Министерство образования Российской Федерации
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

51
Л 442

Б.Ю. ЛЕМЕШКО, С.Н. ПОСТОВАЛОВ

КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ И ИССЛЕДОВАНИЯ СТАТИСТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ

НОВОСИБИРСК
2004

УДК 519.23(075.8)
Л442

Рецензенты: д–р техн. наук, проф. *В.В. Губарев*,
канд. техн. наук, доц. *Д.В. Лисицин*

Работа подготовлена на кафедре
прикладной математики

Лемешко Б.Ю., Постовалов С.Н.

Л 442 Компьютерные технологии анализа данных и исследования статистических закономерностей: Учеб. пособие. – Новосибирск: Изд-во НГТУ, 2004. – 120 с.

В учебном пособии изложена методика численного исследования статистических закономерностей. В основе методики лежит использование метода Монте-Карло для моделирования законов распределений функций от случайных величин.

Методика ориентирована на численное исследование теоретических рекомендаций математической статистики в нестандартных условиях, на исследование свойств различных оценок и статистик, на выявление статистических закономерностей, на исследование распределений статистик различных критериев, используемых для проверки статистических гипотез, на построение аналитических моделей для выявленных закономерностей, на исследование мощности критериев.

Применение методики связано с использованием современной компьютерной техники и развитого программного обеспечения. Методика опирается на использование в учебных и исследовательских целях развиваемой программной системы статистического анализа одномерных наблюдений «ISW 4.0».

Учебное пособие предназначено для студентов, обучающихся по направлению «Прикладная математика и информатика». Пособие будет полезно студентам и аспирантам других направлений, сталкивающимся со статистическим анализом наблюдений и необходимостью выявления вероятностных закономерностей.

УДК 519.23(075.8)

ВВЕДЕНИЕ

Практика применения методов статистического анализа в приложениях богата постановками задач, формулировки которых не укладываются в рамки классических предположений. Использование классических методов математической статистики в таких случаях часто оказывается некорректным. Кроме того, основные классические результаты имеют асимптотический характер, в то время как на практике обычно имеют дело с выборками конечных объемов.

Выявление фундаментальных статистических закономерностей в таких нестандартных условиях аналитическими методами, как правило, является сложной задачей для исследователя. Поэтому в последнее время все большее распространение получают методы компьютерного моделирования и анализа статистических закономерностей.

В настоящем учебном пособии изложена методика компьютерного моделирования фундаментальных статистических закономерностей. В основе этой методики лежит использование метода Монте-Карло для моделирования законов распределений некоторых функций от случайных величин.

Применение компьютерного моделирования возможно для решения следующих задач.

1. Проверка вероятностных закономерностей, вид которых был найден аналитическими методами.
2. Оценка скорости сходимости допредельных распределений к предельным. Определение величин погрешностей, возникающих при использовании предельных законов в случае конечных объемов выборок.
3. Моделирование законов распределений статистик критериев, используемых для проверки статистических гипотез (как в случае справедливости проверяемой гипотезы H_0, так и в случае справедливости альтернативы H_1).
4. Моделирование законов распределений оценок неизвестных параметров.
5. Исследование зависимости закона распределения статистики критерия или оценки параметра от объема выборки.
6. Определение мощности критериев (при заданной конкурирующей гипотезе H_1).
7. Подбор аналитических моделей, наиболее хорошо описывающих выборочные данные (в задачах 3-5).
8. Построение новых более мощных критериев проверки статистических гипотез.
9. Построение новых робастных методов оценивания параметров.

С примерами решения некоторых из этих задач можно ознакомиться в главах 2-6. Отметим, что задачи 8 и 9 находятся на стыке аналитических методов и методов Монте-Карло.

Применение методики связано с использованием современной компьютерной техники и программного обеспечения. Основные научные

результаты по математическим методам математической статистики, полученные авторами на протяжении нескольких десятков лет, реализованы в программной системе статистического анализа одномерных наблюдений «ISW 4.0». Описание программной системы приведено в главе 7. Авторы выражают признательность С.С. Помадину, Е.В. Чимитовой, А.В. Французову, В.М. Пономаренко, Е.П. Миркину, С.Б. Лемешко, внесшим вклад в развитие системы «ISW 4.0».

Учебное пособие предназначено для использования студентами при подготовке к занятиям по курсам «Компьютерные технологии анализа данных и исследования статистических закономерностей», «Методы статистического анализа», «Инструментальные и прикладные средства статистического анализа».

Программная система «ISW 4.0» и основные публикации авторов, дополняющие пособие, на которые приведены ссылки в тексте, доступны студентам в сети INTERNET на сайте факультета прикладной математики и информатики.

Глава 1. Метод монте-карло и компьютерное моделирование

Методы Монте-Карло – это общее название группы методов для решения различных задач с помощью *случайных последовательностей*. Исторически они возникли на базе выборочного метода в статистике и называются также методами *статистических испытаний*. Первоначально метод Монте-Карло использовался главным образом для решения задач математической физики, где традиционные численные методы оказались мало пригодными. Далее его влияние распространилось на теорию массового обслуживания, задачи теории игр и математической экономики, задачи теории передачи сообщений при наличии помех и ряд других [1-3].

1.1. Метод Монте-Карло

Идея метода заключается в следующем. Вместо того чтобы описывать исследуемый случайный процесс аналитически, составляется алгоритм, *имитирующий* этот процесс. В алгоритм включаются специальные процедуры для моделирования случайности. Конкретные вычисления в соответствии с алгоритмом складываются каждый раз по-иному, со своими результатами. Множество реализаций алгоритма используется как некий искусственно полученный статистический материал, обработав который методами *математической статистики*, можно получить любые характеристики: вероятности событий, математические ожидания, дисперсии случайных величин и т.п.

Как правило, программа составляется для осуществления одного случайного испытания. Затем это испытание повторяется N раз, причем каждый опыт не зависит от остальных, и результаты всех опытов усредняются.

Метод Монте-Карло позволяет моделировать любой процесс, на протекание которого влияют случайные факторы. Для многих математических задач, не связанных с какими-либо случайностями, можно искусственно придумать вероятностную модель, которая в некоторых случаях является более выгодной.

Пример 1.1. Вычисление числа π

Впишем в единичный квадрат окружность. Площадь квадрата будет равна 1, а площадь круга $\pi/4$. Тогда вероятность попадания случайно брошенной точки в круг будет равна отношению площади круга к площади квадрата, то есть $\pi/4$.

Смоделируем N равномерно распределенных на квадрате точек. Пусть M точек оказалось внутри круга, а $N-M$ – вне круга. Тогда отношение M/N приближенно оценивает вероятность попадания в круг, то есть в качестве статистической оценки числа π можно взять число $4M/N$.

Для применения методов Монте-Карло достаточно описания вероятностного процесса и не обязательна его формулировка в виде интегрального уравнения. Оценка погрешности метода чрезвычайно проста. Точность слабо зависит от размерности пространства.

Главный недостаток метода Монте-Карло заключается в том, что, являясь в основном численным методом, он не может заменить аналитические методы при расчете существенно новых явлений, где, прежде всего, требуется раскрытие качественных закономерностей.

Аналитические методы исследования позволяют существенно уменьшить погрешность метода Монте-Карло, могут поднять его до уровня получения качественных закономерностей. Синтез аналитических и статистических методов может также существенно уменьшить погрешность.

1.2. Методика компьютерного моделирования статистических закономерностей

Несмотря на то, что впервые метод Монте-Карло был использован для решения задач математической физики, он оказывается весьма плодотворным для выявления *фундаментальных законов теории вероятностей и математической статистики*.

Теория вероятностей – раздел математики, в котором по данным вероятностям одних случайных событий находят вероятности других событий, связанных каким-либо образом с первыми. Теория вероятностей изучает случайные величины и случайные процессы. Одна из основных задач теории вероятностей состоит в выяснении закономерностей, возникающих при взаимодействии большого числа случайных факторов. Так, например, *центральная предельная теорема* устанавливает, что при некоторых предположениях сумма одинаково распределенных случайных величин в пределе подчиняется *нормальному закону* распределения.

Математическая статистика, наоборот, изучает способы получения статистических закономерностей на основании *наблюдений* случайных величин. К таким задачам математической статистики относятся: оценивание параметров статистических моделей; проверка статистических гипотез; выявление статистической зависимости (корреляционный и регрессионный анализ); выявление значимых факторов статистической модели (дисперсионный и факторный анализ).

Методика компьютерного моделирования статистических закономерностей предусматривает статистическое моделирование эмпирических распределений статистик, вычисляемых по выборкам псевдослучайных одномерных и многомерных случайных величин, построение аналитических моделей, наилучшим образом сглаживающих (выравнивающих) полученные эмпирические распределения, уточнение построенных моделей по серии экспериментов.

Рассмотрим использование метода Монте-Карло на примере проверки статистической гипотезы о виде распределения.

Пример 1.2. Вычисление распределения статистики критерия согласия при проверке сложной гипотезы.

Пусть для проверки гипотезы H_0: $F(x) \in \{F(x, \theta), \ \theta \in \Theta\}$ по выборке X_n статистика критерия согласия имеет вид $S(X_n, F(x,\theta))$. Требуется найти распределение $G(S_n|H_0)$ статистики S при условии, что гипотеза H_0 является истинной.

Для этого следует в соответствии с законом $F(x,\theta)$ смоделировать N выборок того же объема n, что и выборка, для которой необходимо проверить гипотезу H_0. Далее, для каждой из N выборок необходимо вычислить оценки тех же параметров закона, а затем вычислить значение статистики S соответствующего критерия согласия. В результате будет сформирована выборка значений статистики $S_1, S_2, \ldots, S_N$ с условным законом распределения $G(S_n|H_0)$ для проверяемой гипотезы H_0. По полученной выборке при достаточно большом N можно построить достаточно гладкую эмпирическую функцию распределения $G_N(S_n|H_0)$, которой можно непосредственно воспользоваться для вывода о том, следует ли принимать гипотезу H_0. При необходимости по эмпирическому распределению $G_N(S_n|H_0)$ можно построить приближенную аналитическую модель, аппроксимирующую $G_N(S_n|H_0)$. И решение относительно проверяемой гипотезы уже принимать, опираясь на такую приближенную модель. В тех случаях, когда распределение статистики $G(S_n|H_0)$ зависит от значения параметра θ закона $F(x,\theta)$, необходимо повторить моделирование при разных значениях этого параметра, а затем, постараться выявить зависимость параметров аналитической модели, аппроксимирующей $G_N(S_n|H_0)$ от параметра θ.

Реализация описанной процедуры компьютерного моделирования распределения статистики в настоящий момент не содержит ни принципиальных, ни практических трудностей. Уровень вычислительной техники позволяет очень быстро получить результаты моделирования, а реализация алгоритма под силу инженеру, владеющему навыками программирования.

1.3. Точность и количество реализаций

Для обеспечения статистической устойчивости результатов моделирования соответствующие оценки вычисляются как средние значения по большому количеству реализаций. Выбор требуемого количества реализаций производится с помощью доверительных интервалов.

Пусть некоторый параметр x, имеющий математическое ожидание μ и дисперсию σ^2, оценивается по результатам моделирования значений x_i. В качестве оценки параметра выбирается выборочное среднее

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i .$$

В силу случайных причин величина $\overline{X}$ будет в общем случае отличаться от μ. Величину ε, такую, что

$$\left|\overline{X} - \mu\right| < \varepsilon , \qquad (1.1)$$

называют *точностью* оценки $\overline{X}$, а вероятность γ того, что неравенство (1.1) выполняется, ее *достоверностью*

$$P\{\left|\overline{X} - \mu\right| < \varepsilon\} = \gamma .$$

Воспользуемся данным принципом для определения точности результатов, получаемых методом статистического моделирования.

1.3.1. Вычисление вероятности появления некоторого случайного события

Пусть требуется вычислить вероятности p появления некоторого случайного события A. В каждой из N реализаций процесса количество наступлений события A является случайной величиной ξ, принимающей значение x_1=1 с вероятностью p, и значение x_2=0 с вероятностью 1- p.

Математическое ожидание и дисперсия случайной величины ξ равны:

$$M\xi = x_1 p + x_2(1-p) = p ,$$
$$D\xi = (x_1 - M\xi)^2 p + (x_2 - M\xi)^2(1-p) = (1-p)^2 p + (0-p)^2(1-p) = p(1-p) .$$

В качестве оценки для искомой вероятности p принимается частота m/N наступлений события A при N реализациях

$$\frac{m}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i ,$$

где x_i – количество наступлений события A в реализации с номером i.

В силу центральной предельной теоремы теории вероятностей частота m/N при достаточно больших N имеет распределение, близкое к нормальному:

$$\sqrt{N}\,\frac{m/N - M\xi}{\sqrt{D\xi}} = \sqrt{N}\,\frac{m/N - p}{\sqrt{p(1-p)}} \to \eta \in N(0,1) .$$

Отсюда

$$\mathrm{P}\left\{\left|\sqrt{N}\,\frac{m/N-p}{\sqrt{p(1-p)}}\right|<t_\gamma\right\}=\Phi(t_\gamma)-\Phi(-t_\gamma)=2\Phi(t_\gamma)-1=\gamma,$$

где $t_\gamma=\Phi^{-1}\left(\frac{\gamma+1}{2}\right)$ – квантиль стандартного нормального распределения,

$$\mathrm{P}\left\{|m/N-p|<t_\gamma\frac{\sqrt{p(1-p)}}{\sqrt{N}}\right\}=\gamma.$$

Таким образом, точность оценки p с достоверностью γ равна

$$\varepsilon=t_\gamma\frac{\sqrt{p(1-p)}}{\sqrt{N}}. \tag{1.2}$$

Отсюда количество реализаций, необходимое для достижения заданной точности ε равно

$$N=t_\gamma^2\frac{p(1-p)}{\varepsilon^2}. \tag{1.3}$$

На практике вероятность p обычно неизвестна. Поэтому для определения количества реализаций поступают следующим образом. Выбирают N_0=50-100, по результатам реализаций определяют m, а затем окончательно назначают N, предполагая, что $p\approx m/N_0$.

Пример 1.3.

Пусть $\gamma=0.99$. Тогда $t_\gamma=2.576$, и, учитывая, что $p(1-p)\le 0.25$, получаем

$$N\ge\frac{1.66}{\varepsilon^2},\quad \varepsilon\ge\frac{1.29}{\sqrt{N}}.$$

При $\varepsilon=0.1$ необходимо смоделировать выборку из $N=166$ наблюдений. При $\varepsilon=0.01$ необходимо смоделировать выборку из $N=16\,600$ наблюдений. С другой стороны, взяв $N=1000$ наблюдений, получаем точность $\varepsilon\approx 0.04$, а при $N=2000$ наблюдений – точность $\varepsilon\approx 0.03$.

Из рассмотренного примера видно, что для увеличения точности на один порядок, необходимо увеличить объем моделирования в 100 раз, поэтому обычно метод Монте-Карло применяется в ситуациях, когда требуемая точность не превышает одного - двух знаков после запятой.

1.3.2. Оценивание распределения случайной величины

Пусть требуется определить функцию распределения $F_\xi(x)$ некоторой случайной величины ξ. В качестве непараметрической оценки распределения можно использовать эмпирическую функцию распределения:

$$F_N(x)=\begin{cases}0, & x<x_{(1)}\\ m/N, & x_{(m)}\le x<x_{(m+1)},\ m=1,2,\ldots,N-1,\\ 1, & x\ge x_{(N)}\end{cases} \qquad (1.4)$$

где $x_{(1)}\le x_{(2)}\le\ldots\le x_{(N)}$ – вариационный ряд, построенный по выборке.

Для любого фиксированного значения x функция $F_n(x)$ представляет собой дискретную случайную величину, принимающую значения $0, 1/N, 2/N, \ldots, 1$. Математическое ожидание и дисперсия $F_N(x)$ равны:

$$M[F_N(x)]=F(x),\ D[F_N(x)]=\frac{F(x)(1-F(x))}{N}.$$

Тогда формулы (1.2) и (1.3) для вычисления значения $p=F(x)$ с помощью случайной величины $m/N=F_N(x)$ примут вид

$$\varepsilon=t_\gamma\frac{\sqrt{F(x)(1-F(x))}}{\sqrt{N}},\ N=t_\gamma^2\frac{F(x)(1-F(x))}{\varepsilon^2}. \qquad (1.5)$$

Отметим, что на практике при проверке статистических гипотез часто опираются на значения *процентных точек* (квантилей) распределения статистики критерия, на такие значения x, при которых $F(x)$=0.85, 0.90, 0.95, 0.99, 0.995, 0.999. Точность моделирования $F(x)$ при $N=1000$ и $N=2000$ приведена в таблице 1.1. Как и следовало ожидать, точность моделирования вероятности в этих точках достаточно высокая. Однако какова при этом будет точность в моделировании самих процентных точек?

Таблица 1.1

Точность моделирования $F(x)$

$F(x)$	$F(x)(1-F(x))$	$\varepsilon_N,\ N=1000$	$\varepsilon_N, N=2000$
0,850	0,128	0,029	0,021
0,900	0,090	0,024	0,017
0,950	0,048	0,018	0,013
0,990	0,010	0,008	0,006
0,995	0,005	0,006	0,004
0,999	0,001	0,003	0,002

В качестве оценки процентной точки уровня p будем брать порядковую статистику $x_{[pN]}$. Известно [4], что данная статистика в асимптотике имеет нормальное распределение с математическим ожиданием и дисперсией

$$M[x_{[pN]}] = F^{-1}(p),\ D[x_{[pN]}] = \frac{p(1-p)}{N \cdot f(F^{-1}(p))},$$

где $f(x)$ – функция плотности. Тогда формулы (1.2) и (1.3) для вычисления значения $F^{-1}(p)$ с помощью случайной величины $x_{[pN]}$ примут вид

$$\varepsilon = t_\gamma \frac{\sqrt{p(1-p)}}{\sqrt{N \cdot f(F^{-1}(p))}},\ N = t_\gamma^2 \frac{p(1-p)}{\varepsilon^2 f(F^{-1}(p))}. \qquad (1.6)$$

Естественно, что на величину погрешности в (1.6) существенный вклад оказывает вид функции распределения случайной величины ξ. Вычислим погрешности при определении процентных точек методом Монте-Карло в предположении, что случайная величина ξ имеет экспоненциальное распределение с параметром масштаба λ:

$$F(x,\lambda) = 1 - \exp(-x/\lambda),\ \ x > 0;$$

$$f(x,\lambda) = \frac{1}{\lambda}\exp(-x/\lambda),\ \ x > 0.$$

В этом случае

$$f(F^{-1}(p)) = \frac{1-p}{\lambda}.$$

Тогда получаем, что

$$\varepsilon = t_\gamma \frac{\sqrt{\lambda p}}{\sqrt{N}},\ \ N = t_\gamma^2 \frac{\lambda p}{\varepsilon^2}. \qquad (1.7)$$

Точность моделирования $F^{-1}(p)$, при $\lambda = 1$, $N = 1000$ и $N = 2000$ приведена в таблице 1.2.

Таблица 1.2

Точность моделирования $F^{-1}(p)$

p	ε_N, $N = 1000$	ε_N, $N = 2000$
0,850	0,075	0,053
0,900	0,077	0,055
0,950	0,079	0,056
0,990	0,081	0,057
0,995	0,081	0,057
0,999	0,081	0,058

При построении статистических закономерностей с помощью метода Монте-Карло желание экспериментатора достичь более точных результатов сталкивается с резким увеличением времени моделирования. Как было показано в примере 1.3, использование объемов моделирования в 2000 наблюдений дает погрешность не более 3-4%. Приемлема ли такая точность на практике? В значительной мере ответ на данный вопрос зависит от того, где будет применяться смоделированный закон.

Так, например, при проверке гипотезы «достигаемый уровень значимости» p будет не фиксированным числом, а 99%-доверительным интервалом $[p-\varepsilon, p+\varepsilon]$. Проблемы нет, если вероятность ошибки первого рода $\alpha > p+\varepsilon$ или $\alpha < p-\varepsilon$, так как погрешность моделирования в этом случае не окажет влияния на статистический вывод. Если же $\alpha \in [p-\varepsilon, p+\varepsilon]$, то фактическая вероятность ошибки первого рода может достигать величины $\alpha+\varepsilon$. В таблице 1.3 показано, на какую величину может отличаться фактическая вероятность ошибки первого рода от задаваемой.

Таблица 1.3

Точность моделирования $F(x)$

α	ε_N, $N=2000$	$\alpha+\varepsilon$
0,150	0,021	0,171
0,100	0,017	0,117
0,050	0,013	0,063
0,010	0,006	0,016
0,005	0,004	0,009
0,001	0,002	0,003

Контрольные вопросы и задачи

1. Предложите алгоритм для вычисления числа e методом Монте-Карло.
2. Сформулируйте достоинства и недостатки метода Монте-Карло.
3. В чем заключается методика компьютерного моделирования статистических закономерностей?
4. Оцените точность вычисления числа π в примере 1.1 при $N=2000$.
5. Какова точность моделирования распределения статистики критерия в примере 1.2?

Глава 2. Оценивание параметров

Пусть в эксперименте наблюдается непрерывная случайная величина ξ с функцией распределения $F(x,\theta)$ и плотностью функции распределения $f(x,\theta)$, где θ – вектор неизвестных параметров. По выборке $X_n = \{x_1, x_2, ..., x_n\}$ требуется оценить неизвестные параметры распределения.

2.1. Методы оценивания

2.1.1. Метод максимального правдоподобия

Оценки максимального правдоподобия (ОМП) вычисляются в результате максимизации по θ функции правдоподобия

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta) \tag{2.1}$$

или её логарифма

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i, \theta). \tag{2.2}$$

Чаще всего в случае скалярного параметра ОМП определяются как решение уравнения, а в случае векторного параметра – как решение системы уравнений правдоподобия вида

$$\frac{\partial \ln L(\theta)}{\partial \theta_l} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \theta)}{\partial \theta_l} = 0, \quad l = \overline{1,m}, \tag{2.3}$$

где m – размерность вектора параметров θ. В общем случае эта система оказывается нелинейной и, за редким исключением, решается только численно.

Оценки максимального правдоподобия при достаточно общих условиях являются асимптотически несмещенными и асимптотически эффективными, что объясняет их популярность на практике. Однако ОМП не всегда являются устойчивыми к наличию аномальных наблюдений в выборке.

2.1.2. Методы минимального расстояния

При вычислении *MD*-оценок (оценок минимального расстояния) по θ минимизируется некоторая мера близости, расстояние $\rho(F(x,\theta), F_n(x))$ между эмпирическим и теоретическим распределениями. *MD*-оценки находятся в процессе решения задачи

$$\hat{\theta} = \arg\min_{\theta} \rho(F(x,\theta), F_n(x))$$

В качестве меры близости можно использовать следующие статистики:

а) Статистика D_n Колмогорова:

$$D_n = \sup_x |F(x,\theta) - F_n(x)|. \qquad (2.4)$$

Расстояние Колмогорова можно вычислить следующим образом:

$$D_n = \max(D_n^+, D_n^-), \qquad (2.5)$$

$$D_n^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F(x_{(i)},\theta)\right\}, \qquad (2.6)$$

$$D_n^+ = \max_{1\le i\le n}\left\{F(x_{(i)},\theta) - \frac{i-1}{n}\right\}, \qquad (2.7)$$

где n – объем выборки, $x_{(1)}, x_{(2)}, ..., x_{(n)}$ – упорядоченные по возрастанию выборочные значения.

б) Статистика ω^2 Мизеса:

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

$$= \frac{2}{n}\sum_{i=1}^{n}\left\{g[F(x_{(i)})] - \frac{2i-1}{2n}f[F(x_{(i)})]\right\} + \int_0^1 (1-t)^2\psi(t)dt, \qquad (2.8)$$

где $M[\cdot]$ – оператор математического ожидания, $\psi(t)$ – заданная на отрезке $0 \le t \le 1$ неотрицательная функция, относительно которой предполагается, что $\psi(t)$, $t\,\psi(t)$, $t^2\psi(t)$ интегрируемы на отрезке $0 \le t \le 1$, и

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

При выборе $\psi(t) \equiv 1$ получается статистика ω^2:

$$\omega_n^2 = \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left\{F(x_{(i)},\theta) - \frac{2i-1}{2n}\right\}^2. \qquad (2.9)$$

При выборе $\psi(t) \equiv 1/t(1-t)$ расстояние задается статистикой Ω^2:

$$\Omega_n^2 = -1 - \frac{2}{n}\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_{(i)},\theta) + \left(1 - \frac{2i-1}{2n}\right)\ln(1 - F(x_{(i)},\theta))\right\}. \qquad (2.10)$$

2.1.3. Оценивание параметров по порядковым статистикам

Для нахождения оценок часто используются линейные комбинации порядковых статистик или выборочных квантилей. Такие оценки называются

L-оценками. L-оценки обладают двумя важными для практического применения качествами: простотой вычислений и хорошими свойствами робастности.

При построении L-оценок по выборочным квантилям $z_1 < z_2 < .... < z_k$ рассматриваемого закона оценки находят в виде:

$$\hat{\theta} = \sum_{i=1}^{k} \alpha_i z(p_i), \;\; z(p) = (x_{([np])} + x_{([np]+1)})/2,$$

где $x_{(i)}$ – i-я порядковая статистика, α_i и p_i – набор коэффициентов и вероятностей, которыми определяется конкретная оценка, n – объем выборки.

Оптимальные несмещенные оценки для параметров *сдвига и масштаба* получены в [5]. В [6] установлено, что оптимальные оценки параметров сдвига и масштаба являются асимптотически эффективными. В [7-10] предложено при построении таких оценок использовать асимптотически оптимальное группирование и получены соответствующие коэффициенты для L-оценок.

2.2. Экспериментальное исследование свойств оценок

Качество оценок, построенных по выборкам конечного объема ($n < \infty$), характеризуется следующими свойствами:

- *Несмещенность*. Оценка $\hat{\theta}$ называется *несмещенной*, если $\mathrm{M}[\hat{\theta}(X_n)] = \theta$.
- *Эффективность*. Несмещенная оценка $\hat{\theta}$ называется *эффективной*, если $\mathrm{D}[\hat{\theta}(X_n)] = J_n^{-1}(\theta)$, где $J_n(\theta)$ – информационная матрица Фишера. Эффективность имеет смысл только для регулярных моделей.

При $n \to \infty$ качество оценок определяется их асимптотическими свойствами:

- *Состоятельность*. Оценка $\hat{\theta}$ называется *состоятельной*, если $\hat{\theta}(X_n) \xrightarrow{p} \theta$, т.е. $\forall \varepsilon > 0 \;\; \lim\limits_{n\to\infty} P\{\left|\hat{\theta}(X_n) - \theta\right| > \varepsilon\} = 0$. Для проверки состоятельности можно использовать следующий критерий: *если оценка является асимптотически несмещенной и дисперсия стремится к нулю с ростом* n*, то оценка состоятельна.*
- *Асимптотическая нормальность.* Оценка $\hat{\theta}$ называется *асимптотически нормальной*, если $F_{\hat{\theta}}(t) \to \Phi(t)$, где $\Phi(t)$ – функция распределения нормального закона.

Свойства оценок обычно проверяются аналитическими методами. Однако в ситуациях, когда оценка не выражается в явном аналитическом виде, а получается в процессе решения сложной оптимизационной задачи, выявле-

ние свойств таких оценок, сравнение их со свойствами других оценок оказывается непростой задачей.

В то же время свойства оценок достаточно просто исследуются методом Монте-Карло по следующей схеме.

1. Моделируется N выборок по n наблюдений в каждой в соответствии с заданным законом распределения и фиксированными значениями вектора параметров θ.
2. По каждой выборке вычисляются оценки скалярных или векторных параметров. В результате получается выборка оценок $T_N = \{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_N\}$.
3. Исследуется (идентифицируется) распределение случайной величины $\hat{\theta}(X_n)$.
4. Пункты 1-3 повторяются при большем значении n.

По серии компьютерных экспериментов делаются выводы об асимптотических свойствах оценок.

Пример 2.1. Найдем распределение *MD*-оценки параметра масштаба нормального закона, получаемой минимизацией расстояния Колмогорова. Для этого в программной системе “ISW 4.0” откроем форму «Моделирование распределений оценок параметров» (см. рис. 2.1) и смоделируем распределения оценок при n=20, 100, 1000. Затем идентифицируем полученные законы (например, n=1000, см. рис. 2.2). В результате идентификации законов, наилучшим образом аппроксимирующих смоделированные выборки, получены следующие результаты (см. таблицу 2.1).

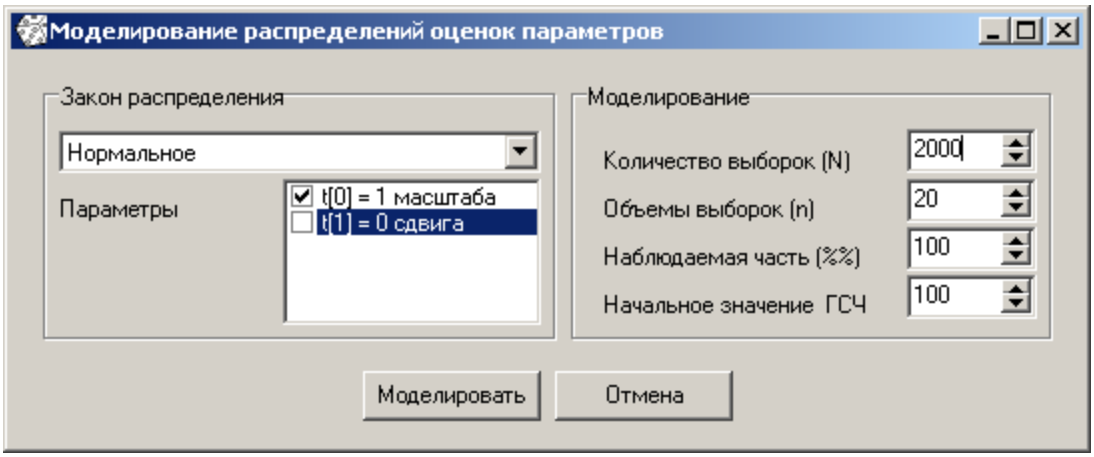

Рис. 2.1. Форма «Моделирование распределений оценок параметров»

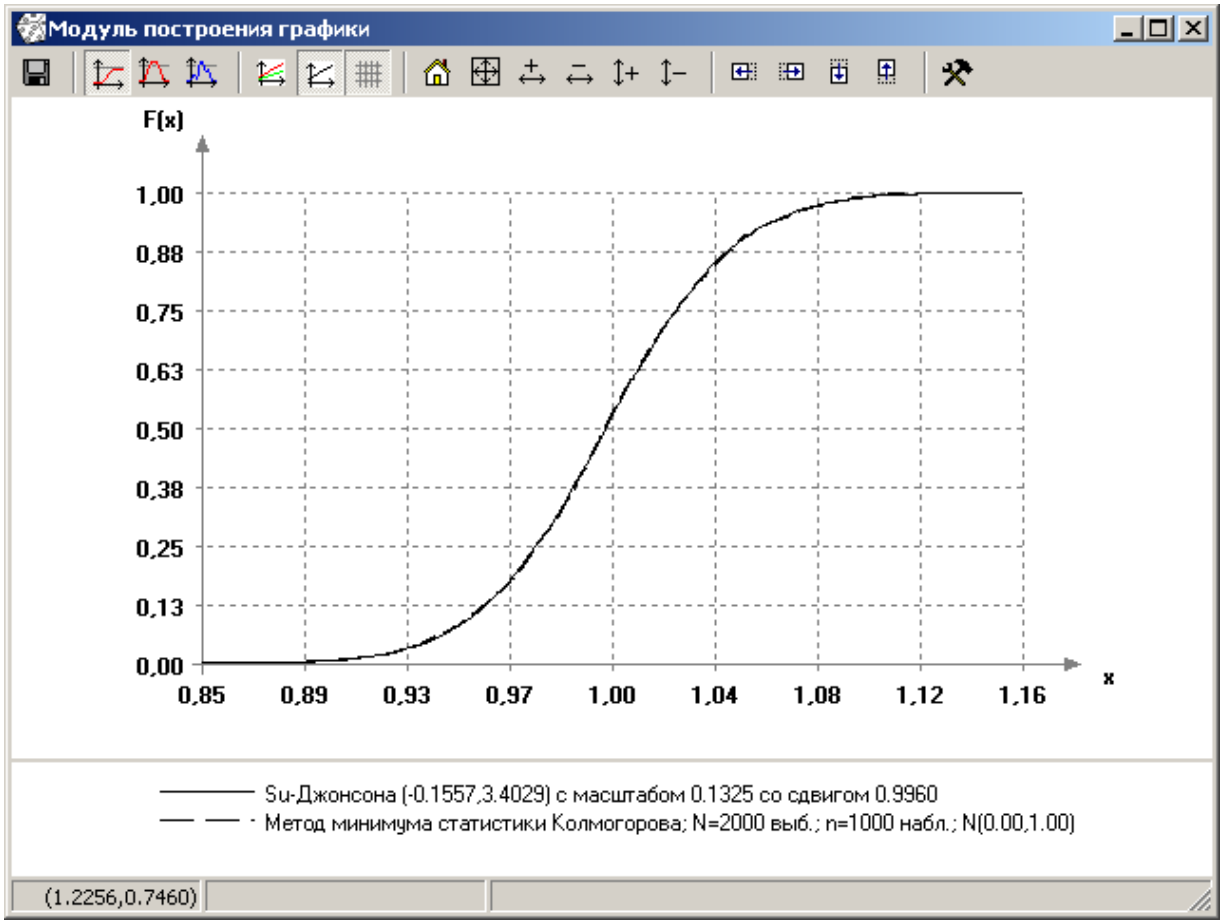

Рис. 2.2. Результаты идентификации закона распределения оценки параметра масштаба нормального закона по методу минимума статистики Колмогорова

Таблица 2.1

Результаты идентификации закона распределения *MD*-оценки параметра масштаба нормального закона, получаемой минимизацией расстояния Колмогорова

Объем выборки	Наилучшее распределение	Достигнутый уровень значимости
n=20	Su-Дж (-1.5443,1.8534) с масш. 0.4038 со сдв. 0.6728	0.18353
n=100	Su-Дж (-1.0870,2.9539) с масш. 0.3485 со сдв. 0.8810	0.12086
n=1000	Su-Дж (-0.1560,3.4024) с масш. 0.1324 со сдв. 0.9960	0.34862
n=5000	Su-Дж (0.0020,1.9071) с масш. 0.0313 со сдв. 0.9998	0.038419

Как мы видим, во всех случаях эмпирические распределения оценок наилучшим образом аппроксимируются распределениям Su-Джонсона, имеющим следующую функцию плотности распределения:

$$f(\theta_0,\theta_1,\theta_2,\theta_3)=\frac{\theta_1\theta_2}{(x-\theta_3)(\theta_2+\theta_3-x)}\exp\left\{-\frac{1}{2}\left[\theta_0-\theta_1\ln\frac{x-\theta_3}{\theta_2+\theta_3-x}\right]^2\right\},$$

где θ_0, θ_1 – параметры формы, θ_2 – параметр масштаба, θ_3 – параметр сдвига.

Проверим, насколько хорошо для аппроксимации подходит нормальное распределение (см. табл. 2.2). При малых объемах выборок (n=20, n=100) гипотеза о согласии отвергается, но уже при n=1000 гипотезу о согласии с нормальным законом можно принять с уровнем значимости $\alpha=0.06$. При этом математическое ожидание стремится к 1 (истинному значению параметра масштаба), а среднеквадратическое отклонение σ при больших объемах выборки хорошо аппроксимируется степенной зависимостью от n (см. рис. 2.3):

$$\sigma(n)=\frac{1.387}{n^{0.5087}}. \qquad (2.11)$$

Тогда дисперсия оценки приблизительно равна

$$D[\hat{\theta}_n]\approx\frac{1,9}{n}.$$

По неравенству Рао-Крамера нижняя граница дисперсии вычисляется по формуле

$$\mathrm{D}[\hat{\theta}(X_n)]=J_n^{-1}(\theta).$$

Для масштабного параметра нормального распределения $J_n(\theta)=\frac{2n}{\theta^2}$. При $\theta=1$ нижняя граница дисперсии равна $\frac{1}{2n}$. Следовательно, оценка не является эффективной.

Таблица 2.2

Проверка согласия распределения оценки с нормальным законом

Объем выборки	Нормальное распределение	Достигнутый уровень значимости
n=20	N(1.1110, 0.3946)	0
n=100	N(1.0201,0.1342)	0.00035976
n=1000	N(1.0023,0.0406)	0.06144
n=5000	N(0.9998,0.0184)	0.019672

Таким образом, по результатам статистического моделирования можно сделать следующие выводы о свойствах *MD*-оценки параметра масштаба нормального закона, получаемой минимизацией расстояния Колмогорова:

1. Оценка обладает малым смещением, в пределах статистической погрешности моделирования.
2. Распределение оценки наилучшим образом описывается распределением Su-Джонсона.
3. Распределение оценки является асимптотически нормальным с математическим ожиданием 0 и среднеквадратическим отклонением (2.11).
4. Оценка является состоятельной, но не является эффективной.

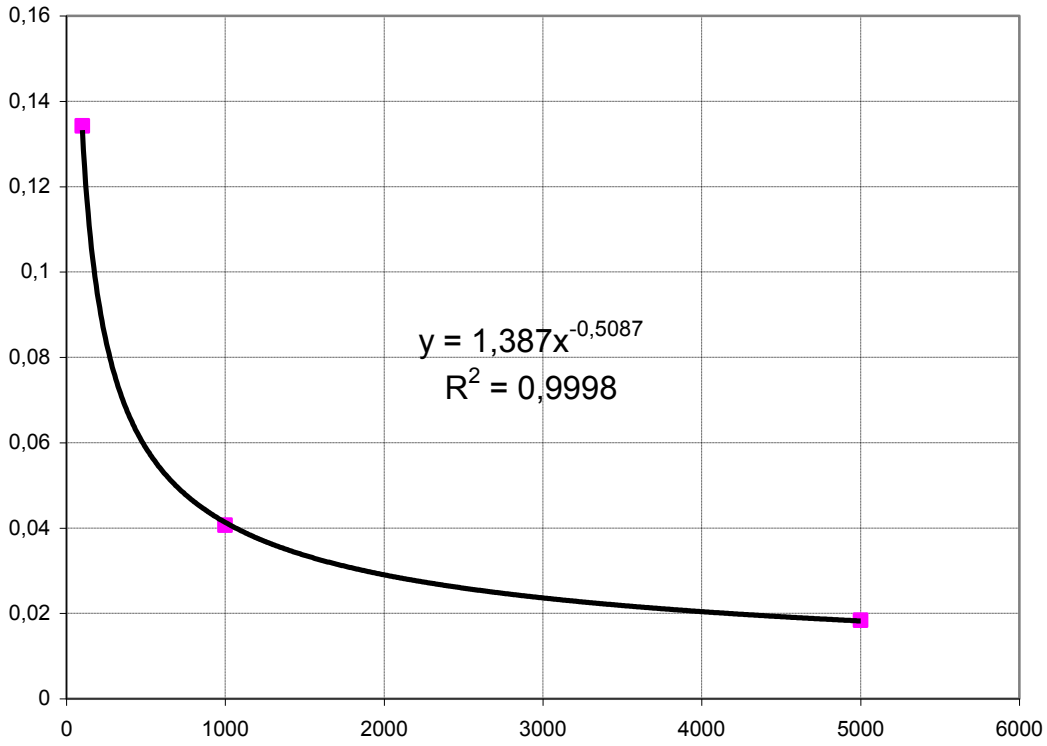

Рис. 2.3. Зависимость среднеквадратического отклонения $\sigma(n)$ от n

2.3. Робастность

Под *робастностью* в статистике понимают нечувствительность к малым отклонениям от предположений.

Для исследования робастности рассмотрим *модель с засорением выборки* [11]. Пусть в эксперименте наблюдается непрерывная случайная величина ξ с функцией распределения

$$F_\xi(x) = (1-\nu)F(x,\theta) + \nu F_1(x,\theta_1), \qquad (2.12)$$

где ν – доля засорения выборки аномальными (с точки зрения закона $F(x,\theta)$) наблюдениями, подчиняющимися закону $F_1(x,\theta_1)$. По выборке $X_n = \{x_1, x_2, ..., x_n\}$ требуется оценить неизвестные параметры распределения $F(x,\theta)$. Будем считать, что доля засорения ν относительно невелика.

Экспериментальное исследование робастности может проводиться методом Монте-Карло по следующей схеме.

1. Моделируется N выборок по n наблюдений в каждой в соответствии с (2.12).
2. По каждой выборке вычисляются оценки параметров закона $F(x,\theta)$. В результате получается выборка оценок $T_N = \{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_N\}$.
3. Исследуется (идентифицируется) распределение случайной величины $\hat{\theta}(X_n)$.
4. Пункты 1-3 повторяются при разных значениях ν.

По серии компьютерных экспериментов делаются выводы о влиянии доли засорения ν на распределение оценки.

Робастность метода оценивания можно увеличить с помощью процедуры группирования [12-14]. При группировании выборки теряется информация об индивидуальных наблюдениях, а фиксируется только количество наблюдений, попавших в интервалы группирования. В результате, небольшие отклонения от предполагаемого закона и аномальные выбросы не оказывают существенного влияния на оценки.

Для исследования робастности оценок аналитическими методами часто используется функция влияния Хампеля [15], которая определяется следующим образом

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

где δ_x – единичная масса в точке x, $T(F)$ – статистика.

Если функция влияния неограничена, то резко выделяющиеся наблюдения могут приводить к существенным изменениям оценок или статистик. Чувствительность к большой ошибке характеризуется величиной:

$$\gamma^* = \sup_x \left| IF(x,F,T) \right|. \qquad (2.13)$$

Функция влияния для асимптотически эффективных оценок, к которым относятся и ОМП, имеет вид:

$$IF(x,F,T)=J^{-1}(F(x,\theta))\frac{\partial \ln f(x,\theta)}{\partial\theta}, \qquad (2.14)$$

где $J^{-1}(F(x,\theta))=\int\limits_{-\infty}^{+\infty}\left(\frac{\partial \ln f(x,\theta)}{\partial\theta}\right)^2 f(x,\theta)dx$ – информационное количество Фишера.

2.4. Параметрическая процедура отбраковки аномальных наблюдений

Пусть гипотеза H_0 заключается в том, что все наблюдения подчинены закону с функцией распределения $F(x)$. Альтернативная гипотеза H_1 состоит в том, что в выборке присутствует наблюдение, подчиненное закону распределения, существенно сдвинутому вправо относительно закона $F(x,\theta)$.

Пусть d – критическое значение. Если $x_i < d$, $i=1,...,n$, то принимается гипотеза H_0, в противном случае принимается гипотеза H_1. Если вероятность ошибки первого рода α (вероятность отвергнуть гипотезу H_0, когда она верна) задана, можно из условия

$$\alpha = P\{\max_i x_i > d \mid H_0\} = 1-F^n(d)$$

найти критическое значение d:

$$d = F^{-1}(\sqrt[n]{1-\alpha}).$$

Аналогично, если гипотеза H_1 состоит в том, что в выборке присутствует наблюдение, подчиненное закону распределения, существенно сдвинутому влево относительно закона $F(x,\theta)$, то критическое значение определяется по формуле:

$$d = F^{-1}(1-\sqrt[n]{1-\alpha}).$$

Таким образом, процедура отбраковки аномального наблюдения состоит из двух этапов. Сначала одним из робастных методов находим оценки параметров распределения $F(x,\hat{\theta})$. Затем отбрасываем все наблюдения x: $x_i < \underline{d} \vee x_i > \overline{d}$. Пороговые значения определяются по формулам:

$$\underline{d} = F^{-1}(1-\sqrt[n]{1-\alpha},\hat{\theta}), \quad \overline{d} = F^{-1}(\sqrt[n]{1-\alpha},\hat{\theta}), \qquad (2.15)$$

где n – объем выборки, α – уровень значимости критерия.

2.5. Исследование оценок максимального правдоподобия по цензурированным данным

В исследованиях на надежность используется следующая схема эксперимента. Имеется партия из n изделий. В эксперименте фиксируются моменты выхода изделий из строя. Очевидно, что если проводить испытания бесконечно долго, то все изделия рано или поздно выйдут из строя. Однако реально время проведения эксперимента ограничено интервалом $[0,\alpha]$, где α – момент прекращения испытаний.

Для определения параметров надежности по незавершенным испытаниям характерна ситуация, когда к моменту прекращения испытаний большой партии изделий наблюдается выход из строя лишь части из них, обычно достаточно малой по сравнению с объемом всей партии. В этом случае анализ приходится проводить по выборке, сильно цензурированной справа. Особенно часто приходится сталкиваться с задачей обработки цензурированных выборок, когда наблюдению оказывается доступной только часть области определения случайной величины, а для выборочных значений, попавших левее или правее этой области, фиксируется лишь сам факт этого попадания.

Очевидно, что в такой неполной (цензурированной) выборке содержится меньше информации, чем в полной и это, естественно, отражается на точности оценивания параметров аппроксимирующего закона распределения. В этой связи оказывается интересным, насколько точно можно оценить параметры наблюдаемого закона в зависимости от объема всей выборки n (объема партии) и величины наблюдаемой ее части.

Наиболее универсальным методом по отношению к форме представления выборочных данных является метод максимального правдоподобия. В отличие от других метод позволяет находить ОМП параметров по негруппированным, частично группированным и группированным данным.

С точки зрения структуры данных цензурированные выборки являются частным случаем понятия частично группированной выборки, которую можно определить следующим образом [2].

Выборка называется *частично группированной*, если имеющаяся в нашем распоряжении информация связана с множеством непересекающихся интервалов, которые делят область определения случайной величины на k непересекающихся интервалов граничными точками

$$x_{(0)} < x_{(1)} < \ldots < x_{(k-1)} < x_{(k)},$$

где $x_{(0)}$ - нижняя грань области определения случайной величины X, $x_{(k)}$ - верхняя грань области определения случайной величины X, так, что каждый интервал принадлежит к одному из двух типов:

а) i-й интервал принадлежит к первому типу, если число n_i известно, но индивидуальные значения x_{ij}, $j = \overline{1,n_i}$ неизвестны;

б) i-й интервал принадлежит ко второму типу, если известно не только число n_i, но и все индивидуальные значения x_{ij}, $j=\overline{1,n_i}$.

В дальнейшем суммирование по интервалам первого и второго типов (аналогично умножение) обозначается соответственно, как $(\sum\limits_{(1)})$ и $(\sum\limits_{(2)})$.

ОМП неизвестного параметра по частично группированным наблюдениям называется такое значение параметра, при котором функция правдоподобия

$$L(\theta)=\prod_{(1)}P_i^{n_i}(\theta)\prod_{(2)}\prod_{j=1}^{n_i}f(x_{ij},\theta), \qquad (2.16)$$

где $f(x,\theta)$ – функция плотности случайной величины; $P_i(\theta)=\int\limits_{x_{(i-1)}}^{x_{(i)}}f(x,\theta)dx$ – вероятность попадания наблюдения в i-й интервал значений, достигает максимума на множестве возможных значений параметра. При вычислении ОМП максимизируют (2.16) или решают систему уравнений правдоподобия

$$\sum_{(1)}n_i\frac{\partial\ln P_i(\theta)}{\partial\theta_l}+\sum_{(2)}\sum_{j=1}^{n_i}\frac{\partial\ln f(x_{ij},\theta)}{\partial\theta_l}=0,\ l=\overline{1,m}, \qquad (2.17)$$

где m – размерность вектора параметров $\theta=(\theta_1,\theta_2,\ldots,\theta_m)^{\mathrm{T}}$. В случае группированных или частично группированных данных система (2.17), за редким исключением, решается только численно.

При выборке, цензурированной с двух сторон, являющейся частным случаем частично группированной выборки, область определения случайной величины разбита на 3 интервала граничными точками $x_{(1)}<x_{(2)}$ так, что значения левее $x_{(1)}$ и правее $x_{(2)}$ не наблюдаются. И система (2.17) принимает вид

$$n_1\frac{\partial\ln P_1(\theta)}{\partial\theta_l}+\sum_{j=1}^{n_2}\frac{\partial\ln f(x_{2j},\theta)}{\partial\theta_l}+n_3\frac{\partial\ln P_3(\theta)}{\partial\theta_l}=0,\ l=\overline{1,m}. \qquad (2.18)$$

Если оценивается скалярный параметр, то асимптотическая дисперсия его ОМП определяется соотношением

$$D(\hat{\theta})=n^{-1}J_c^{-1}(\hat{\theta}), \qquad (2.19)$$

где информационное количество Фишера определяется выражением

$$J_c(\theta)=\frac{1}{P_1(\theta)}\left[\frac{\partial P_1(\theta)}{\partial\theta}\right]^2+\int_{x_{(1)}}^{x_{(2)}}\left[\frac{\partial\ln f(x,\theta)}{\partial\theta}\right]^2 f(x,\theta)dx+\frac{1}{P_3(\theta)}\left[\frac{\partial P_3(\theta)}{\partial\theta}\right]^2. \qquad (2.20)$$

Если выборка цензурирована только справа, то в выражении (2.20) исчезает левое слагаемое, только слева - правое слагаемое. Это соотношение позволяет судить о потерях информации о параметре распределения в зависимости от степени цензурирования слева или справа.

Об эффективности оценивания параметров по цензурированной выборке по отношению к оцениванию по полной выборке можно судить по величине $J_c(\theta)/J(\theta)$, где $J(\theta)$ – количество информации Фишера в полной выборке [16,17].

2.6. Непараметрическое оценивание плотности распределения вероятностей

Пусть в эксперименте наблюдается непрерывная случайная величина ξ с плотностью распределения вероятностей $f(x)>0$. По выборке $X_n=\{x_1,x_2,...,x_n\}$ требуется найти оценку функции плотности $\tilde{f}_n(x)$.

В качестве оценки $\tilde{f}_n(x)$ в непараметрической статистике используется смесь из n ядерных функций [18]:

$$\tilde{f}_n(x)=\frac{1}{n\lambda_n}\sum_{i=1}^{n} g\left(\frac{x-x_i}{\lambda_n}\right), \qquad (2.21)$$

где n – объем выборки, $g(x)$ – "ядро", функция, удовлетворяющая условиям регулярности:

1) $g(x)=g(-x)$;
2) $0\le g(x)<\infty$;
3) $\int_{-\infty}^{+\infty} g(x)\,dx=1$;
4) $\int_{-\infty}^{+\infty} x^2 g(x)\,dx=1$;
5) $\int_{-\infty}^{+\infty} x^m g(x)\,dx<\infty$,

а параметр масштаба ("размытости") такой, что $\lim\limits_{n\to\infty}\lambda_n=0$, $\lim\limits_{n\to\infty} n\lambda_n=\infty$.

Асимптотические свойства оценки (2.21), такие как несмещенность, состоятельность, сходимость почти наверное к плотности $f(x)$, подробно исследованы в работах [19-21].

Среднеквадратическая ошибка аппроксимации (2.21) равна:

$$M\left[\int_{-\infty}^{+\infty}(f(x)-\tilde{f}_n(x))^2dx\right]\sim\frac{1}{n\lambda_n}\int_{-\infty}^{+\infty}(g(x))^2dx+\frac{\lambda_n^4}{4}\int_{-\infty}^{+\infty}(f''(x))^2dx. \quad (2.22)$$

Минимизируя (2.22) по λ_n и $g(x)$, можно найти оптимальное значение параметра размытости и функции ядра:

$$\lambda_n^*=\left[\frac{\int_{-\infty}^{+\infty}(g(x))^2dx}{n\int_{-\infty}^{+\infty}(f''(x))^2dx}\right]^{1/5}. \quad (2.23)$$

$$g^*(x)=\begin{cases}\frac{3}{4\sqrt{5}}-\frac{3}{20\sqrt{5}}x^2, \forall|x|<\sqrt{5},\\ 0, \forall|x|<\sqrt{5}.\end{cases} \quad (2.24)$$

Однако, в (2.23) присутствует вторая производная функции плотности, поэтому вычислить оптимальное значение параметра размытости возможно лишь при априорно известной функции плотности распределения.

Оценку параметра λ_n можно найти также по методу максимального правдоподобия:

$$\hat{\lambda}=\arg\min_{\lambda}\prod_{j=1}^{n}\frac{1}{(n-1)\lambda}\sum_{\substack{i=1\\ i\neq j}}^{n}g\left(\frac{x_j-x_i}{\lambda}\right). \quad (2.24)$$

Если случайная величина ξ определена на интервале $(0,\infty)$, то можно перейти к случайной величине $\eta=\ln\xi$, определенной на всем множестве R. Если случайная величина ξ определена на интервале (a,b), тогда можно перейти к случайной величине $\eta=\ln\frac{\xi-a}{b-\xi}$, определенной на всем множестве R.

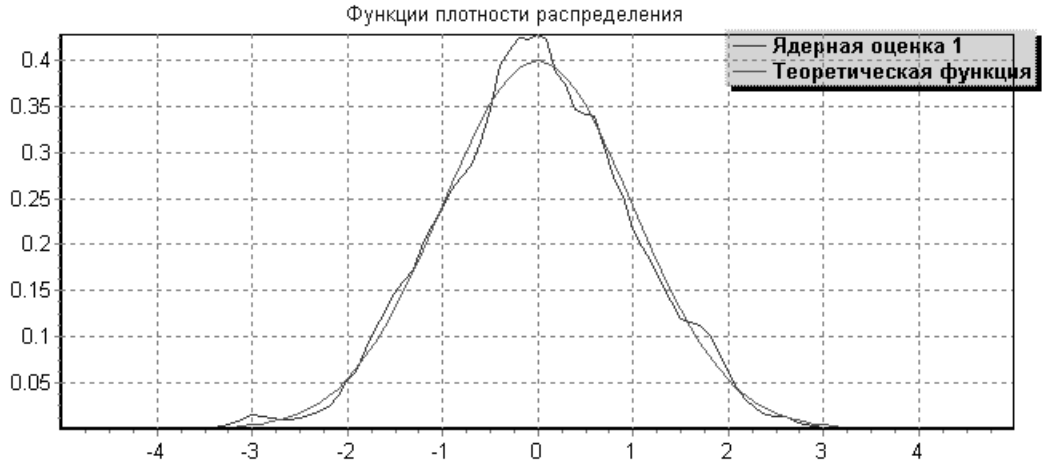

Рис.2.4. График непараметрической оценки плотности с параметром λ = 0.186193 (n=500) и плотности стандартного нормального закона

Пример 2.2. На рис. 2.4 приведена «ядерная» оценка функции плотности по выборке нормального распределения с масштабом 1.0 со сдвигом 0.0 при размерности выборки $n=500$.

Контрольные вопросы и задачи

1. Сравните методы вычисления оценок параметров.
2. Перечислите свойства оценок параметров. Каким образом можно исследовать свойства оценок?
3. Что такое робастность? Каким образом можно исследовать робастность оценок?
4. В чем заключается процедура параметрической отбраковки аномальных наблюдений? Могут ли в выборке оставаться аномальные наблюдения после проведения процедуры отбраковки?
5. Что такое цензурирование? Какие типы цензурирования Вы знаете? Какие проблемы возникают при оценивании параметров по цензурированным наблюдениям?
6. Какие функции $g(x)$ можно использовать в (2.21)? Можно ли в качестве ядерной функции использовать плотность распределения Коши? Нормального распределения? Распределения Вейбулла?
7. Можно ли использовать аппроксимацию (2.21) для законов распределения неотрицательных случайных величин?

Глава 3. Критерии согласия

Целью первичной обработки экспериментальных наблюдений обычно является выбор закона распределения, наиболее хорошо описывающего случайную величину, выборку которой мы наблюдали. Проверка того, насколько хорошо наблюдаемая выборка описывается теоретическим законом, осуществляется с использованием различных *критериев согласия*. Целью проверки гипотезы о согласии опытного распределения с теоретическим является стремление удостовериться в том, что данная модель теоретического закона не противоречит наблюдаемым данным, и использование ее не приведет к существенным ошибкам при вероятностных расчетах. Некорректное использование критериев согласия может приводить к необоснованному принятию (*чаще всего*) или необоснованному отклонению проверяемой гипотезы.

При проверке согласия различают ***простые*** и ***сложные*** гипотезы. **Простая** проверяемая гипотеза имеет вид H_0: $f(x,\theta) = f(x,\theta_0)$, где $f(\cdot)$ – функция плотности, а θ_0 – известный скалярный или векторный параметр теоретического распределения, с которым проверяется согласие. **Сложная** гипотеза имеет вид H_0: $f(x) \in \{f(x,\theta),\ \theta \in \Theta\}$, где Θ – пространство параметров и оценка скалярного или векторного параметра $\hat{\theta}$ вычисляется по той же самой выборке, по которой проверяется гипотеза о согласии.

Сама процедура проверки гипотезы осуществляется по следующей схеме. В соответствии с применяемым критерием согласия вычисляется значение S^* статистики S как некоторой функции от выборки и теоретического закона распределения с плотностью $f(x,\theta_0)$ (или $f(x,\hat{\theta})$ при сложной гипотезе). Для используемых на практике критериев асимптотические (предельные) распределения $g(s|H_0)$ соответствующих статистик при условии истинности гипотезы H_0 обычно известны. В общем случае *для простых и сложных гипотез эти распределения отличаются.* Далее в принятой практике статистического анализа обычно полученное значение статистики S^* сравнивают с критическим значением S_α при заданном уровне значимости α. Нулевую гипотезу отвергают, если $S^* > S_\alpha$ (см. рис. 3.1). Критическое значение S_α, определяемое в случае одномерной статистики из уравнения

$$\alpha = \int_{S_\alpha}^{\infty} g(s)ds = 1 - G_S(S_\alpha),\ S_\alpha = G_S^{-1}(1-\alpha),$$

обычно берётся из соответствующей статистической таблицы или вычисляется.

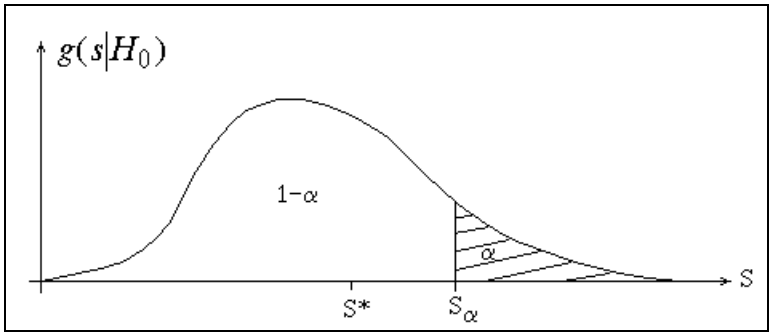

Рис. 3.1. Распределение статистики при истинной гипотезе H_0

Естественно, что больше информации о степени согласия можно почерпнуть из "достигаемого уровня значимости": вероятности возможного превышения полученного значения статистики при истинности нулевой гипотезы $P\{S > S^*\} = \int_{S^*}^{\infty} g(s|H_0)ds$. Именно эта вероятность позволяет судить о том, насколько хорошо выборка согласуется с теоретическим распределением (рис. 3.2). Гипотеза о согласии не отвергается, если $P\{S > S^*\} > \alpha$.

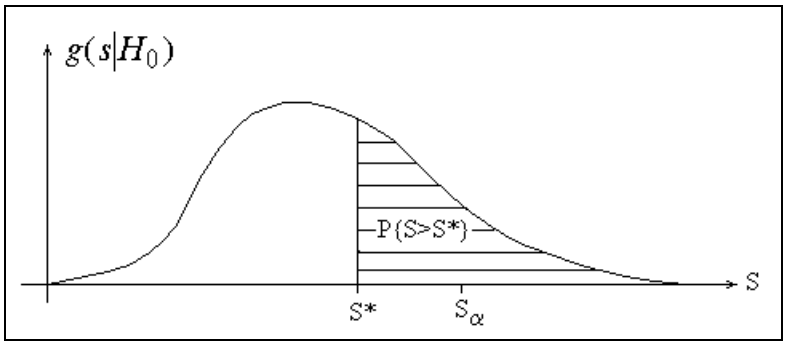

Рис. 3.2. Распределение статистики при истинной гипотезе H_0

Задачи оценивания параметров и проверки гипотез опираются на выборки независимых случайных величин. Случайность самой выборки предопределяет, что возможны и ошибки в результатах статистических выводов. С результатами проверки гипотез связывают ошибки 2 видов: ошибка 1-го рода состоит в том, что отклоняется гипотеза H_0, когда она верна; ошибка 2-го рода – в том, что принимается гипотеза H_0, в то время как справедлива альтернатива (конкурирующая гипотеза) H_1. Величина α задает вероятность ошибки 1-го рода. Обычно в критериях согласия не рассматривают конкретную альтернативу, и тогда конкурирующая гипотеза имеет вид H_1: $f(x,\theta) \neq f(x,\theta_0)$. Если гипотеза H_1 определена и имеет, например, вид H_1: $f(x,\theta) = f_1(x,\theta_1)$, то задание α определяет для используемого критерия проверки гипотез и вероятность ошибки 2-го рода β. На рис. 3.3

$g(s|H_0)$ отображает плотность распределения статистики S при истинности гипотезы H_0, а $g(s|H_1)$ – плотность распределения при справедливости гипотезы H_1.

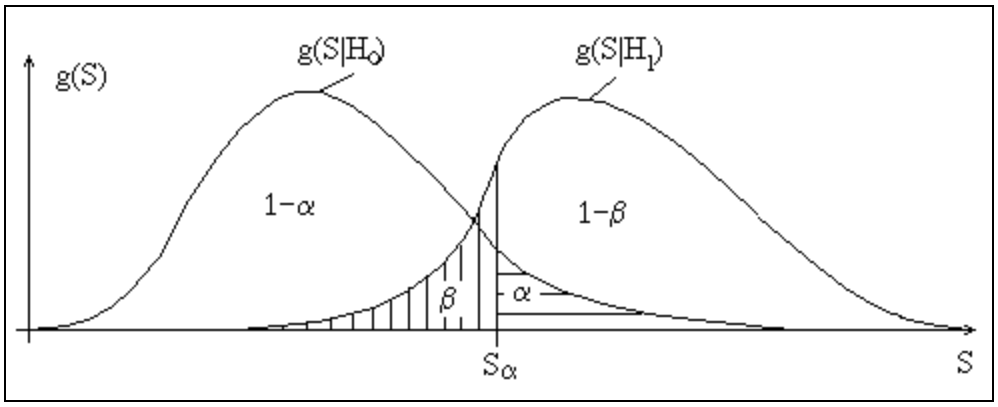

Рис 3.3. Распределения статистик при справедливости гипотез H_0 и H_1

Мощность критерия представляет собой величину $1-\beta$. Очевидно, что *чем выше мощность* используемого критерия при заданном значении α, *тем лучше он различает гипотезы* H_0 и H_1. Особенно важно, чтобы используемый критерий хорошо различал близкие альтернативы. Графически требование максимальной мощности критерия означает, что на рис 3.3 плотности $g(s|H_0)$ и $g(s|H_1)$ должны быть максимально "раздвинуты". Построение *наиболее мощного критерия* при проверке простой гипотезы H_0 против простой альтернативы H_1 основывается на лемме Неймана-Пирсона [4]. Однако в случае сложной гипотезы *равномерно наиболее мощного критерия* (т.е. наиболее мощного при любой альтернативе) в общем случае не существует.

Критерий называется *состоятельным*, если при заданном значении α мощность $(1-\beta_n)\to 1$, при $n\to\infty$. Критерий называется *несмещенным*, если $1-\beta>\alpha$. Одной из важных исследовательских задач является нахождение такого объема выборки n, при котором достигается требуемая мощность критерия $1-\beta_n$.

3.1. Непараметрические критерии согласия

3.1.1. Критерий Колмогорова

В случае простых гипотез предельные распределения статистик рассматриваемых критериев согласия Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса известны и не зависят от вида наблюдаемого закона распределения и, в частности, от его параметров. Говорят, что эти критерии являются "свободными

от распределения". Это достоинство предопределило широкое использование данных критериев в приложениях.

Распределение статистики

$$D_n = \sup_{|x|<\infty} |F_n(x) - F(x,\theta)|, \tag{3.1}$$

где $F_n(x)$ – эмпирическая функция распределения, $F(x,\theta)$ – теоретическая функция распределения, n – объём выборки, было получено Колмогоровым в [22]. При $n \to \infty$ распределение статистики $\sqrt{n}D_n$ сходится равномерно к распределению Колмогорова

$$K(S) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2} . \tag{3.2}$$

Наиболее часто в критерии Колмогорова (Колмогорова-Смирнова) используется статистика вида [23]

$$S_K = \frac{6nD_n + 1}{6\sqrt{n}}, \tag{3.3}$$

где

$$D_n = \max(D_n^+, D_n^-) , \tag{3.4}$$

$$D_n^+ = \max_{1\le i\le n}\left\{\frac{i}{n} - F(x_{(i)},\theta)\right\}, \tag{3.5}$$

$$D_n^- = \max_{1\le i\le n}\left\{F(x_{(i)},\theta) - \frac{i-1}{n}\right\}, \tag{3.6}$$

n – объем выборки, $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ – упорядоченные по возрастанию выборочные значения, $F(x,\theta)$ – функция закона распределения, согласие с которым проверяется. Распределение величины S_K при простой гипотезе в пределе подчиняется закону Колмогорова $K(S)$.

Если для вычисленного по выборке значения статистики S_K^* выполняется неравенство

$$P\{S > S_K^*\} = 1 - K(S_K^*) > \alpha,$$

то нет оснований для отклонения гипотезы H_0.

3.1.2. Критерий Смирнова

В критерии Смирнова используется статистика

$$D_n^+ = \sup_{|x|<\infty}(F_n(x) - F(x,\theta)) \tag{3.7}$$

или статистика

$$D_n^- = -\inf_{|x|<\infty}(F_n(x) - F(x,\theta)), \tag{3.8}$$

значения которых вычисляются по эквивалентным соотношениям (3.5),(3.6).

Реально в критерии обычно используется статистика [23]

$$S_m = \frac{(6nD_n^+ + 1)^2}{9n}, \tag{3.9}$$

которая при простой гипотезе в пределе подчиняется распределению χ^2 с числом степеней свободы, равным 2.

Гипотеза H_0 не отвергается, если для вычисленного по выборке значения статистики S_m^*

$$P\{S_m > S_m^*\} = \int_{S_m^*}^{\infty} \frac{1}{2} e^{-x/2} dx = e^{-S_m^*/2} > \alpha .$$

3.1.3. Критерии ω^2

В критериях типа ω^2 расстояние между гипотетическим и истинным распределениями рассматривается в квадратичной метрике.

Проверяемая гипотеза H_0 имеет вид [23]

$$H_0 : \int_{-\infty}^{\infty} \{M[F_n(x)] - F(x)\}^2 \psi(F(x)) dF(x) = 0 \tag{3.10}$$

при альтернативной гипотезе

$$H_1 : \int_{-\infty}^{\infty} \{M[F_n(x)] - F(x)\}^2 \psi(F(x)) dF(x) > 0, \tag{3.11}$$

где $M[\cdot]$ – оператор математического ожидания, $\psi(t)$ – заданная на отрезке $0 \le t \le 1$ неотрицательная функция, относительно которой предполагается, что $\psi(t)$, $t\,\psi(t)$, $t^2\psi(t)$ интегрируемы на отрезке $0 \le t \le 1$ [24]. Статистика критерия выражается соотношением [23]

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

$$= \frac{2}{n} \sum_{i=1}^{n} \left\{ g[F(x_i)] - \frac{2i-1}{2n} f[F(x_i)] \right\} + \int_0^1 (1-t)^2 \psi(t) dt, \tag{3.12}$$

где

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

При выборе $\psi(t) \equiv 1$ для критерия ω^2 Мизеса получают статистику вида (статистику **Крамера-Мизеса-Смирнова**)

$$S_\omega = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left\{F(x_{(i)},\theta) - \frac{2i-1}{2n}\right\}^2, \qquad (3.13)$$

которая при простой гипотезе подчиняется распределению $a1(S)$, имеющему вид [23]

$$a1(s) = \frac{1}{\sqrt{2s}}\sum_{j=0}^{\infty}\frac{\Gamma(j+1/2)\sqrt{4j+1}}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2}{16s}\right\}\times$$

$$\times\left\{I_{-\frac{1}{4}}\left[\frac{(4j+1)^2}{16s}\right] - I_{\frac{1}{4}}\left[\frac{(4j+1)^2}{16s}\right]\right\}, \qquad (3.14)$$

где $I_{-\frac{1}{4}}(\cdot),\ I_{\frac{1}{4}}(\cdot)$ – модифицированные функции Бесселя,

$$I_\nu(z) = \sum_{k=0}^{\infty}\frac{\left(\frac{z}{2}\right)^{\nu+2k}}{\Gamma(k+1)\Gamma(k+\nu+1)},\ |z|<\infty,\ |\arg z|<\pi. \qquad (3.15)$$

При выборе $\psi(t) \equiv 1/t(1-t)$ для критерия Ω^2 Мизеса статистика приобретает вид (статистика **Андерсона-Дарлинга**)

$$S_\Omega = n\Omega_n^2 = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_{(i)},\theta) + \left(1-\frac{2i-1}{2n}\right)\ln(1-F(x_{(i)},\theta))\right\}. \qquad (3.16)$$

В пределе эта статистика подчиняется распределению $a2(S)$, имеющему вид [23]

$$a2(s) = \frac{\sqrt{2\pi}}{s}\sum_{j=0}^{\infty}(-1)^j\frac{\Gamma(j+1/2)(4j+1)}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2\pi^2}{8s}\right\}\times$$

$$\times\int_0^{\infty}\exp\left\{\frac{s}{8(y^2+1)} - \frac{(4j+1)^2\pi^2y^2}{8s}\right\}dy. \qquad (3.17)$$

Гипотезы о согласии не отвергаются, если выполняются неравенства

$$P\{S_\omega > S_\omega^*\} = 1 - a1(S_\omega^*) > \alpha \quad \text{и} \quad P\{S_\Omega > S_\Omega^*\} = 1 - a2(S_\Omega^*) > \alpha.$$

Распределения статистик непараметрических критериев согласия при проверке сложных гипотез зависят от характера этой сложной гипотезы. На закон распределения статистики $G(S|H_0)$ влияет целый ряд факторов, определяющих "сложность" гипотезы [25-32]:

- вид наблюдаемого закона распределения $F(x,\theta)$, соответствующего истинной гипотезе H_0;

- тип оцениваемого параметра и количество оцениваемых параметров;
- в некоторых ситуациях конкретное значение параметра (например, в случае гамма-распределения);
- используемый метод оценивания параметров.

При малых объемах выборки n распределение $G(S_n|H_0)$ зависит от n. Однако существенная зависимость распределения статистики от n наблюдается только при небольших объемах выборки. Уже при $n \geq 15 \div 20$ распределение $G(S_n|H_0)$ достаточно близко к предельному $G(S|H_0)$ и зависимостью от n можно пренебречь.

В случае задания конкретной альтернативы (конкурирующей гипотезы H_1, которой соответствует распределение $F_1(x,\theta)$), функция распределения статистики $G(S|H_1)$ также зависит от всех перечисленных факторов. Но в отличие от $G(S|H_0)$ распределение статистики $G(S|H_1)$ при справедливой гипотезе H_1 очень сильно зависит от объема выборки n. Именно благодаря этому с ростом n повышается способность критериев различать гипотезы, возрастает мощность критериев.

3.2. Критерии типа χ^2

3.2.1. Критерий типа χ^2 Пирсона

Пусть $\xi_1, \xi_2, \ldots, \xi_N$ – выборка значений наблюдаемой случайной величины объемом N. Процедура проверки гипотез с применением критериев типа χ^2 предусматривает группирование наблюдений. Область определения случайной величины разбивается на k непересекающихся интервалов граничными точками

$$x_0 < x_1 < \ldots < x_{k-1} < x_k,$$

где x_0 – нижняя грань области определения случайной величины, x_k – верхняя грань. В соответствии с заданным разбиением подсчитывают количества выборочных значений n_i, попавших в i-й интервал, и вычисляют вероятности попадания в интервал $P_i(\theta) = \int_{x_{i-1}}^{x_i} f(x,\theta)dx$, соответствующие теоретическому закону с функцией плотности $f(x,\theta)$. При проверке **простой** гипотезы известны как вид функции плотности, так и все параметры закона (известен скалярный или векторный параметр θ). При этом $\sum_{i=1}^{k} n_i = N$,

$\sum_{i=1}^{k} P_i(\theta) = 1$. В основе статистик, используемых в критериях согласия типа χ^2, лежит измерение отклонений n_i / N от $P_i(\theta)$.

К критериям такого рода, в частности, относятся критерий χ^2 Пирсона, критерий отношения правдоподобия [33] и критерии типа χ^2 [34-36].

Статистика *критерия согласия* χ^2 *Пирсона* вычисляется в соответствии с соотношением

$$S_{\chi^2} = N \sum_{i=1}^{k} \frac{(n_i / N - P_i(\theta))^2}{P_i(\theta)}. \qquad (3.18)$$

В случае проверки **простой** гипотезы в пределе при $N \to \infty$ она подчиняется χ_r^2-распределению с $r = k - 1$ степенями свободы, если верна нулевая гипотеза. Плотность χ_r^2-распределения описывается выражением

$$g(s) = \frac{1}{2^{r/2}\Gamma(r/2)} s^{r/2-1} e^{-S/2}. \qquad (3.19)$$

Если верна конкурирующая гипотеза H_1 и выборка соответствует распределению с плотностью $f_1(x, \theta_1)$ с параметром θ_1, то эта же статистика в пределе подчиняется нецентральному χ_r^2-распределению с тем же числом степеней свободы и параметром нецентральности

$$\lambda_N = N \sum_{i=1}^{k} \frac{(P_i^1(\theta_1) - P_i(\theta))^2}{P_i(\theta)}, \qquad (3.20)$$

где $P_i^1(\theta_1)$ вероятность попадания в интервал при справедливой гипотезе H_1. Плотность нецентрального χ_r^2-распределения имеет вид [31]

$$g(s, \lambda) = \frac{e^{-(s+\lambda)/2} s^{(r-2)/2}}{2^{r/2}\Gamma[(r-1)/2)] \cdot \Gamma(1/2)} \sum_{k=0}^{\infty} \frac{\lambda^k s^k}{(2k)!} B\left\{\frac{1}{2}(r-1), \frac{1}{2} + k\right\}, \qquad (3.21)$$

где $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta) / \Gamma(\alpha + \beta)$ – бета-функция.

При заданном уровне значимости α нулевая гипотеза о согласии не должна отвергаться, если

$$P\{S_{\chi^2} > S_{\chi^2}^*\} = \frac{1}{2^{r/2}\Gamma(r/2)} \int_{S_{\chi^2}^*}^{\infty} s^{r/2-1} e^{-s/2} ds > \alpha, \qquad (3.22)$$

где $S_{\chi^2}^*$ – вычисленное в соответствии с (3.18) значение статистики.

В критерии *отношения правдоподобия* используется статистика [31]

$$S_{\text{оп}} = -2\ln l = -2\sum_{i=1}^{k} n_i \ln\left(\frac{P_i(\theta)}{n_i / N}\right), \tag{3.23}$$

которая при верной нулевой гипотезе также асимптотически распределена как χ_r^2 с $r = k - 1$ степенями свободы. Если верна конкурирующая гипотеза H_1 и выборка соответствует распределению с плотностью $f_1(x, \theta_1)$ с параметром θ_1, мерой близости сравниваемых законов является величина

$$-2\ln l = 2N\sum_{i=1}^{k} P_i^1(\theta_1)\ln\left(\frac{P_i^1(\theta_1)}{P_i(\theta)}\right). \tag{3.24}$$

При справедливости H_0 в случае проверки **сложной** гипотезы и при условии, что оценки параметров находятся в результате минимизации статистики S_{χ^2} по этой же самой выборке, статистика S_{χ^2} асимптотически распределена как χ_r^2 с числом степеней свободы $r = k - m - 1$, где m – количество оцененных параметров. Статистика S_{χ^2} имеет это же распределение, если в качестве метода оценивания выбирается метод максимального правдоподобия и оценки вычисляются по сгруппированным данным в результате максимизации по θ функции правдоподобия

$$L(\theta) = \prod_{i=1}^{k} P_i^{n_i}(\theta), \tag{3.25}$$

где $P_i(\theta) = \int_{x_{i-1}}^{x_i} f(x,\theta)dx$ – вероятность попадания наблюдения в i-й интервал значений, зависящая от θ.

При вычислении ОМП по негруппированным данным эта же статистика распределена в пределе как сумма независимых слагаемых $\chi_{k-m-1}^2 + \sum_{j=1}^{m} \lambda_j \xi_j^2$, где $\xi_1, \ldots, \xi_m$ – стандартные нормальные случайные величины, независимые между собой и с χ_{k-m-1}^2, а $\lambda_1, \ldots, \lambda_m$ – некоторые числа между 0 и 1 [33,37,38], представляющие собой корни уравнения

$$\left|(1-\lambda)\mathbf{J}(\theta) - \mathbf{J}_\Gamma(\theta)\right| = 0.$$

В данном уравнении $\mathbf{J}(\theta)$ – информационная матрица Фишера по негруппированным наблюдениям с элементами, определяемыми соотношением

$$J(\theta_l,\theta_j)=\int_{-\infty}^{+\infty}\left(\frac{\partial f(x,\theta)}{\partial\theta_l}\frac{\partial f(x,\theta)}{\partial\theta_j}\right)f(x,\theta)dx, \qquad (3.26)$$

а $\mathbf{J}_\Gamma(\theta)$ – информационная матрица по группированным наблюдениям

$$\mathbf{J}_\Gamma(\theta)=\sum_{i=1}^{k}\frac{\nabla P_i(\theta)\nabla^{\mathrm{T}}P_i(\theta)}{P_i(\theta)}. \qquad (3.27)$$

Функция распределения статистики лежит между χ^2_{k-1}- и χ^2_{k-m-1}-распределениями. В этом случае, принимая нулевую гипотезу, мы должны удостовериться, что статистика S_{χ^2} не превышает критических значений $\chi^2_{k-m-1,\alpha}$ и $\chi^2_{k-1,\alpha}$, где α – задаваемый уровень значимости. И если $\chi^2_{k-m-1,\alpha}<S^*_{\chi^2}<\chi^2_{k-1,\alpha}$, то, принимая или отклоняя гипотезу о согласии, мы можем с одинаковым риском совершить ошибку.

Вышесказанное относится и к критерию отношения правдоподобия.

С зависимостью мощности критериев типа χ^2 от способа группирования данных подробно можно ознакомиться в [39-41], с влиянием способа группирования на распределения этих статистик при использовании ОМП по негруппированным данным – в [42], о влиянии числа интервалов на мощность критериев типа χ^2 при различных способах группирования говорится в работах [43-45], общие рекомендации по правилам применения критериев типа χ^2 даются в [46-47].

3.2.2. Критерий типа χ^2 Никулина

В работах [43-45] предложено видоизменение стандартной статистики S_{χ^2}, при котором предельное распределение есть обычное χ^2_{k-1}-распределение (количество степеней свободы не зависит от числа оцениваемых параметров). Неизвестные параметры распределения $F(x,\theta)$ в этом случае должны оцениваться по негруппированным данным методом максимального правдоподобия. При этом вектор $\mathbf{P}=(P_1,\ldots,P_k)^{\mathrm{T}}$ предполагается заданным, и граничные точки интервалов определяются соотношениями $x_i(\theta)=F^{-1}(P_1+\ldots+P_i)$, $i=\overline{1,(k-1)}$. Предложенная статистика имеет вид [44]

$$Y^2_N(\theta)=S_{\chi^2}+N^{-1}a^{\mathrm{T}}(\theta)\mathbf{\Lambda}(\theta)a(\theta), \qquad (3.28)$$

где S_{χ^2} определяется по (1), матрица $\mathbf{\Lambda}(\theta) = \left\| J(\theta_l, \theta_j) - \sum_{i=1}^{k} \frac{w_{\theta_l i} w_{\theta_j i}}{p_i} \right\|^{-1}$, элементы и размерность которой определяются оцениваемыми компонентами вектора параметров θ, $J(\theta_l, \theta_j)$ – элементы информационной матрицы $\mathbf{J}(\theta)$ по негруппированным данным, $a_{\theta_l} = w_{\theta_l 1} n_1 / P_1 + \ldots + w_{\theta_l k} n_k / P_k$ – элементы вектора $a(\theta)$, и

$$w_{\theta_l i} = -f[x_i(\theta), \theta] \frac{\partial x_i(\theta)}{\partial \theta_l} + f[x_{i-1}(\theta), \theta] \frac{\partial x_{i-1}(\theta)}{\partial \theta_l}. \qquad (3.29)$$

Для распределений, определяемых только параметрами сдвига и масштаба, справедливо соотношение

$$\left\| \sum_{i=1}^{k} \frac{w_{\theta_l i} w_{\theta_j i}}{P_i} \right\| = \sum_{i=1}^{k} \frac{\nabla^{\mathrm{T}} P_i(\theta) \nabla P_i(\theta)}{P_i(\theta)} = \mathbf{J}_{\Gamma}(\theta) \qquad (3.30)$$

и, следовательно,

$$\mathbf{\Lambda}(\theta) = [\mathbf{J}(\theta) - \mathbf{J}_{\Gamma}(\theta)]^{-1}. \qquad (3.31)$$

Действительно, для законов с параметром сдвига θ_1 и масштаба θ_2 с функцией распределения $F((x-\theta_1)/\theta_2)$ и плотностью $\frac{1}{\theta_2} f((x-\theta_1)/\theta_2)$ элементы информационной матрицы $\mathbf{J}_{\Gamma}(\theta)$ имеют вид:

$$J_{\Gamma}(\theta_1, \theta_1) = \sum_{i=1}^{k} \frac{1}{\theta_2^2 P_i(\theta)} (-f(t_i) + f(t_{i-1}))^2,$$

$$J_{\Gamma}(\theta_2, \theta_2) = \sum_{i=1}^{k} \frac{1}{\theta_2^2 P_i(\theta)} (-t_i f(t_i) + t_{i-1} f(t_{i-1}))^2,$$

$$J_{\Gamma}(\theta_1, \theta_2) = \sum_{i=1}^{k} \frac{1}{\theta_2^2 P_i(\theta)} (-f(t_i) + f(t_{i-1})) \times (-t_i f(t_i) + t_{i-1} f(t_{i-1})),$$

где $t_i = (x_i - \theta_1)/\theta_2$. Тогда нетрудно заметить, что

$$w_{\theta_1 i} = \frac{1}{\theta_2} (-f(t_i) + f(t_{i-1})),$$

$$w_{\theta_2 i} = \frac{1}{\theta_2} (-t_i f(t_i) + t_{i-1} f(t_{i-1})).$$

Если проверяемая гипотеза H_0 о принадлежности закона распределению параметрическому семейству $f(x, \theta)$ неверна, и на самом деле справедлива конкурирующая гипотеза H_1, которой соответствует распределение с

плотностью $f_1(x,\theta) = f(x,\theta) + \delta(x,\theta)/\sqrt{N}$, статистика $Y_N^2(\theta)$ в пределе подчиняется нецентральному χ^2_{k-1}-распределению с параметром нецентральности [35]

$$\lambda_N(\theta) = \sum_{i=1}^{k} \frac{c_i^2(\theta)}{P_i(\theta)} + \mathbf{d}^{\mathrm{T}}(\theta)\mathbf{\Lambda}(\theta)\mathbf{d}(\theta), \qquad (3.32)$$

где $c_i(\theta) = \int_{x_{i-1}}^{x_i} \delta(x,\theta)dx$, $d_{\theta_l} = w_{\theta_l 1}c_1(\theta)/P_1 + \ldots + w_{\theta_l k}c_k(\theta)/P_k$ – элементы вектора $\mathbf{d}(\theta)$, соответствующие оцениваемым компонентам вектора θ, а размерность вектора равна числу оцениваемых параметров.

Как отражается на распределениях статистики и мощности критерия типа χ^2 Никулина способы группирования и выбор числа интервалов, показано в работах [47, 48]

3.3. Экспериментальное исследование распределений статистик критериев согласия в системе ISW

В практике статистического анализа с необходимостью использования критериев согласия приходится сталкиваться как при проверке простой гипотезы H_0: $f(x) = f(x,\theta_u)$, где $f(\cdot)$ – плотность распределения наблюдаемого закона, θ_u – известное истинное значение параметра (вектора параметров) закона, так и при проверке сложной гипотезы H_0: $F(x) \in \{F(x,\theta), \theta \in \Theta\}$. В последнем случае оценка параметра предполагаемого закона распределения $\hat{\theta}$ вычисляется по той же самой выборке, по которой проверяется согласие. Если оценка $\hat{\theta}$ вычисляется по другой выборке, то гипотеза простая. В дальнейшем будем обозначать сложную гипотезу следующим образом H_0: $F(x) = F(x,\hat{\theta})$, где $\hat{\theta}$ – оценка параметра, вычисляемая по этой же выборке.

В случае проверки сложных гипотез предельные распределения статистик непараметрических критериев согласия типа Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса, при справедливости нулевой гипотезы H_0: $f(x) = f(x,\hat{\theta})$ отличаются от предельных распределений классических статистик (когда по выборке не оцениваются параметры). В случае сложной гипотезы предельные распределения статистик зависит: от вида наблюдаемого закона, от количества и типа оцениваемых параметров этого распределения, от используемого метода оценивания параметров. А при ограниченных объемах выборок распределение статистики существенно зависит и от объема выборки.

Знание распределения статистики при проверке одной и той же гипотезы, но при различных истинных гипотезах (H_0 или H_1) позволяет опреде-

лить мощность критерия, т.е. его способность различать эти гипотезы. Задавая конкретную альтернативу и имея возможность построить распределения статистик при истинности нулевой гипотезы H_0 ($g(S|H_0)$) и истинности альтернативы H_1 ($g(S|H_1)$), можно при заданном уровне значимости α вычислить мощность критерия $1-\beta$, которая определяет способность различения этих альтернатив.

Для построения распределения $g(S|H_0)$ (распределения статистики при справедливости H_0) следует моделировать псевдослучайные величины, соответствующие наблюдаемому закону, и оценивать параметры этого закона, после чего вычислять значение требуемой статистики S критерия. А для построения распределения $g(S|H_1)$ (распределения той же статистики при проверке той же самой сложной гипотезы H_0, но при справедливой гипотезе H_1) следует моделировать псевдослучайные величины по закону, соответствующему гипотезе H_1, а оценивать параметры закона, соответствующего гипотезе H_0.

На рис. 3.5 приведена форма для моделирования распределений статистик критериев согласия.

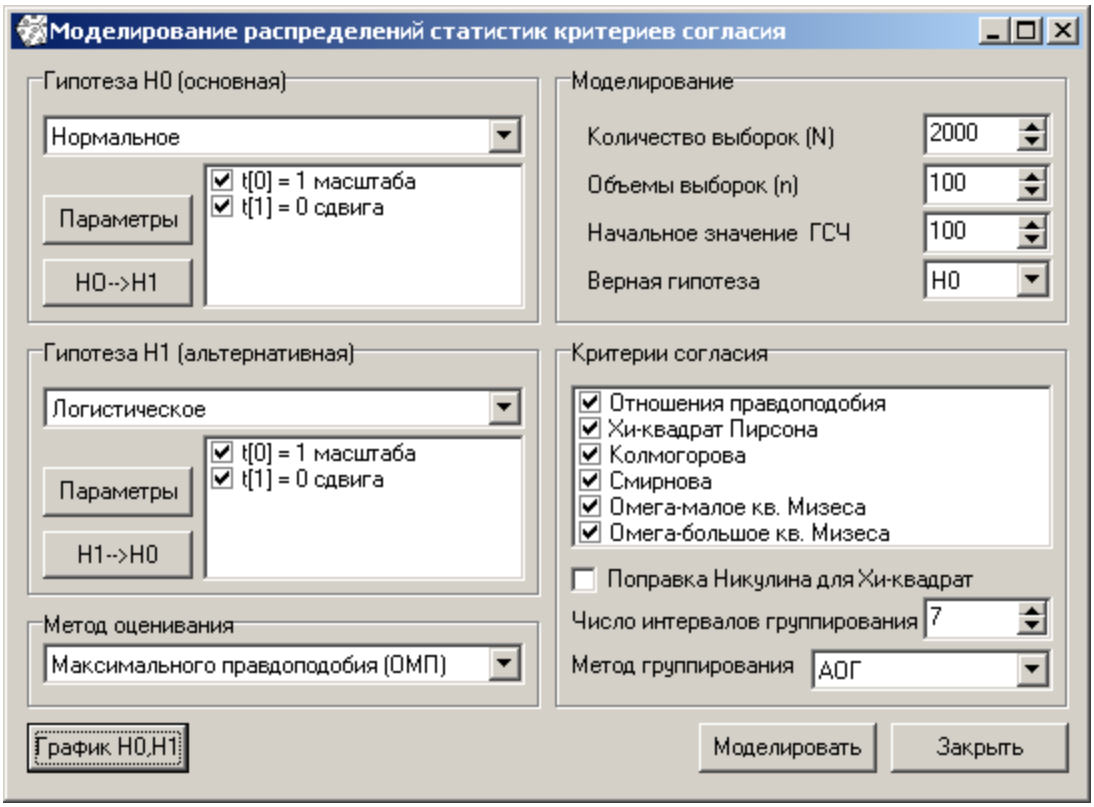

Рис. 3.5. Форма для моделирования распределений статистик критериев согласия

Для моделирования распределения $G_n(S|H_0)$ при заданном объеме выборок n необходимо задать гипотезу H_0, выбрав закон из списка возможных законов распределений, отметить галочками параметры, которые необходимо оценивать при проверке согласия, в случае сложной гипотезы, и установить переключатель «верная гипотеза» в положение «H_0». Метод оценивания необходимо выбрать из списка возможных значений, так как в случае сложной гипотезы метод оценивания влияет на распределение статистики критерия.

На качество моделирования влияет количество моделируемых выборок – N (объем выборки статистик). Для получения точности моделирования в пределах 0.01-0.04 необходимо выбирать $N \geq 2000$ (см. пункт 1.3.2). Отметим также, что получаемое в результате эмпирическое распределение статистики является случайным, и если изменить начальное значение генератора случайных чисел (ГСЧ), то мы получим несколько другое эмпирическое распределение. Однако при одном и том же начальном значении ГСЧ, мы будем получать идентичные эмпирические распределения статистик.

Для моделирования распределения $G_n(S|H_1)$ необходимо также задать альтернативную гипотезу H_1, и установить переключатель «верная гипотеза» в положение «H_1». Обычно распределение $G_n(S|H_1)$ необходимо для вычисления мощности критерия, причем при близких гипотезах. Чтобы не подбирать вручную параметры альтернативного распределения, приближающие его к основному распределению, можно нажать на кнопку «H1→H0», и получить оценки параметров распределения H_1 по неслучайной выборке из распределения H_0, в соответствии с выбранным методом оценивания. Неслучайная выборка получается по формуле

$$x_i = F^{-1}\left(\frac{2i-1}{2\cdot 100}\right),\ i = 1,\ldots,100.$$

Пример 3.1.

На рис. 3.6. показана подгонка параметров логистического распределения к стандартному нормальному закону. Как видно на графике стандартное нормальное распределение и логистическое распределение с параметром масштаба 0.5557 представляют собой очень близкие законы.

При моделировании по одной и той же выборке можно одновременно вычислять статистики нескольких критериев. Чтобы моделировать не все статистики, а только некоторые из них, следует отметить галочками те критерии согласия, которые нас интересуют. Для критериев типа χ^2 необходимо задать также число интервалов группирования и метод группирования (способ разбиения выборки на интервалы).

После того, как все параметры заданы, при нажатии кнопки «Моделировать» запускается процедура моделирования. Результаты моделирования

записываются в файлы с именами, указанными в таблице 3.1. Формат файлов соответствует формату негруппированной выборки (см. главу 7).

По полученным выборкам можно построить графики эмпирических функций распределения статистик соответствующих критериев, а также подобрать аналитическую модель, наиболее близко описывающую закон распределения статистики.

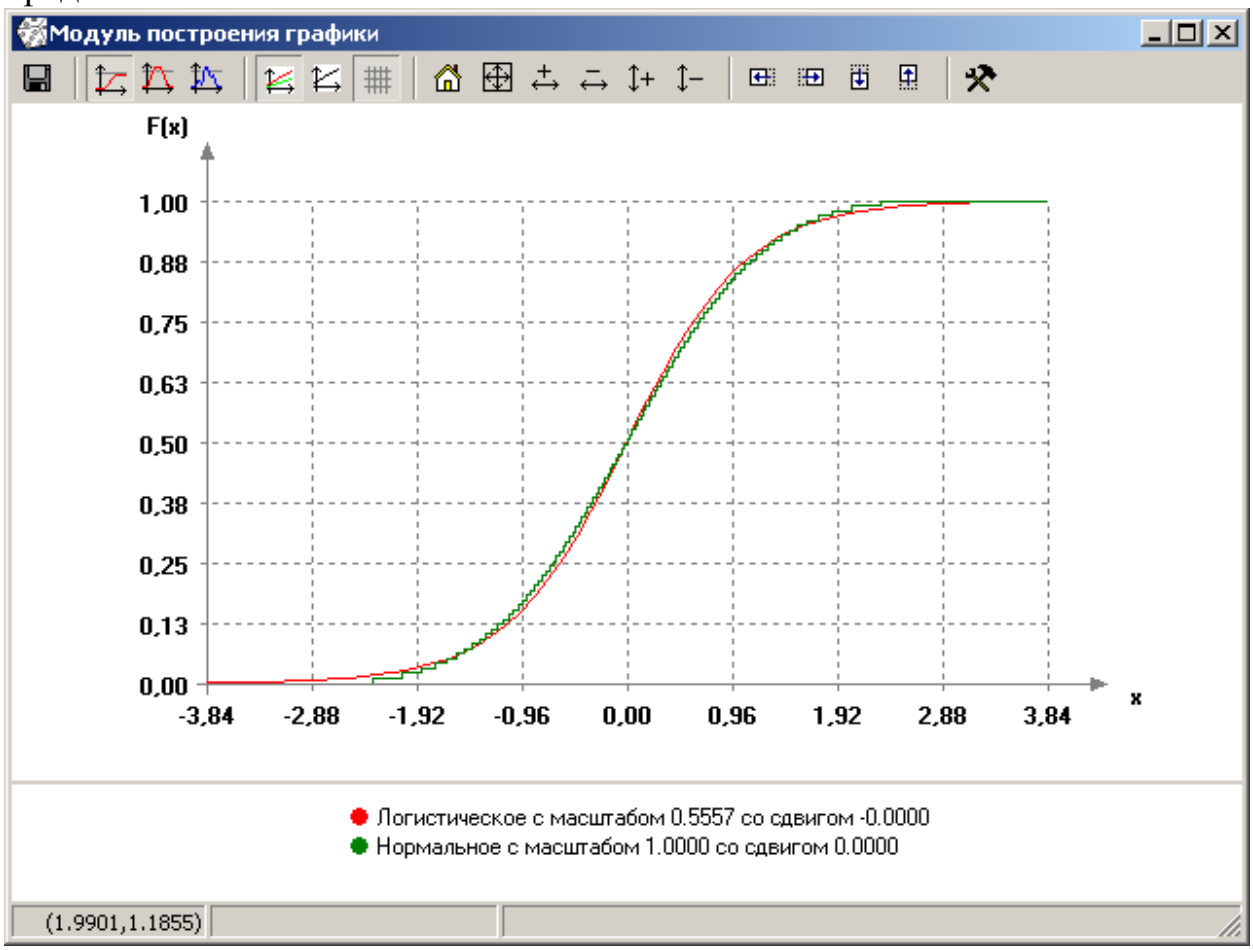

Рис. 3.6. Форма для моделирования распределений статистик критериев согласия

Таблица 3.1

Имена файлов с выборками статистик критериев согласия

Верная гипотеза H_0	Верная гипотеза H_1	Критерий
G(S_lr|H0).dat	G(S_lr|H1).dat	Отношения правдоподобия
G(S_chi2|H0).dat	G(S_ chi2|H1).dat	χ^2 Пирсона
G(S_chi2_Nik|H0).dat	G(S_ chi2_ Nik |H1).dat	χ^2 с поправкой Никулина
G(S_kolm|H0).dat	G(S_kolm|H1).dat	Колмогорова
G(S_smir|H0).dat	G(S_smir|H1).dat	Смирнова
G(S_om_s|H0).dat	G(S_om_s|H1).dat	ω^2 Мизеса
G(S_om_b|H0).dat	G(S_om_b|H1).dat	Ω^2 Мизеса

Пример 3.2.

Попробуем идентифицировать закон распределения статистики Колмогорова при проверке сложной гипотезы о согласии с нормальным распределением с параметрами сдвига и масштаба, оцениваемыми по методу максимального правдоподобия. Открываем форму «Статистический анализ», выбираем файл с выборкой G(S_kolm|H0).dat.

Известно [32], что эмпирические законы распределения статистик непараметрических критериев согласия наиболее хорошо описываются одним из следующих законов распределения: логарифмически нормальным, гамма-распределением, распределением Sl-Джонсона или распределением Su-Джонсона. Поэтому для определения закона распределения статистики нажмем на флажок «Идентификация», отметим галочками перечисленные выше модели и нажмем на кнопку «Оценить и проверить». Получаем следующие результаты, приведенные на рис. 3.7 и в таблице 3.2. Таким образом, распределение статистики Колмогорова наилучшим образом описывает распределение Su-Джонсона.

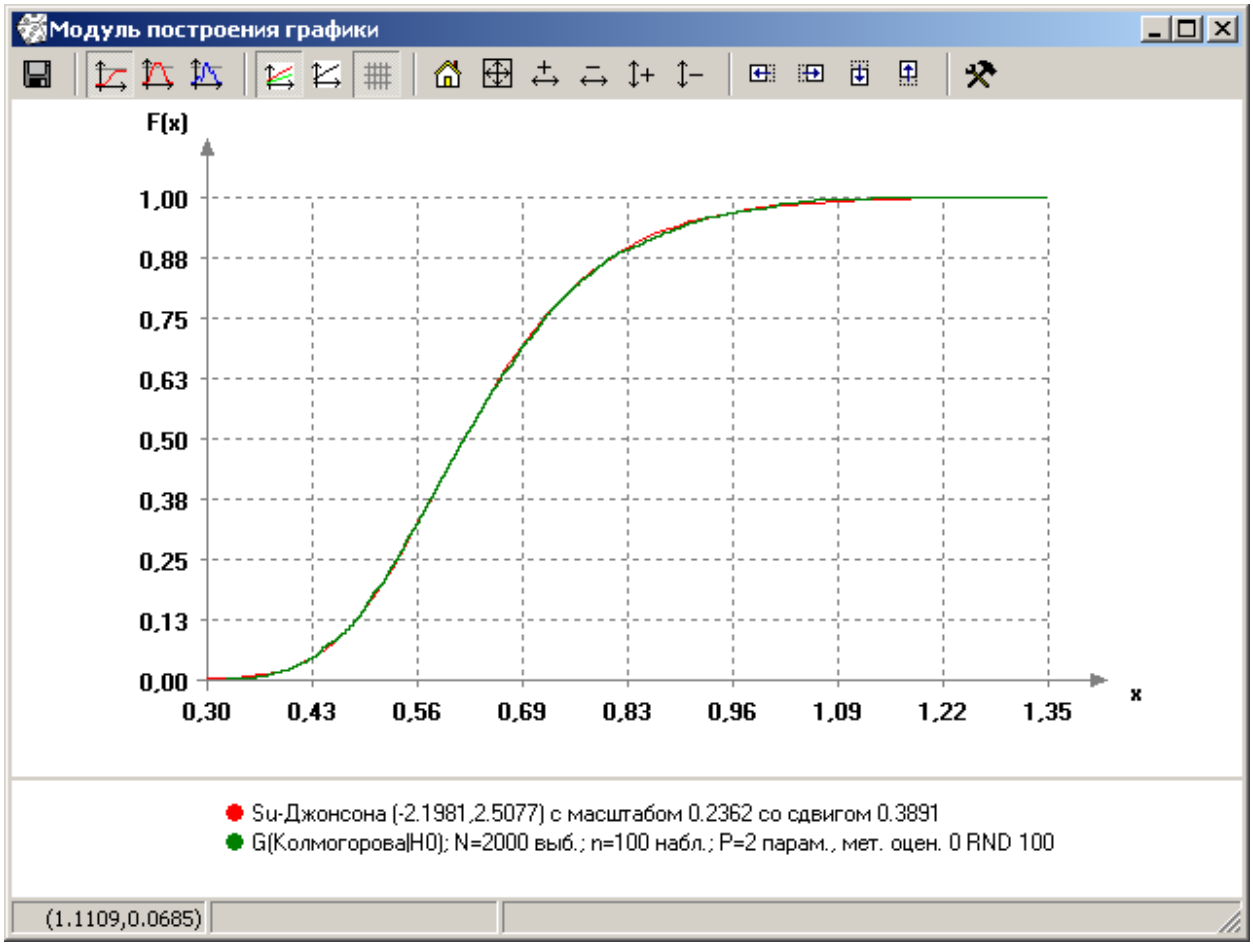

Рис. 3.7. Идентификация распределения статистики Колмогорова при проверке сложной гипотезы о нормальном распределении

Таблица 3.2.

Идентификация распределения статистики Колмогорова при проверке сложной гипотезы о нормальном распределении

Распределение	Параметры	Достигнутый уровень значимости
lnN	-0.47, 0.22	0.045638
Гамма	4.9272, 0.0663, 0.3182	0
Sl-Дж	-2.1319, 4.5357, 0.3924, 0.0000	0.18527
Su-Дж	-2.1981,2.5077,0.2362,0.3891	0.29639

3.4. Экспериментальное исследование мощности критериев согласия

После того, как получены эмпирические распределения статистики при верной гипотезе H_0 и при верной гипотезе H_1 (файлы «G(S|H0).dat» и «G(S|H1).dat»), можно вычислить мощность критерия согласия при заданной альтернативе. Для этого построим эмпирические функции распределения на одном графике (см. рис. 3.8).

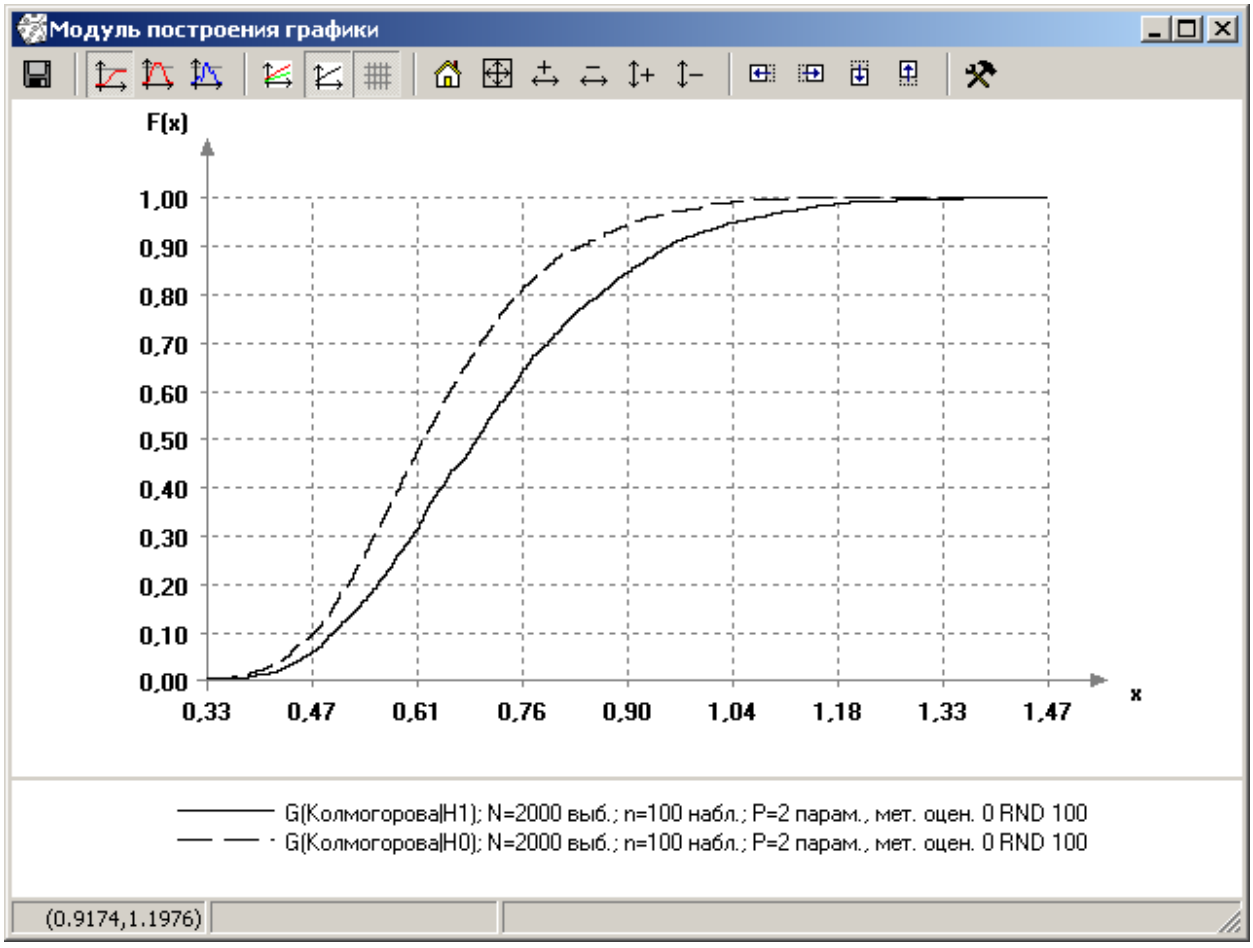

Рис. 3.8. Вычисление мощности критерия согласия

По графику определяем значение мощности $1-\beta$, варьируя α. Например, «на глаз» можно определить, что при α=0.1 мощность $1-\beta\approx$0.2. Более точно эти значения можно определить следующим образом. Открываем форму «Статистический анализ», выбираем выборку «G(S_Kolm|H0).dat» и нажимаем кнопку «Q», чтобы вычислить выборочные квантили $S_{1-\alpha}$ (см. рис. 3.9). Затем выбираем выборку «G(S_Kolm|H1).dat» и нажимаем кнопку «P», чтобы вычислить вероятности вида $\beta = G(S_{1-\alpha} \mid H_1)$. В результате получаем следующие значения мощности критерия (см. таблицу 3.3).

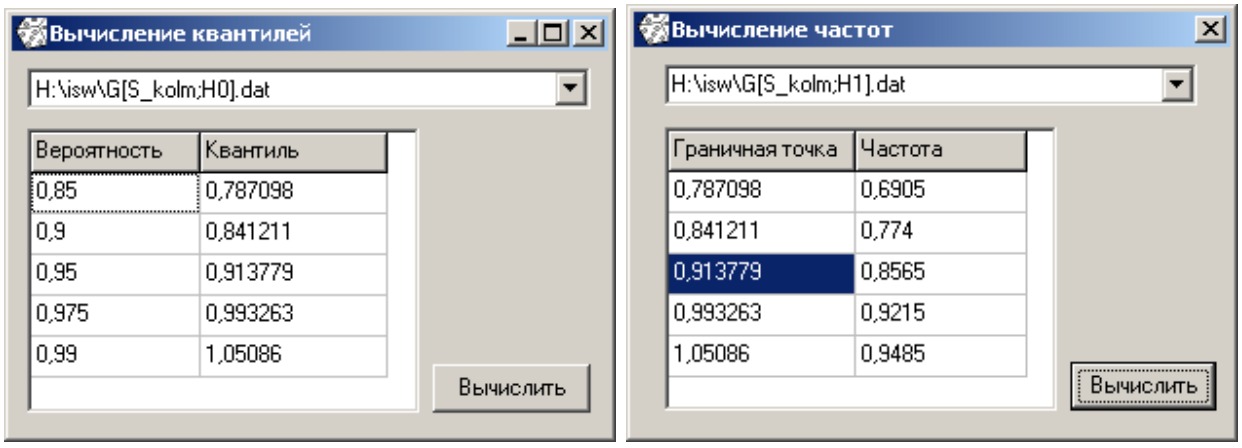

Рис. 3.9. Вычисление мощности критерия согласия

Таблица 3.3

Мощность критерия Колмогорова при проверке сложной гипотезы, $n=100$

α	$1-\beta$
0.15	0,3095
0.10	0,226
0.05	0,1435
0.025	0,0785
0.01	0,0515

Контрольные вопросы и задачи

1. Что такое статистическая гипотеза? Какие виды статистических гипотез Вы знаете?
2. Что такое критерий согласия? Оперативные характеристики критерия?

3. Какой критерий называется равномерно наиболее мощным? Состоятельным? Несмещенным?
4. Непараметрические критерии согласия. Особенности проверки сложных гипотез.
5. Критерии согласия типа χ^2 Пирсона. Влияние способа группирования данных на мощность критерия χ^2.
6. Сравните критерий согласия типа χ^2 Никулина с критерием χ^2 Пирсона.
7. Как можно экспериментально исследовать свойства критериев согласия?

Глава 4. Регрессионный анализ

4.1. Линейная регрессия

Линейной регрессией назовем зависимость

$$M[y \mid x] = \eta(x, \boldsymbol{\theta}) = f^T(x)\boldsymbol{\theta} = \sum_{i=1}^{m} f_i(x)\boldsymbol{\theta}. \qquad (4.1)$$

Пусть проведено n измерений величины y при некоторых значениях вектора x. Подставив результаты измерений в уравнение (4.1), получаем:

$$y_1 = f_1(x_1)\theta_1 + \ldots + f_m(x_1)\theta_m + e_1,$$
$$\ldots$$
$$y_n = f_1(x_n)\theta_1 + \ldots + f_m(x_n)\theta_m + e_n,$$

где $e_1, e_2, \ldots e_n$ - ошибки измерений.

Обозначим

$$\mathbf{Y} = (y_1, \ldots y_n)^T, \quad \mathbf{X} = \begin{bmatrix} f_1(x_1) & \cdots & f_m(x_1) \\ & \cdots & \\ f_1(x_n) & \cdots & f_m(x_n) \end{bmatrix}, \quad \mathbf{e} = (e_1, \cdots e_n)^T.$$

Тогда уравнение (4.1) принимает вид:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}. \qquad (4.2)$$

Здесь $\mathbf{Y}$ – вектор наблюдений размерности $(n \times 1)$, который называют также откликом системы, $\mathbf{X}$ – матрица независимых переменных размерности $(n \times m)$, называемая регрессором системы, $\boldsymbol{\theta}$ - вектор оцениваемых параметров размерности $(m \times 1)$, $\mathbf{e}$ – вектор случайных отклонений системы размерности $(n \times 1)$. Наличие в уравнении случайных ошибок говорит о том, что зависимость (4.2) является стохастической.

Предположим, что:

1) На вектор неизвестных параметров регрессии (4.2) не наложено ограничений, т.е. $\Theta = R^m$, где Θ - множество априорных значений вектора параметров $\boldsymbol{\theta}$.

2) Вектор $\mathbf{e} = (e_1 \cdots e_n)^T$ – случайный, отсюда следует, что $\mathbf{Y} = (y_1, \ldots y_n)^T$ – тоже случайный вектор.

3) Математическое ожидание вектора случайных отклонений равно нулю, т.е. $M[e_i] = 0, \quad i = \overline{1,n}$.

4) Для любых $i_1 \neq i_2$ $M[e_{i_1} \times e_{i_2}] = 0$, $M[e_i^2] = \sigma^2$ для всех $i = \overline{1,n}$. Другими словами, $\mathrm{cov}(\mathbf{e}) = \sigma^2 I_n$, здесь σ^2 – дисперсия отклонений, $\mathrm{cov}(\mathbf{e})$ - матрица ковариаций отклонений размерности $(n \times n)$, I_n-единичная матрица размерности $(n \times n)$, т.е.

$$\mathrm{cov}(e) = \begin{bmatrix} \sigma^2 & & O \\ & \ddots & \\ O & & \sigma^2 \end{bmatrix}_{n \times n}.$$

5) Матрица $\mathbf{X}$ является детерминированной, т.е. x_{ij} не являются случайными переменными.

6) Ранг матрицы регрессоров $rg(\mathbf{X}) = m$.

Будем считать, что *модель линейной регрессии* задана, если нам известны $\mathbf{Y}, \mathbf{X}$ и закон распределения вектора ошибок.

4.2. Оценивание параметров линейной регрессии методом максимального правдоподобия

Предположим, что известно распределение вектора случайных отклонений $\mathbf{e}$, а, следовательно, и распределение отклика системы. Пусть отклик системы имеет плотность распределения $f(\mathbf{Y}, \boldsymbol{\theta})$, где $\boldsymbol{\theta} = (\theta_1, \ldots \theta_m, \sigma^2)$ – вектор неизвестных параметров системы, значение которых необходимо оценить.

Определение. *Функцией правдоподобия* случайной величины $\mathbf{Y} = (y_1, \ldots y_n)^T$ назовем функцию вида:

$$f(\mathbf{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}).$$

Определение. *Логарифмической функцией правдоподобия* случайной величины $\mathbf{Y} = (y_1, \ldots y_n)^T$ назовем функцию следующего вида:

$$L(\mathbf{Y}, \boldsymbol{\theta}) = \ln \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(y_i, \boldsymbol{\theta}).$$

Определение. *Оценкой максимального правдоподобия* (ОМП) вектора параметров θ назовем такое значение вектора параметров $\hat{\boldsymbol{\theta}} \in \Theta$ (здесь Θ – множество возможных значений параметра θ), при котором достигается максимум функции правдоподобия или логарифмической функции правдоподобия (что совершенно идентично, поскольку функции $\varphi(x)>0$ и $ln\varphi(x)$ достигают экстремумов в одних и тех же точках). Вычислить значение ОМП можно, воспользовавшись одним из методов поиска максимума функции многих переменных.

4.3. Проверка гипотез в линейном регрессионном анализе. Критерий отношения правдоподобия

Существует общий способ построения критериев проверки статистических гипотез, предложенный Нейманом и Пирсоном. Он аналогичен методу максимального правдоподобия в статистическом оценивании и называется *критерием отношения правдоподобия*. Суть его заключается в следующем. Пусть плотность распределения $\mathbf{Y}$ равна $f(\mathbf{Y},\boldsymbol{\theta})$, т.е. зависит от некоторого неизвестного вектора параметров $\boldsymbol{\theta}\in\Theta$. Для каждого $\mathbf{Y}$ найдем $\max\limits_{\theta\in\Theta_H} f(\mathbf{Y},\boldsymbol{\theta})$ и $\max\limits_{\theta\in\Theta} f(\mathbf{Y},\boldsymbol{\theta})$ (считаем, что максимум достигается при $\hat{\boldsymbol{\theta}}$).

В качестве критического множества при проверке гипотезы $H:\boldsymbol{\theta}\in\Theta_H$ выбираем

$$W_K=\left\{\mathbf{Y}\in R^n:\max_{\theta\in\Theta_H} f(\mathbf{Y},\boldsymbol{\theta})\Big/\max_{\theta\in\Theta} f(\mathbf{Y},\boldsymbol{\theta})<\varphi_\alpha\right\}, \tag{4.3}$$

где φ_α – фиксированное число, которое задает верхнюю границу вероятности совершения ошибки первого рода $P_\theta(W_K)\le\alpha$ для всех $\boldsymbol{\theta}\in\Theta_H$. Статистика

$$\max_{\theta\in\Theta_H} f(\mathbf{Y},\boldsymbol{\theta})/\max_{\theta\in\Theta} f(\mathbf{Y},\boldsymbol{\theta}) \tag{4.4}$$

называется *статистикой критерия отношения правдоподобия*.

Рассмотрим линейную гипотезу в самом общем виде

$$H:\mathbf{R}\boldsymbol{\theta}=\mathbf{r}, \tag{4.5}$$

где $\mathbf{R}$ – известная матрица $k\times m$, $rg(\mathbf{R})=k\le m$; $\mathbf{r}$ – заданный вектор $k\times 1$.Таким образом, в гипотезе H на $\boldsymbol{\theta}$ накладывается k независимых линейных ограничений.

Для проверки гипотез необходимо знать вид распределения $\mathbf{Y}$. Положим, что $\mathbf{e}\sim N(0,\sigma^2\mathbf{I}_n)$, тогда $\mathbf{Y}\sim N(\mathbf{X}\boldsymbol{\theta},\sigma^2\mathbf{I}_n)$, и в этом случае статистика критерия отношения правдоподобия имеет вид:

$$Q=\frac{(\mathbf{r}-\mathbf{R}\hat{\boldsymbol{\theta}})^T(\mathbf{R}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{R}^T)^{-1}(\mathbf{r}-\mathbf{R}\hat{\boldsymbol{\theta}})}{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}})^T(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}})}\cdot\frac{n-m}{k} \tag{4.6}$$

Отметим частный случай гипотезы (4.5). Если матрица $\mathbf{R} = \mathbf{I}$, т.е. мы проверяем гипотезу обо всех параметрах регрессии $(k = m)$, то вид статистики (4.6) несколько упростится:

$$Q = \frac{(\mathbf{r} - \hat{\boldsymbol{\theta}})^T (\mathbf{X}^T \mathbf{X})(\mathbf{r} - \hat{\boldsymbol{\theta}})}{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})^T (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})} \cdot \frac{n-m}{m}. \tag{4.7}$$

Доказано [49], что статистика (4.7) критерия отношения правдоподобия при верной гипотезе о параметрах линейной регрессии с отклонениями, **подчиненными нормальному закону**, имеет распределение Фишера со степенями свободы k и $n-m$ соответственно, т.е. $F_{k,n-m}(Q)$.

Таким образом, если вычисленное значение $Q < F_{\alpha}(k, n-m)$, где α – заданный уровень значимости критерия, то гипотезу (4.5) принимаем, иначе – отклоняем.

Распределение Фишера является частным случаем бета-распределения II рода $Be_{II}(a, \sigma, \alpha, \beta)$:

$$F(k, n-m) = Be_{II}\left(0, \frac{n-m}{k}, \frac{k}{2}, \frac{n-m}{2}\right),$$

где плотность бета-распределения II-го рода задается выражением

$$f(x) = Be_{II}(a, \sigma, \alpha, \beta) = \frac{1}{\sigma \cdot \mathrm{B}(\alpha, \beta)} \left(\frac{x-a}{\sigma}\right)^{\alpha-1} \left(1 + \frac{x-a}{\sigma}\right)^{-\alpha-\beta}$$

В случае проверки гипотезы относительно параметров регрессии с отклонениями, подчиненными закону распределения, отличному от нормального, вид предельного закона распределения статистики (4.6) не известен.

4.4. Экспериментальное исследование распределений статистики критерия отношения правдоподобия

В [50] методами компьютерного моделирования были проведены исследования статистики (4.7) для случаев, когда распределения ошибок подчинялись логистическому закону, распределению Коши и экспоненциальному семейству распределений. Исследования показали, что в большинстве рассмотренных случаев эмпирические функции распределения статистики (4.7), полученные в результате моделирования при использовании для оценивания вектора параметров регрессии метода максимального

правдоподобия, хорошо описываются бета-распределением II рода. Для различных значений числа наблюдений n и количества m оцениваемых параметров линейной регрессии найдены значения параметров бета-распределений II рода, аппроксимирующих в соответствующих случаях распределение статистики (4.7). Найденные аппроксимации могут выступать в качестве моделей предельных распределений статистики(4.7) при ошибках наблюдений отклика, подчиняющихся законам распределения Коши и экспоненциальному семейству (ЭС) с параметрами формы λ=0.5, λ=1.0 (соответствует распределению Лапласа), λ=2.0 (соответствует нормальному закону), λ=3.0, λ=10.0. В случае логистического закона ошибок наблюдений в качестве предельного распределения статистики (4.7), как и в «классическом» случае, может использоваться распределение Фишера $F_{m,n-m}(Q)$.

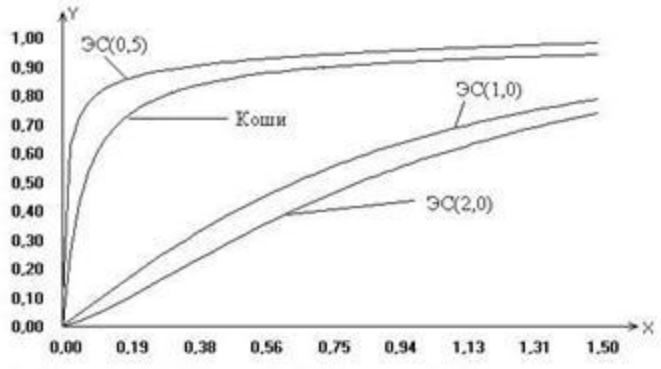

Рис. 4.1

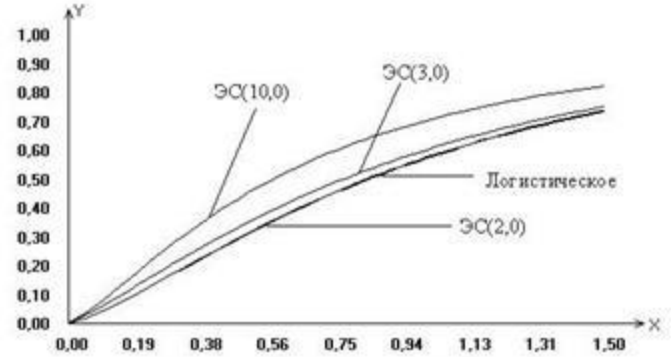

Рис. 4.2

На рисунках 4.1 и 4.2 представлены построенные в результате исследований модели предельных распределений статистики (4.7), соответствующие различным законам распределения ошибок наблюдений. На рисунках указано, каким законам распределения ошибок соответствуют приводимые распределения статистики (4.7).

При моделировании эмпирических распределений статистики (4.7) в случае линейной регрессии задается матрица $\mathbf{X}$ размерности $(n \times m)$, вектор параметров θ размерности $(m \times 1)$ и в соответствии с заданным законом распределения генерируется вектор случайных отклонений **e** размерности $(n \times 1)$. В результате получаем уравнение (4.2). Далее в соответствии с выбранным методом вычисляется вектор оценок $\hat{\theta}$ и значение статистики (4.7). При повторении данной процедуры N раз получаем смоделированную выборку значений статистики Q_1, $Q_2, \ldots, Q_N$, на основании которой при достаточно большом N можно делать надежные выводы о законе распределения статистики.

Контрольные вопросы

1. Что такое линейная регрессия?
2. Методы определения параметров линейной регрессии.
3. Проверка гипотезы о равенстве вектора параметров заданному вектору при нормальном законе распределения ошибок наблюдений.
4. Проверка гипотезы о равенстве вектора параметров заданному вектору при отклонении закона распределения ошибок наблюдений от нормального.

Глава 5. Корреляционный анализ

Введем для дальнейшего использования следующие обозначения:

- $X_1, X_2, \ldots, X_n$ – выборка m-мерного случайного вектора объема n;
- $M = [m_i]_{i=1}^{m}$ – вектор математического ожидания случайного вектора X_i;
- $\Sigma = [\sigma_{ij}]_{i,j=1}^{m}$ – ковариационная матрица случайного вектора X_i;
- $\hat{M}$ и $\hat{\Sigma}$ – оценки максимального правдоподобия (ОМП) для вектора математического ожидания и ковариационной матрицы, вычисляемые по негруппированным данным:

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{и} \quad \hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{M})(X_i - \hat{M})^T .$$

В процессе корреляционного анализа выборок многомерных случайных величин осуществляется оценка параметров многомерного закона и проверка различных статистических гипотез [51,52]. В основе классического корреляционного анализа лежит предположение о принадлежности наблюдаемых случайных величин многомерному нормальному закону.

5.1. Проверка гипотез о равенстве математического ожидания некоторому известному вектору

Практически важной проблемой является задача проверки гипотезы о равенстве вектора математических ожиданий нормального распределения некоторому заданному значению. Такая задача очень часто возникает на практике, когда, например, на основании наблюдений некоторого технологического процесса желают убедиться, что эти показатели равны номинальному значению, т.е. процесс протекает нормально, а отклонения наблюдаемых значений от номинальных объясняются лишь ошибками наблюдений (измерений). Рассмотрим эту проблему сначала в предположении, что ковариационная матрица известна, а затем – когда неизвестна.

Проверяемая гипотеза имеет вид $H_0 : M = M_0$. Здесь возможны две ситуации.

а) В случае, когда нам известна ковариационная матрица Σ (например, из ранее проводимых экспериментов или предположений), то для вывода предельных распределений статистик, используемых при проверке данной гипотезы, в многомерном случае используется факт, что разность между векторами среднего значения выборки и среднего значения генеральной совокупности распределена нормально с вектором математических ожиданий, равным нулевому, и известной ковариационной матрицей.

В соответствии с критерием вычисляется статистика:

$$X_m^2 = n(\hat{M} - M_0)^T \Sigma^{-1} (\hat{M} - M_0). \qquad (5.1)$$

При справедливой гипотезе H_0 предельным распределением статистики (5.1) является $G(X_m^2 \mid H_0) = \chi_m^2$ - распределение, с числом степеней свободы m. Проверяемая гипотеза H_0 принимается, если

$$X_m^2 < \chi_{m,\alpha}^2,$$

где α - уровень значимости и

$$1-\alpha = P\{X_m^2 > \chi_{m,\alpha}^2\} = \frac{1}{2^{m/2}\Gamma(m/2)} \int_0^{\chi_{m,\alpha}^2} s^{m/2-1} e^{-s/2} ds.$$

б) Ковариационная матрица неизвестна. В одномерном случае используется статистика, являющаяся частным от деления разности между выборочным средним значением и гипотетическим математическим ожиданием генеральной совокупности на среднее квадратичное отклонение. В многомерном аналоге данной статистики используются математическое ожидание и матрица, обратная к оценке ковариационной матрицы.

В соответствии с используемым критерием вычисляется статистика:

$$T^2 = \frac{n(n-m)}{m(n-1)} (\hat{M} - M_0)^T \hat{\Sigma}^{-1} (\hat{M} - M_0). \qquad (5.2)$$

При справедливости проверяемой гипотезы H_0 предельным распределением статистики (5.2) является $G(T^2 \mid H_0) = F_{m,n-m}$ – распределение Фишера с параметрами m и n-m. Гипотеза H_0 не отклоняется, если выполняется условие

$$T^2 < F_{m,n-m,\alpha}.$$

Критическое значение $F_{m,n-m,\alpha}$ определятся из равенства

$$1-a = P\{T^2 > F_{m,n-m,\alpha}\} = \left(\frac{m}{n-m}\right) \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n-m}{2}\right)} \int_0^{F_{m,n-m,\alpha}} s^{m/2-1} \left(1 + \frac{m}{n-m} s\right)^{-n/2} ds.$$

5.2. Проверка гипотез о коэффициенте парной корреляции

При анализе совокупности случайных величин нас может интересовать взаимосвязь между несколькими случайными величинами, зависимость одной или большего числа величин от остальных и т.д. Когда рассматривается взаимосвязь двух величин, то речь идет о парной корреляции. Взаимозависимость же величин при устранении влияния некоторой совокупности других – характеризуется частной корреляцией, а

вот зависимость одной величины от группы величин – множественной корреляцией.

Под парной корреляцией понимается обычная корреляция между двумя величинами. Если оценка ковариационной матрицы $\hat{\Sigma}$ уже известна, то оценка парного коэффициента корреляции может быть найдена в соответствии с выражением

$$\hat{r}_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}.$$

Здесь можно решать следующие задачи: определения оценки парного коэффициента корреляции, проверки гипотезы о его значимости (гипотеза вида: $H_0 : r_{ij} = 0$), проверки гипотезы о равенстве его определенному значению (гипотеза вида: $H_0 : r_{ij} = r_0$).

Для проверки гипотезы $H_0 : r_{ij} = 0$ согласно [51] вычисляется статистика:

$$t = \frac{\hat{r}_{ij}\sqrt{n-2}}{\sqrt{1-\hat{r}_{ij}^2}}. \qquad (5.3)$$

При этом предельным распределением статистики (5.3) является $G(t \mid H_0) = t_{n-2}$ -распределение Стьюдента с числом степеней свободы *n*-2. При конкурирующей гипотезе $H_1 : r_{ij} \neq 0$ гипотеза H_0 принимается, если

$$|t| < t_{n-2,\alpha/2},$$

где α - уровень значимости. Величина $t_{n-2,\alpha/2}$ при $k = n-2$ определяется равенством

$$1-\alpha = P\{|t| < t_{k,\alpha/2}\} = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \int_{-t_{k,\alpha/2}}^{t_{k,\alpha/2}} \left(1+\frac{s^2}{k}\right)^{-\frac{k+1}{2}} ds.$$

Если же проверяемая гипотеза имеет вид $H_0 : r_{ij} = r_0$, то используется статистика вида:

$$z_0 = \sqrt{n-3}\left(\frac{1}{2}\ln\left(\frac{1+\hat{r}_{ij}}{1-\hat{r}_{ij}}\right) - \frac{1}{2}\ln\left(\frac{1+r_0}{1-r_0}\right) - \left(\frac{r_0}{2(n-1)}\right)\right). \qquad (5.4)$$

При этом предельным распределением статистики (5.4) является стандартное нормальное распределение $G(z_0 \mid H_0) = N_{0,1}$. Гипотеза H_0 принимается, если

$$z < t_{\alpha/2},$$

где $t_{\alpha/2}$ – квантиль стандартного нормального распределения и

$$1-a = P\{t < t_{\alpha/2}\} = \frac{1}{\sqrt{2\pi}} \int_{-t_{\alpha/2}}^{t_{\alpha/2}} e^{-s^2/2} ds .$$

5.3. Проверка гипотез о коэффициенте частной корреляции

В случае двух нормальных или почти нормальных величин коэффициент корреляции между ними может быть использован в качестве меры взаимозависимости. Однако на практике при интерпретации «взаимозависимости» часто встречаются определенные трудности, так как, если одна величина коррелирована с другой, то это может быть всего лишь отражением того факта, что они обе коррелированы с некоторой третьей величиной или с совокупностью величин. Это приводит к необходимости рассмотрения условных корреляций между двумя величинами при фиксированных значениях остальных величин. Это, так называемые, *частные корреляции*.

Если корреляция между двумя величинами уменьшается при фиксировании некоторой другой случайной величины, то это означает, что их взаимозависимость возникает частично через воздействие этой величины. Если частная корреляция равна нулю или очень мала, то делается вывод, что их взаимозависимость целиком обусловлена этим воздействием. Напротив, когда частная корреляция больше первоначальной корреляции между двумя величинами, то это означает, что другие величины ослабляли связь, или, можно сказать, «маскировали» корреляцию. Но следует помнить, что даже в последнем случае нельзя предполагать наличие причинной связи, так как некоторая, совершенно отличная от рассматриваемых при анализе, величина может быть источником этой корреляции. Как при обычной корреляции, так и при частных корреляциях предположение о причинности должно всегда иметь внестатистические основания.

Представим случайный вектор X в следующем виде:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

где $X_1 = (x_1, x_2, \ldots, x_l)^T$, $X_2 = (x_{l+1}, x_{l+2}, \ldots, x_m)^T$, а вектор математических ожиданий и ковариационную матрицу соответственно в виде:

$$M = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Если случайный вектор X подчиняется нормальному закону с вектором средних M и ковариационной матрицей Σ, то условное распределение подвектора X_1 при известном X_2 является нормальным с математическим

ожиданием $M_1 + B(X_2 - M_2)$ и ковариационной матрицей $\Sigma_{11\bullet 2}$, где $B = \Sigma_{12}\Sigma_{22}^{-1}$, $\Sigma_{11\bullet 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

ОМП для частного коэффициента корреляции определяется следующим соотношением:

$$\hat{r}_{ij;l+1,\ldots,m} = \frac{\hat{\sigma}_{ij;l+1,\ldots,m}}{\sqrt{\hat{\sigma}_{ii;l+1,\ldots,m}\hat{\sigma}_{jj;l+1,\ldots,m}}},$$

где $\hat{\sigma}_{ij\bullet l+1,\ldots,m}$ - элемент i-й строки и j-го столбца матрицы $\Sigma_{11\bullet 2}$, l – число компонент в условном распределении, $2 \le l \le m$.

В данном случае при оценке взаимозависимости между компонентами x_i и x_j случайной величины X исключается влияние компонент $x_{l+1}, x_{l+2}, \ldots, x_m$.

При проверке гипотез вида $H_0 : r_{ij\bullet l+1,\ldots,m} = 0$ и $H_0 : r_{ij\bullet l+1,\ldots,m} = r_0$ используются те же самые статистики, что и для парного коэффициента корреляции. При этом в соответствующих соотношениях n заменяется на $n - m + l$.

Для проверки гипотезы $H_0 : r_{ij;l+1,\ldots,m} = 0$ вычисляется статистика:

$$t = \frac{\hat{r}_{ij\bullet l+1,\ldots,m}\sqrt{n-m+l-2}}{\sqrt{1-\hat{r}_{ij\bullet l+1,\ldots,m}}}. \qquad (5.5)$$

При этом предельным распределением статистики (5.5) оказывается $G(t \mid H_0) = t_{n-m+l-2}$ -распределение Стьюдента с числом степеней свободы n-m+l-2.

Если же проверяемая гипотеза имеет вид $H_0 : r_{ij;l+1,\ldots,m} = r_0$, тогда используется статистика:

$$z_0 = \sqrt{n-3}\left(\frac{1}{2}\ln\left(\frac{1+\hat{r}_{ij\bullet l+1,\ldots,m}}{1-\hat{r}_{ij\bullet l+1,\ldots,m}}\right) - \frac{1}{2}\ln\left(\frac{1+r_0}{1-r_0}\right) - \left(\frac{r_0}{2(n-1)}\right)\right). \qquad (5.6)$$

При этом предельным распределением статистики (5.6) оказывается $G(z_0 \mid H_0) = N(0,1)$ – стандартное нормальное распределение.

5.4. Проверка гипотез о коэффициенте множественной корреляции

Множественный коэффициент корреляции является мерой зависимости компоненты многомерной случайной величины от некоторого множества компонент.

Можно рассматривать корреляцию между одной компонентой случайного вектора и множеством всех остальных или каким-то подмножеством.

Следует отметить, что множественный коэффициент корреляции r_i случайной величины x_i относительно некоторого множества других случайных величин всегда не меньше, чем абсолютная величина любого парного коэффициента корреляции r_{ij} с таким же первичным индексом. Более того, множественный коэффициент корреляции никогда нельзя уменьшить путем расширения множества величин, относительно которых измеряется зависимость x_i.

Если коэффициент корреляции между x_i и множеством всех остальных компонент многомерной случайной величины равен нулю ($r_i = 0$), то все коэффициенты корреляции этой величины относительного любого подмножества также равны 0, т.е. величина x_i полностью некоррелирована со всеми остальными величинами.

С другой стороны, если r_i относительно множества всех остальных компонент равен единице $r_i = 1$, то, по крайней мере, один из коэффициентов корреляции относительно некоторого подмножества компонент должен быть равен 1.

Надо отметить, что коэффициент корреляции, например, между x_1 и множеством всех остальных компонент является обычным коэффициентом корреляции между x_1 и условным математическим ожиданием $E[x_1 \mid x_2, x_3, \ldots, x_n]$.

Если представить случайный вектор X в том виде, как это было показано в разделе частной корреляции, то ОМП множественного коэффициента корреляции между x_i, $i \le l$ и множеством компонент $x_{l+1}, x_{l+2}, \ldots, x_m$ определяется соотношением

$$\hat{r}_{i;l+1,\ldots,m} = \sqrt{\frac{\hat{\sigma}_{(i)}\Sigma_{22}^{-1}\hat{\sigma}_{(i)}^{T}}{\hat{\sigma}_{ii}}},$$

где $\sigma_{(i)} - i^{ая}$ строка матрицы Σ_{12}, σ_{ii} – элемент матрицы Σ_{11}.

Для проверки гипотезы $H_0 : r_{i\bullet l+1,\ldots,m} = 0$ вычисляется статистика:

$$F = \frac{n-m+l-1}{m-l}\frac{\hat{r}^2_{i\bullet l+1,\ldots,m}}{1-\hat{r}^2_{i\bullet l+1,\ldots,m}}. \qquad (5.7)$$

При этом предельное распределение статистики $G(F \mid H_0) = F_{m-l,n-m+l-1}$ – распределение Фишера с параметрами m-l и n-m+l-1. Гипотеза H_0 принимается, если

$$F < F_{m-l,n-m+l-1,\alpha},$$

где α - уровень значимости, а $F_{m-l,n-m+l-1,\alpha}$ – критическое значение с уровнем значимости α.

5.5. Экспериментальное исследование распределений статистик корреляционного анализа

Выше рассмотрена только часть критериев проверки гипотез, используемых в классическом корреляционном анализе. Подчеркнем, что *все предельные распределения статистик указанных критериев имеют место, если наблюдается многомерный нормальный закон*.

Очевидно, что многомерный нормальный закон далеко не всегда является наилучшей моделью для описания реально наблюдаемых многомерных случайных величин. Что произойдет с предельными распределениями этих статистик, насколько могут быть справедливы выводы, формулируемые на основании решения задач классического корреляционного анализа, если наблюдаемый закон отличается от многомерного нормального, заранее сказать нельзя.

Исследования, проведенные в [53], показали, что распределения некоторых статистик корреляционного анализа устойчивы к отклонениям от нормальности (критерии проверки гипотез о математических ожиданиях и коэффициентах корреляции), других же очень чувствительны (о ковариационных матрицах).

Ключевым моментом для исследования распределений статистик корреляционного анализа при некоторых произвольных многомерных законах (отличающихся от нормального) является необходимость моделирования псевдослучайных векторов в соответствии с такими законами. Причем желательно иметь возможность моделирования псевдослучайных векторов по законам с «регулируемым удалением» от многомерного нормального, чтобы проследить соответствующие изменения распределений исследуемых статистик корреляционного анализа.

Моделирование псевдослучайных нормальных векторов. Многомерное нормальное распределение случайного вектора $\overline{X} = \|X_1, X_2, ..., X_m\|^T$ размерности m полностью определяется вектором математических ожиданий $\overline{M} = \|M_1, M_2, ..., M_m\|^T$ и ковариационной матрицей

$$\Sigma = \| \sigma_{ij} \| = E[(X_i - M_i)(X_j - M_j)] .$$

Функция плотности многомерного нормального закона имеет вид

$$f(\overline{X}) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} e^{-\frac{1}{2}(\overline{X}-\overline{M})^T \Sigma^{-1} (\overline{X}-\overline{M})} .$$

Хорошо зарекомендовавший себя алгоритм генерирования псевдослучайных нормальных векторов подробно изложен в [54]. Пусть мы имеем совокупность случайных величин $\{Z_i\}$, $i = \overline{1, m}$, где Z_i – подчиняется стандартному нормальному закону с параметрами (0, 1). Тогда вектор $\overline{X}$, распределенный по многомерному нормальному закону с параметрами $\overline{M}$ и Σ, получается через линейное преобразование вида

$$\overline{X} = A\overline{Z} + \overline{M} . \qquad (5.8)$$

Обычно полагают, что A является нижней треугольной матрицей

$$A=\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix},$$

коэффициенты a_{ij} которой определяются рекуррентной процедурой:

$$a_{ij}=\frac{\sigma_{ij}-\sum_{k=1}^{j-1} a_{ik}a_{jk}}{\sqrt{\sigma_{jj}-\sum_{k=1}^{j-1} a_{jk}^2}},\ 1\le j\le i\le m. \qquad (5.9)$$

Моделирование многомерных законов, отличных от нормального. Процедуру моделирования многомерных величин, распределенных по законам, отличным от нормального, с заданными математическим ожиданием и ковариационной матрицей, предложено [55] реализовать в соответствии с описанным выше алгоритмом. При этом совокупность $\{Z_i\}$, $i=\overline{1,m}$, формируется уже не по стандартному нормальному закону, а в соответствии с некоторым одномерным законом распределения с нулевым математическим ожиданием и единичной дисперсией. Затем заданная матрица Σ раскладывается по формуле (5.9) и осуществляется преобразование (5.8). На выходе мы имеем некоторый многомерный закон, отличный от нормального закона, с известным математическим ожиданием, но, вообще говоря, с неизвестной ковариационной матрицей, так как ковариационная матрица смоделированного закона не совпадает с используемой при моделировании матрицей Σ.

Для моделирования различных совокупностей $\{Z_i\}$, $i=\overline{1,m}$, удобно использовать экспоненциальное семейство распределений с плотностью

$$f(x)=\frac{\lambda}{2\sqrt{2}\theta_1\Gamma(1/\lambda)}\exp\left(-\left(\frac{|x-\theta_0|}{\sqrt{2}\theta_1}\right)^{\lambda}\right),$$

где λ – параметр формы, так как оно охватывает целый класс симметричных распределений. Частными случаями данного закона являются распределение Лапласа (при $\lambda = 1$), нормальное ($\lambda = 2$), предельными – распределение Коши ($\lambda \to 0$) и равномерное ($\lambda \to +\infty$). С помощью параметра формы λ мы можем задавать непрерывное «удаление» моделируемого (наблюдаемого) многомерного закона от нормального, делая его более плосковершинным по сравнению с нормальным при $\lambda > 2$ или более островершинным при $0 < \lambda < 2$. При $\lambda = 2$ будут формироваться псевдослучайные векторы $\overline{X}$ в соответствии с нормальным законом.

К сожалению, такая процедура не позволяет нам моделировать многомерный закон с некоторой произвольной функцией распределения, с заданными математическим ожиданием и ковариационной матрицей и

который находится на «заданном» расстоянии (определяемом в смысле некоторой меры) от многомерного нормального закона. Однако мы можем построить датчик, генерирующий псевдослучайные векторы по закону, отличающемуся от нормального (в соответствии с процессом моделирования), с известными математическим ожиданием и ковариационной матрицей. При этом вектор математического ожидания и ковариационная матрица определяются на основании исследования свойств полученного датчика (при заданных $\overline{M}$, Σ и λ). Для определения «истинной» ковариационной матрицы моделируемого многомерного закона можно использовать оценки максимального правдоподобия, усредняемые по множеству проведенных экспериментов.

Для моделирования распределений статистик корреляционного анализа основным является разработка датчика, генерирующего псевдослучайные векторы по заданному закону. Если такой датчик есть, процедура моделирования эмпирических распределений всех рассмотренных статистик становится очевидной.

Контрольные вопросы и задачи

1. Корреляционный анализ.
2. Проверка гипотез о равенстве математического ожидания заданному вектору.
3. Проверка гипотез о коэффициенте парной корреляции.
4. Проверка гипотез о коэффициенте частной корреляции.
5. Проверка гипотез о коэффициенте множественной корреляции.
6. Моделирование псевдослучайных нормальных векторов
7. Моделирование многомерных законов, отличных от нормального.
8. Моделирование законов распределений статистик критериев для проверки гипотез корреляционного анализа.

Глава 6. Статистический анализ интервальных наблюдений

Интервальная статистика – раздел математики, возникший на границе между интервальной математикой и математической статистикой.

Объектом исследования интервальной статистики являются интервальные наблюдения, т.е. наблюдения, заданные интервалом значений. Основной задачей интервальной статистики (также как и математической статистики) является восстановление статистических зависимостей (закономерностей).

6.1. Интервальная арифметика

Основная идея интервального анализа [56,57] состоит в том, что вещественное число представляется не одним, а двумя числами – оценкой снизу и оценкой сверху, образующими интервальное число.

Арифметические операции над интервальными числами выполняются следующим образом:

$[a_1, a_2] = [b_1, b_2] \circ [c_1, c_2]$, если $b \in [b_1, b_2]$, и $c \in [c_1, c_2]$, то $b \circ c \in [a_1, a_2]$,

где "$\circ$" – обычная арифметическая операция над вещественными числами (<+>,<->,<*>,</>).

Множество всех интервалов на R обозначается через IR.

Если $r(x)$ – непрерывная унарная операция на R, то $r(X) = [\min\limits_{x \in X} r(x), \max\limits_{x \in X} r(x)]$, $X \in IR$, определяет соответствующую ей операцию на множестве IR.

Особенностью такого определения интервальных чисел является то, что произвольный невырожденный интервал из IR не имеет обратного ни по сложению, ни по умножению. Вместо *дистрибутивности* вещественных чисел для интервальных чисел выполняется свойство *субдистрибутивности*:

$$A(B + C) \subseteq AB + AC,$$

которое лежит в основе "*интервального расширения*". Когда интервальные числа стали использовать в обычных алгоритмах, оказалось, что небольшие погрешности в исходных данных приводили к очень большим интервалам в результате вычислений. Этот эффект стали называть "интервальным расширением".

6.2. Интервальная выборка

6.2.1. Абсолютная погрешность

Пусть в результате эксперимента наблюдается случайная величина $\xi + \eta$. Первая случайная величина задаёт статистическую неопределенность, а вторая η – измерительную погрешность, действующую аддитивно на

результат измерения. Про погрешность измерения известно, что $|\eta| < \Delta$, где $\Delta > 0$ – максимальная абсолютная погрешность измерения.

Нас интересует распределение случайной величины ξ, при неизвестном распределении ошибки η.

Теорема 6.1. При сделанных выше предположениях

$$F_\xi(x-\Delta) \le F_{\xi+\eta}(x) \le F_\xi(x+\Delta).$$

Доказательство. Воспользуемся свойством монотонности функции распределения: если $x_1 \le x_2$, то $F(x_1) \le F(x_2)$. Имеем:

$$F_{\xi+\eta}(x) = P\{\xi+\eta < x\} = P\{\xi < x-\eta\} = F_\xi(x-\eta).$$

Так как $x-\eta \ge x-\Delta$, то $F_\xi(x-\eta) \ge F_\xi(x-\Delta)$. Аналогично, так как $x-\eta \le x+\Delta$, то $F_\xi(x-\eta) \le F_\xi(x+\Delta)$. Теорема доказана.

Таким образом, когда наблюдается случайная величина с аддитивной измерительной погрешностью, то в результате может получиться любое распределение в полосе от $F_\xi(x-\Delta)$, до $F_\xi(x+\Delta)$, как показано на рис. 6.1.

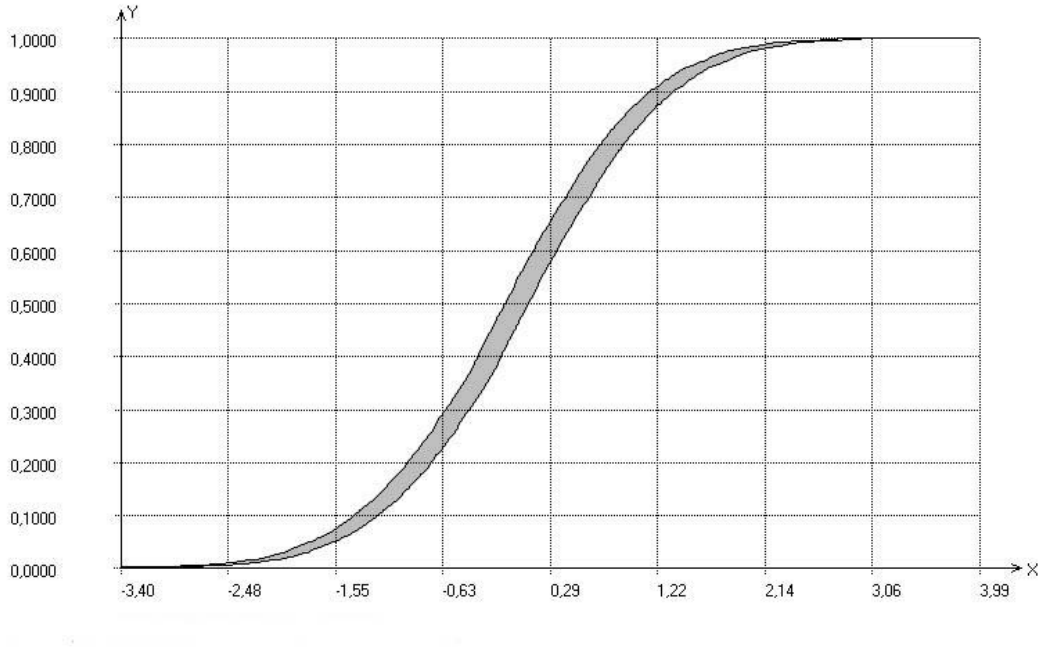

Рис. 6.1. Полоса распределений случайной величины с аддитивной погрешностью

6.2.1. Относительная погрешность

Пусть в результате эксперимента наблюдается случайная величина $\xi(1+\eta)$. Первая случайная величина задаёт статистическую неопределенность, а вторая η – измерительную погрешность, действующую мультипликативно на результат измерения. Про погрешность измерения

известно, что $|\eta| \le \Delta < 1$, где $\Delta > 0$ – максимальная относительная погрешность измерения. Нас интересует распределение случайной величины ξ, при неизвестном распределении ошибки η.

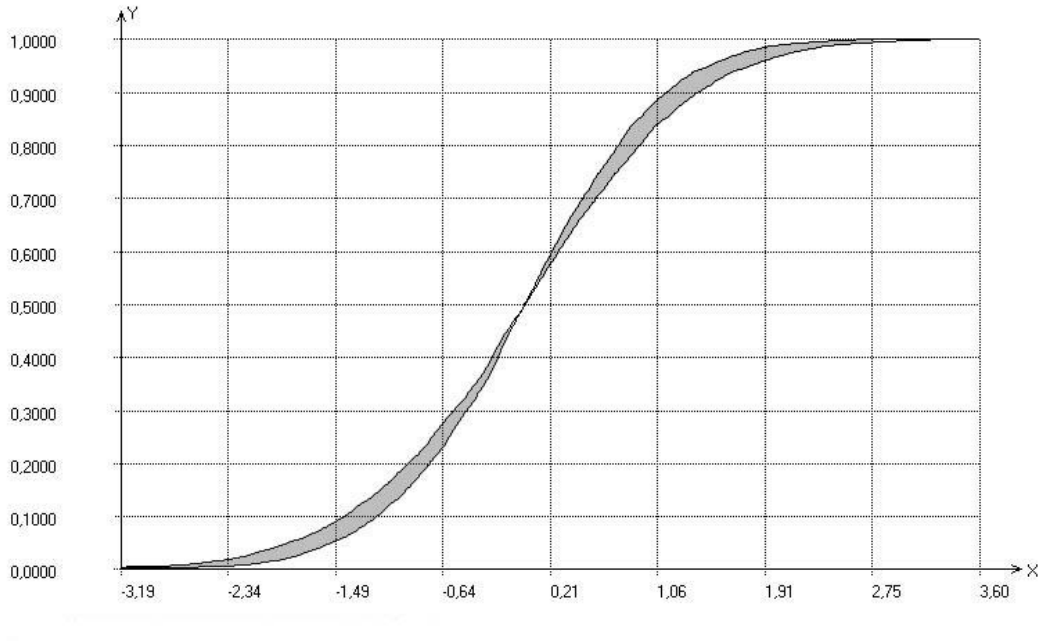

Рис. 6.2. Полоса распределений случайной величины с мультипликативной погрешностью

Теорема 6.2. При сделанных выше предположениях

$$F_\xi\left(\frac{x}{1+\Delta}\right) \le F_{\xi(1+\eta)}(x) \le F_\xi\left(\frac{x}{1-\Delta}\right),\ x > 0;$$

$$F_\xi\left(\frac{x}{1-\Delta}\right) \le F_{\xi(1+\eta)}(x) \le F_\xi\left(\frac{x}{1+\Delta}\right),\ x < 0.$$

Доказательство. Воспользуемся свойством монотонности функции распределения: если $x_1 \le x_2$, то $F(x_1) \le F(x_2)$. Имеем:

$$F_{\xi(1+\eta)}(x) = P\{\xi(1+\eta) < x\} = P\left\{\xi < \frac{x}{1+\eta}\right\} = F_\xi\left(\frac{x}{1+\eta}\right).$$

Пусть $x > 0$. Тогда $\frac{x}{1+\Delta} \le \frac{x}{1+\eta} \le \frac{x}{1-\Delta}$ и, следовательно,

$$F_\xi\left(\frac{x}{1+\Delta}\right) \le F_\xi\left(\frac{x}{1+\eta}\right) \le F_\xi\left(\frac{x}{1-\Delta}\right).$$

Пусть $x < 0$. Тогда $\frac{x}{1-\Delta} \le \frac{x}{1+\eta} \le \frac{x}{1+\Delta}$ и, следовательно,

$$F_\xi\left(\frac{x}{1-\Delta}\right) \le F_\xi\left(\frac{x}{1+\eta}\right) \le F_\xi\left(\frac{x}{1+\Delta}\right).$$

Если же $x = 0$, то неравенства обратятся в равенство. Теорема доказана.

На рисунке 6.2 показана полоса распределений в случае мультипликативной погрешности измерений.

На практике же, скорее всего, будут наблюдаться оба вида погрешностей. Если объединить результаты теорем 6.1 и 6.2, то при максимальной аддитивной погрешности Δ_1 и максимальной мультипликативной погрешности Δ_2 распределение случайной величины $\xi(1+\eta_1)+\eta_2$ будет находиться в полосе, изображенной на рисунке 6.3.

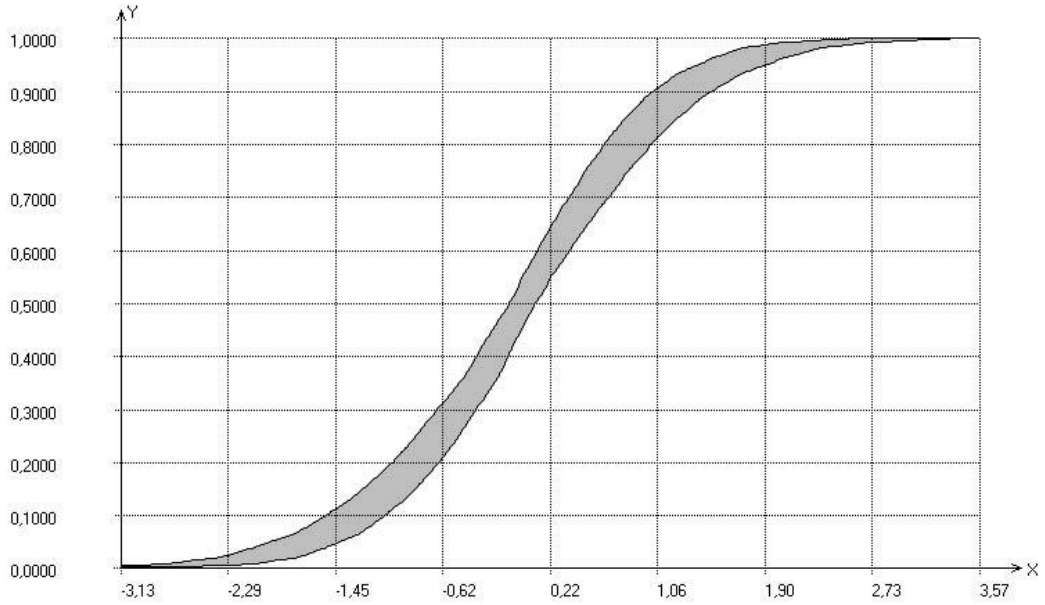

Рис. 6.3. Полоса распределений случайной величины с аддитивной и мультипликативной погрешностями

6.2.3. Интервальные наблюдения

Пусть было произведено измерение x какой-либо случайной величины ξ. Естественно, что в результате измерений допущена погрешность и, на самом деле, $x-$ это реализация случайной величины $\xi(1+\eta_1)+\eta_2$. Причем распределения η_1 и η_2 не только неизвестны, но и могут меняться от эксперимента к эксперименту (например, при смене прибора, которым выполняются измерения). Пусть мы знаем максимально возможные

погрешности Δ_1 и Δ_2 для η_1 и η_2 соответственно. Тогда мы можем утверждать, что при положительном x

$$x \in [x_{ист}(1-\Delta_1)-\Delta_2,\ \ x_{ист}(1+\Delta_1)+\Delta_2],$$

где $x_{ист}$ – истинное значение реализации случайной величины ξ. В то же время можно утверждать, что

$$x_{ист} \in [x\ (1-\Delta_1)-\Delta_2,\ \ x\ (1+\Delta_1)+\Delta_2].$$

Естественно, что эти интервалы пересекаются. Но, если первый интервал для нас неизвестен (т.к. неизвестно $x_{ист}$), то второй интервал известен и позволяет оценить $x_{ист}$ сверху и снизу (см. рис. 6.4).

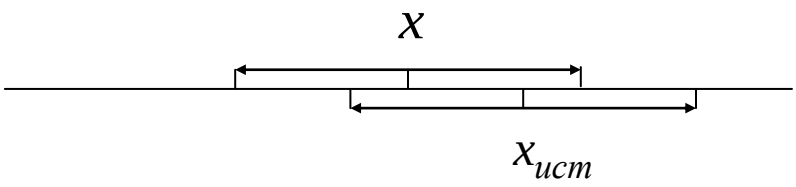

Рис. 6.4. Истинное значение случайной величины и наблюдение

Интервальным наблюдением называется интервал, содержащий значение реализации случайной величины.

Интервальной выборкой объема n называется множество из n интервальных наблюдений:

$$\mathbf{X}_n = \{[a_i, b_i] \in IR \mid\ a_i \le x_i \le b_i,\ a_i \in R,\ b_i \in R,\ \ i = 1,\ldots,n\}$$

К интервальной выборке могут привести процедуры группирования и цензурирования. Отличие заключается в том, что интервалы группирования задаются априори, а в модели с погрешностями измерений границы интервалов порождаются самими наблюдениями и, таким образом, также являются случайными.

Интервалы $[a_i, b_i]$ могут быть бесконечными. Эта ситуация может возникнуть, например:

а) в случае, когда стрелка измерительного прибора зашкаливает и, поэтому установить точное значение границы невозможно;

б) при испытаниях на надежность фиксируется момент выхода прибора из строя. На момент окончания испытаний часть приборов все еще работает, поэтому время их поломки неизвестно.

6.3. Геометрическая интерпретация интервальной выборки

В пространстве R^n выборки, традиционно рассматриваемые в математической статистике $X_n = \{x_i,\ \ i = 1,\ldots,n\}$, представляют собой точку. Будем называть такие выборки *точечными*. Интервальная выборка в

пространстве R^n задает $n-$мерный параллелепипед $[a_1,b_1]\times[a_2,b_2]\times\ldots\times[a_n,b_n]$. Будем говорить, что точечная выборка принадлежит интервальной, $X_n \in \mathbf{X}_n$, если $a_i \le x_i \le b_i,\ i=1,\ldots,n$.

6.4. Эмпирическая функция распределения и гистограмма

Основную информацию о распределении случайной величины ξ исследователь получает по эмпирической функции распределения и гистограмме, на которые опираются статистические методы анализа.

6.4.1. Интервальная гистограмма

Разобьём область определения случайной величины на k интервалов точками $X_0 < X_1 <\ldots< X_k$ (X_0 – левая граница области определения, X_k – правая граница области определения) и подсчитаем число наблюдений, попавших в каждый интервал $(X_j, X_{j+1}]$, $j=1,\ldots,k-1$. Если интервальное наблюдение $[a_i,b_i]$ покрывает точку разбиения X_j (т.е. $X_j \in [a_i,b_i]$), то точечное значение можно отнести как к интервалу $(X_{j-1}, X_j]$, так и к интервалу $(X_j, X_{j+1}]$. Таким образом можно получить 2^p гистограмм, где p – число наблюдений, попавших на границы разбиения. Совокупность всех гистограмм дает нам интервальную гистограмму (см. рис. 6.5).

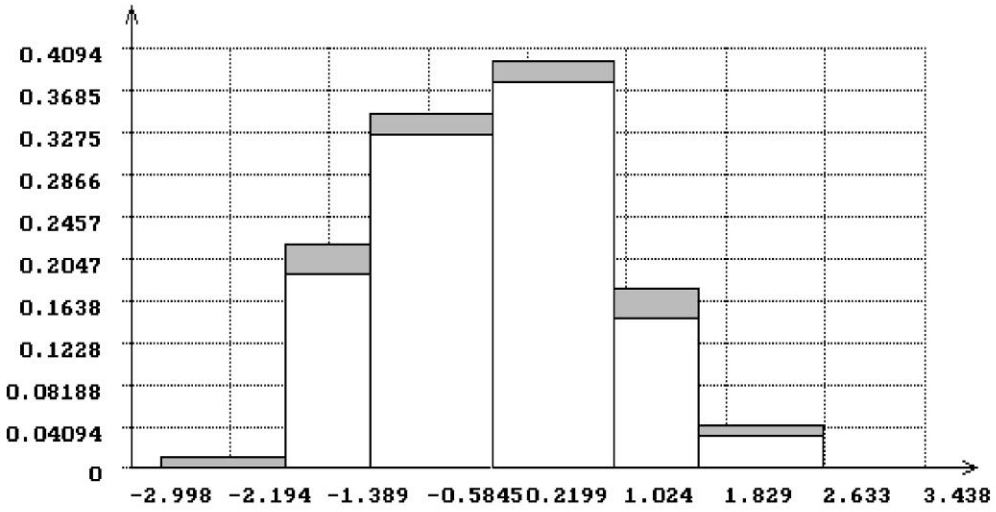

Рис. 6.5. Интервальная гистограмма

6.4.2. Интервальная эмпирическая функция распределения

Упорядочим граничные точки интервалов:

$$a_{(1)} \le a_{(2)} \le \ldots \le a_{(n)} \text{ и } b_{(1)} \le b_{(2)} \le \ldots \le b_{(n)}$$

Предположим, что все точечные наблюдения x_i совпали с левыми границами интервалов. Тогда эмпирическая функция распределения будет иметь вид

$$\overline{F_n}(x) = \begin{cases} 0, & x < a_{(1)}; \\ i/n, & a_{(i)} \le x < a_{(i+1)}, \ i = 1,2,\ldots,n-1; \\ 1, & x \ge a_{(n)}. \end{cases}$$

Аналогично, если все точечные наблюдения x_i совпали с правыми границами интервалов, то эмпирическая функция распределения будет иметь вид

$$\underline{F_n}(x) = \begin{cases} 0, & x < b_{(1)}; \\ i/n, & b_{(i)} \le x < b_{(i+1)}, \ i = 1,2,\ldots,n-1; \\ 1, & x \ge b_{(n)}. \end{cases}$$

Пример 6.1. Участникам статистического эксперимента предлагали оценить свой рост и вес. Были получены следующие интервальные наблюдения (см. таблицу 6.1). Из таблицы хорошо видно, что интервалы неопределенности имеют разную длину, причем, чем выше значение наблюдения, тем больше величина погрешности. Соответствующие интервальные эмпирические функции распределения приведены на рис. 6.6 и 6.7.

Таблица 6.1

Оценки роста и веса группы студентов 5-го курса

№	Рост		Вес	
	оценка снизу	оценка сверху	оценка снизу	оценка сверху
1.	179	181	72	78
2.	180	185	60	70
3.	182	187	80	90
4.	179	182	71	75
5.	155	157	46	48
6.	160	165	51	52,5

7.	174	176	80	90
8.	178	184	85	95
9.	177	181	55	65
10.	174	176	55	60

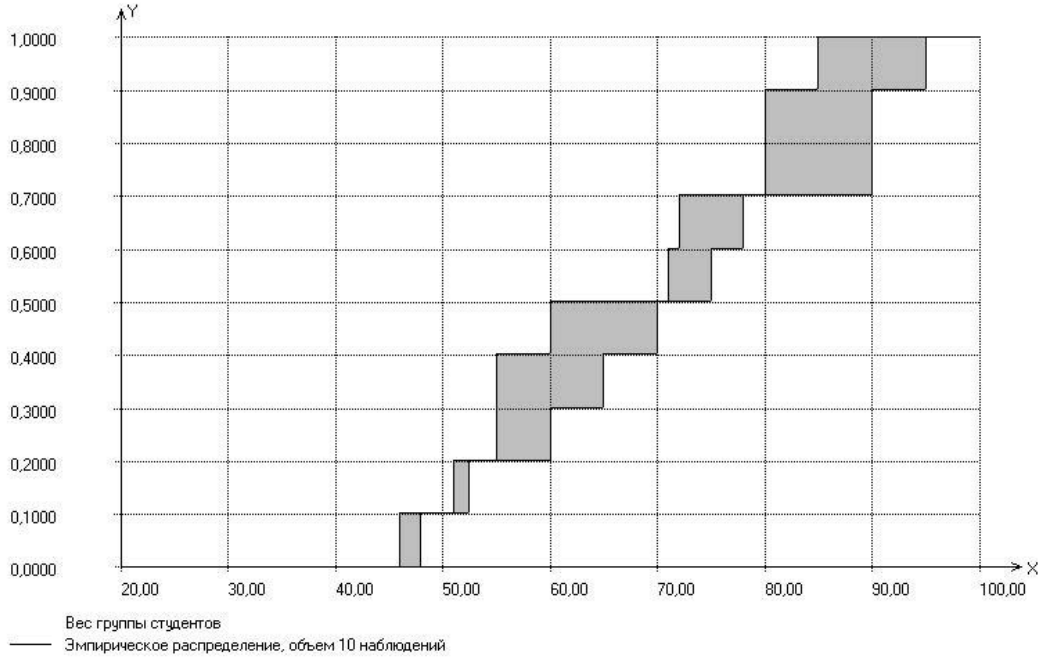

Рис. 6.6. Распределение группы студентов по весу

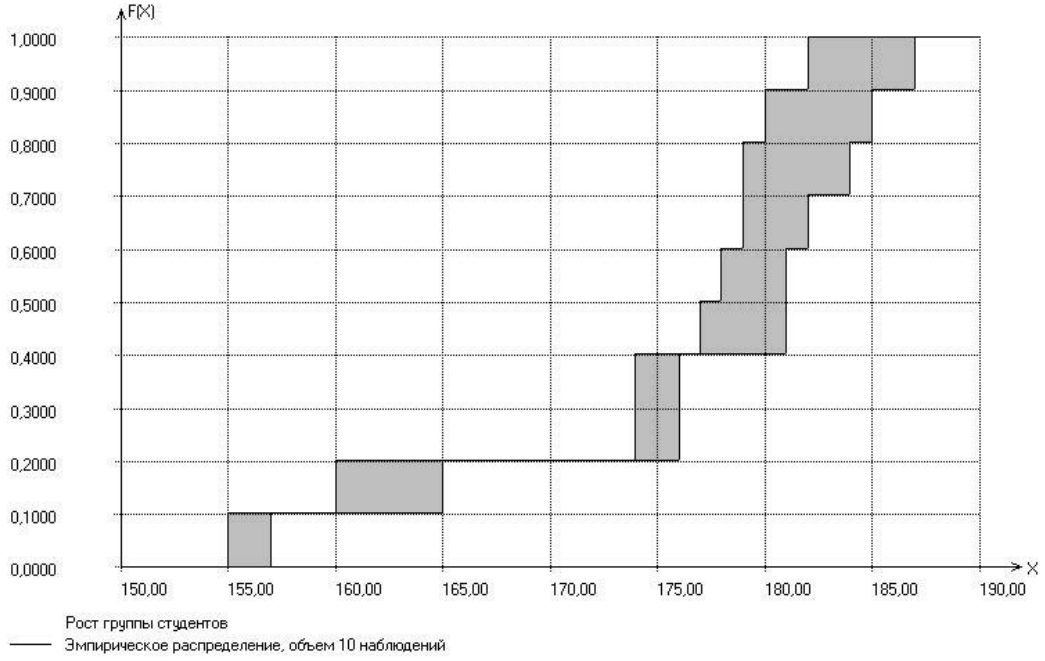

Рис. 6.7. Распределение группы студентов по росту

6.5. Проверка простых гипотез о согласии по интервальной выборке

Пусть дана интервальная выборка $\mathbf{X}_n$. Тогда мы можем определить границы для статистики критерия:

$$\underline{S_n}(\mathbf{X}_n, F) = \inf_{X_n \in \mathbf{X}_n} S(X_n, F) \le S_n(\mathbf{X}_n, F) \le \sup_{X_n \in \mathbf{X}_n} S(X_n, F) = \overline{S_n}(\mathbf{X}_n, F).$$

Тогда достигаемый уровень значимости будет лежать в интервале $[p_{\min}, p_{\max}]$, где $p_{\min} = \int_{\overline{S_n}(X_n,F)}^{+\infty} g(S|H_0)dS$, $p_{\max} = \int_{\underline{S_n}(X_n,F)}^{+\infty} g(S|H_0)dS$

В результате проверки гипотезы о согласии можно сделать следующие выводы:

- $p_{\max} < \alpha$ – гипотеза H_0 отвергается;
- $p_{\min} > \alpha$ – гипотеза H_0 не отвергается;
- $p_{\min} \le \alpha \le p_{\max}$ – гипотеза H_0 может быть либо отвергнута, либо не отвергнута (зона нечувствительности критерия).

В последнем случае возможны разные варианты принятия решения: можно задать степень доверия исследователя к наблюдаемым данным, выполнить процедуру рандомизации... Однако из описываемой далее теоремы об асимптотических свойствах границ статистики Колмогорова по интервальной выборке следует, что для истинной модели и для любой модели, близкой к истинной в пределах погрешности измерений, интервал $[p_{\min}, p_{\max}]$ стремится к интервалу $[0, 1]$ с ростом объема выборки. Таким образом, если считать, что $p = p_{\min}$, то рано или поздно истинная гипотеза будет отвергнута, а если считать, что $p = p_{\max}$, то рано или поздно может быть принята любая близкая конкурирующая гипотеза.

6.5.1. Критерий согласия Колмогорова

Для статистики Колмогорова

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

найдем верхнюю и нижнюю границу

$$\underline{F_n}(x) \le F_n(x) \le \overline{F_n}(x).$$

Отсюда

$$\underline{F_n}(x) - F(x) \le F_n(x) - F(x) \le \overline{F_n}(x) - F(x),$$
$$F(x) - \overline{F_n}(x) \le F(x) - F_n(x) \le F(x) - \underline{F_n}(x).$$

Эти неравенства выполняются для всех x, следовательно, они будут выполняться и для супремумов:

$$\sup_x(\underline{F_n}(x)-F(x)) \le \sup_x(F_n(x)-F(x)) \le \sup_x(\overline{F_n}(x)-F(x)),$$

$$\sup_x(F(x)-\overline{F_n}(x)) \le \sup_x(F(x)-F_n(x)) \le \sup_x(F(x)-\underline{F_n}(x)).$$

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

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

$$\le D_n = \max\ \sup_x(F_n(x)-F(x)), \sup_x(F(x)-F_n(x))\ \le \qquad (6.1)$$

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

6.5.2. Асимптотические свойства критерия Колмогорова по интервальной выборке

Естественно, что, чем меньше длина интервала $[p_{\min}, p_{\max}]$, тем более определенные выводы можно сделать.

На величину $\Delta p = p_{\max} - p_{\min}$ в случае верной основной гипотезы H_0 влияют:

* диаметр множества $\mathbf{X}_n$ ($d(\mathbf{X}_n)=0 \Rightarrow \Delta p = 0$);
* закон распределения $F(x)$;
* критерий согласия;
* количество наблюдений.

Теорема 6.3. [58] Пусть дана последовательность интервальных выборок $\{\mathbf{X}_n, n=1,2,\ldots\}$ и $\exists \underline{F}(x) \ne \overline{F}(x)$: $\forall \varepsilon > 0$

$$P\left\{\sup_x\left|\underline{F_n}(x)-\underline{F}(x)\right| > \varepsilon\right\} = O(1/n),$$

$$P\left\{\sup_x\left|\overline{F_n}(x)-\overline{F}(x)\right| > \varepsilon\right\} = O(1/n).$$

Тогда, если $\forall x\ \underline{F}(x) \le F(x) \le \overline{F}(x)$, то $\Delta p \to 1$, иначе $\Delta p \to 0$.

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

$$P\{S > S^*\} = 1 - K(\sqrt{S^*/2}),$$

где $K(y) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2y^2}$ – функция распределения Колмогорова.

Для оценок границ $\underline{D_n}$ и $\overline{D_n}$ статистики D_n имеем:

$$p_{\min} = 1 - K\left(\frac{6n\overline{D_n} + 1}{6\sqrt{n}}\right), \quad p_{\max} = 1 - K\left(\frac{6n\underline{D_n} + 1}{6\sqrt{n}}\right)$$

Тогда $\Delta p \to 1$, если $p_{\max} \to 1$, $p_{\min} \to 0$; и $\Delta p \to 0$, если $p_{\max} \to 0$.

В свою очередь, $p_{\min} \to 0$, если статистика $\overline{D_n}$ не будет стремиться к нулю; $p_{\max} \to 0$, если статистика $\underline{D_n}$ также не будет стремиться к нулю; и $p_{\max} \to 1$, если $\underline{D_n}$ стремится к нулю со скоростью $O(1/n)$.

Рассмотрим теперь два случая: 1) $\forall x \quad \underline{F}(x) \le F(x) \le \overline{F}(x)$ и 2) $\exists x_0 \, F(x_0) \notin [\underline{F}(x_0), \overline{F}(x)]$.

1) Пусть $\forall x \quad \underline{F}(x) \le F(x) \le \overline{F}(x)$. Нижняя граница статистики вычисляется по формуле

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

Если неравенство строгое, $\underline{F}(x) < F(x) < \overline{F}(x)$, то первые две величины в фигурных скобках, становятся отрицательными с вероятностью 1 при достаточно большом n, и поэтому максимум будет равен нулю. Если же $F(x)$ совпадает с $\underline{F}(x)$ или $\overline{F}(x)$, то, сделав соответствующую замену, мы получим, что $\forall \varepsilon > 0 \quad P\{\underline{D_n} > \varepsilon\} = O(1/n)$. Таким образом, мы доказали, что верхняя граница достигаемого уровня значимости (вероятности согласия) стремится к единице.

Верхняя граница статистики вычисляется по формуле

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

Возьмем любую точку x_0, в которой $\overline{F}(x_0) - \underline{F}(x_0) = c > 0$. Тогда $P\{\overline{D_n} > c/2\} \to 1$. Значит $p_{\min} \to 0$ и $\Delta p \to 1$.

2) Пусть x_0 – точка, в которой $F(x) > \overline{F}(x)$ (аналогично можно рассмотреть случай, когда $F(x) < \underline{F}(x)$). Обозначим $d = F(x_0) - \overline{F}(x_0)$.

Тогда

$$P\{\underline{D_n} \ge d/2\} = P\left\{\max\left\{\sup_x(\underline{F_n}(x) - F(x)), \sup_x(F(x) - \overline{F_n}(x)), 0\right\} \ge d/2\right\} \ge$$

$$\geq P\{F(x_0)-\overline{F_n}(x_0)\geq d/2\}=P\{F(x_0)-\overline{F}(x_0)+\overline{F}(x_0)-\overline{F_n}(x_0)\geq d/2\}=$$

$$=P\{d+\overline{F}(x_0)-\overline{F_n}(x_0)\geq d/2\}=P\{\overline{F}(x_0)-\overline{F_n}(x_0)\geq -d/2\}=$$

$$=1-P\{\overline{F}(x_0)-\overline{F_n}(x_0)<-d/2\}\geq 1-P\ \sup_x\left|\overline{F}(x)-\overline{F_n}(x)\right|>d/2\ =$$

$$=1-O(1/n).$$

Следовательно, $p_{\max}\to 0$ и $\Delta p\to 0$. Теорема доказана.

Таким образом, с ростом объема выборки зона нечувствительности критерия растёт.

6.6. Экспериментальное исследование критериев согласия по интервальным наблюдениям

Очевидно, что при интервальном характере наблюдений распределения статистик критериев согласия, рассмотренных в главе 4, также будут представлять собой интервальные распределения. Найти и идентифицировать их можно, используя методику компьютерного моделирования статистических закономерностей.

5. Моделируется N интервальных выборок по n наблюдений в каждой в соответствии с гипотезой H_0 и фиксированными значениями параметров θ.
6. По каждой интервальной выборке $\mathbf{X}_n$ вычисляется пара статистик – $\underline{S}$ и $\overline{S}$ (например, для критерия Колмогорова по (6.1)).
7. По выборкам статистик $\underline{S}$ и $\overline{S}$ идентифицируется верхняя и нижняя граница распределения статистики критерия $\underline{G}(S\,|\,H_0)$ и $\overline{G}(S\,|\,H_0)$.
8. Исследуется зависимость $\underline{G}(S\,|\,H_0)$ и $\overline{G}(S\,|\,H_0)$ от объема выборки n, от величины абсолютной и относительной погрешности.

В случае проверки сложной гипотезы о согласии в зависимости от вида регистрируемых наблюдений возможно применение двух видов оценок – точечных и интервальных [59], свойства которых также можно исследовать по предложенной методике.

Контрольные вопросы и задачи

1. Пусть $A=[1,3], B=[2,4]$. Найти $A+B, A-B, AB, A/B$.

2. Вывести формулы для вычисления арифметических операций над интервальными числами через арифметические операции над вещественными числами и операции максимума и минимума.
3. Доказать свойство субдистрибутивности интервальных чисел. Привести пример, когда нарушается дистрибутивность интервальных чисел.
4. Что такое абсолютная и относительная погрешность?
5. Что такое интервальное наблюдение и интервальная выборка?
6. Как влияет интервальная неопределенность данных на статистические выводы при проверке гипотезы о согласии?
7. Как можно смоделировать интервальную выборку?
8. Опишите процедуру моделирования распределений статистики критерия согласия по интервальным данным.
9. Опишите процедуру исследования свойств оценок параметров по интервальным данным.

Глава 7. Программная система статистического анализа одномерных наблюдений ISW

Программная система «ISW» версия 4.0 является следующим поколением системы статистического анализа одномерных наблюдений [60-62].

7.1. Возможности системы

Программная система обладает рядом достоинств, которые выгодно отличают ее от конкурирующих систем.

1. Широкий выбор моделей теоретических законов распределения, включает более 30 стандартных законов и распределений, получаемых с помощью операций над этими стандартными моделями: операций сдвига, масштаба, смеси законов, произведения, усечения, логарифмирования.
2. Универсальное представление входных данных и возможность обрабатывать негруппированные (точечные), группированные, цензурированные, частично группированные и интервальные выборки.
3. Группирование наблюдений в задачах робастного оценивания и проверки статистических гипотез может осуществляться четырьмя различными способами: в соответствии с асимптотически оптимальным (минимизирует потери в информации Фишера), равновероятным, равночастотным и равномерным группированием.
4. Для проверки согласия эмпирического распределения с теоретическим используется восемь критериев: отношения правдоподобия, χ^2 Пирсона, χ^2 Пирсона с поправкой Никулина, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса, Реньи. На базе результатов авторов программной системы корректность применения критериев согласия гарантируется как при проверке простых, так и при проверке сложных гипотез.
5. Оценивание параметров может осуществляться различными методами: максимального правдоподобия, максимального правдоподобия с предварительной группировкой наблюдений, MD-оценивания с минимизаций расстояний, измеряемых статистиками типа Колмогорова, статистиками типа ω^2 и Ω^2 Мизеса, с использованием предложенных авторами оптимальных L-оценок по выборочным квантилям.
6. На базе включенных в систему робастных методов оценивания реализована эффективная параметрическая процедура отбраковки аномальных наблюдений.
7. Графическая подсистема позволяет просматривать функции распределения, плотности, гистограммы, ядерные оценки плотности.
8. Разработаны средства для моделирования распределений статистик критериев согласия при различных сложных гипотезах и различных альтернативах. Это позволяет исследовать распределения статистик при

различных сложных проверяемых гипотезах, строить приближенные математические модели этих распределений, исследовать мощность критериев относительно различных близких альтернатив.
9. На базе системы возможна организация исследований законов распределений различных одномерных статистик, вычисляемых при статистическом анализе наблюдений.

7.2. Настройка параметров системы

Параметры системы можно задать как в режиме диалога (Кнопка  на панели инструментов), так и в файле инициализации «is.ini».

7.2.1. Структура файла инициализации «is.ini»

В файле содержатся ключевые слова разделов, команды инициализации и комментарии.

- Ключевые слова разделов

[Distributions]
<Список распределений>
[Samples]
<Список выборок>
[Options]
<Параметры>
[Job]
<Задание на выполнение>

- 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 не более чем из 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.

- Комментарии

Комментарием считается любая строчка, начинающаяся с символа "*" или "//".

7.2.2. Разделы

7.2.2.1. Раздел [Distributions]

В этом разделе происходит инициализация списка распределений. Распределение инициализируется командой

<распределение> = {D0 | D1 | D2 | ... | D38} ([<список параметров>]),

где D0, D1, ... , D38 – это зарезервированные в системе идентификаторы распределений. Список возможных законов приведен в таблице 7.1.

Таблица 7.1

Законы распределений встроенных в программную систему

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	Модуля *n*-мерного нормального распределения	0
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 мин. 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
D30	SIN	Синуса	0
D31	SQR	Квадратичное	0
D32	KOLM	Распределение статистики Колмогорова	0
D33	KOLM_CENS	Распределение статистики Колмогорова по цензурированной выборке	0
D34	RENI	Распределение статистики Реньи	0
D35	SMIR	Распределение статистики Смирнова	0
D36	A1	Распределение статистики ω^2	0
D37	A2	Распределение статистики Ω^2	0
D38	DStudent	Распределение Стьюдента	1

Вместо идентификаторов "Dxx" можно использовать их синонимы, указанные в таблице.

Над стандартными распределениями можно применять операции преобразования:

<распределение> = <операция> (<список распределений>, [<параметр>])

где <операция> = {Shift | Scale | Reflection | Left | Right | Mixt | Mult}
Список возможных операций приведен в таблице 7.2.

Таблица 7.2

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

Примечание 1. Новые распределения рекомендуется обозначать строчными буквами, чтобы они отличались от стандартных.
Примечание 2. При наборе нужно учитывать регистр, т.е., например, нельзя набирать «MIXT» вместо «Mixt».
Примечание 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.

Пример 7.1.

```
[Distributions]
d1=D9() // инициализируем d1 стандартным нормальным распределением
d2=Scale(d1,2) // добавляем параметр масштаба, равный 2
d3=Shift(d2,1) // добавляем параметр сдвига, равный 1
d4=Shift(Scale(D9(),2),1) // то же, что и d3, но в одной строке
d5=Shift(Scale(D9())) // то же, что и d4, но параметры неизвестны
d6=Mixt(d4,d5,0.1) // смесь двух нормальных распределения с параметром смеси 0.1 (параметр смеси задает долю распределения d4)
```

7.2.2.2 Раздел [Samples]

В этом разделе происходит инициализация списка выборок. Выборка инициализируется командой
<выборка> = <имя файла>

Имя файла не должно содержать знаки препинания и круглые скобки.

7.2.2.3 Раздел [Options]

В разделе [Options] задаются основные параметры системы. В таблице 7.3 приведены основные опции системы.

Таблица 7.3

Опции программной системы

Константа	*Допустимые значения*	*Содержание*
EstimateMethod	OMP	Метод максимального правдоподобия
	KOLM	Метод минимума статистики Колмогорова
	MISES	Метод минимума статистики ω^2 Мизеса
	MISES_B	Метод минимума статистики Ω^2 Мизеса
	QUANTIL	L-оценки
Robust	ON | OFF	Группирование перед оцениванием по методу максимального правдоподобия
NuclearEstimate	ON | OFF	Применять непараметрическое оценивание
NuclearFunction	D9() | D31()	Тип ядерных функций

W0	0 | 1	Использовать критерий отношения правдоподобия
W1	0 | 1	Использовать критерий Хи-квадрат Пирсона
W2	0 | 1	Использовать критерий Колмогорова
W3	0 | 1	Использовать критерий Смирнова
W4	0 | 1	Использовать критерий ω^2 Мизеса
W5	0 | 1	Использовать критерий Ω^2 Мизеса
W6	0 | 1	Использовать критерий Реньи
Nik	ON | OFF	Поправка Никулина
GenTestDistr	ON | OFF	Моделирование распределения статистик (использовать метод Монте-Карло при проверке гипотезы о согласии)
NumberSamples	10 .. 2000	Число выборок
MaxSampleSize	100 .. 2000	Максимальный объем выборки
SignLevel	0 .. 1	Уровень значимости критерия
SearchMeth	RND	Случайный поиск
	DIR	Покоординатный спуск
	MHG	Метод Хука-Дживса
	SG1	Метод сопряженных градиентов (Флетчера-Ривса)
	SG2	Метод сопряженных градиентов (Пшеничного)
	MGS	Метод градиентного спуска
EpsilonVariable	10^{-10} .. 10^{-2}	Точность поиска по переменным
EpsilonFunction	10^{-10} .. 10^{-2}	Точность поиска по функции
MaxIteration	50 .. 1000	Максимальное число итераций
Newton	ON | OFF	Использовать метод Ньютона при одномерном поиске
Approximate	ON | OFF	Аппроксимировать производные функций конечными разностями
Trace	0	Не делать запись сообщений в журнал
	1	кратко
	2	подробно
	3	очень подробно
	4	Выводить все сообщения в журнал
GroupNumb	5..50	Число интервалов
GroupType	GR_AOG	Асимптотически оптимальное
	GR_EPG	Равновероятное
	GR_EFG	Равночастотное
	GR_EG0	Равномерное

	GR_MIN	Минимальное (область определения разбивается на интервалы так, чтобы в каждом интервале было только одно наблюдение)
GrLeft	Число	Граница слева при равномерном группировании
GrRight	Число	Граница справа при равномерном группировании
GrDelta	Число	Длина интервала при равномерном группировании

7.2.2.4 Раздел [Job]

В разделе [Job] содержится задание на выполнение. Список доступных команд приведен в таблице 7.4.

Таблица 7.4

Команды пакетного задания

Команда	*Содержание*
Set(<опция>,<значение>)	Установить опцию системы
Estimate(<выборка>, <распределение>)	Оценить параметры
Test(<выборка>, <распределение>)	Проверить согласие
Estimate(<выборка>, <распределение>)	Оценить параметры
Test(<выборка>, <распределение>)	Проверить согласие
Ident()	Идентификация всех выборок и распределений
IdentSample(<выборка>)	Идентификация выборки
IdentDistr(<распределение>)	Идентификация распределения
Ident(<выборка>, <распределение>)	Оценить параметры и проверить согласие
Anomalous(<выборка>, <распределение>)	Выделить аномальные наблюдения
Shell()	Запуск оболочки
Exit()	Выход

Пример 7.2.

```
[Job]
Set(EstimateMethod,KOLM) // Устанавливаем метод оценивания
Estimate(s1, d5) // Оценить параметры нормального распределения
```

Test(s1, d5) // Проверить согласие
Anomalous(s1, d5) // Выделить аномальные наблюдения

7.2.3. Настройка параметров в режиме диалога

Для вызова окна настройки параметров можно нажать кнопку на панели инструментов или в меню “Действия” выбрать “Параметры” (см. рис. 7.1).

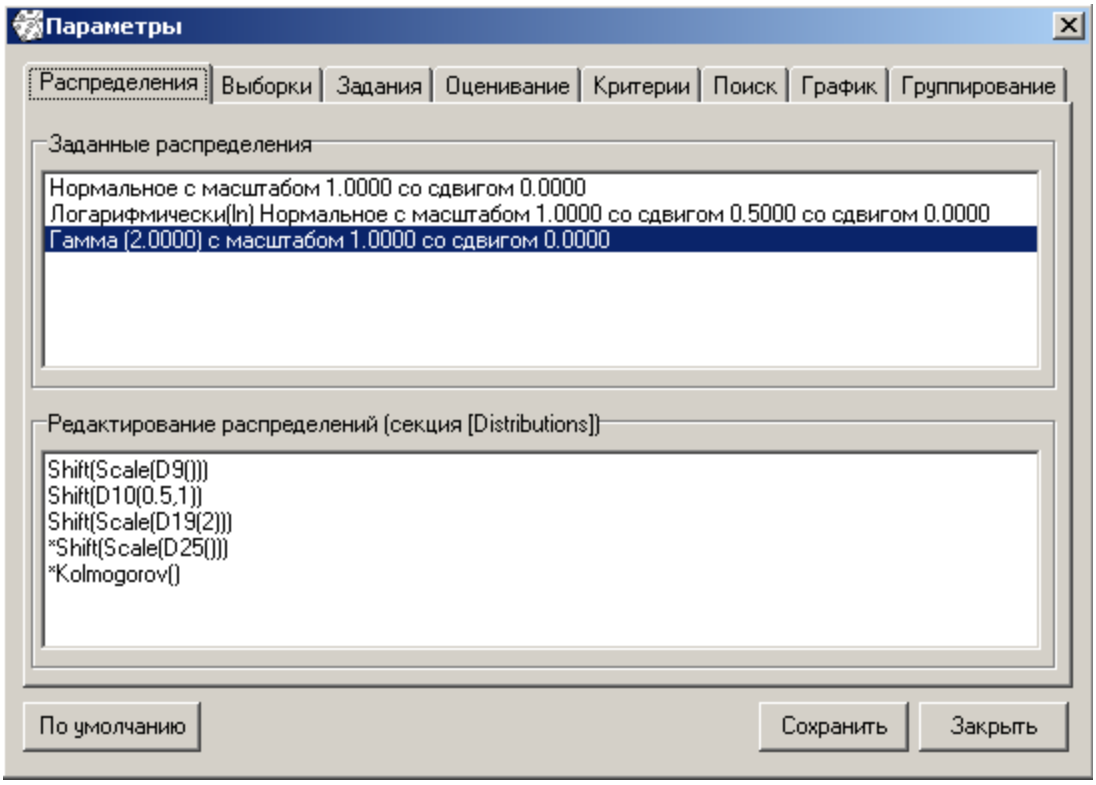

Рис. 7.1. Настройка параметров системы в режиме диалога

В форме "Параметры" основные параметры системы распределены по закладкам. Закладки "Распределения", "Выборки" и "Задания" соответствуют разделам [Distributions], [Samples] и [Jobs] файла «is.ini». Кнопка "По умолчанию" восстанавливает исходные значения параметров по умолчанию. Кнопка "Сохранить" сохраняет сделанные изменения в файле «is.ini».

На закладке "Оценивание" выбирается метод оценивания параметров распределения.

На закладке "Критерии" можно задать критерии согласия, по которым будет проверяться согласие выборочного распределения с теоретическим законом. Уровень значимости определяет вероятность ошибки первого рода (вероятность отвергнуть истинную гипотезу). Проверяться может как простая (когда параметры законов распределений не оцениваются), так и сложная

гипотеза о согласии (когда перед проверкой согласия находят оценки параметров распределения по этой же выборке). Если поставить флажок "Использовать метод Монте-Карло", то при вычислении вероятности согласия (достигаемого уровня значимости) будет проводиться моделирование выборок по основной гипотезе и подсчет числа случаев, когда статистика критерия была меньше либо равна значению статистики по проверяемой выборке. Число выборок и объем задаются в соответствующих полях формы.

На закладке "Поиск" задается метод поиска, используемый при нахождении оценок параметров распределений.

На закладках “График” и “Группирование”, соответственно, настраиваются параметры графиков и указываются тип группирования и число интервалов.

7.3. Формат входных данных

Выборка с наблюдениями хранится в текстовом формате (файл с расширением "dat"). Структура файла зависит от типа выборки.

- Точечная выборка

Точечная выборка объемом *n* наблюдений имеет следующий формат:
<название выборки>
0 n
<наблюдение 1>
<наблюдение 2>
...
<наблюдение n>

- Интервальная выборка с абсолютной и относительной погрешностью

Интервальная выборка объемом *n* наблюдений с абсолютной погрешностью *a* и относительной погрешностью *r* имеет следующий формат:

<название выборки>
1 n a r
<наблюдение 1>
<наблюдение 2>
...
<наблюдение n>

- 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

- 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>
где <n_i> - количество наблюдений в i-м интервале и <x_i> - i-я граничная точка

В системе предусмотрена возможность группирования точечной выборки. Кнопка  на панели инструментов выполняет группирование выборки одним из методов: асимптотически оптимальным, равновероятным, равночастотным, равномерным или минимальным.

- Цензурированная слева выборка I-го типа

Цензурированная выборка из *n* точечных наблюдений и интервала цензурирования слева имеет формат:
<название выборки>
4 n
<n_c>
<x_c>
<наблюдение 1>
<наблюдение 2>
...
<наблюдение n>
где <n_c> - количество наблюдений в интервале цензурирования
и <x_c> - точка цензурирования

- Цензурированная справа выборка I-го типа

Цензурированная выборка из *n* точечных наблюдений и интервала цензурирования справа имеет формат:
<название выборки>

5 n
<n_c>
<x_c>
<наблюдение 1>
<наблюдение 2>
...
<наблюдение n>
где <n_c> - количество наблюдений в интервале цензурирования
и <x_c> - точка цензурирования

- Цензурированная с двух сторон выборка I-го типа

Цензурированная выборка из *n* точечных наблюдений и интервалов цензурирования слева и справа имеет формат:
<название выборки>
6 n
<n_l><n_r>
<x_l><x_r>
<наблюдение 1>
<наблюдение 2>
...
<наблюдение n>
где <n_l> - количество наблюдений в интервале цензурирования слева
и <n_r> - количество наблюдений в интервале цензурирования справа
и <x_l> - точка цензурирования слева
и <x_r> - точка цензурирования справа

- Интервальная выборка

Интервальная выборка из *n* интервальных наблюдений

<название выборки>
10
<a_1> <b_1>
<a_2> <b_2>
<a_3> <b_3>
….
<a_n> <b_n>

где <a_i> - левая граница интервального наблюдения
и <b_i> - правая граница интервального наблюдения.

7.4. Статистический анализ

Статистический анализ выборки производится в форме "Оценивание параметров и проверка согласия" (кнопка F→F̂ на панели инструментов), как показано на рис. 7.2. Необходимо выбрать выборку, закон распределения, метод оценивания и критерии согласия.

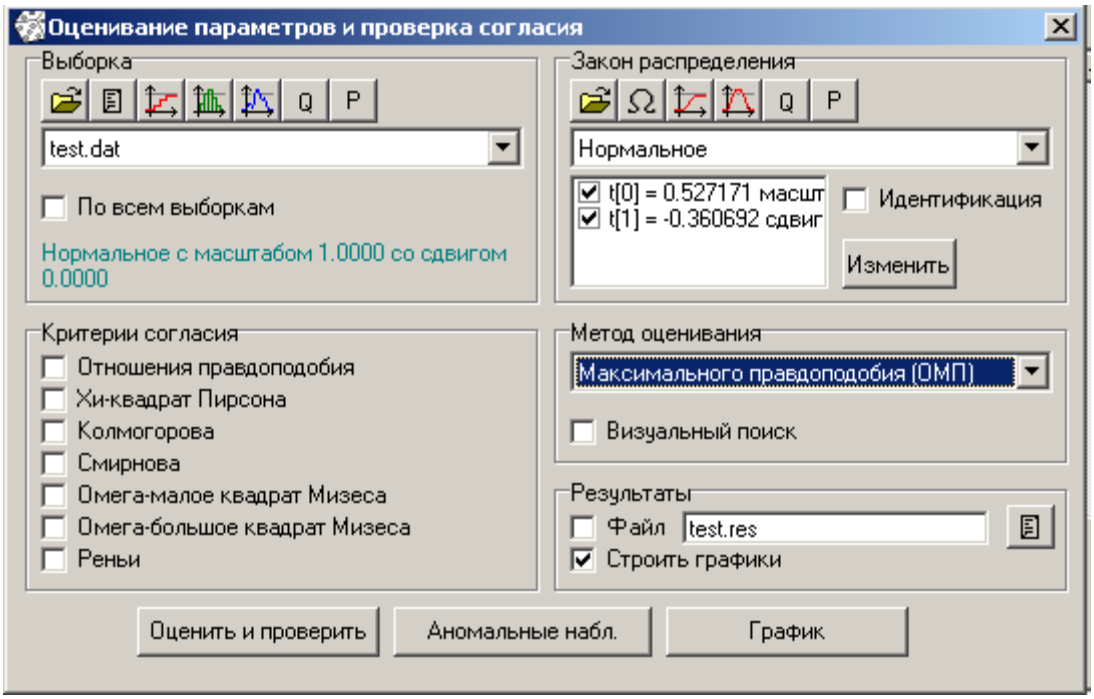

Рис. 7.2. Форма «Статистический анализ»

- Выборка

Выборку можно выбрать из списка, либо открыть файл с выборкой. В списке отображаются только те выборки, которые перечислены в разделе [Samples] в файле инициализации «is.ini». Здесь можно просмотреть саму выборку, эмпирическую функцию распределения по этой выборке, гистограмму (если выборка группированная), ядерную оценку плотности; а также вычислить выборочные квантили по заданным вероятностям - кнопка Q или по заданным точкам вычислить частоты (т.е. отношение количества наблюдений, попавших левее точки к объему выборки) – кнопка P.

- Закон распределения

В системе заложено более 30 стандартных распределений и возможность добавлять новые распределения, получаемые из стандартных с помощью операций сдвига, масштабирования, смеси, произведения, зеркального отображения, усечения.

В списке отображаются те распределения, которые перечислены в разделе [Distributions] в файле инициализации «is.ini». Можно открыть другой (подготовленный ранее) список распределений, он задается в файле

с расширением «dst». В форме "Параметры распределений" (кнопка Ω) выдается информация о распределениях списка: идентификатор, наименование, тип, область определения, граница слева, граница справа, число параметров, параметры и их значения. Здесь также предусмотрена возможность просмотра графиков функции распределения и функции плотности, а также возможность вычисления квантилей распределения по заданным вероятностям - кнопка Q и по заданным точкам *x* вычисления вероятностей P{x<X} – кнопка P.

Кнопка "График" выводит функцию распределения выбранного закона и эмпирическую функцию распределения выбранной выборки на одном рисунке.

- Оценка параметров и проверка гипотез

При нажатии кнопки "Оценить и проверить" производится поиск оценок параметров закона распределения выбранным методом оценивания и выполняется проверка согласия выбранной выборки с выбранным законом распределения. При этом вычисляются оценки тех параметров, напротив которых стоит флажок. Если не выбран ни один из критериев согласия, то производится только оценивание параметров. Проверяется простая гипотеза, если ни один из параметров не оценивается.

Если стоит флажок "По всем выборкам", то действия будут выполняться последовательно по каждой выборке. Если стоит флажок "Идентификация", то по совокупности критериев согласия будет найден наилучший закон (из тех распределений, которые представлены в списке), описывающий конкретную выборку.

- Аномальные наблюдения

При нажатии кнопки "Аномальные наблюдения" производится отбраковка аномальных наблюдений по выбранному закону распределения.

7.5. Графики

На один рисунок можно вывести:

- графики эмпирических функций распределения по всем выборкам, перечисленным в разделе [Samples] инициализационного файла «is.ini» – кнопка на панели инструментов (или в меню “Графики” выбрать “Все выборки”).
- графики всех функций распределения, перечисленных в разделе [Distributions] файла is.ini – кнопка на панели инструментов (или в меню “Графики” выбрать “Все распределения”).
- графики эмпирических функций распределения по всем выборкам, перечисленным в разделе [Samples] и графики всех функций

распределения, перечисленных в разделе [Distributions] файла is.ini – кнопка на панели инструментов (или в меню “Графики” выбрать “Все графики”).

На рис. 7.3 показан пример формирования графика – функция распределения статистики Колмогорова.

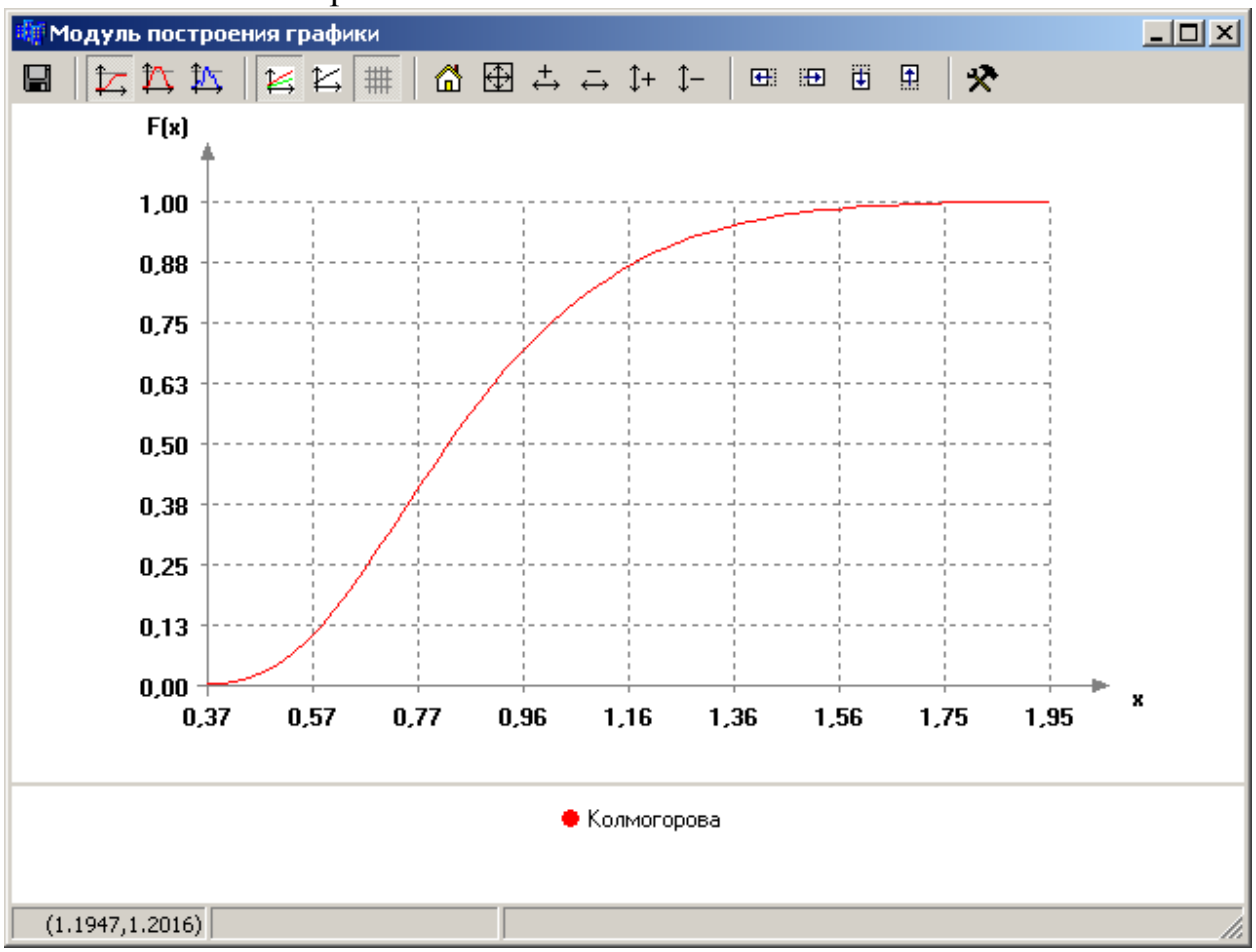

Рис. 7.3. Модуль построения графики

В окне “График” можно менять настройки графика:

– выводить/не выводить сетку
– изменить палитру
– задать границы по X и Y
– сжать график по горизонтали
– растянуть по горизонтали
– растянуть по вертикали
– сжать по вертикали
– сохранить рисунок в формате bmp или jpg
– отобразить график(и) функции распределения
– отобразить график(и) функции плотности
– ядерная оценка плотности распределения

7.6. Моделирование

7.6.1. Создание новой выборки

Для создания новой выборки необходимо выбрать в меню “Моделирование” пункт “Выборка”. Тип создаваемой выборки задается выбором одной из четырех закладок: "Точечная", "Группированная", "Интервальная", "Цензурированная".

Для моделирования выборки задается закон распределения, которому подчиняется выборка и количество наблюдений. Полученная выборка записывается в файле в текстовом формате. Начальное значение генератора случайных чисел позволяет получать одинаковые выборки (т.е. чтобы получить ту же самую выборку второй раз, можно просто запомнить начальное значение ГСЧ).

Для моделирования группированной выборки необходимо задать интервалы группирования. Количество и граничные точки можно вводить вручную, либо используя процедуру асимптотического группирования (кнопка "АОГ") равновероятного группирования (кнопка "РВГ") либо равномерного группирования (кнопка "РГ").

При создании цензурированной выборки определяется вид группирования (слева, справа или с обеих сторон), тип группирования (первый или второй) и, в зависимости от типа группирования – количество наблюдений в интервалах, или точки цензурирования.

7.6.2. Моделирование распределений оценок параметров

В системе заложена возможность моделирования статистических закономерностей методом Монте-Карло. Форма "Моделирование распределений оценок параметров" (в меню “Моделирование” выбрать “Распределения оценок параметров”) позволяет сгенерировать распределения оценок параметров по всем методам оценивания, имеющимся в системе. Для моделирования задается закон распределения, параметры которого оцениваются, флажками отмечаются параметры, которые нужно оценивать, количество выборок, объемы выборок, наблюдаемая часть (если меньше 100%, то производится цензурирование), начальное значение генератора случайных чисел.

7.6.3. Моделирование распределений статистик критериев согласия

Форма "Моделирование распределений статистик критериев согласия" (для запуска в меню “Моделирование” выбрать “Распределения статистик критериев”) позволяет сгенерировать распределения статистик критериев

согласия. Для моделирования задается закон распределения при верной нулевой гипотезе, закон распределения при верной альтернативной гипотезе. Флажками отмечаются параметры, которые нужно оценивать, количество выборок, объемы выборок, начальное значение генератора случайных чисел, верная гипотеза (т.е. закон, в соответствии с которым моделируются выборки). Для критериев типа χ^2 задается число интервалов группирования и тип группирования. Кнопка "H–>H0" находит параметры распределения H1, наиболее близкие к распределению H0. Кнопка "H–>H1" находит параметры распределения H0, наиболее близкие к распределению H1.

Контрольные вопросы и задачи

1. Какими достоинствами обладает программная система ISW 4.0?
2. Из каких разделов состоит инициализационный файл «is.ini»?
3. Как задать усеченное слева в точке 0 распределение Стьюдента?
4. В каком формате хранятся исходные данные? Может ли имя файла с данными содержать круглые скобки?
5. Создайте интервальную выборку роста, приведенную в главе 6 и нарисуйте гистограмму и эмпирическую функцию распределения.

ЗАКЛЮЧЕНИЕ

Задачи, возникающие в различных приложениях, зачастую не укладываются в русло классических результатов, так как не выполняются предположения, в которых корректно применение классических методов.

Форма регистрации наблюдений, их представление часто оказываются такими (группированные, поразрядно группированные, интервальные, цензурированные данные), что исключается возможность применения классических процедур. Применяемые на практике статистические методы контроля качества, обработки измерений в основном опираются на нормальное распределение или узкий класс моделей распределений, что далеко не всегда обосновано.

Большинство наиболее весомых результатов в математической статистике имеет асимптотический характер. На практике же всегда имеют дело с ограниченными объемами наблюдений. И свойства используемых статистик в таких ситуациях зачастую существенно отличаются от асимптотических.

Разработка аппарата математической статистики для таких нестандартных условий чисто аналитическими методами оказывается чрезвычайно сложной задачей. В то же время накопленный опыт показывает, что использование вычислительных технологий, статистического моделирования и компьютерного анализа позволяет выявлять фундаментальные статистические закономерности, исследовать их и строить для них математические модели, применение которых обеспечивает корректность статистических выводов в тех ситуациях, когда использование классических процедур и методов неправомерно.

Программная поддержка разделов учебного пособия обеспечивает проведение исследований и анализ свойств различных оценок и статистик. В программном обеспечении реализован аппарат применения критериев согласия, в полном объеме включающий результаты рекомендаций по стандартизации Р 50.1.033-2001 и Р 50.1.037-2002. Программная система «ISW 4.0» позволяет моделировать и исследовать более широкий круг статистик, чем упомянуто в тексте данного пособия. В частности, в систему заложены средства моделирования и исследования распределений статистик ряда критериев проверки нормальности, как вошедших в ГОСТ Р ИСО 5479-2002 (Шапиро-Уилка, Эппса-Палли и т.д.), так и несправедливо не включенных в него, критериев проверки гипотез о математических ожиданиях и дисперсиях случайных величин, критериев Бартлетта, Кохрена, *F*–критерия, критерия Граббса (ГОСТ Р ИСО 5725–5–2002) и ряда других, возможности исследования критериев в нестандартных условиях.

Освоение соответствующих учебных дисциплин на базе данного пособия предполагает обязательное изучение рекомендуемых материалов, дополняющих пособие и доступных студентам на сайте факультета.

Литература

1. Ермаков С.М. Метод Монте-Карло и смежные вопросы. М.: Наука, 1975. – 471 с.
2. Соболь И.М. Численные методы Монте-Карло. М.: Наука, 1973. – 312 с.
3. Бусленко Н.П., Шрейдер Ю.А. Метод статистических испытаний Монте-Карло и его реализация в цифровых машинах. М.: Физматгиз, 1961. – 266 с.
4. Ивченко Г.И., Медведев Ю.Я. Математическая статистика: Учебное пособие для втузов. – М.: ВШ, 1994. – 248с.
5. Ogawa J. Contributions to the theory of systematic statistics. I. Osaka Math. J. 3 (1951). – P. 175-213.
6. Сархан А.Е., Гринберг Б.Г. Введение в теорию порядковых статистик. – М.: Статистика, 1970. – 414 с.
7. Лемешко Б.Ю. Оптимальные оценки параметров сдвига и масштаба по выборочным квантилям для больших выборок // Труды третьей МНТК "Актуальные проблемы электронного приборостроения АПЭП-96". – Новосибирск, 1996. – Т. 6. – Ч.1. – С.37-44.
8. Лемешко Б.Ю., Постовалов С.Н. Оптимальные оценки параметров сдвига и масштаба по выборочным квантилям // Мат. межд. научн.-практ. конференции "САКС-2001". – Красноярск: САА. – Ч.2. 2001. – С.302-304.
9. Лемешко Б.Ю., Чимитова Е.В. Построение оптимальных L-оценок параметров сдвига и масштаба распределений по выборочным квантилям // Сибирский журнал индустриальной математики. 2001. –Т.4. – № 2. – С. 166-183.
10. Лемешко Б.Ю., Чимитова Е.В. Оптимальные L-оценки параметров сдвига и масштаба распределений по выборочным квантилям // Заводская лаборатория. Диагностика материалов. 2004. – Т.70. – №1.
11. Хьюбер П. Робастность в статистике. М.: Мир, 1984. – 303 с.
12. Лемешко Б.Ю., Постовалов С.Н. К вопросу о робастности оценок по группированным данным // Сб. научных трудов НГТУ. – Новосибирск: Изд-во НГТУ. – 1996. – № 2(4). – С. 9-18.
13. Лемешко Б.Ю. Робастные методы оценивания и отбраковка аномальных измерений // Заводская лаборатория. - 1997. - Т.63. - № 5. - С. 43-49.
14. Лемешко Б.Ю. Группирование наблюдений как способ получения робастных оценок // Надежность и контроль качества. - 1997. - № 5. - С. 26-35.
15. Hampel F.R. The influence curve and its role in robust estimation // J. Amer. Statist. Ass., 1974. - V. 69, № 346. - P. 383-393.
16. Лемешко Б.Ю., Гильдебрант С.Я., Постовалов С.Н. К оцениванию параметров надежности по цензурированным выборкам // Заводская лаборатория. Диагностика материалов. 2001. Т. 67. - № 1. - С. 52-64.
17. Лемешко Б.Ю. О некоторых вопросах оценивания параметров распределений и проверки гипотез по цензурированным выборкам // Методы менеджмента качества. 2001. - № 4. - С.32-38.

18. Parzen E. On the estimation of probability density function and the mode // Ann. Math. Stat., 1962. – Vol. 33. – P.1065-1076.
19. Надарая Э.А. Об оценке плотности распределения случайных величин // Сообщ. АН ГССР. – 1964. – Т.34. – № 2. – С. 277-280.
20. Надарая Э.А. Непараметрическое оценивание плотности вероятности и кривой регрессии. – Тбилиси: Изд-во ТГУ, 1983. – 194 с.
21. Епанечников В.А. Непараметрическая оценка многомерной плотности вероятности. Теория вероятностей и ее применения, 1969. – Т.14. – № 1. – с. 156-161.
22. Kolmogorov A.N. Sulla determinazione empirico di una legge di distribuzione // Giornale Instit. Ital. Attuari. 1933. – № 4. – P.83-91.
23. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1983. - 416 с.
24. Anderson T.W., Darling D.A. Asymptotic theory of certain ‘goodness of fit’ criteria based on stochastic processes // Ann. Math. Stat., 1952. V.23. - P.193-212.
25. Лемешко Б.Ю., Постовалов С.Н. О распределениях статистик непараметрических критериев согласия при оценивании по выборкам параметров наблюдаемых законов // Заводская лаборатория. 1998. Т. 64. - № 3. - С. 61-72.
26. Лемешко Б.Ю., Постовалов С.Н. О правилах проверки согласия опытного распределения с теоретическим // Методы менеджмента качества. Надежность и контроль качества. - 1999. № 11. - С. 34-43.
27. Лемешко Б.Ю., Постовалов С.Н. Применение непараметрических критериев согласия при проверке сложных гипотез // Автометрия. 2001. - № 2. - С. 88-102.
28. Лемешко Б.Ю., Постовалов С.Н. О зависимости распределений статистик непараметрических критериев и их мощности от метода оценивания параметров // Заводская лаборатория. Диагностика материалов. 2001. Т. 67. – № 7. – С. 62-71.
29. Лемешко Б.Ю., Постовалов С.Н. Непараметрические критерии при проверке сложных гипотез о согласии с распределениями Джонсона // Доклады СО АН ВШ. 2002. – № 1(5). – С.65-74.
30. Лемешко Б.Ю., Постовалов С.Н., Французов А.В. К применению непараметрических критериев согласия для проверки адекватности непараметрических моделей // Автометрия. 2002. – № 2. – С.3-14.
31. Лемешко Б.Ю., Постовалов С.Н. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Часть II. Непараметрические критерии. - Новосибирск: Изд-во НГТУ, 1999. - 85 с.
32. Р 50.1.037-2002. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть II. Непараметрические критерии. - М.: Изд-во стандартов. 2002. - 64 с.

33. Кендалл М., Стьюарт А. Статистические выводы и связи. - М.: Наука, 1973. - 900 с.
34. Никулин М.С. Критерий хи-квадрат для непрерывных распределений с параметрами сдвига и масштаба / Теория вероятностей и ее применение. 1973. Т. XVIII. № 3. С.583-591.
35. Никулин М.С. О критерии хи-квадрат для непрерывных распределений // Теория вероятностей и ее применение. 1973. Т. XVIII. – № 3. – С.675-676.
36. Мирвалиев М., Никулин М.С. Критерии согласия типа хи-квадрат / Заводская лаборатория. 1992. Т. 58. № 3. С.52-58.
37. Chernoff H., Lehmann E.L. The use of maximum likelihood estimates in χ^2 test for goodness of fit // Ann. Math. Stat., 1954. V. 25. - P. 579-586.
38. Чибисов Д.М. Некоторые критерии типа хи-квадрат для непрерывных распределений // Теория вероятностей и ее применение. 1971. – Т. XVI. – № 1. – С. 3-20.
39. Лемешко Б.Ю. Асимптотически оптимальное группирование наблюдений – это обеспечение максимальной мощности критериев // Надежность и контроль качества. - 1997. - № 8. - С. 3-14.
40. Лемешко Б.Ю., Постовалов С.Н. Прикладные аспекты использования критериев согласия в случае проверки сложных гипотез // Надежность и контроль качества. – 1997. – № 11. - С. 3-17.
41. Лемешко Б.Ю. Асимптотически оптимальное группирование наблюдений в критериях согласия // Заводская лаборатория, 1998. Т. 64. – №1. – С.56-64.
42. Лемешко Б.Ю., Постовалов С.Н. О зависимости предельных распределений статистик χ^2 Пирсона и отношения правдоподобия от способа группирования данных // Заводская лаборатория. 1998. Т. 64. – № 5. – С.56-63.
43. Лемешко Б.Ю., Чимитова Е.В. Максимизация мощности критериев типа χ^2 // Доклады Сибирского отделения Академии наук высшей школы. Новосибирск, 2000. – № 2. – С. 53-61.
44. Лемешко Б.Ю., Чимитова Е.В. О выборе числа интервалов в критериях согласия типа χ^2 // Заводская лаборатория. Диагностика материалов. 2003. – Т.69. – № 1. – С.61-67.
45. Лемешко Б.Ю., Чимитова Е.В. Об ошибках и неверных действиях, совершаемых при использовании критериев согласия типа χ^2 // Измерительная техника. 2002. - № 6. - С. 5-11.
46. Денисов В.И., Лемешко Б.Ю., Постовалов С.Н. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Часть I. Критерии типа χ^2. – Новосибирск: Изд-во НГТУ, 1998. - 126 с.
47. Р 50.1.033-2001. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теорети-

ческим. Часть I. Критерии типа хи-квадрат. - М.: Изд-во стандартов. 2002. - 87 с.
48. Лемешко Б.Ю., Постовалов С.Н., Чимитова Е.В. О распределениях статистики и мощности критерия типа χ^2 Никулина // Заводская лаборатория. Диагностика материалов. 2001. Т. 67. - № 3. - С. 52-58.
49. Демиденко Е.З. Линейная и нелинейная регрессия. – М.: Финансы и статистика, 1981. – 302 с.
50. Лемешко Б.Ю., Пономаренко В.М., Трушина Е.В. К проверке статистических гипотез в регрессионном и дисперсионном анализах при нарушении предположений о нормальности ошибок // Мат. 6-й всероссийской НТК "Информационные технологии в науке, проектировании и производстве". Н.Новгород, 2002. - С.1-5.
51. Андерсон Т. Введение в многомерный статистический анализ. – Москва: Физматгиз, 1963. – 500 с.
52. Кендалл М., Стьюарт А.. Многомерный статистический анализ и временные ряды. – М.: Наука, 1976. – 736 с.
53. Лемешко Б.Ю., Помадин С.С. Корреляционный анализ наблюдений многомерных случайных величин при нарушении предположений о нормальности // Сибирский журнал индустриальной математики. 2002. – Т.5. – № 3. – С.115-130.
54. Ермаков С.М., Михайлов Г.А. Статистическое моделирование. – Москва: Наука, 1982. – 296 с.
55. Лемешко Б.Ю., Помадин С.С. Один подход к моделированию псевдослучайных векторов с "заданными" числовыми характеристиками по законам, отличным от нормального // Материалы международной НТК "Информатика и проблемы телекоммуникаций". – Новосибирск, 2002. – С. 121-122.
56. Шокин Ю.И. Интервальный анализ. – Новосибирск: Наука, 1981. – 112 с.
57. Алефельд Г. Херцбергер Ю. Введение в интервальные вычисления /Пер. с англ. – М.: Мир, 1987 – 356 с.
58. Лемешко Б.Ю., Постовалов С.Н. О решении задач статистического анализа интервальных наблюдений // Вычислительные технологии. - 1997. - Т.2. - № 1. - С. 28-36.
59. Лемешко Б.Ю., Постовалов С.Н. Об оценивании параметров распределений по интервальным наблюдениям // Вычислительные технологии. 1998. Т.3. - № 2. - С. 31-38.
60. Лемешко Б.Ю. Статистический анализ одномерных наблюдений случайных величин: Программная система. - Новосибирск: Изд-во НГТУ, 1995. – 125 с.
61. Лемешко Б.Ю., Постовалов С.Н. Система статистического анализа одномерных непрерывных распределений случайных величин (версия 3.0) // Мат. III международной НТК "Микропроцессорные системы автоматики", Новосибирск, 1996. - С. С-16 - С-17.

62. Лемешко Б.Ю., Постовалов С.Н. Система статистического анализа наблюдений и исследования статистических закономерностей // Материалы международной НТК "Информатика и проблемы телекоммуникаций". - Новосибирск, 2001. - С. 80-81.