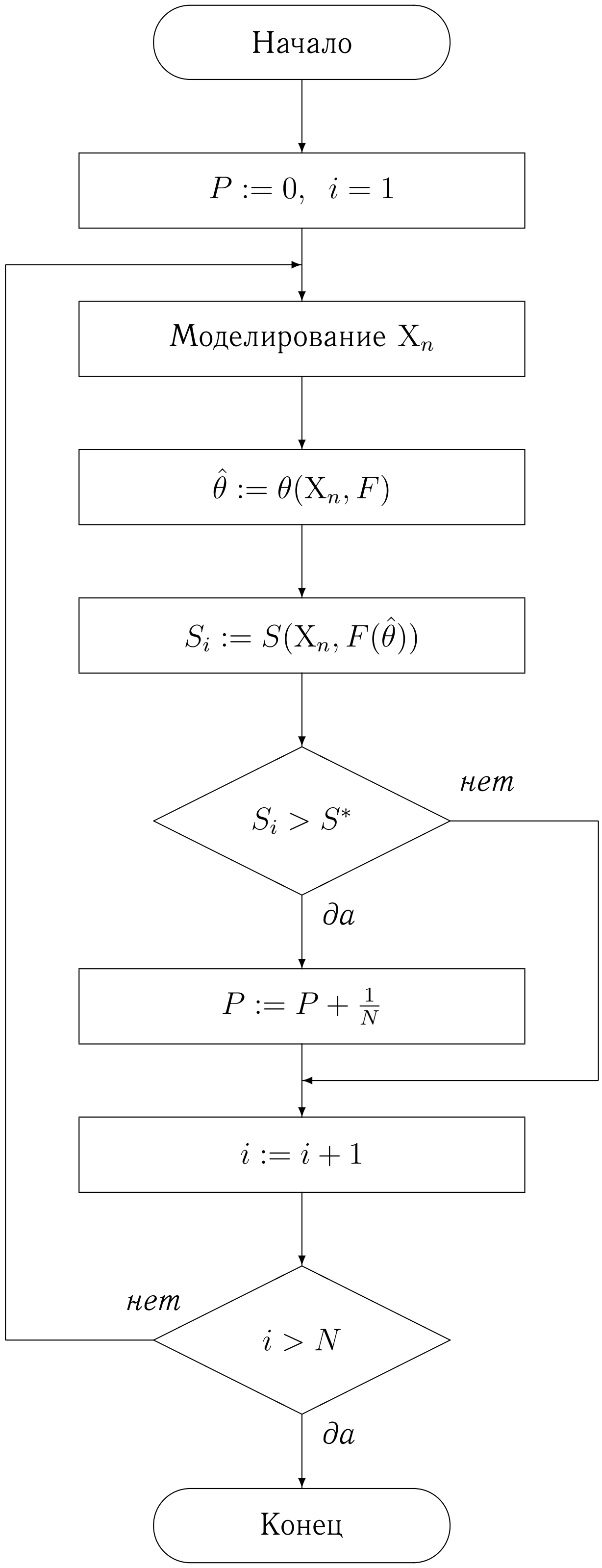


Рис. 3.38. Вычисление вероятности согласия

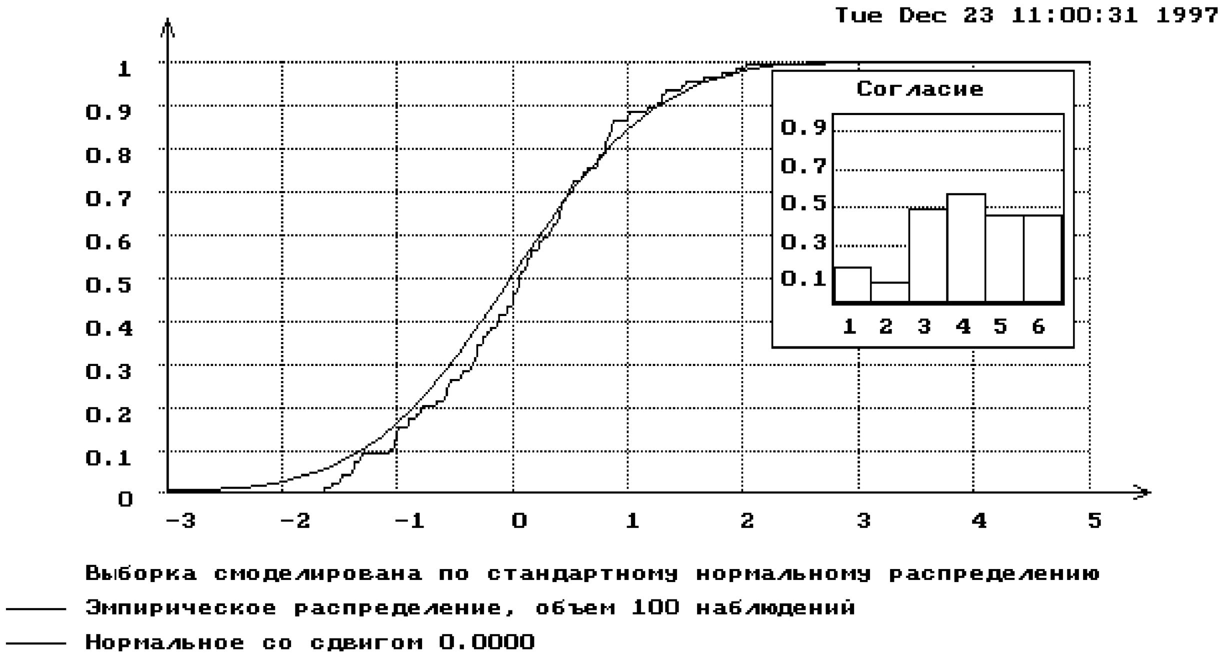


Рис. 3.39. Проверка согласия выборки, смоделированной по стандартному нормальному закону, с нормальным распределением с параметром $\mu = 0$

Таблица 3.11

Сравнение ошибок в вычислении вероятности согласия, допускаемых при использовании предельных законов, не учитывающих факта оценивания параметров, и ошибок, допускаемых из-за недостаточного объема моделирования распределений статистик по методу Монте-Карло

Критерий	без оцен.	$N = 10$	$N = 50$	$N = 100$	$N = 1000$
Отношения правдоподобия	19.1%	20.2%	12.8%	3.8%	1.3%
χ^2 Пирсона	14.1%	11.6%	7.9%	3.7%	0.6%
Колмогорова	45.4%	21.2%	10.6%	8.0%	2.0%
Смирнова	19.3%	7.8%	1.4%	1.4%	0.9%
ω^2 Мизеса	56.8%	3.3%	3.3%	8.4%	1.7%
Ω^2 Мизеса	54.0%	24.5%	12.1%	7.4%	3.4%

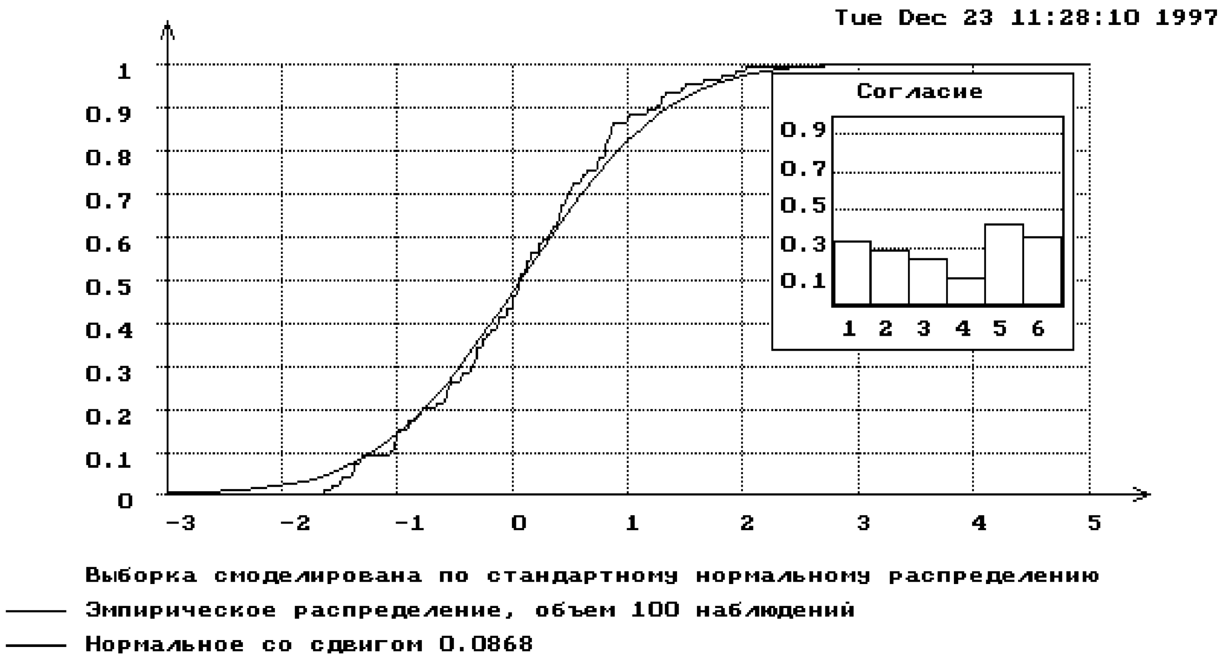


Рис. 3.40. Проверка согласия выборки, смоделированной по стандартному нормальному закону, с нормальным распределением с параметром сдвига $\mu = 0.0868$ полученным по методу максимального правдоподобия

3.5.3. Выделение аномальных наблюдений

Процедура выделения аномальных наблюдений из выборки состоит из двух этапов [101, 102]. Сначала одним из робастных (т.е. устойчивых к аномальным наблюдениям) методов идентифицируют выборочное распределение $F(x)$. Затем используется параметрическая процедура отбраковки наблюдений распределенных по какому-то другому закону $G(x)$.

Основная гипотеза H_0 процедуры отбраковки аномальных наблюдений состоит в том, что наблюдения рассматриваются как реализации независимых одинаково распределенных случайных величин $X_1, X_2, \dots, X_n$ с функцией распределения $F(x)$. При альтернативной гипотезе H_1 случайные величины $X_1, X_2, \dots, X_n$ также независимы, но часть из них имеет распределение $G(x)$, которое существенно сдвинуто относительно $F(x)$, например, $G(x) = F(x - A)$, где $|A|$ существенно велико. Если основная гипотеза верна, то

$$P\{\max_{1 \le i \le n} X_i \le \overline{d}\} = [F(\overline{d})]^n = 1 - \alpha$$

и

$$P\{\min_{1 \le i \le n} X_i \le \underline{d}\} = [1 - F(\underline{d})]^n = 1 - \alpha,$$

где α — вероятность ошибки первого рода.

Основная гипотеза отвергается, если максимальное выборочное наблюдение больше верхнего критического значения $\overline{d} = F^{-1}(\sqrt[n]{1-\alpha})$ или минимальное меньше нижнего критического значения $\underline{d} = F^{-1}(1 - \sqrt[n]{1-\alpha})$. Если основная гипотеза отвергнута, то все наблюдения, которые меньше $\underline{d}$ или больше $\overline{d}$, помечаются как аномальные.

Надежность процедуры отбраковки зависит от "правильной" идентификации закона $F(x)$. Рассмотрим параметрический случай, когда распределение одномерной случайной величины известно с точностью до параметров. В этом случае задача "правильной" идентификации закона $F(x)$ в основном определяется задачей *робастного оценивания* [1].

Для анализа робастности того или иного метода оценивания часто используются функции влияния, предложенные Хампелем [2]:

$$IF(x; F, T) = \lim_{s \to 0} \frac{T((1-s)F + s\delta_x) - T(F)}{s}, \tag{3.6}$$

где δ_x — единичная масса в точке x; F — функция распределения, которому принадлежит выборка; $T(F)$ — вычисляемая статистика.

Функция влияния позволяет оценить относительное влияние отдельного наблюдения на значение статистики критерия или оценки параметров. Если функция влияния неограничена, то резко выделяющиеся наблюдения могут приводить к существенным изменениям оценок или статистик. Чувствительность к большой ошибке может характеризоваться величиной

$$\gamma^* = \sup_x |IF(x; F, T)|. \tag{3.7}$$

В [60] было показано, что

- за редким исключением оценки максимального правдоподобия по негруппированным данным являются неробастными, что следует из неограниченности функция влияния;
- оценки максимального правдоподобия по группированным данным являются робастными, так как соответствующие им функции влияния являются ограниченными.

Вообще, оценки, построенные по методу, минимизирующему расстояние между эмпирической и теоретической функцией распределения (MD-оценки), являются робастными [3], и, в частности, хорошими свойствами робастности обладают оценки, получаемые в результате минимизации статистики Колмогорова. Подставим статистику Колмогорова

$$T(F) = \sup_y |F_n(y) - F(y)|$$

в (3.6) и найдем верхнюю и нижнюю границу функции влияния. Получим

$$\begin{aligned}
IF(x) &= \lim_{s\to 0} \frac{\sup\limits_y |F_n(y) - (1-s)F(y) - s\delta_x(y)| - \sup\limits_y |F_n(y) - F(y)|}{s} \le \\
&\le \lim_{s\to 0} \frac{\sup\limits_y |F_n(y) - F(y)| + s\sup\limits_y |F(y) - \delta_x(y)| - \sup\limits_y (|F_n(y) - F(y)|}{s} = \\
&= \sup_y |F(y) - \delta_x(y)| = \max\{F(x), 1 - F(x)\} \le 1,
\end{aligned}$$

и

$$IF(x) = \lim_{s\to 0} \frac{\sup\limits_y |F_n(y) - (1-s)F(y) - s\delta_x(y)| - \sup\limits_y |F_n(y) - F(y)|}{s} \ge$$

$$
\geq \lim_{s\to 0} \frac{\sup_y |F_n(y) - F(y)| - s\sup_y |F(y) - \delta_x(y)| - \sup_y(|F_n(y) - F(y)|}{s} =
$$
$$
= -\sup_y |F(y) - \delta_x(y)| = -\max\{F(x), 1 - F(x)\} \geq -1.
$$

Таким образом, для статистики Колмогорова $\gamma^* = 1$, что свидетельствует о малой чувствительности к аномальным наблюдениям оценок по методу минимума статистики Колмогорова.

3.5.4. Группирование

В разрабатываемой системе можно выполнять группирование четырех видов: равномерное (на интервалы равной длины), равновероятное (на интервалы равной вероятности), равночастотное (на интервалы равной частоты попадания наблюдений) и асимптотически оптимальное (интервалы, максимимзирующие информацию по Фишеру относительно оцениваемых параметров).

Для равномерного группирования необходимо задать левую и правую границу и количество интервалов. По умолчанию в качестве границ используются минимальная и максимальная порядковая статистики.

При равночастотном группировании разбиение происходит таким образом, чтобы в каждом интервале было одинаковое количество наблюдений. Это возможно, если количество точечных наблюдений делится нацело на число интервалов.

При равновероятном группировании граничные точки вычисляются следующим образом:

$$
X_i = F^{-1}(i/k), i = 1, \ldots, k-1,
$$

где k — число интервалов группирования, $F^{-1}(y)$ — функция, обратная к функции распределения наблюдаемой случайной величины.

При асимптотически оптимальном группировании граничные точки вычисляются в результате максимизации функционала от информационной матрицы Фишера по группированным данным

$$
\mathcal{F}(I) \to \max_{X_1,\ldots,X_{k-1}}
$$

где $\mathcal{F}$ — функционал от матрицы, I — информационная матрица Фишера, элементы которой определяются следующим образом:

$$I^{\cdot}_{ml} = \sum_{i=1}^{k} \frac{\partial \ln P_i}{\partial \theta_m} \frac{\partial \ln P_i}{\partial \theta_l} P_i,$$

где $P_i = F(X_i) - F(X_{i-1})$ — вероятность попадания в i-й интервал, θ_l и θ_m — параметры распределения.

В пределе при увеличении числа интервалов группирования информационное количество Фишера по группированным данным стремится к информационному количеству Фишера по негруппированным данным:

$$I^{\cdot}_{ml} \to I^{\cdot}_{ml} = \int_{-\infty}^{+\infty} \frac{\partial \ln f}{\partial \theta_l} \frac{\partial \ln f}{\partial \theta_m} f(x) dx.$$

Таким образом, асимптотически оптимальное группирование минимизирует потери информации по Фишеру о параметрах, вызванные группированием данных.

В [50] для ряда распределений получены таблицы асимптотически оптимального группирования, инвариантные относительно параметров. Однако в общем случае (гамма-распределение, семейство бета-распределений, смеси распределений и т.д.) такие таблицы получить невозможно, поэтому в разрабатываемой системе при отстуствии таблиц АОГ задача нахождения граничных точек асимптотически оптимального группирования решается непосредственно перед группированием.

3.6. Идентификация

Пусть задано множество выборок $\mathbf{S} = \{S_1, S_2, \ldots, S_m\}$ и множество распределений $\mathbf{D} = \{D_1, D_2, \ldots, D_n\}$.

Задача статистической идентификации заключается в установлении такого соответствия между множеством выборок и множеством распределений, при котором каждой выборке из $\mathbf{S}$ ставится в соответствие "наилучшее" распределение из $\mathbf{D}$.

Выбор "наилучшего" распределения зависит от критерия, по которому производится идентификация. Так, например, в соответствии с принципом максимального правдоподобия в качестве наилучшего может быть выбрано распределение, которое максимизирует функцию правдоподобия:

$$L(S_i, D_j) \to \max_{D_j}, i = 1, 2, \ldots, m,$$

где $L(S_i, D_j)$ — функция правдоподобия. Однако, недостатком такого критерия идентификации является то, что нельзя сразу сказать насколько хорошо выбранное в результате идентификации распределение соответствует выборке. На самом деле может оказаться, что выборка распределена по закону отличному от любого распределения из $\mathbf{D}$. Процедура идентификации должна "улавливать" такие ситуации.

Рассмотрим использование критериев согласия в качестве критерия идентификации. Чем меньше значение статистики критерия, тем выше вероятность согласия. Если все параметры у распределений фиксированы, то процедура идентификации состоит в том, чтобы выбрать то распределение, для которого вероятность согласия максимальна. Если некоторые параметры неизвестны, то идентификация состоит из двух этапов:

- оценивание параметров распределений;
- проверка сложных гипотез о согласии.

Если гипотезы о согласии отвергаются для всех распределений, то делается вывод о том, что распределение выборки не идентифицировано. Если не противоречат гипотезе о согласии сразу несколько распределений, то из них выбирается то, у которого вероятность согласия максимальна.

Неоднозначность процедуры идентификации возникает тогда, когда имеется несколько методов оценивания параметров и несколько критериев согласия. В этом случае может получиться так, что по разным критериям "наилучшими" окажутся различные распределения (или одни и те же распределения, но с разными оценками параметров). Проиллюстрируем это на примере.

Пример 3.11. *По выборке, смоделированной в примере 3.10 были построены кривые согласия по критериям отношения правдоподобия, χ^2 Пирсона, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса, когда в качестве основной гипотезы выступали различные значения параметра сдвига от -0.2 до 0.3. На рис. 3.41 хорошо видно, что максимумы кривых согласия достигаются при разных значениях параметра. Причина такого "разногласия" объясняется различием мер, используемых в критериях.*

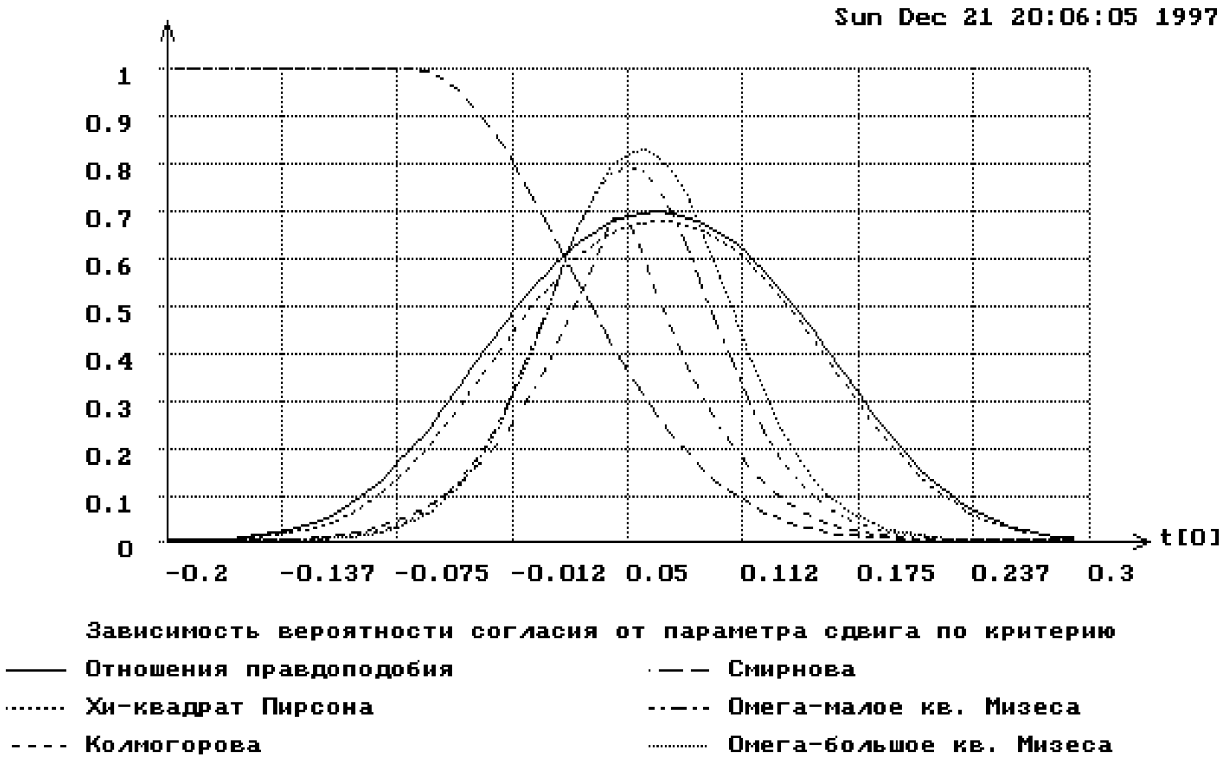


Рис. 3.41. Вероятности согласия по критериям отношения правдоподобия, χ^2 Пирсона, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса при различных значениях параметра сдвига

В [51] предложено принимать решение на основании компромиссного критерия вида

$$\max_{D_j} \sum_{r=1}^{k} \omega_r \alpha_{rj},$$

где α_{rj} — вероятность согласия по r-му критерию, ω_r — весовой коэффициент r-го критерия, k — число критериев согласия.

Очевидно, что вероятность согласия α_{rj} будет максимальна, если оценки параметров выбирать из условия минимума статистики r-го критерия. Поэтому желательно, чтобы оценивание параметров проводилось по методу минимума статистики того критерия, по которому затем проверяется согласие.

3.7. Статистический анализ геодезических наблюдений

Качественная техническая эксплуатация оборудования современных предприятий атомной энергетики была бы невозможна без правильной постановки технической диагностики, позволяющей установить причины и признаки повреждения конструкций. Одним из важнейших параметров технической диагностики являются осадки, перемещения, деформации, отклонения от проектного положения конструкций или всего объекта в целом, которые называют эксплуатационными отклонениями геометрических параметров.

Контроль многих эксплуатационных отклонений геометрических параметров осуществляется геодезическими методами и средствами измерений. Разности осадок, прогибов конструкции, отклонений по вертикали и т.п. позволяют дать оценку технического состояния объекта.

Применение статистического анализа результатов геодезических наблюдений является одним из путей совершенствования методов сбора информации о недостатках конструктивных элементов зданий в процессе ее эксплуатации, состояния геометрической формы сооружения.

На Ленинградской АЭС научными сотрудниками Сибирской государственной геодезической академии были выполнены пять циклов геодезических наблюдений по определению деформации стеновых панелей зданий турбинных цехов [103]. Анализ геометрических параметров стеновых панелей позволяет дать качественную характеристику состояния исследуемых сооружений.

Нами проведены исследования закона распределения разности отметок 1-го и 2-го циклов наблюдений на панели корпуса 410 по выборке из 1580 разностей $\Delta = H_2 - H_1$.

3.7.1. Наблюдения за деформациями стеновых панелей на АЭС

Исходные данные для статистического анализа были заданы в виде негруппированной (точечной) выборки. Предварительный анализ данных показал, что количество различающихся между собой точечных наблюдений невелико, и, следовательно, данные были получены с очень большой погрешностью, связанной либо с ошибками измерений, либо с округлениями при регистрации выборочных наблюдений. Чтобы уменьшить влияние погрешностей в задании исходных данных, выборка была сгруппирована на 17 интервалов равной длины. Количество точечных наблюдений, попавших в соответствующие интервалы группирования, приведено в таблице 3.12. Гистограмма приведена на рис. 3.42.

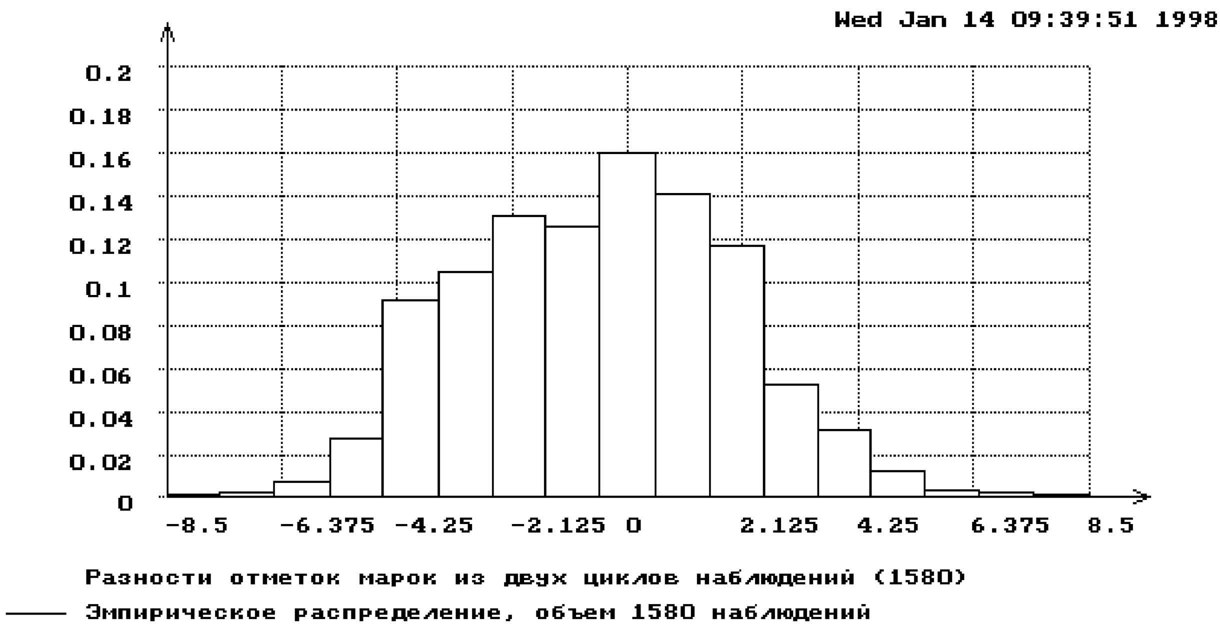


Рис. 3.42. Разности отметок марок из двух циклов наблюдений

Таблица 3.12

Разности отметок марок из двух циклов наблюдений

b	Левая граница	Правая граница	Количество	Частота
1	-8.5	-7.5	0	0.00%?
?2	-7.5	-6.5	3	0.19%?
?3	-6.5	-5.5	10	0.63%?
?4	-5.5	-4.5	42	2.66%?
?5	-4.5	-3.5	143	9.05%?
?6	-3.5	-2.5	165	10.44%?
?7	-2.5	-1.5	205	12.97%?
?8	-1.5	-0.5	198	12.53%?
?9	-0.5	0.5	252	15.95%?
?10	0.5	1.5	222	14.05%?
?11	1.5	2.5	184	11.65%?
?12	2.5	3.5	81	5.13%?
?13	3.5	4.5	49	3.10%?
?14	4.5	5.5	19	1.20%?
?15	5.5	6.5	4	0.25%?
?16	6.5	7.5	3	0.19%?
?17	7.5	8.5	0	0.00%?

3.7.2. Идентификация распределения разности отметок двух циклов наблюдений на АЭС

В качестве критерия идентификации был выбран компромиссный критерий согласия с уровнем значимости 0.05, с весами для критериев отношения правдоподобия и χ^2 Пирсона по 0.48 и для непараметрических критериев по 0.01. Такой выбор весов объясняется тем, что данные являются группированными. В качестве метода оценивания был выбран метод максимального правдоподобия по группированным данным.

Уже по гистограмме видно, что распределение разности отметок явно отличается от нормального и, вообще говоря, является двухмодальным и не является симметричным. Этот вывод подтверждается тем, что распределение не удалось идентифицировать как на множестве стандартных распределений с параметрами сдвига и масштаба (см. таблицу 3.13), так и на множестве усеченных слева и справа стандартных распределений с параметрами сдвига и масштаба (см. таблицу 3.14), хотя во втором случае значение компромиссного критерия было немножко больше почти для всех распределений.

При идентификации распределения разности отметок на множестве смесей стандартных симметричных распределений (двустороннее экспоненциальное, нормальное, логистическое, Лапласа и Коши) с параметрами сдвига и масштаба (см. таблицу 3.15), по компромиссному критерию не отвергались только три распределения:

- смесь двустороннего экспоненциального (44%) с параметром формы 14.9723, с параметром сдвига -1.1010, с параметром масштаба 3.7244 и нормального (56%) с параметром сдвига 0.00 и параметром масштаба 2.48 (см. рис. 3.43);

- смесь нормального (15%) с параметром сдвига -3.52, с параметром масштаба 0.92 и нормального (85%) с параметром сдвига 0.03, с параметром масштаба 2.17 (см. рис. 3.43);

- смесь нормального (32%) с параметром сдвига -2.90, с параметром масштаба 1.28 и логистического (68%) с параметром сдвига 0.64, с параметром масштаба 1.06 (см. рис. 3.43).

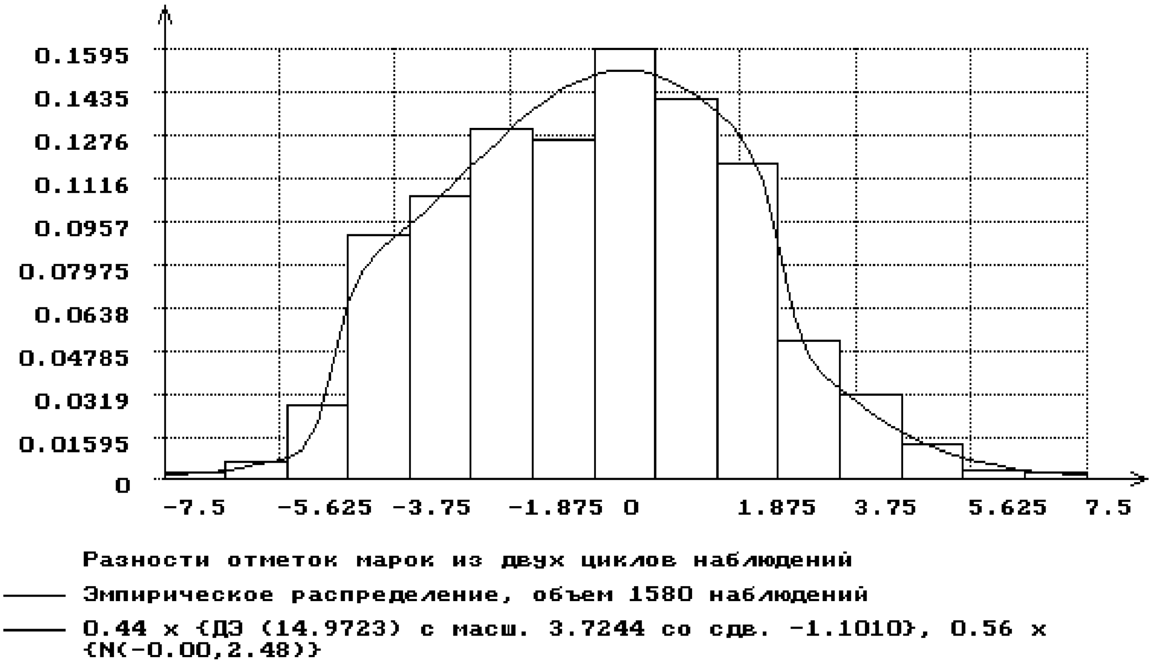


Рис. 3.43. Выравнивание наблюдений с помощью смеси двустороннего экспоненциального и нормального распределений

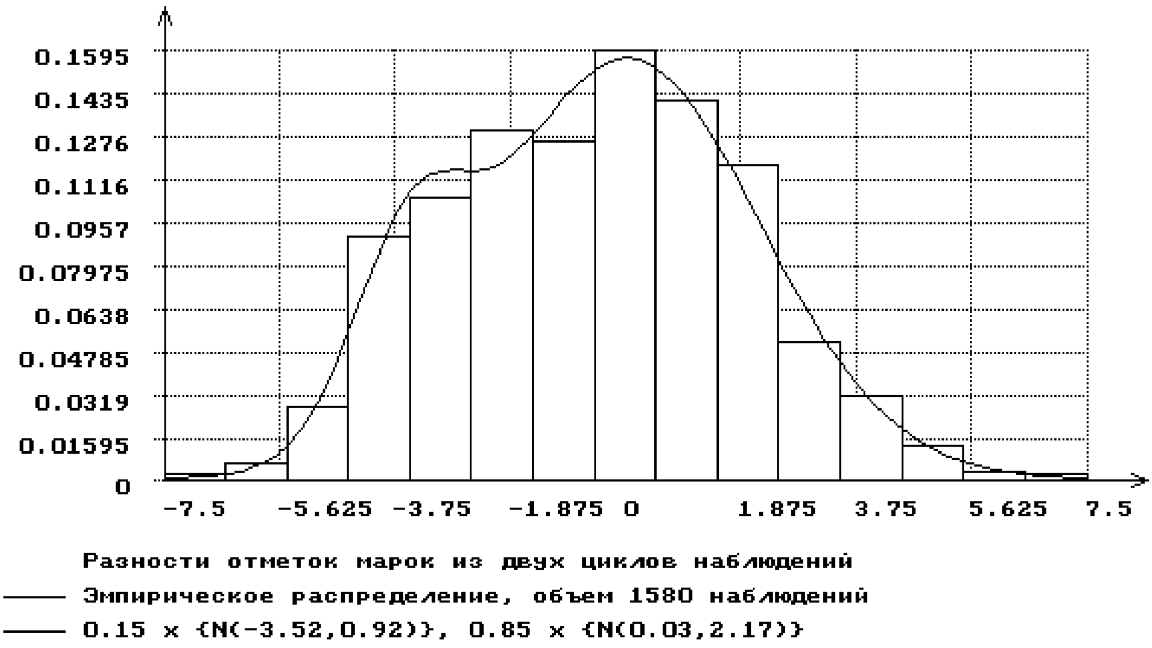


Рис. 3.44. Выравнивание наблюдений с помощью смеси двух нормальных распределений

Наилучшим из этих трех законов по компромиссному критерию является смесь двустороннего экспоненциального и нормального распределений.

Наконец, была проведена идентификация закона распределения разности отметок на множестве смесей из нескольких нормальных распределений (см. таблицу 3.16). Гипотеза о согласии проходит для смеси из двух (см. рис. 3.44), трех (см. рис. 3.46) и четырех (см. рис. 3.47) нормальных распределений, но уже для смеси из пяти нормальных распределений гипотеза о согласии отвергается.

Наилучшим на множестве из смесей из нескольких нормальных распределений оказалась смесь из четырех нормальных распределений. Однако, даже смесь из четырех нормальных распределений существенно проигрывает смеси двустороннего экспоненциального и нормального распределений.

Таким образом, на основании проведенных исследований разности отметок двух циклов наблюдений на АЭС можно сдедать вывод, что в выборке содержатся наблюдения из смеси двух распределений: нормального и

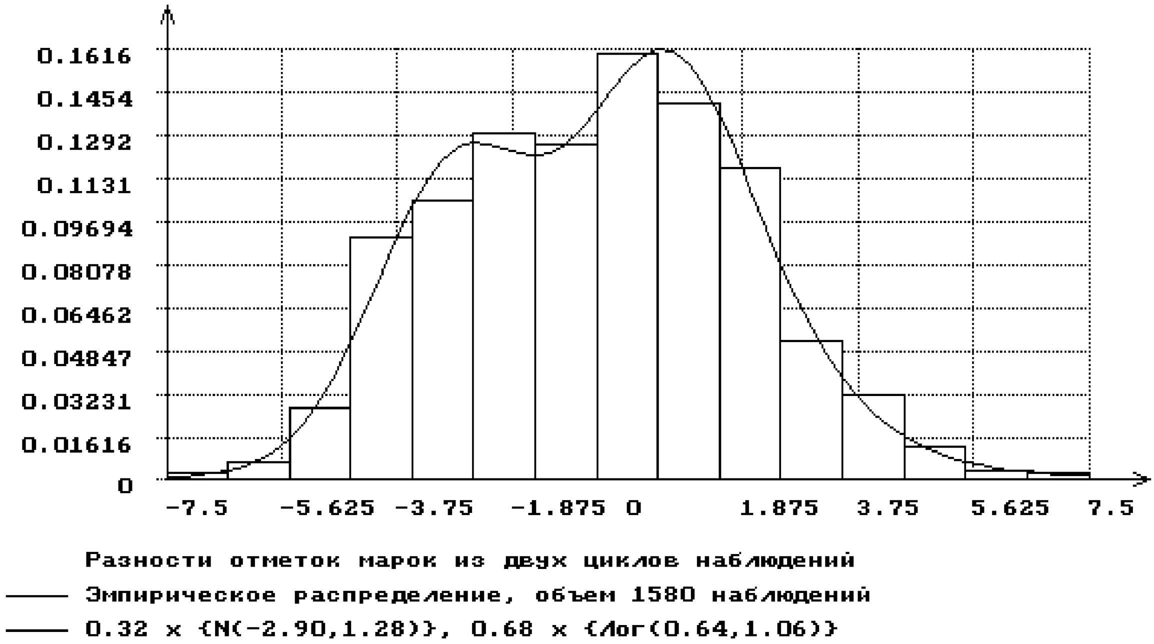


Рис. 3.45. Выравнивание наблюдений с помощью смеси нормального и логистического распределений

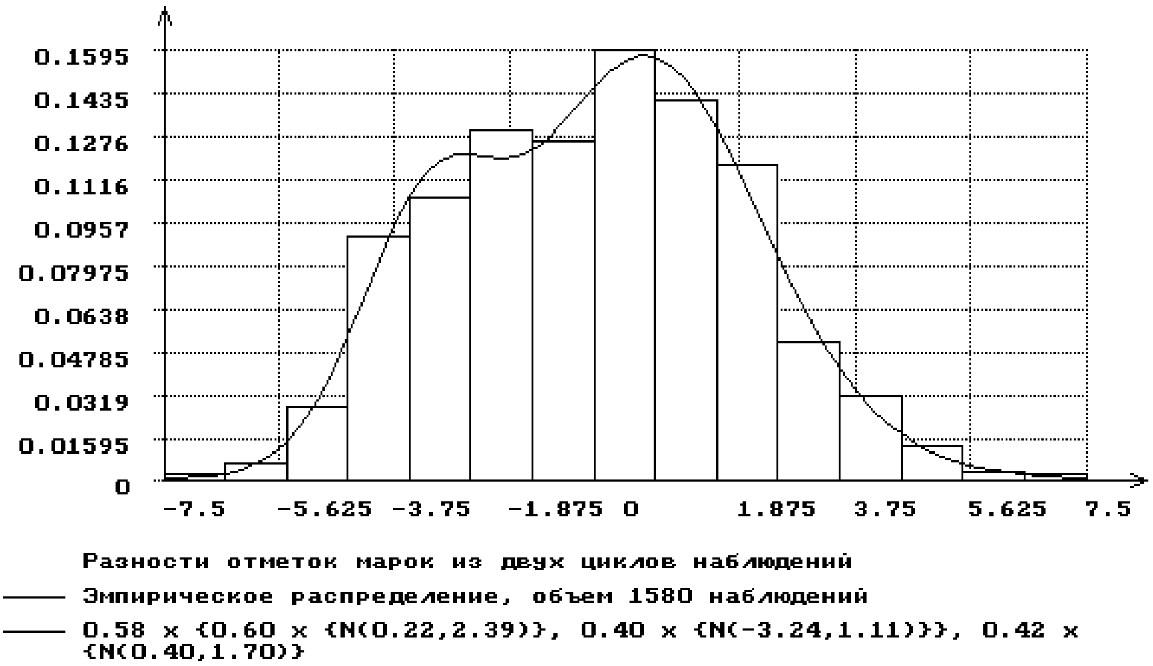


Рис. 3.46. Выравнивание наблюдений с помощью смеси трех нормальных распределений

двустороннего экспоненциального, причем последнее очень близко к равномерному, так как его параметр формы α достаточно велик.

Выводы

В главе дано описание объектно-ориентированной программной системы статистического анализа интервальных наблюдений (негруппированных, группированных, цензурированных, частично группированных, интервальных), которая

- включает все возможности программной системы [51], в том числе применение асимптотически оптимального группирования в задачах оценивания и проверки гипотез;
- обладает существенно большим множеством параметрических моделей законов распределения для описания наблюдаемых случайных величин за счет использования групповых семейств и семейств гамма-распределений, бета-распределений и распределений Джонсона;
- обеспечивает возможность вычисления оценок параметров распреде-

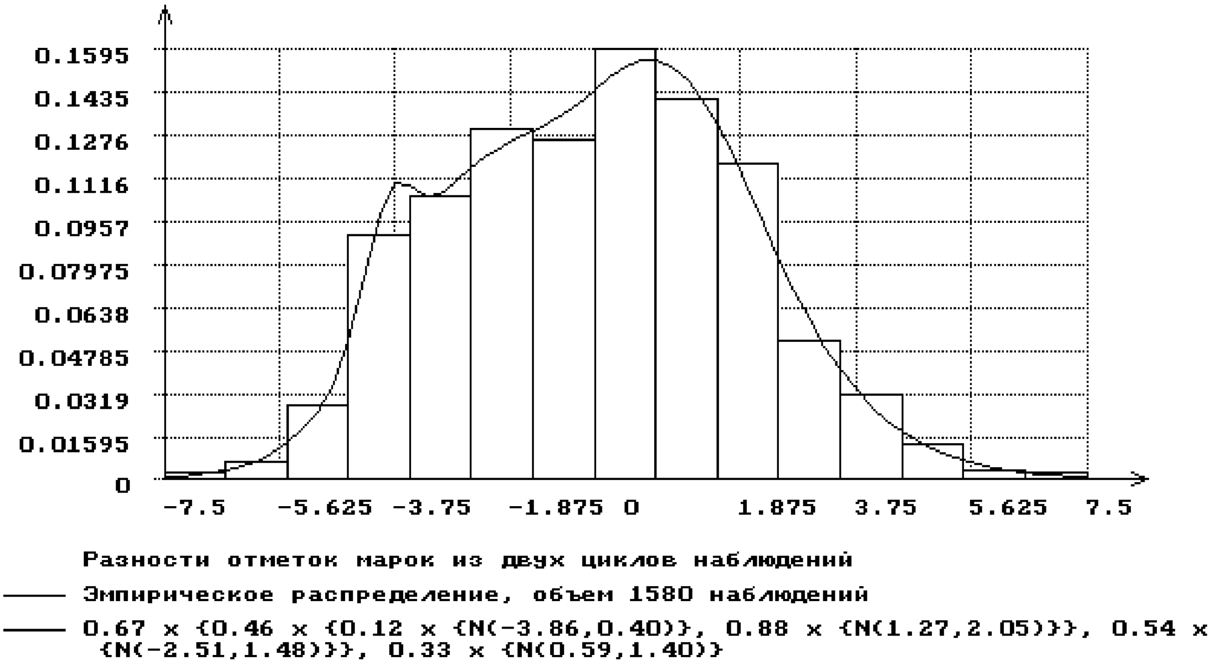


Рис. 3.47. Выравнивание наблюдений с помощью смеси четырех нормальных распределений

Таблица 3.13

Результаты статистического анализа на множестве стандартных распределений с параметрами сдвига и масштаба

Распределение	Значение критерия	Гипотеза о согласии
Равномерное	0.005	отвергается
Релея	0.005	отвергается
Максвелла	0.005104	отвергается
Эрланга	0.0067001	отвергается
Лапласа	0.0053736	отвергается
Нормальное	0.011242	отвергается
Логарифмически нормальное	0.0051212	отвергается
Коши	0.0050024	отвергается
Логистическое	0.0089105	отвергается
Вейбулла	0.014637	отвергается
Минимального значения	0.0054004	отвергается
Максимального значения	0.0053165	отвергается
Обобщенное мин. значения	0.010495	отвергается
Накагами	0.0097451	отвергается
Гамма-распределение	0.0067001	отвергается
Бета-I	0.014117	отвергается
Бета-II	0.0050026	отвергается
Бета-III	0.013461	отвергается
Sb-Джонсона	0.012575	отвергается
Sl-Джонсона	0.0051212	отвергается
Su-Джонсона	0.011363	отвергается
ДЭ	0.017174	отвергается
H-распределение	0.016098	отвергается
Г-распределение	0.01493	отвергается

Таблица 3.14

Результаты статистического анализа на множестве усеченных слева и справа стандартных распределений с параметрами сдвига и масштаба

Распределение	Значение критерия	Гипотеза о согласии
Равномерноое	0.005	отвергается
Экспоненциальное	0.005	отвергается
Полунормальное	0.005	отвергается
Релея	0.005	отвергается
Максвелла	0.0052281	отвергается
Хи	0.005	отвергается
Эрланга	0.0068611	отвергается
Лапласа	0.0061116	отвергается
Нормальное	0.012044	отвергается
Логарифмически нормальное	0.0053426	отвергается
Коши	0.0062609	отвергается
Логистическое	0.010417	отвергается
Вейбулла	0.01475	отвергается
Минимального значения	0.0069514	отвергается
Максимального значения	0.005857	отвергается
Обобщенное мин. значения	0.011551	отвергается
Накагами	0.010278	отвергается
Гамма	0.0068617	отвергается
Бета-I	0.014117	отвергается
Бета-II	0.0067001	отвергается
Бета-III	0.013468	отвергается
Sb-Джонсона	0.012575	отвергается
Sl-Джонсона	0.0053426	отвергается
Su-Джонсона	0.012067	отвергается
ДЭ	0.017149	отвергается
Н-распределение	0.016173	отвергается
Г-распределение	0.01507	отвергается

Таблица 3.15

Результаты статистического анализа на множестве смесей стандартных симметричных распределений с параметрами сдвига и масштаба

Распределение	Значение критерия	Гипотеза о согласии
ДЭ и ДЭ	0.04338	отвергается
ДЭ и Нормальное	0.17003	
ДЭ и логистическое	0.014382	отвергается
ДЭ и Лапласа	0.011679	отвергается
ДЭ и Коши	0.016325	отвергается
Нормальное и нормальное	0.11893	
Нормальное и логистическое	0.081603	
Нормальное и Лапласа	0.015657	отвергается
Нормальное и Коши	0.012311	отвергается
Логистическое и логистическое	0.033166	отвергается
Логистическое и Лапласа	0.013734	отвергается
Лоистическое и Коши	0.010951	отвергается
Лапласа и Лапласа	0.011565	отвергается
Лапласа и Коши	0.0078598	отвергается
Коши и Коши	0.0054779	отвергается

Таблица 3.16

Результаты статистического анализа на множестве смесей нормальных распределений

Распределение	Значение критерия	Гипотеза о согласии
Нормальное	0.011242	отвергается
Смесь из 2-х нормальных	0.057435	
Смесь из 3-х нормальных	0.061278	
Смесь из 4-х нормальных	0.079749	
Смесь из 5-и нормальных	0.027413	отвергается

лений несколькими методами: получаемых по методу максимального правдоподобия (M-оценок); получаемых минимизацией статистик критериев согласия Колмогорова и ω^2 Мизеса (MD-оценок).

- обеспечивает проверку *сложных* гипотез о согласии непараметрическими критериями Колмогорова, Смирнова, ω^2 и Ω^2;
- позволяет проводить отбраковку аномальных наблюдений с использованием робастных оценок, получаемых при минимизации статистики Колмогорова;
- позволяет идентифицировать закон распределения выборки на заданном множестве распределений по совокупности нескольких критериев согласия;
- позволяет генерировать псевдослучайные выборки, подчиненные заданному закону распределения.

При создании программной системы были проведены следующие исследования:

1. Методами статистического моделирования исследованы распределения статистик непараметрических критериев согласия при проверке сложных гипотез и вычислении оценок параметров методом максимального правдоподобия [80, 83].
2. Методами статистического моделирования исследована зависимость распределений статистик критериев согласия χ^2 Пирсона и отношения правдоподобия от способа группирования [81, 83].
3. Установлена существенная зависимость распределений статистик критериев согласия при проверке сложных гипотез от метода оценивания параметров и от объема наблюдаемой выборки. В частности, практический интерес вызывает то, что статистика Колмогорова при использовании MD-оценки, построенной по минимуму статистики Колмогорова, имеет один и тот же закон распределения при наблюдении случайных величин, подчиняющихся различным по форме законам: нормальному, Лапласа, Коши.

4. На основе функций влияния Хампеля исследована робастность оценок, получаемых при минимизации статистики Колмогорова.

Программная система обеспечивает получение более корректных статистических выводов по сравнению с существующими программными средствами статистического анализа, представляет собой инструмент для исследования статистических свойств оценок и критериев проверки гипотез.

Система может использоваться при решении задач контроля качества и исследованиях надежности, при обработке результатов наблюдений в любой прикладной области.

ЗАКЛЮЧЕНИЕ

В диссертации получены следующие результаты.

1. Предложена классификация наблюдений одномерных непрерывных случайных величин. Рассмотрены нечеткие, размытые, интервальные и точечные наблюдения. Найдены условия, при которых размытые и нечеткие наблюдения являются эквивалентными.

2. По интервальной выборке построены функции, определяющие коридор для эмпирической функции распределения.

3. Найдены аналитические выражения для вычисления границ интервалов, содержащих (все) возможные значения статистик непараметрических критериев согласия Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса по интервальной выборке. Показано, что процедура принятия решения при проверке гипотез о согласии по интервальной выборке становится неопределенной, так как значение статистики критерия известно с точностью до интервала.

4. На примере использования критерия согласия Колмогорова в случае интервальной выборки показано, что при достаточно большом объеме выборки существует целое множество моделей (значений параметра), не отвергаемых критерием с точки зрения “крайнего оптимиста”, так как из-за погрешностей средств измерения они являются неразличимыми. И в то же время, с точки зрения “крайнего пессимиста”, при

достаточно большом числе наблюдений можно отвергнуть любую модель.

5. Предложен подход для получения интервальных оценок параметров распределений по интервальной выборке. Предложены и реализованы процедуры для вычисления интервальных оценок по интервальной выборке в классах L-, M-, MD-оценок.

6. Показано, что функция влияния Хампеля для статистики Колмогорова является ограниченной, откуда следует робастность оценок, получаемых при минимизации статистики Колмогорова. Предложено и реализовано использование этих оценок в процедуре выделения аномальных наблюдений.

7. Методами статистического моделирования показано, что при проверке сложных гипотез распределения статистик непараметрических критериев согласия существенно зависят не только от вида наблюдаемого закона распределения, типа и количества оцениваемых параметров, но и от метода оценивания. Предложен и реализован алгоритм проверки согласия по непараметрическим критериям в случае сложных гипотез, учитывающий вид наблюдаемого закона распределения, тип и количество оцениваемых параметров, метод оценивания параметров и ограниченный объем выборки.

8. Разработана объектно-ориентированная программная система статистического анализа, обеспечивающая большой выбор параметрических вероятностных моделей, получаемых за счет применения следующих операций над законами распределений: сдвига, масштабирования, зеркального отражения, усечения, смешивания, произведения. Система обеспечивает решение задач статистического анализа одномерных наблюдений непрерывных случайных величин (оценивания параметров распределений, проверки гипотез о согласии, идентификации закона распределения) по интервальным выборкам, в том числе по точечным, цензурированным, частично группированным и группированным. В системе реализована процедура автоматической

идентификации закона распределения выборки на множестве заданных законов.

СПИСОК ЛИТЕРАТУРЫ

1. Хьюбер П. Робастность в статистике. — М.: Мир, 1984. — 303 с.
2. Робастность в статистике. Подход на основе функций влияния / Хампель Ф., Рончетти Э., Рауссей П., Штаэль В. — М.: Мир, 1989. — 512 с.
3. Шуленин В.П. Введение в робастную статистику. — Томск: Изд-во Том. ун-та, 1993. — 227 с.
4. Тарасенко Ф.П. Непараметрическая статистика. — Томск: изд-во Томского ун-та, 1976. — 291 с.
5. Орлов А.И. О влиянии погрешностей наблюдений на свойства статистических процедур (на примере гамма-распределения) // Статистические методы оценивания и проверки гипотез: Межвуз. сб. науч. тр. / Перм. ун-т. — Пермь, 1988. — С.45–55.
6. Орлов А.И. Комментарий IV к статье Вощинина А.П., Бочкова А.Ф., Сотирова Г.Р. “Метод анализа данных при интервальной нестатистической ошибке” // Заводская лаборатория. — 1990. Т.56. *ь 7. — С. 86–89.*
7. *Орлов А.И. Некоторые алгоритмы реалистической статистики // Статистические методы оценивания и проверки гипотез: Межвуз. сб. науч. тр. / Перм. ун-т. — Пермь, 1991. — С.77–86.*
8. *Ибрагимов И.А., Хасьминский Р.З. Асимптотическая теория оценивания. — М.: Наука, 1979. — 528 с.*
9. *Орлов А.И. Устойчивость в социально-экономических моделях. — М.: Наука, 1979. — 296 с.*
10. *Заде Л. Понятие лингвистической переменной и ее применения. — М.: Мир, 1976. — 168 с.*

11. *Обработка нечеткой информации в системах принятия решения / Борисов А.Н., Алексеев А.В. и др. — — М.: Радио и связь, 1989. — 304 с.*

12. *Шошин П.Б. Размытые числа как средство описания субъективных величин // Статистические методы анализа экспертных оценок. М.: Наука, 1977.*

13. *Viertl R. Statistical inference for fuzzy data in environmetrics // Environmetrics* **1** *(1990), p.37–42.*

14. *Kruse R., Meier K.D. Statistics with Vague Data // Theory and Decision Library. Series B, Mathematical and Statistical Methods, Dordecht-Boston, 1987.*

15. *Fruhwirth-Schnatter S. On statistical inference for fuzzy data with applications to descriptive statistics // Fuzzy Sets and Systems. — 1992. Vol. 50. — P.143–165.*

16. *Orlov A.I. Interval Statistics // Interval computations. — 1992. ь 1. — P.44–52.*

17. *Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. — Л.: Энергоатомиздат, 1991. — 303 с.*

18. *Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? // Заводская лаборатория. — 1990. Т.57. ь 7. — С. 64–66.*

19. *Новицкий П.В., Зограф И.А., Лабунец В.С. Динамика погрешностей средств измерений. — Л.: Энергоатомиздат, 1990. — 192 с.*

20. *Новицкий П.В. Об особых свойствах* 95%*-й квантили большого класса распределений и предпочтительном значении доверительной вероятности при указании погрешностей приборов и измерений. // Метрология. — 1979. — ь 2. — С. 18–24.*

21. *Кузнецов В.П. Интервальные статистические модели. — М.: Радио и связь, 1991. — 352 с.*

22. *R.E. Moore. Inteval Analysis. — Prentice-Hall, Englewood Cliffs, New-Jersey, 1966.*

23. *Шокин Ю.И. Интервальный анализ. — Новосибирск: Наука, 1981. — 284 с.*

24. *Алефельд Г., Херцбергер Ю. Введение в интервальные вычисления. — М.: Мир, 1987.*

25. *Орлов А.И. О развитии реалистической статистики // Статистические методы оценивания и проверки гипотез: Межвуз. сб. науч. тр. / Перм. ун-т. — Пермь, 1990. — С.89–99.*

26. *Орлов А.И. Интервальный статистический анализ // Статистические методы оценивания и проверки гипотез: Межвуз. сб. науч. тр. / Перм. ун-т. — Пермь, 1993. — С.149–158.*

27. *Orlov A.I. Invariance Leads to the Interval Character of Ordinal Statistical Characteristics // Proceedings of "Applications of Interval Computations" (APIC'95). — El Paso, 1995.*

28. *Кузнецов В.П. Комментарий VI к статье Вощинина А.П., Бочкова А.Ф., Сотирова Г.Р. "Метод анализа данных при интервальной нестатистической ошибке" // Заводская лаборатория. — 1990. Т.56. ь 7. — С. 93–95.*

29. *Kuznetsov V.P. Auxiliary Problems of Statistical Data Processing: Interval Approach // Proceedings of "Applications of Interval Computations"*
(APIC'95). — El Paso, 1995.

30. *Kuznetsov V.P. Interval Methods For Processing Statistical Characteristics // Proceedings of "Applications of Interval Computations" (APIC'95). — El Paso, 1995.*

31. *Вощинин А.П., Бочков А.Ф., Сотиров Г.Р. Метод анализа данных при интервальной нестатистической ошибке // Заводская лаборатория. — 1990. Т.56. ь 7. — С. 75–81.*

32. *Бородюк В.П. Комментарий I к статье Вощинина А.П., Бочкова А.Ф., Сотирова Г.Р. "Метод анализа данных при интервальной нестатистической ошибке" // Заводская лаборатория. — 1990. Т.56. ь 7. — С. 81–83.*

33. *Демиденко Е.З. Комментарий II к статье Вощинина А.П., Бочкова А.Ф., Сотирова Г.Р. "Метод анализа данных при интервальной нестатистической ошибке" // Заводская лаборатория. — 1990. Т.56. ь 7. — С. 83–84.*

34. *Лецкий Э.К. Комментарий III к статье Вощинина А.П., Бочкова А.Ф., Сотирова Г.Р. "Метод анализа данных при интервальной нестатистической ошибке" // Заводская лаборатория. — 1990. Т.56. ь 7. — С. 84–86.*

35. *Легостаева И.Л. Комментарий V к статье Вощинина А.П., Бочкова А.Ф., Сотирова Г.Р. "Метод анализа данных при интервальной нестатистической ошибке" // Заводская лаборатория. — 1990. Т.56. ь 7. — С. 90–93.*

36. *Родионов А.С. Оценивание параметра экспоненциального распределения по группированным выборкам с пропусками между интервалов // Тр. ВЦ СО РАН. Сер. Системное моделирование. — 1993. Вып 1. — С. 81–91.*

37. *Родионов А.С., Романов Д.В. Об одной задаче оценивания параметров распределений // Тр. ВЦ СО РАН. Сер. Системное моделирование. — 1995. Вып 3. — С. 115–122.*

38. *Теоретические и прикладные аспекты метода центра неопределенности / Белов В.М., Суханов В.А., Унгер Ф.Г. — Новосибирск: Наука, 1995. — 144 с.*

39. *Kreinovich V. Data Processing Beyond Traditional Statistics: Applications of Interval Computations. A Brief Introduction // Proceedings of "Applications of Interval Computations" (APIC'95). — El Paso, 1995.*

40. *Объектно-ориентированная программная система статистического анализа. Таблицы коэффициентов для оптимальных L-оценок параметров сдвига и масштаба по выборочным квантилям больших выборок и таблицы асимптотически оптимального группирования наблюдений / Отчет по НИР, НГТУ, 1996г. ь гос. рег. 01.9.70 000190 инв. ь 02.9.60 008130. Лемешко Б.Ю., Постовалов С.Н. — 129 с.*

41. *Теория, методы и программное обеспечение задач статистического анализа независимых и зависимых случайных величин в геодезии / Отчет по НИР, НГТУ, 1994г. ь гос. рег. 01.9.50 001519 инв. ь 02.9.50 001199. Денисов В.И., Лемешко Б.Ю., Лесных Н.Б., Постовалов С.Н. — 40 с.*

42. *Теория, методы и программное обеспечение задач статистического анализа независимых и зависимых случайных величин в геодезии. Вероятностные модели, непараметрические критерии, анализ / Отчет по НИР, НГТУ, 1995г. ь гос. рег. 01.9.60 003119 инв. ь 02.9.60 005332. Денисов В.И., Лемешко Б.Ю., Лесных Н.Б., Постовалов С.Н. — 42 с.*

43. *Статистический анализ независимых и зависимых случайных величин в геодезии / Отчет по НИР, НГТУ, 1996г. ь гос. рег. 01.9.60 003120 инв. ь 02.9.60 008130. Лемешко Б.Ю., Лесных Н.Б., Мизина Г.И., Постовалов С.Н. — 69 с.*

44. *Parzen E. On the estimation of a probability density function and the mode // Ann. Math. Stat., 1962. — Vol. 33. — P. 1065–1076.*

45. *Канторович Л.В. О некоторых новых подходах к вычислительным методам и обработке наблюдений // Сиб. мат. журн., 1962. ь 3(5). — С. 701–709.*

46. *Artbauer O. Application of interval, statistical, and fuzzy methods to the evaluation of measurements / Metrologia, 1988. — Vol. 25. — P. 81–86.*

47. *Рао С.Р. Линейные статистические методы и их применения. — М.: Наука, 1968. — 548 с.*

48. *Куллдорф Г. Введение в теорию оценивания по группированным и частично группированным выборкам. — М.: Наука, 1966. — 176 с.*

49. *Бодин Н.А. Оценка параметров распределений по группированным выборкам // Тр. Мат. ин-та АН СССР. — 1970. — Т.111. — С. 100–154.*

50. *Денисов В.В., Лемешко Б.Ю, Цой Е.Б. Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов. В 2-х ч. / Новосиб. гос. техн. ун-т — Новосибирск, 1993. — 347 с.*

51. *Лемешко Б.Ю. Статистический анализ одномерных наблюдений случайных величин: Программная система. — Новосибирск: Изд-во НГТУ, 1995. — 125 с.*

52. *Лемешко Б.Ю., Постовалов С.Н. Проверка непараметрических гипотез по группированным данным // Материалы международной НТК "Информатика и проблемы телекоммуникаций". — Новосибирск, 1995. — Т.1. — С. 63–65.*

53. *Лемешко Б.Ю., Постовалов С.Н. К использованию непараметрических критериев по частично группированным данным // Сб. научных трудов НГТУ. — Новосибирск, 1995. — ь 2. — С. 21–30.*

54. *Gastaldi T. A Kolmogorov-Smirnov Test Procedure Involving a Possibly Censored or Truncated Sample // Communication in Statistics. Theory and Methods. — 1993. ь 22 (1). — P.31–39.*

55. *Лемешко Б.Ю., Постовалов С.Н. Статистический анализ одномерных наблюдений по частично группированным данным // Изв. вузов. Физика. — Томск, 1995. — ь 9. — С. 39–45.*

56. *Лемешко Б.Ю., Постовалов С.Н. Статистический анализ наблюдений, имеющих интервальное представление // Сб. научных трудов НГТУ. — Новосибирск, 1996. — ь 1. — С. 3–12.*

57. *Лемешко Б.Ю., Постовалов С.Н. О решении задач статистического анализа интервальных наблюдений // Вычислительные технологии. — Новосибирск, 1997. — Т.2. — ь 1. — С. 28–36.*

58. *Крамер Г. Математические методы статистики. — М.: Мир, 1975. — 648 с.*

59. *Леман Э. Теория точечного оценивания. — М.: Наука, 1991. — 448 с.*

60. *Лемешко Б.Ю., Постовалов С.Н. К вопросу о робастности оценок по группированным данным // Сб. научных трудов НГТУ. — Новосибирск, 1996. — ь 2. — С. 9–18.*

61. *Айвазян С.А. Программное обеспечение персональных ЭВМ по статистическому анализу данных (проблемы, тенденции, перспективы отечественных разработок) // Заводская лаборатория, 1991. Т.57. ь 1. — С. 54-58.*

62. *Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ. М.: Финансы и статистика, 1989. — 191 с.*

63. *Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. М.: ИНФРА-М, Финансы и статистика, 1995. — 384 с.*

64. *Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. — М.: Мир, 1982. — 488 с.*

65. *Статистические методы для ЭВМ / Под ред. К.Эйнслейна, Э.Рэлстона, Г.С. Уолфа. — М.: Наука, 1986. — 459 с.*

66. *Сильвестров Д.С. Программное состояние прикладной статистики: обзор состояния. Тенденция развития. — М.: Финансы и статистика, 1988. — 232 с.*

67. *Сильвестров Д.С., Семенов Н.А., Марищер В.В. Пакеты прикладных программ статистического анализа. — Киев: Техника, 1990. — 176 с.*

68. *Кощеев В.А. Автоматизация статистического анализа данных. — М.: Наука, 1988. — 232 с.*

69. *Лемешко Б.Ю., Постовалов С.Н. Вопросы обработки выборок одномерных случайных величин // Научный вестник НГТУ, — Новосибирск, 1996. — ь 2. — С. 3–24.*

70. *Лемешко Б.Ю., Постовалов С.Н. Некоторые вопросы статистического анализа одномерных распределений // Материалы международной научной конференции "Всесибирские чтения по математике и механике". — Томск, 1997. — С.*

71. *Лемешко Б.Ю., Постовалов С.Н. Система статистического анализа одномерных непрерывных распределений случайных величин (версия 3.0) // Материалы III международной НТК "Микропроцессорные системы автоматики". — Новосибирск, 1996. — Т.1. — С. С-16–С-17.*

72. *Лемешко Б.Ю., Постовалов С.Н. Система статистической обработки случайных наблюдений // Тезисы докладов международной НТК "Информационные технологии в моделировании и управлении". — Санкт-Петербург, 1996. — С. 155–159.*

73. *Лемешко Б.Ю., Постовалов С.Н. Программное обеспечение статистического анализа смесей случайных величин, представленных частично группированными и интервальными выборками // Тр. третьей международной НТК "Актуальные проблемы в электронного приборостроения" (АПЭП-96). — Новосибирск, 1996. — Т.6. Ч.1. — С. 50–53.*

74. *Лемешко Б.Ю., Постовалов С.Н. Программное обеспечение задач статистического анализа одномерных непрерывных случайных величин // Материалы международной научно-методической конференции "Новые информационные технологии в университетском образовании". — Новосибирск, НИИМИОО, 1997. — С. 44–45.*

75. *Denisov V.I., Lemeshko B.You., Tsoi Ye.B., Tishkovskaya S.V., Postovalov S.N. Software for statistical analysis of grouped data // Proceedings the*

First Korea-Russia International Symposium of Science and Technology, 1997. — P. 239–243

76. *Лемешко Б.Ю., Постовалов С.Н. Статистический анализ смесей распределений по частично группированным данным // Сб. научных трудов НГТУ. — Новосибирск, 1995. — ь 1. — С. 25–31.*

77. *Лемешко Б.Ю., Постовалов С.Н. Статистический анализ смесей распределений по группированным данным // Материалы международной НТК "Информатика и проблемы телекоммуникаций". — Новосибирск, 1995. — Т.1. — С. 83–85.*

78. *Лемешко Б.Ю., Постовалов С.Н. Статистический анализ смесей и усеченных распределений случайных величин // Тезисы российской НТК "Информатика и проблемы телекоммуникаций". — Новосибирск, 1996. — Т.2. — С. 38–39.*

79. *Лемешко Б.Ю., Постовалов С.Н. К вопросу о распределениях статистик непараметрических критериев согласия // Сб. научных трудов НГТУ, — Новосибирск, 1997. — ь 1. — С. 23–32.*

80. *Лемешко Б.Ю., Постовалов С.Н. О распределениях статистик непараметрических критериев при потере свойства "свободы от распределения" // Материалы международной НТК "Информатика и проблемы телекоммуникаций". — Новосибирск, 1997. — С. 117–120.*

81. *Лемешко Б.Ю., Постовалов С.Н. О влиянии способа группирования данных на распределения статистик χ^2 Пирсона и отношения правдоподобия // Материалы международной НТК "Информатика и проблемы телекоммуникаций". — Новосибирск, 1997. — С. 120–123.*

82. *Лемешко Б.Ю., Постовалов С.Н. Моделирование распределений статистик непараметрических критериев согласия при потере свойства "свободы от распределения"// Труды международной НТК "Научные основы высоких технологий" (НОВТ-97). — Новосибирск, 1997. Т.6.— С. 159–163*

83. *Лемешко Б.Ю., Постовалов С.Н. Прикладные аспекты использования критериев согласия в случае проверки сложных гипотез // Надежность и контроль качества. — Москва, 1997. — ь 11. — С. 3–17.*

84. *Страуструп Б. Язык программирования С++. — М.: Радио и связь, 1991. — 352 с.*

85. *Ермаков С.М., Михайлов Г.А. Статистическое моделирование. — М.: Наука, 1982. — 296 с.*

86. *Исаенко О.К., Урбах В.Ю. Разделение смеси распределений вероятностей на их составляющие // Итоги науки и техники. Сер. ТВ.МС.ТК — ВИНИТИ, 1976. — Т.13. — С. 37–58.*

87. *Губарев В.В. Вероятностные модели: Справочник. В 2-х ч. / Новосиб. электротехн. ин-т. — Новосибирск, 1992. — 422 с.*

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

89. *Воинов В.Г., Никулин М.С. Несмещенные оценки и их применения. — М.: Наука, 1989. — 440 с.*

90. *Статистическая классификация, основанная на выборочных распределениях / Иголкин В.Н., Ковригин А.Б., Старшинов А.И., Хохлов В.П. // Л.: изд-во ЛГУ, 1978. — 104 с.*

91. *Durbin J. Kolmogoriv-Smirnov test when parameters are estimated // Lect. Notes Math. 1976. V. 566. P. 33–44.*

92. *Мартынов Г.В. Критерии омега-квадрат. — М.: Наука, 1978. — 80 с.*

93. *Pearson E.S., Hartley H.O. Biometrica tables for Statistics. V.2. — Cambridge: University Press, 1972. —634 p.*

94. *Stephens M.A. Use of Kolmogorov-Smirnov, Cramer - von Mises and related statistics — vithout extensive table // J. R. Stat. Soc., 1970, B. 32. — P. 115–122.*

95. *Stephens M.A. EDF statistics for goodness of fit and some comparisons // J. Am. Statist. Assoc., 1974, v.69. — P. 730–737.*

96. *Chandra M., Singpurwalla N.D., Stephens M.A. Statistics for Test of Fit for the Extrem-Value and Weibull Distribution // J. Am. Statist. Assoc., 1981, v.76. — P. 375.*

97. *Тюрин Ю.Н. О предельном распределении статистик Колмогорова-Смирнова для сложной гипотезы // Изв. АН СССР. Сер. Матем., 1984, т. 48, ь 6. — С. 1314–1343.*

98. *Тюрин Ю.Н., Саввушкина Н.Е. Критерии согласия для распределения Вейбулла-Гнеденко. // Изв. АН СССР. Сер. Техн. кибернетика, 1984, ь 3. — С. 109–112.*

99. *Тюрин Ю.Н. Исследования по непараметрической статистике (непараметрические методы и линейная модель). Автореф. дисс. на соиск. учен. степени д-ра физ.-мат. наук. — М., 1985. – 33 с. — (МГУ).*

100. *Саввушкина Н.Е. Критерий Колмогорова-Смирнова для логистического и гамма-распределения // Сб. тр. ВНИИ систем. исслед. — 1990, ь 8.*

101. *Лемешко Б.Ю., Постовалов С.Н. Статистический анализ одномерных непрерывных распределений случайных величин // Тезисы докладов Второго Сибирского Конгресса по Прикладной и Индустриальной Математике (ИНПРИМ-96). — Новосибирск, 1996. — С. 178–179.*

102. *Лемешко Б.Ю., Постовалов С.Н. Робастные алгоритмы оценивания и параметрические методы отбраковки аномальных наблюдений // Тр. третьей международной НТК "Актуальные проблемы в электронного приборостроения" (АПЭП-96). — Новосибирск, 1996. — Т.6. Ч.1. — С. 45–49.*

103. *Отчет о научно-исследовательской работе "Исследование деформаций наружных строительных элементов (панелей) Ленинград-*

ской АЭС" / Лесных И.В., Жарников В.В. и др.

ПРИЛОЖЕНИЕ 1

Руководство пользователя программной системы “Статистический анализ интервальных наблюдений одномерных непрерывных случайных величин”

П.1.1. Компоненты системы

Система состоит из основных файлов и файлов, появляющихся в результате работы

П.1.1.1. Основные файлы

`is.exe`	исполняемый файл
`title.pcx`	заставка
`is.ini`	опции и задание на выполнение
`read.me`	описание системы
`obs2dat.exe`	конвертор данных из двоичного формата данных в текстовый
`dat2obs.exe`	конвертор данных из текстового формата данных в двоичный

П.1.1.2. Вспомогательные файлы

`is.log`	протокол работы системы
`*.dat`	файлы с данными в текстовом формате
`*.obs`	файлы с данными в двоичном формате
`*.res`	файлы результатов

П.1.2. Инициализационный файл “`is.ini`”

В файле инициализации можно задать основные параметры системы. Система может работать в пакетном режиме, в режиме диалога и в специальной оболочке с оконным интерфейсом.

П.1.2.1. Структура файла инициализации

В файле содержатся ключевые слова разделов, команды инициализации и комментарии.

П.1.2.1.1. Ключевые слова разделов

Разделы начинаются с ключевых слов и идут в следующем порядке:

```
[Distributions]
<Список распределений>

[Samples]
<Список выборок>

[Options]
<Параметры>

[Job]
<Задание на выполнение>
```

П.1.2.1.2. Команды

Разделы состоят из наборов команд, причем в одной строке может быть только одна команда. Каждая команда имеет следующий формат:

```
[<идентификатор> =] <процедура> (<список параметров>)
```

<идентификатор> — это уникальное имя объекта, инициализируемого процедурой <процедура>, состоит из не более чем 15 букв и цифр без пробелов и управляющих символов. Идентификатор может использоваться в качестве параметров других процедур.

<список параметров> — это набор параметров процедуры <процедура>, разделенных запятой.

П.1.2.1.3. Комментарии

Комментарием считается любая строка, начинающаяся с символа “*”, а также любой текст после символов “//” до конца строки.

П.1.2.2. Разделы

Файл инициализации состоит из четырех разделов, каждый из которых начинается с ключевого слова, заключенного в квадратные скобки. Выполнение каждой команды производится по порядку следования.

П.1.2.2.1. Раздел `[Distributions]`

В этом разделе происходит инициализация списка распределений. Распределение инициализируется командой

```
<распределение> = {D0|D1|D2|...|D29} ([<список параметров>])
```

`D0`, `D1`, ... , `D29` — зарезервированные в системе идентификаторы распределений (таблица П.1.1). Вместо идентификаторов “Dxx” можно использовать их синонимы, указанные в таблице.

Над стандартными распределениями можно применять операции преобразования:

```
<распределение> = <операция> (<список распределений>, [<параме
<операция> = {Shift|Scale|Reflection|Left|Right|Mixt|Mult}
```

Операции преобразования распределений перечислены в таблице П.1.2.

Примечание П.1. Новые распределения рекомендуется обозначать строчными буквами, чтобы они отличались от стандартных.

Примечание П.2. Максимальное количество распределений определяется константой `MaxDistrNum` в разделе `[Options]`.

Примечание П.3. Если параметр распределения не указан явно, то он инициализируется по умолчанию с флагом “неизвестный” и допускает возможность оценивания. В противном случае, параметр инициализируется с флагом “известный” и оцениваться не может.

Следующий пример иллюстрирует инициализацию нескольких распределений.

Пример П.1.

```
[Distributions]
d1=D9()                   // инициализируем d1 стандартным
                          // нормальным распределением
d2=Scale(d1,2)            // добавляем параметр масштаба, равный
d3=Shift(d2,1)            // добавляем параметр сдвига, равный 1
d4=Shift(Scale(D9(),2),1) // то же, что и d3, но в одной строк
d5=Shift(Scale(D9()))     // то же, что и d4, но параметры неизве
```

Таблица П.1.1

Идентификаторы стандартных распределений

Идентификатор	Синоним	Название распределения	Число параметров
D0	UNIFORME	Равномерное	0
D1	EXP	Экспоненциальное	0
D2	SEMI_NORM	Полунормальное	0
D3	RELEY	Релея	0
D4	MAXWELL	Максвелла	0
D5	CHI	χ-распределения	1
D6	PARETO	Парето	1
D7	ERL	Эрланга	1
D8	LAPLACE	Лапласа	0
D9	NORM	Нормальное	0
D10	LN_NORM	Логарифмически (ln) нормальное	2
D11	LG_NORM	Логарифмически (lg) нормальное	2
D12	CAUCHIE	Коши	0
D13	LOGIST	Логистическое	0
D14	VEI	Вейбулла	1
D15	MIN	Минимального значения	0
D16	MAX	Максимального значения	0
D17	G_MIN	Обобщенное минимального значения	1
D18	NAK	Накагами	1
D19	GAMMA	Гамма	1
D20	BETA_I	Бета 1-го рода	2
D21	BETA_II	Бета 2-го рода	2
D22	BETA_III	Бета 3-го рода	3
D23	SB_J	Sb-Джонсона	2
D24	SL_J	Sl-Джонсона	2
D25	SU_J	Su-Джонсона	2
D26	DEXP	Двустороннее экспоненциальное	1
D27	H	H-распределение	2
D28	G	Г-распределение	2
D29	L	L-распределение	2
D31	CHI_SQR	χ^2-распределение	1
D30	KOLM	Колмогорова	0
D31	A1	Распределение статистики ω^2	0
D32	A2	Распределение статистики Ω^2	0

Таблица П.1.2

Операции над распределениями

Операция	Название	Число распределений	Число параметров
`Shift`	Сдвиг	1	1
`Scale`	Масштаб	1	1
`Reflection`	Зеркальное отражение	1	0
`Left`	Усечение слева	1	1
`Right`	Усечение справа	1	1
`Mixt`	Смесь	2	1
`Mult`	Произведение	2	0

```
d6=Mixt(d4,d5,0.1)      // смесь двух нормальных распределения
                        // с параметром смеси 0.1
```

П.1.2.2.2 Раздел `[Samples]`

В этом разделе происходит инициализация списка выборок. Выборка инициализируется командой

```
<выборка> = { Open | Read } (<имя файла>)
```

Команда “`Open`” отличается от команды “`Read`” тем, что выборка не считывается в оперативную память, что позволяет обрабатывать достаточно большие объемы данных. Имя файла задается в кавычках.

Новая выборка создается командой “`Generate`”:

```
<выборка> = Generate (<имя файла>, <название>, <распределение>
               <объем>, <абс. погрешность>, <отн. погрешность>
```

Над выборками также допустимы операции преобразования:

```
<выборка> = <операция> (<список выборок>, [<параметр>])
```

Виды возможных преобразований перечислены в таблице П.1.3.

Примечание П.4. В команде преобразования вместо параметра `<выборка>` можно указывать непосредственно имя файла.

Примечание П.5. Максимальное количество выборок определяется константой `MaxSampleNum` в разделе `[Options]`.

Таблица П.1.3

Операции над выборками

Операция	Название	Число выборок	Число параметров
`Shift`	Сдвиг	1	1
`Scale`	Масштаб	1	1
`Reflection`	Зеркальное отражение	1	0
`Left`	Усечение слева	1	1
`Right`	Усечение справа	1	1
`Mixt`	Смесь выборок	2	0
`CensLeft`	Цензурирование слева	1	1
`CensRight`	Цензурирование справа	1	1
`Group`	Группирование на `k` интервалов	1	`k`-1

Примечание П.6. Если параметр преобразования не указан явно, то он инициализируется по умолчанию с флагом "неизвестный" и допускает возможность оценивания. В противном случае параметр инициализируется с флагом "известный" и оцениваться не может.

Примечание П.7. Преобразования цензурирования и группирования допустимы только над точечными выборками.

Следующий пример иллюстрирует описание раздела `[Samples]`.

Пример П.2.

```
[Samples]
s1=Read("100.dat")      // Выборка считывается в оперативную п
s2=Shift(Scale(s1,0.5),1) // Преобразование выборки
s3=Open("5000.dat")       // Выборка открывается и индексирует
s4=Group(s3, 5, 0.0, 1.0, 2.0, 3.0) // Выборка группируется на
                                    // интервалов
// Создается интервальная выборка подчиненная
// нормальному распределению
// объемом 100 наблюдений с абсолютной и
// относительной погрешностью
s5=Generate("test.dat", "Новая выборка", d1, 100, 0.01, 0.01)
```

П.1.2.2.3 Раздел `[Options]`

В разделе `[Options]` задаются основные параметры системы (таблица П.1.4)

П.1.2.2.4 Раздел `[Job]`

В разделе `[Job]` содержится задание на выполнение, допустимые команды которого перечислены в таблице П.1.5.

В следующем примере иллюстрируется стандартная статистическая обработка выборки.

Пример П.3.

```
[Job]
Estimate(s1, d5)    // Оценить параметры нормального распредел
Test(s1, d5)        // Проверить согласие
Anomalous(s1, d5)   // Выделить аномальные наблюдения
GraphIdent(s1, d5)  // График результатов идентификации
```

П.1.2.3. Ошибки инициализации

Если строка имеет неправильный формат, то она игнорируется, а сообщение об ошибке заносится файл `is.log`

П.1.3. Форматы входных данных

Выборка может находиться либо в двоичном (файл с расширением “`obs`”) либо в текстовом формате (файл с расширением “`dat`”).

П.1.3.1. Текстовый формат

Структура файла зависит от типа выборки:

П.1.3.1.1. Точечная выборка

Точечная выборка объемом `n` наблюдений имеет следующий формат:

```
<название_выборки>
0 n
<наблюдение_1>
<наблюдение_2>
...
<наблюдение_n>
```

Таблица П.1.4

Опции системы

Константа	Название	Значения	Содержание
Echo	Трассировка	ON | OFF	
IdentCrit	Критерий идентификации	GOODNESS_OF_FIT	Критерий согласия
		STATISTICS	Статистика
EstimateMethod	Метод оценивания	OMP	Макс. правд.
		KOLM	Мин. ст. Колм.
		MISES	Мин. ст. Мизеса
EstimateType	Тип оценки	POINT	Точечная
		INTERVAL	Интервальная
EstPointType	Тип точечной оценки по интервальным данным	CENTER	По центрам
		RULE	В среднем
		PESSIMIST	"кр. пессимиста"
		OPTIMIST	"кр. оптимиста"
GroupType	Тип группирования	GR_AOG	оптимальное
		GR_EPG	равновероятное
		GR_EFG	равночастотное
		GR_EG0	равномерное I
		GR_EG1	равномерное II
GroupNumb	Число интервалов	5..50	
Robust	Группирование перед оцениванием	ON | OFF	
Hypothesis	Гипотеза	SIMPLE	простая
		COMPLEX	сложная
W0..W5	Веса критериев	[0,1]	
Nik	Поправка Никулина	ON | OFF	
GenTestDistr	Моделирование распределения статистик	ON | OFF	
NumberSamples	Число выборок	10..2000	
MaxSampleSize	Максимальный объем выборки	200..2000	
SignLevel	Уровень значимости критерия	(0,1)	
SearchMeth	Метод поиска	SG1	Флетчера-Ривса
		SG2	Пшеничного
		MGS	МГС
GraphDriver	Тип монитора	EGA | VGA	
GraphMode	Видеорежим	EGAHI | EGALO	

Продолжение табл. П.1.4

Константа	Название	Значения	Содержание
`Palette`	Палитра	`BLACK_WHITE`	черно-белая
		`COLOR`	цветная
		`WHITE_BLACK`	бело-черная
`Nets`	Масштабная сетка	`ON | OFF`	
`Times`	Время	`ON | OFF`	
`NX`	Число делений по горизонтали	`1..8`	
`NY`	Число делений по вертикали	`1..10`	

Таблица П.1.5

Команды системы

Команда	Содержание
`Estimate(<выборка>,<распределение>)`	Оценить параметры
`Test(<выборка>, <распределение>)`	Проверить согласие
`Ident()`	Идентификация всех выборок и распределений
`IdentSample(<выборка>)`	Идентификация выборки
`IdentDistr(<распределение>)`	Идентификация распределения
`Anomalous(<выборка>, <распределение>)`	Выделить аномальные наблюдения
`SaveSample(<выборка>)`	Сохранить выборку
`SaveSampleAs(<выборка>, <имя файла>)`	Сохранить под другим именем
`GraphDistr([<распределение>])`	График распределения
`GraphSample([<выборка>])`	График выборочного распределения
`GraphIdent(<выборка>,<распределение>)`	График результатов идентификации
`Menu()`	Запуск меню
`Shell()`	Запуск оболочки

П.1.3.1.2. Интервальная выборка

Интервальная выборка объемом n наблюдений с абсолютной погрешностью a и относительной погрешностью r имеет формат:

```
<название_выборки>
1 n a r
<наблюдение_1>
<наблюдение_2>
...
<наблюдение_n>
```

П.1.3.1.3. Частично группированная выборка

Частично группированная выборка из n точечных наблюдений и k интервальных наблюдений имеет формат:

```
<название_выборки>
2 k n
<n_1> <n_2> ... <n_k>
<x_1> <x_2> ... <x_k-1>
<наблюдение_1>
<наблюдение_2>
...
<наблюдение_n>
```

где <n_i> — количество наблюдений в i-м интервале и <x_i> — i-я граничная точка.

П.1.3.1.4. Группированная выборка

Группированная выборка k интервальных наблюдений имеет формат:

```
<название_выборки>
3 k
<n_1> <n_2> ... <n_k>
<x_1> <x_2> ... <x_k-1>
```

Таблица П.1.6

Формат заголовка файла данных

Байт	Длина	Тип	Значение
0	2	int	версия
2	127	char[127]	заголовок
129	2	int	тип выборки
131	2	int	количество наблюдений
133	4	long int	объем выборки
137	8	double	левая граница выборки
145	8	double	правая граница выборки
151	8	double	абсолютная погрешность
159	8	double	относительная погрешность
167	20	Observation	<данные>

Таблица П.1.7

Структура наблюдения

Число байт	Тип	Содержание
2	int	Порядковый номер
8	double	Левая граница
8	double	Правая граница
2	int	Количество наблюдений или вес
Всего 20 байт		

где `<n_i>` — количество наблюдений в i-м интервале и `<x_i>` — i-я граничная точка.

П.1.3.2. Двоичный формат

В двоичном формате файл состоит из заголовка и данных (таблица П.1.6). Структура каждого наблюдения приведена в таблице П.1.7. Наблюдения в файле идут последовательно друг за другом, начиная со смещения в 167 байт от начала файла.