Министерство образования и науки Российской Федерации

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Компьютерные технологии анализа данных и исследования статистических закономерностей: проверка гипотезы о виде распределения

Методические указания
к курсовому проекту за первый семестр
для магистрантов I-го курса ФПМИ
по направлению 010400
дневного отделения

Новосибирск, 2012

Методические указания предназначены магистрантам направления 010400 «Прикладная математика и информатика» для выполнения курсового проекта по дисциплине «Компьютерные технологии анализа данных и исследования статистических закономерностей» в первом семестре. Указания содержат необходимые сведения для выполнения заданий, варианты заданий и список литературы.

Составители: доктор техн. наук., проф. *Б.Ю. Лемешко*,
канд. техн. наук, доц. *С.Н. Постовалов*,
канд. техн. наук, доц. *Е.В. Чимитова*

Работа подготовлена на кафедре
прикладной математики

Оглавление

1. Проверка статистической гипотезы о виде распределения

В практике статистического анализа с необходимостью использования критериев согласия приходится сталкиваться 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, или при проверке сложной гипотезы, когда по этой же выборке оцениваются параметры предполагаемого закона распределения H_0: $f(x) = f(x, \hat{\theta})$, где $\hat{\theta}$ - оценка параметра, вычисленная по выборке.

Рассмотрим критерии, которые обычно применяются для проверки гипотез о виде распределения. Для проверки гипотезы H_0: можно использовать критерии согласия и критерии проверки нормальности.

1.1. Критерии типа χ^2

1.1.1. Критерий согласия χ^2 Пирсона

Статистика χ^2 [1,2] Пирсона вычисляется в соответствии с соотношением

$$X^2 = N \sum_{i=1}^{k} \frac{(n_i / N - P_i(\theta))^2}{P_i(\theta)}, \tag{1}$$

где n_i – количество наблюдений, попавших в интервал, $P_i(\theta)$ - вероятность попадания наблюдения в i-й интервал. При справедливой (простой) гипотезе H_0 ее предельное распределение $g(S|H_0)$ есть χ_r^2-распределение с числом степеней свободы $r = k - 1$. Если по выборке оценивалось p параметров закона в результате минимизации статистики X^2, статистика подчиняется χ_r^2-распределению с $r = k - p - 1$ степеней свободы. При справедливой альтернативной гипотезе H_1 предельное распределение $g(S|H_1)$ представляет собой нецентральное χ_r^2-распределение с тем же числом степеней свободы и параметром нецентральности

$$\lambda = n \sum_{i=1}^{k} \frac{(P_i(\theta_1) - P_i(\theta))^2}{P_i(\theta)},$$

где $P_i(\theta_1)$ - вероятности попадания наблюдения в i-й интервал при альтернативной гипотезе.

В случае проверки **сложных** гипотез и оценивании по выборке параметров распределений использование в качестве предельных χ^2_{k-p-1}-распределений справедливо лишь при определении оценок параметров по сгруппированным данным и использовании для оценивания статистики X^2:

$$\hat{\theta} = \arg \min_{\theta} X^2 .$$

При использовании критериев согласия конкурирующая гипотеза H_1 (альтернатива) обычно не задается. 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 (α - вероятность ошибки первого рода) вычислить мощность критерия $1-\beta$, которая определяет способность различения этих гипотез (β - вероятность ошибки второго рода).

1.1.2. Критерий Рао-Робсона-Никулина

Никулиным [3] предложено такое видоизменение стандартной статистики (1), при котором предельное распределение есть обычное распределение χ^2_{k-1} (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 вероятностей попадания в интервал $p=(p_1,\ldots,p_k)^{\mathrm{T}}$ 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 отличается от X_n^2 только при сложных гипотезах и имеет вид

$$Y_n^2(\theta) = X_n^2 + n^{-1} a^{\mathrm{T}}(\theta)\Lambda(\theta)a(\theta),$$

где X_n^2 вычисляется в соответствии с (1). Элементы и размерность матрицы

$$\Lambda(\theta) = \left[J(\theta_l,\theta_j) - \sum_{i=1}^{k} \frac{w_{\theta_l i} w_{\theta_j i}}{p_i} \right]^{-1}_{m\times m}$$

определяются оцениваемыми компонентами вектора параметров θ, $J(\theta_l,\theta_j)$ - элементы информационной матрицы $\mathbf{J}(\theta)$, $a(\theta_l) = w_{\theta_l 1} n_1 / p_1 + \ldots + w_{\theta_l k} n_k / p_k$ - элементы вектора $a(\theta)$, величины $w_{\theta_l i}$ определяются соотношением

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

1.2. Непараметрические критерии

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

$$D_n = \sup_{|x|<\infty} | F_n(x) - F(x, \theta) |, \tag{2}$$

где n – объем выборки.

Наиболее часто в критерии Колмогорова (Колмогорова-Смирнова) используют статистику вида [1]

$$S_k = \frac{6nD_n + 1}{6\sqrt{n}}, \tag{3}$$

где

$$D_n = \max(D_n^+, D_n^-),\ D_n^+ = \max_{1 \le i \le n} \left\{ \frac{i}{n} - F(x_i, \theta) \right\},\ D_n^- = \max_{1 \le i \le n} \left\{ F(x_i, \theta) - \frac{i-1}{n} \right\}, \tag{4}$$

и $x_1, x_2, ..., x_n$ – упорядоченные по возрастанию выборочные значения.

Распределение статистики S_k при простой гипотезе в пределе подчиняется закону Колмогорова, а в случае сложной гипотезы – различным законам, в зависимости от вида распределения, оцениваемых параметров. Статистические модели распределений статистик $G(S_k | H_0)$ для наиболее распространенных семейств законов распределений приведены в [4].

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

1.2.2. Критерии типа ω^2. В критериях типа ω^2 расстояние между гипотетическим и истинным распределениями рассматривают в квадратичной метрике. Статистика критерия [1] выражается соотношением

$$\omega_n^2 [\psi(F)] = \int_{-\infty}^{\infty} \left\{ E[F_n(x)] - F(x) \right\}^2 \psi(F(x)) dF(x) =$$
$$= \frac{2}{n} \sum_{i=1}^{n} \left\{ g[F(x_i)] - \frac{2i-1}{2n} f[F(x_i)] \right\} + \int_0^1 (1-t)^2 \psi(t) dt, \tag{5}$$

где

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

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

$$S_\omega = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left\{ F(x_i, \theta) - \frac{2i-1}{2n} \right\}^2 . \qquad (6)$$

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

$$S_\Omega = n\Omega_n^2 = -n - 2\sum_{i=1}^{n} \left\{ \frac{2i-1}{2n} \ln F(x_i, \theta) + \left(1 - \frac{2i-1}{2n}\right) \ln(1 - F(x_i, \theta)) \right\}. \quad (7)$$

Распределение статистик S_ω и S_Ω при простой гипотезе в пределе подчиняется законам a_1 и a_2, а в случае сложной гипотезы – различным законам, в зависимости от вида распределения и оцениваемых параметров. Статистические модели распределений статистик $G(S_\omega|H_0)$ и $G(S_\Omega|H_0)$ для наиболее распространенных семейств законов распределений приведены в [4].

1.3. Критерии проверки нормальности

1.3.1. Критерий симметричности предназначен для проверки гипотез о симметричности наблюдаемого закона (против наличия асимметрии) при объемах выборки $8 \le n \le 5000$. Статистика критерия имеет вид

$$\sqrt{\beta_1} = \frac{\hat{\mu}_3}{\hat{\sigma}^3}, \qquad (8)$$

Проверяется гипотеза H_0: $\sqrt{\beta_1}$ =0 против альтернативы $\sqrt{\beta_1}$ >0 (положительная асимметрия) или $\sqrt{\beta_1}$ <0 (отрицательная асимметрия).

1.3.2. Критерий проверки на эксцесс рассматривается при объемах выборок $8 \le n \le 5000$. Статистика критерия проверки на значение эксцесса имеет вид

$$\hat{\beta}_2 = \frac{\hat{\mu}_4}{\hat{\sigma}^4}. \qquad (9)$$

Проверяется гипотеза вида H_0: β_2=3 против альтернативы β_2>3 (больший эксцесс) или β_2<3 (меньший эксцесс).

3.3. В критерии Шапиро-Уилка для вариационного ряда $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$, полученного по наблюдаемой выборке $X_1, X_2, ..., X_n$, вычисляют величину

$$S = \sum_k a_k \left[X_{(n+1-k)} - X_{(k)} \right],$$

где индекс k изменяется от 1 до $n/2$ или от 1 до $(n-1)/2$ при четном и нечетном n соответственно. Коэффициенты a_k приведены в стандарте и первоисточниках. Статистика критерия имеет вид

$$W = S^2 \Big/ \sum_{i-1}^{n} \left(X_i - \bar{X} \right)^2 . \qquad (10)$$

Гипотеза о нормальности отвергается при малых значениях статистики W.

1.3.4. Статистика критерия Эппса-Палли для наблюдаемой выборки $X_1, X_2, ..., X_n$ имеет вид

$$T_{EP} = 1 + \frac{n}{\sqrt{3}} + \frac{2}{n} \sum_{k=2}^{n} \sum_{j=1}^{k-1} \exp\left\{ -\frac{\left(X_j - X_k \right)^2}{2\hat{\mu}_2} \right\} - \sqrt{2} \sum_{j=1}^{n} \exp\left\{ -\frac{\left(X_j - \bar{X} \right)^2}{4\hat{\mu}_2} \right\}, \qquad (11)$$

где $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $\hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$. Выборка может быть неупорядочена, порядок наблюдений произволен, но он должен быть неизменным в течение всех проводимых вычислений. Гипотезу о нормальности отвергают при больших значениях статистики.

1.3.5. Модификация D’Agostino критерия проверки на симметричность. В данной модификации на основании следующих соотношений статистика (8) преобразуется в статистику z_1, приближенно подчиняющуюся стандартному нормальному закону:

$$b = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}, \qquad \omega^2 = -1 + \left\{ 2(b-1) \right\}^{1/2},$$

$$\delta = \frac{1}{\left\{ \log(\sqrt{\omega^2}) \right\}^{1/2}}, \qquad y = \sqrt{\hat{\beta}_1} \left\{ \frac{\omega^2 - 1}{2} \cdot \frac{(n+1)(n+3)}{6(n-2)} \right\}^{1/2},$$

$$z_1 = \delta \log\left\{ y + (y^2 + 1)^{1/2} \right\}. \qquad (12).$$

1.3.6. Модификация D’Agostino критерия одновременной проверки на симметричность и значение эксцесса. Здесь предложено преобразование статистик (8) и (9) к статистике z_2, приближенно распределенной в соответствии со стандартным нормальным законом. Преобразование осуществляется с помощью следующих соотношений:

$$\delta = (n-3)(n+1)(n^2+15n-4), \quad a = \frac{(n-2)(n+5)(n+7)(n^2+27n-70)}{6\delta}$$

$$c = \frac{(n-7)(n+5)(n+7)(n^2+2n-5)}{6\delta}, \quad k = \frac{(n+5)(n+7)(n^3+37n^2+11n-313)}{12\delta},$$

$$\alpha = a + \beta_1 c, \qquad \chi = (\hat{\beta}_2 - 1 - \hat{\beta}_1)2k,$$

$$z_2 = \left\{\left(\frac{\chi}{2\alpha}\right)^{1/3} - 1 + \frac{1}{9\alpha}\right\}(9\alpha)^{1/2}. \qquad (13)$$

1.4. Применение программы ISW для проверки гипотез о согласии

Проверку гипотез о согласии проиллюстрируем на примере программы ISW 4.4. Для запуска системы нужно запустить на выполнение файл **isw.exe**. После запуска открывается окно, как показано на рис. 1.1.

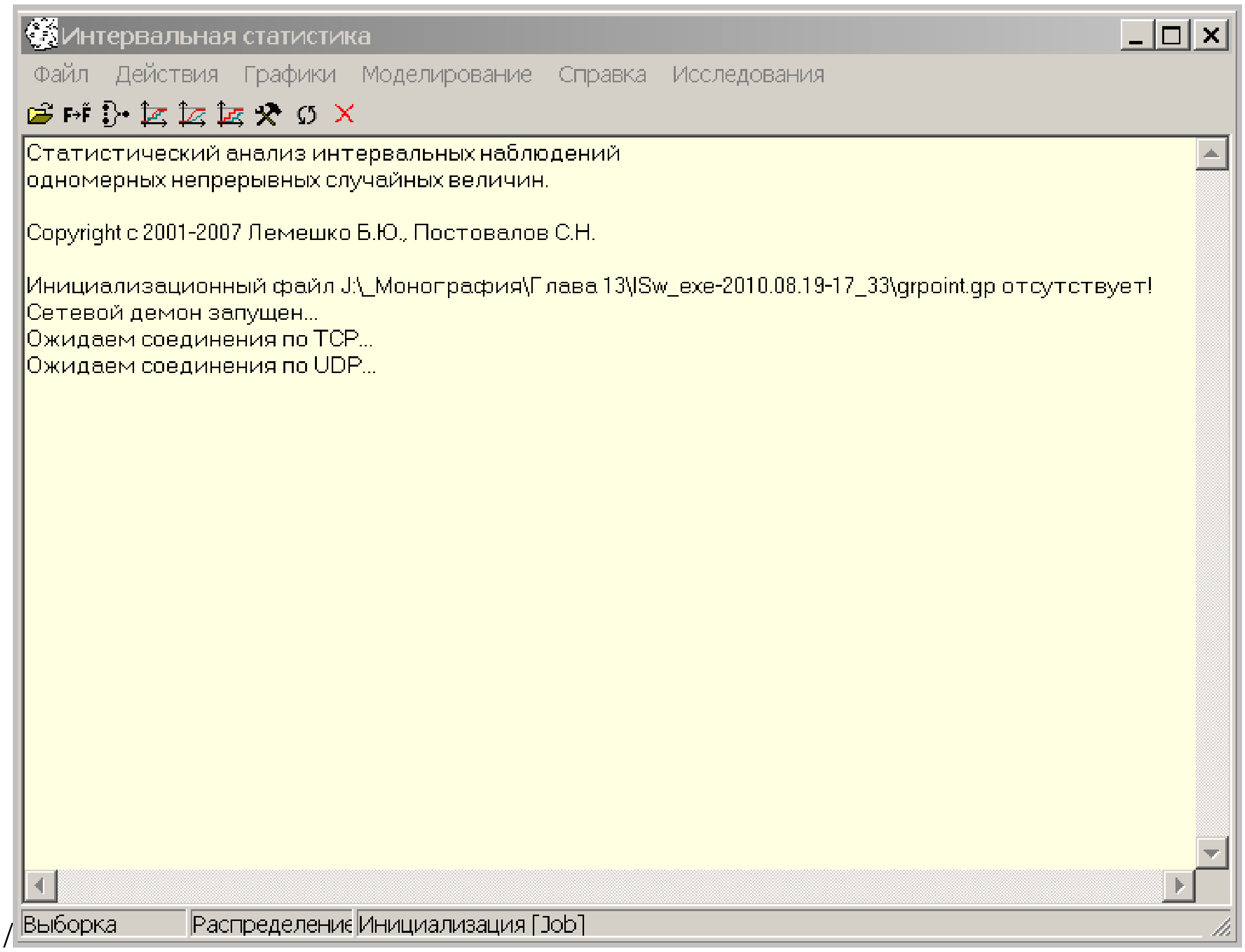
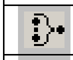
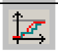
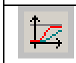
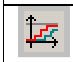
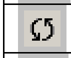
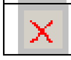


Рис. 1.1. Главное окно системы ISW

Для выполнения основных функций системы нужно выбрать пункт меню системы, либо нажать на кнопку панели инструментов. Поясним назначение отдельных кнопок.

	Открыть из файла: выборку (*.dat) или список законов распределений (*.dst) или инициализационный файл (*.ini)
	Открыть форму статистического анализа, которая позволяет провести оценивание параметров и проверку гипотезы о согласии, выявить аномальные наблюдения.
	Группирование выборки
	Отображение в одном окне всех графиков: для законов распределения – это либо функция распределения, либо функция плотности; а для выборок – это эмпирическая функция распределения либо гистограмма.
	Отображение в одном окне графиков для законов распределения – это либо функция распределения, либо функция плотности.
	Отображение в одном окне всех графиков для выборок – это либо эмпирическая функция распределения, либо гистограмма.
	Настройка параметров системы
	Чтение параметров системы из файла is.ini
	Очистка окна сообщений

1.4.1. Открытие выборки из файла

Система работает только с одномерными выборками, хранящимися в файлах с расширением “**dat**”. Файл можно создать с помощью любого текстового редактора, либо сгенерировать программно по заданному формату.

1.4.1.1. Формат входных данных

Первая строка файла содержит название выборки, в этой строке может быть произвольная информация, но мы рекомендуем вводить в этой строке источник этой выборки, информацию о случайной величине, условия проведения эксперимента. Информация из этой строки используется при построении графиков.

Во второй строке файла содержится информация о типе выборки. Остальные строки содержат информацию в зависимости от типа выборки.

- **Тип выборки 0. Точечная выборка**

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

<наблюдение n>

- **Тип выборки 1. 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>

- **Тип выборки 2. 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

- **Тип выборки 3. 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

- **Тип выборки 4. 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

- **Тип выборки 5. 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

- **Тип выборки 6. 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

- **Тип выборки 10. Интервальная выборка**

Интервальная выборка из *n* интервальных наблюдений

<название выборки>

10 n
<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.

1.4.1.2. Создание выборки в текстовом редакторе

Рассмотрим пример, как можно создать выборку с использованием текстового редактора, например Notepad (Блокнот).

Пример 1.1.

А) Время ремиссии (в неделях) 42 пациентов с острой лейкемией было приведено в отчете [5] *о клинических испытаниях препарата 6-mercaptopurine (6-MP). Каждый пациент случайным образом получал 6-MP или плацебо. Изучение было закончено через один год.*
Были получены следующие выборки, в неделях:
Выборка с применением плацебо (21 пациент) содержит следующие наблюдения: 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.
Выборка с применением препарата 6-MP (21 пациент): 6, 6, 6, 7, 10, 13, 16, 22, 23, 7, 10, 11, 13, 19, 20, 24, 27, 33, 35, 37, 42.
Создать выборки «Выборка плацебо.dat» и «Выборка 6-MP.dat» в текстовом редакторе Notepad.

Б) В таблице приведено распределение толщины 12 000 бобов.

Толщина, мм	До 7.00	7.00-7.25	7.25-7.5	7.5-7.75	7.75-8.00	8.00-8.25	8.25-8.5	8.5-8.75
Количество бобов	32	103	239	624	1187	1650	1883	1930
Толщина, мм	8.75-9.00	9.00-9.25	9.25-9.5	9.25-9.75	9.75-10.00	10.00-10.25	10.25-10.5	Свыше 10.5
Количество бобов	1638	1130	737	427	221	110	57	32

Создать выборку «Толщина бобов.dat» в текстовом редакторе Notepad.

Так как все наблюдения выборки с применением плацебо являются точками, то тип первой выборки – точечный. Чтобы ввести эту выборку, открываем в программе Notepad новый файл и вводим в него данные, как показано на рис. 1.2. Затем сохраняем этот файл, например, с именем “Выборка плацебо.dat”.
Аналогично вводим вторую выборку “Выборка 6-MP.dat”.

Выборка с толщиной бобов является группированной, поэтому вводим граничные точки и количества по формату «3» (рис. 1.3).

Теперь можно открыть эти выборки в системе. Для этого выбираем в меню **Файл** пункт **Открыть**. Открывается стандартное окно Windows выбора файла. Допускается выбрать не один файл, а несколько, используя клавиши <Ctrl> или <Shift> (рис. 1.4). Список открытых выборок можно посмотреть по кнопке на вкладке **Выборки** (рис. 1.5).

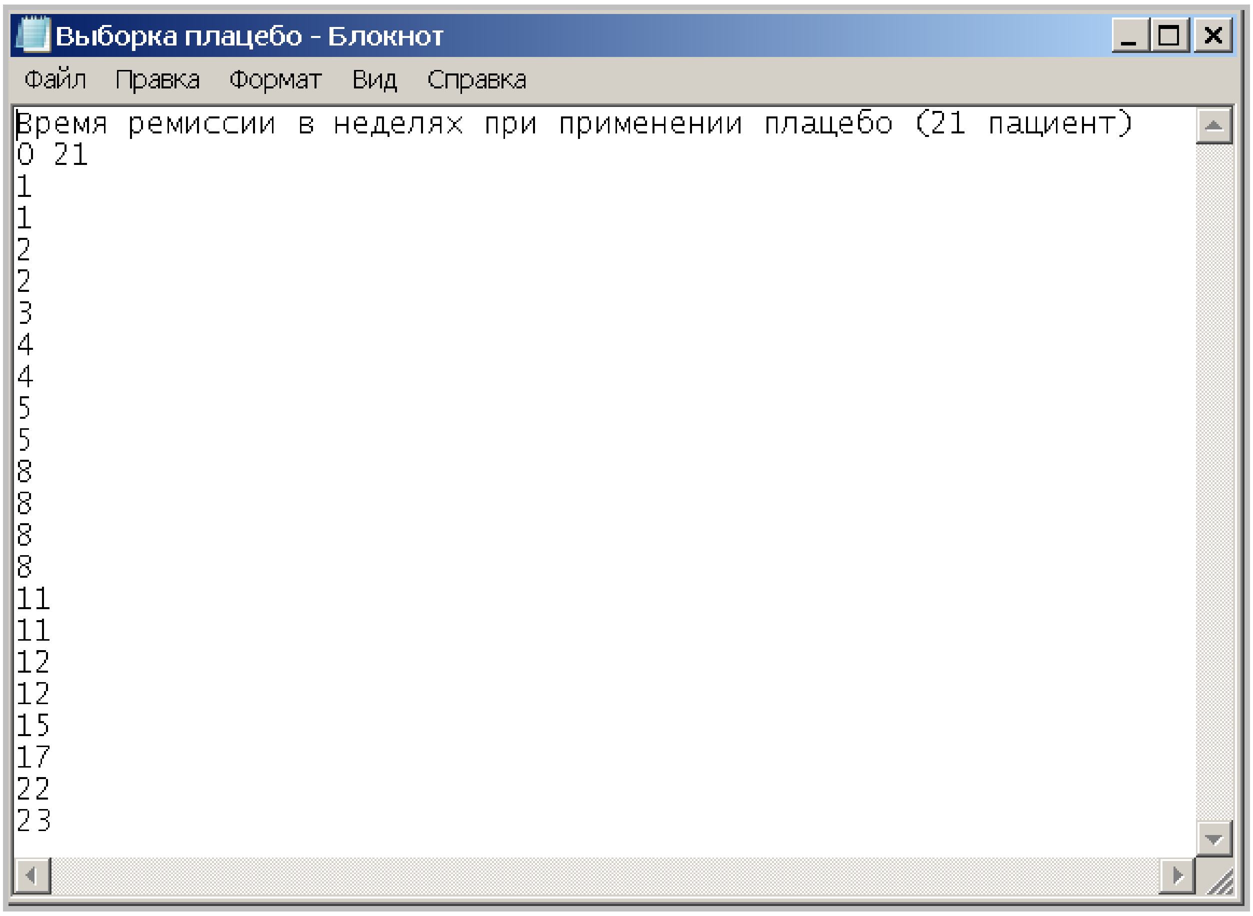


Рис. 1.2. Создание выборки «Выборка плацебо» в текстовом редакторе

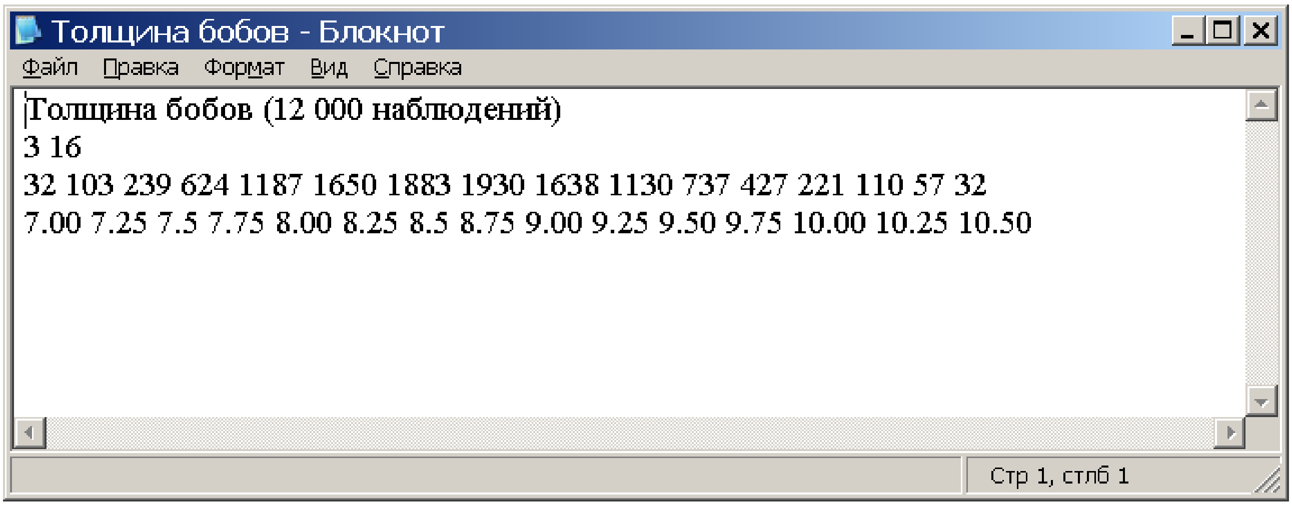


Рис. 1.3. Создание группированной выборки

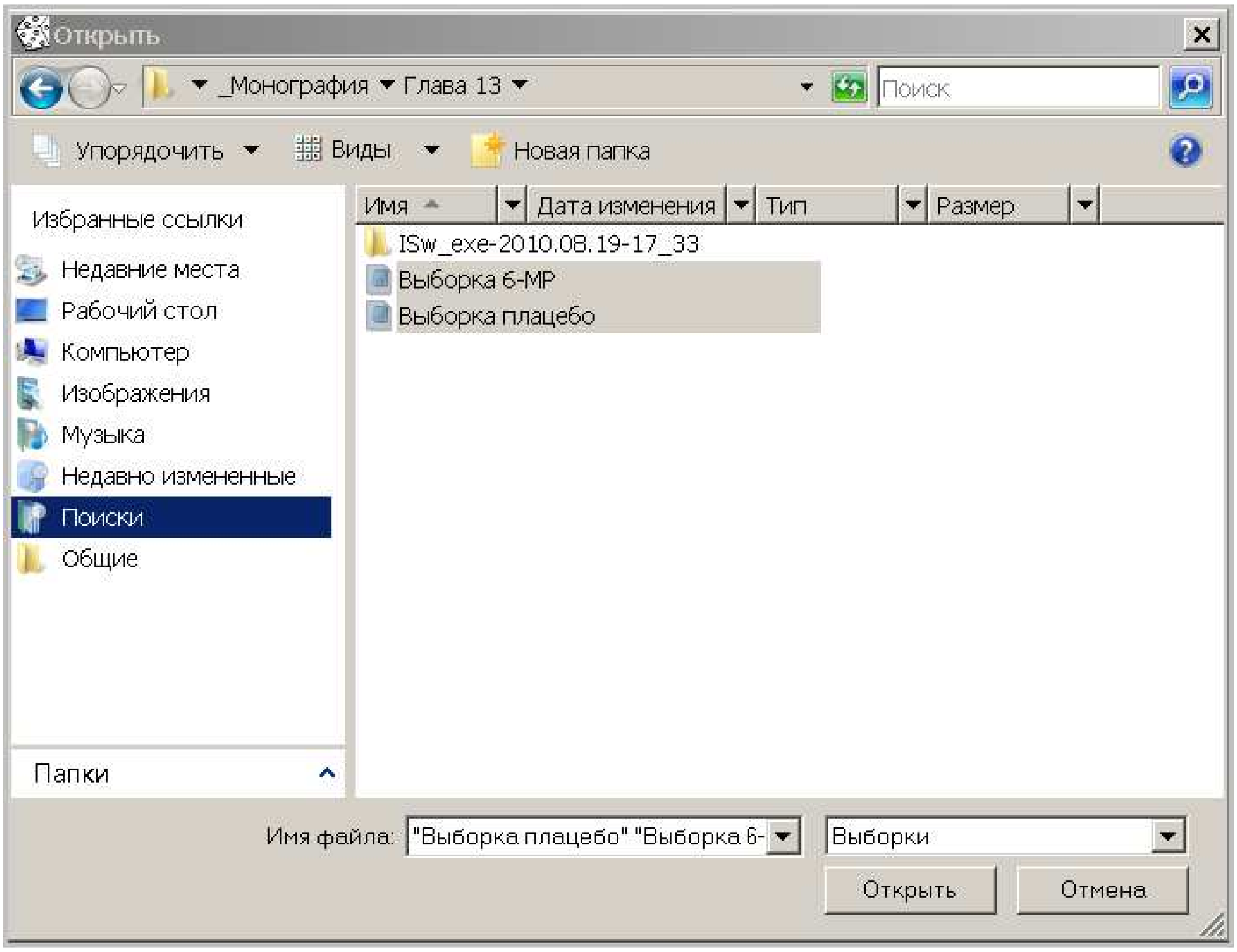


Рис. 1.4. Открытие выборки из файла

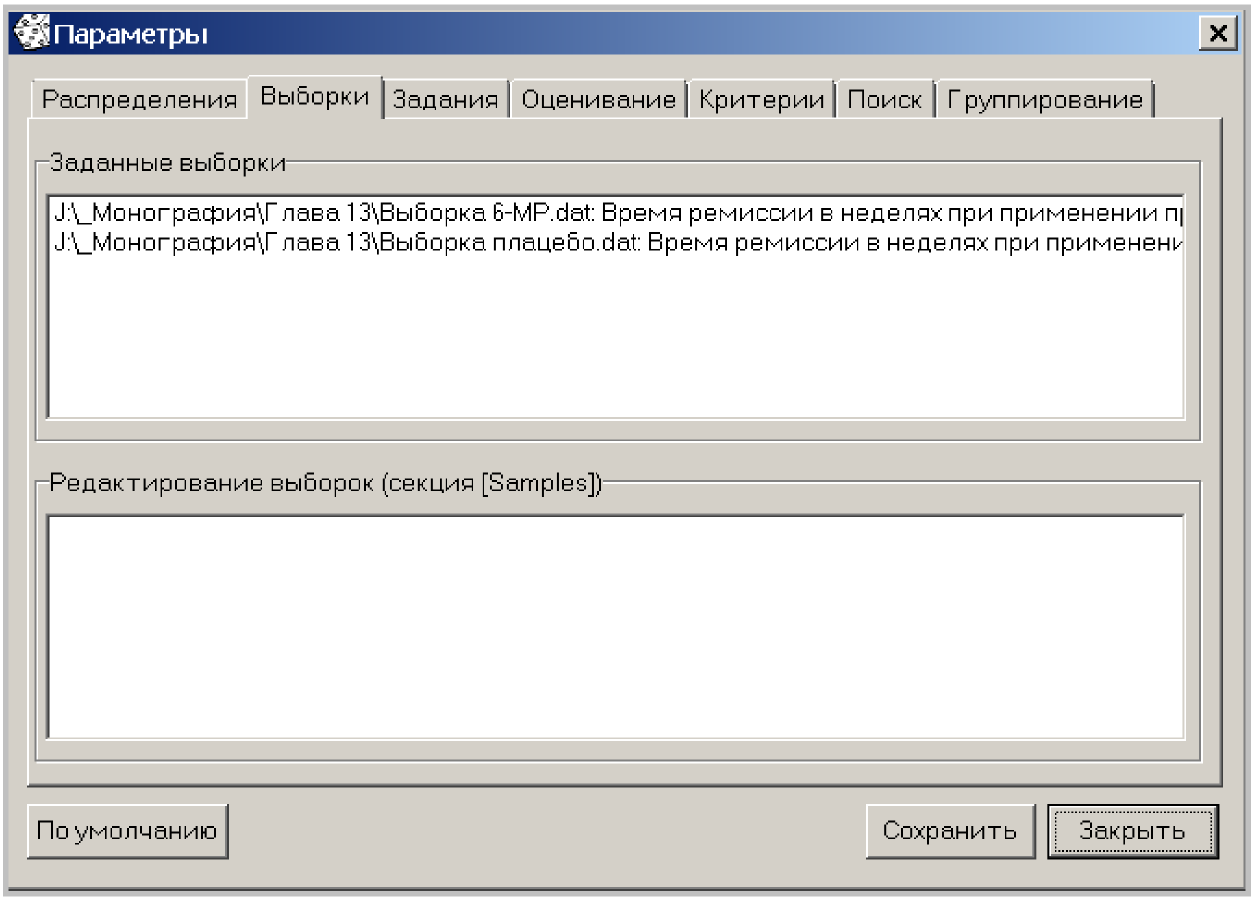


Рис. 1.5.Список открытых выборок

4.2. Открытие списка распределений

В системе встроено большое количество законов распределений, а также есть возможность создавать новые распределения, используя различные операции над распределениями. Но для того, чтобы работать с законами, мы должны либо описать их файле is.ini, либо открыть список распределений, который находится в файле с расширением “dst”.

В дистрибутиве системы уже имеется несколько готовых списков распределений:

- Стандартные
- Симметричные
- Положительные
- Специальные

При желании пользователь системы может создать свой список распределений. Откроем, например, список положительных распределений. Для этого в меню **Файл** выбираем пункт **Открыть**. Далее указываем, что мы хотим открыть список распределений (рис. 1.6) и выбираем файл «Положительные.dst» (рис. 1.7). После чего нажимаем кнопку **Открыть**. Список открытых законов распределения можно увидеть в параметрах системы по кнопке  на вкладке **Распределения** (рис. 1.8).

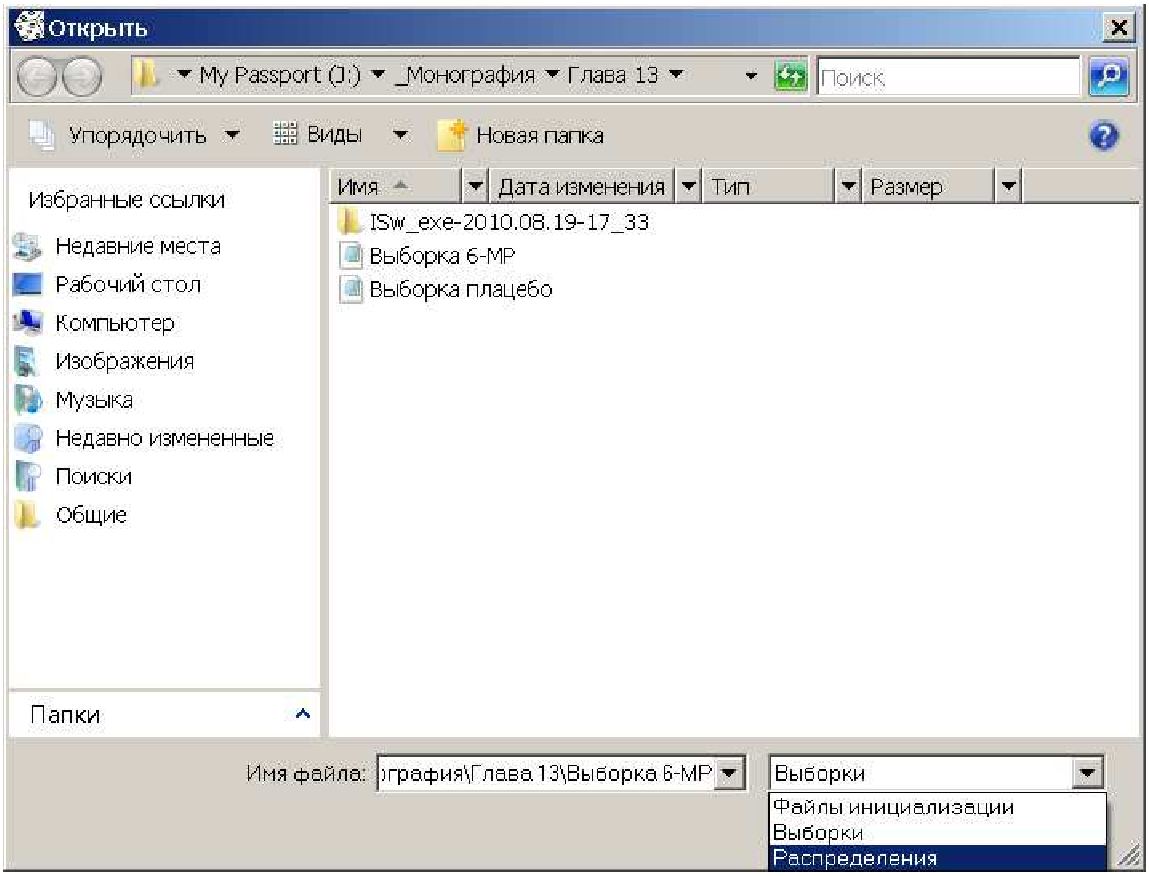


Рис. 1.6.Открытие списка законов распределения

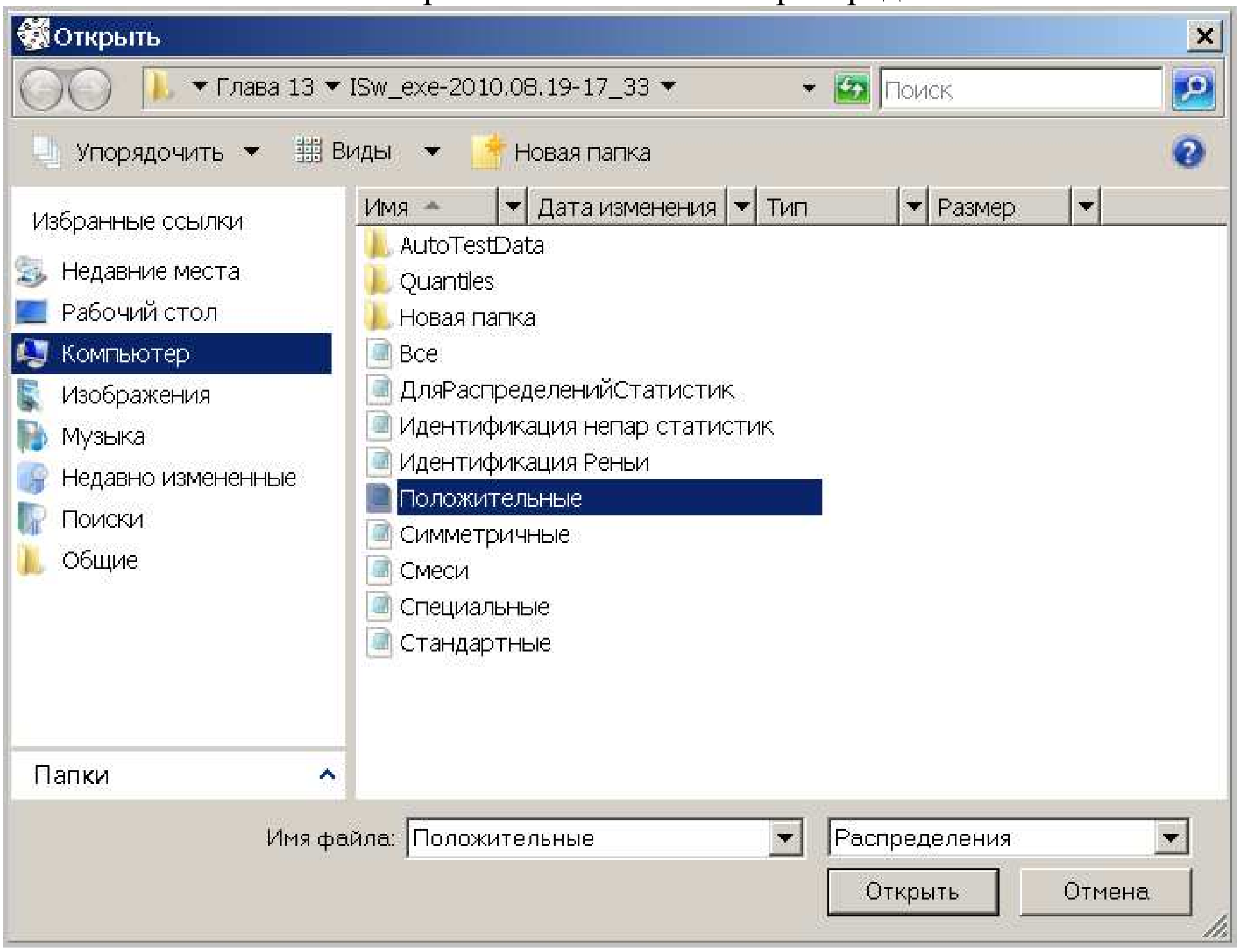


Рис. 1.7.Открытие списка положительных законов распределения

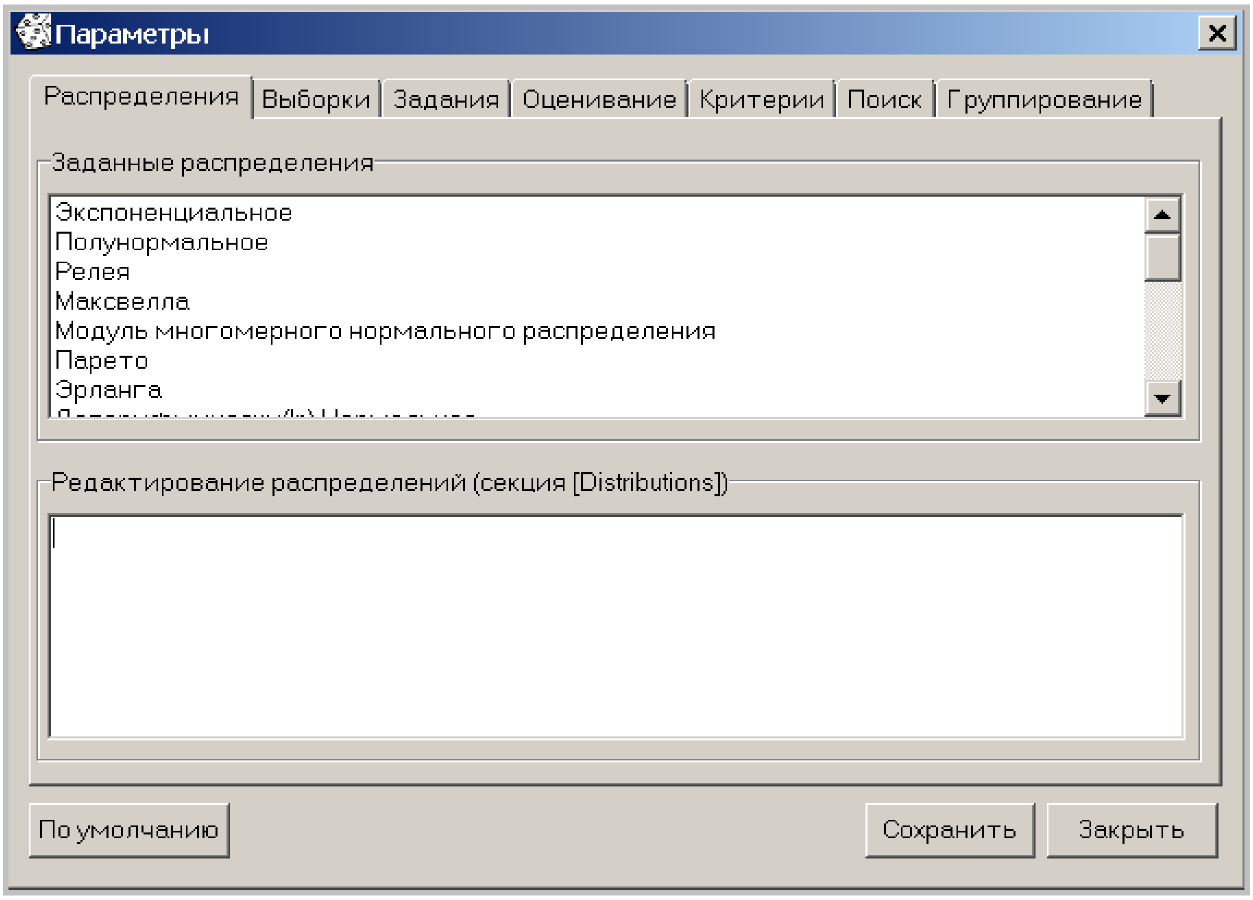


Рис. 1.8. Список открытых законов распределения

1.4.3. Построение графиков

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. 1.9 показаны графики функций распределения положительных случайных величин.

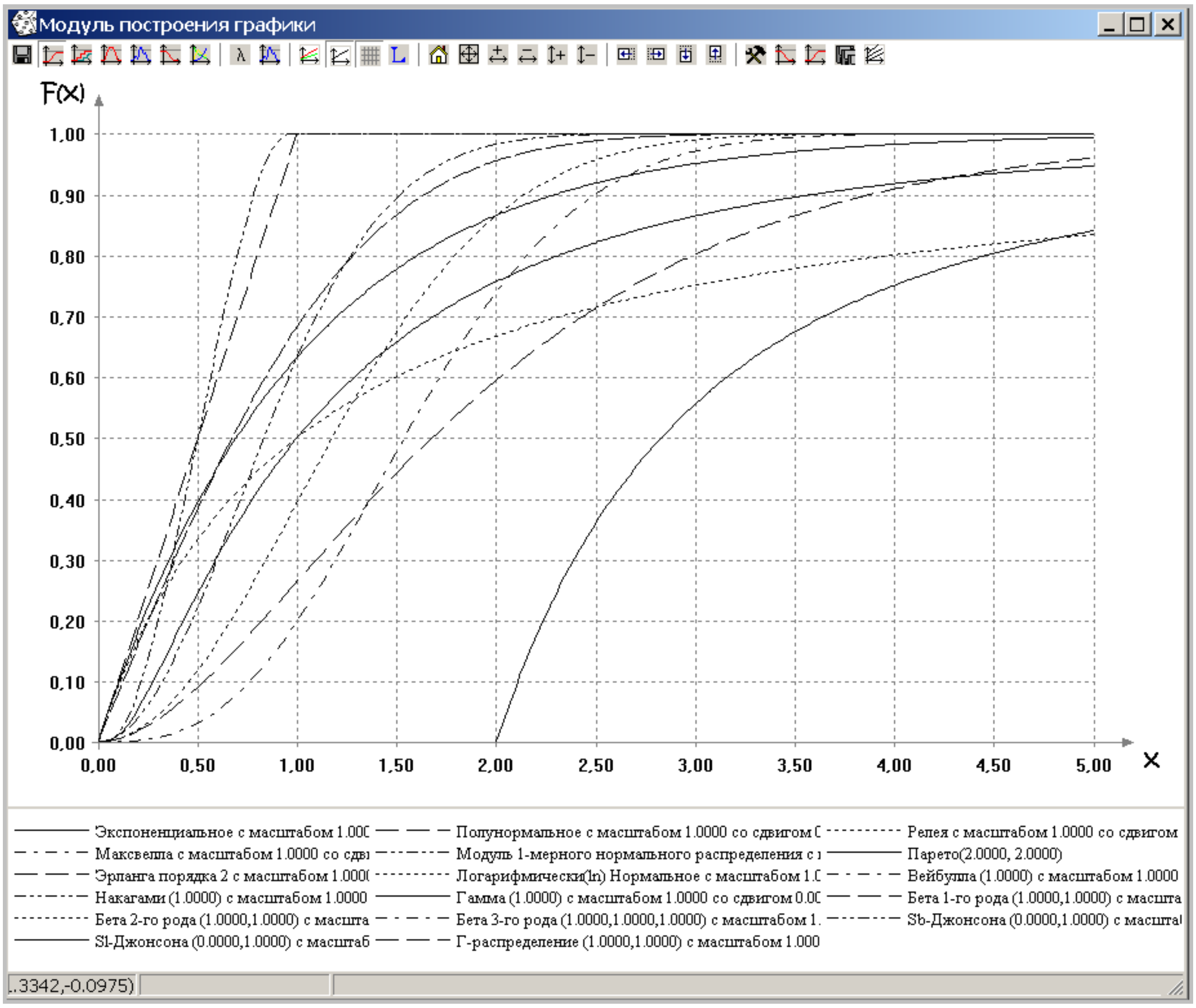


Рис. 1.9. Функции распределения положительных случайных величин

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

На что следует обратить внимание при построении графиков функций? Для того чтобы график хорошо выглядел, был удобен для анализа, для подготовки

отчетов, надо задать область построения. Делается это с помощью кнопки на панели инструментов окна графика (рис. 1.10).
На графике, на этом рисунке выведены функции распределения положительных случайных величин. В этом случае имеет смысл отображать только значения по оси ординат в интервале от 0 до 1. Поэтому устанавливаем, как показано на рис. 13.10 нижнюю границу выводимой области в 0, а верхнюю границу в 1.
Далее задаем границы области слева и справа. Если посмотреть на рис. 13.10, то мы увидим, что правая граница установлена равной 3.81, а область разбита на 6 интервалов. В этом случае цена деления равна 0.635. Поскольку в шкале отображается только 2 знака после запятой (но это можно изменить, см. ниже), то у пользователя создается иллюзия, что цена деления неодинакова!
В нашем примере график будет выглядеть лучше, если левую границу задать равной 0 (положительные случайные величины не могут быть меньше нуля!), а правую задать равной 5, т.к. в этом случае будут лучше видны «хвосты» распределений.

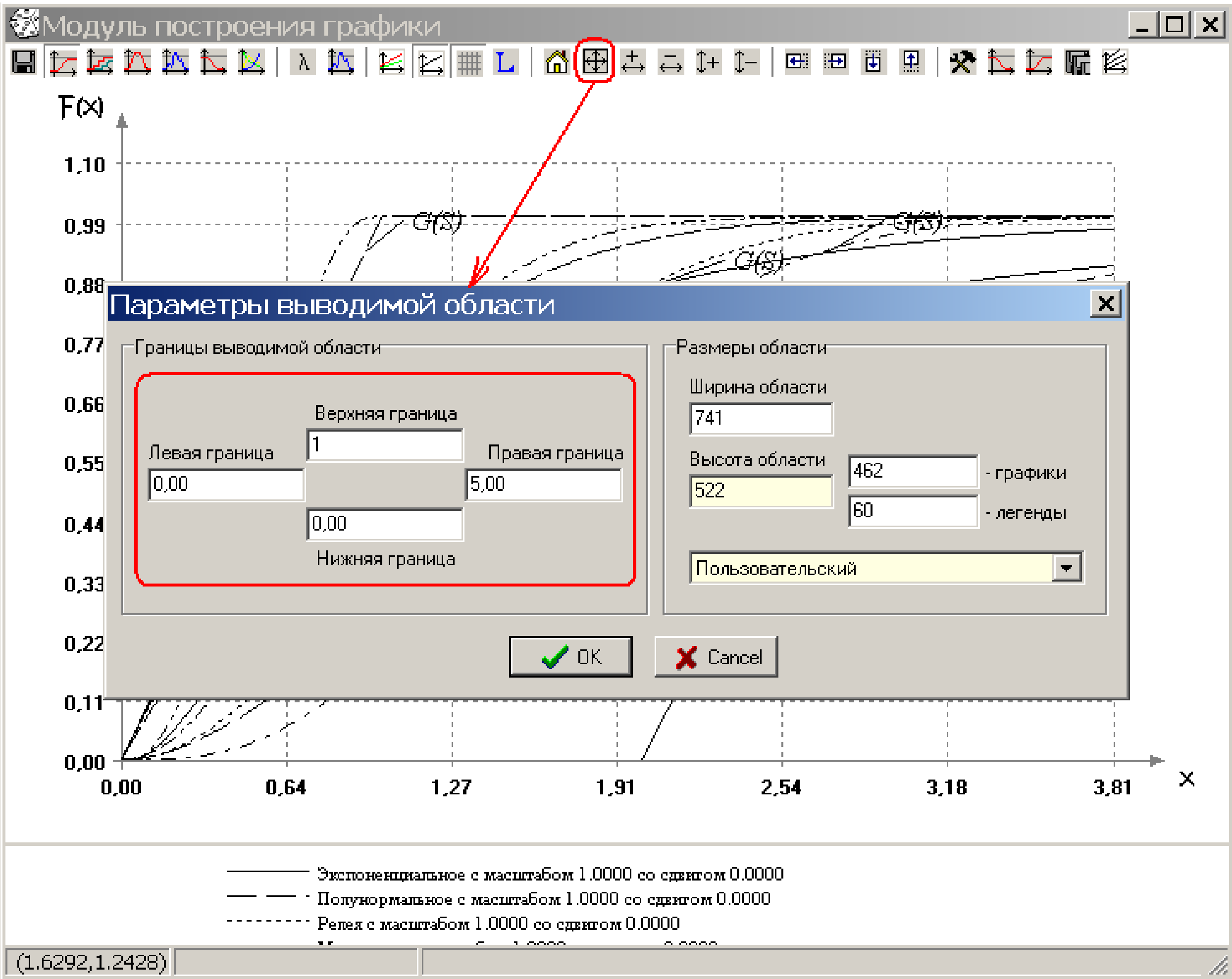


Рис. 1.10. Настройка области построения графика

Однако в этом случае цена деления будет равна 5/6 = 0.8(3), что также будет не очень наглядно. Лучше увеличить число интервалов до 10, нажав на кнопку в панели инструментов (рис. 1.10). В этом случае цена деления будет равна 5/10 = 0.5, и график будет выглядеть как на рис. 1.9.

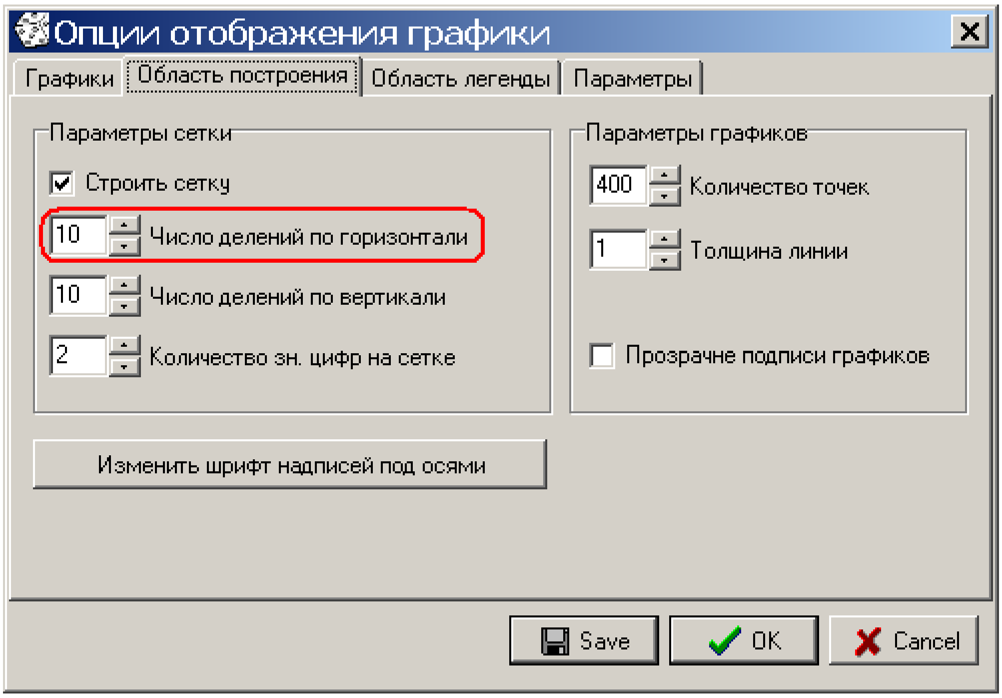


Рис. 13.10. Настройка параметров сетки

Если количество графиков достаточно большое, то удобно выводить подписи графиков к легенде в несколько столбцов, как это показано на рис.1.9. Задать эти параметры можно на вкладке **Область легенды** (рис. 1.11).

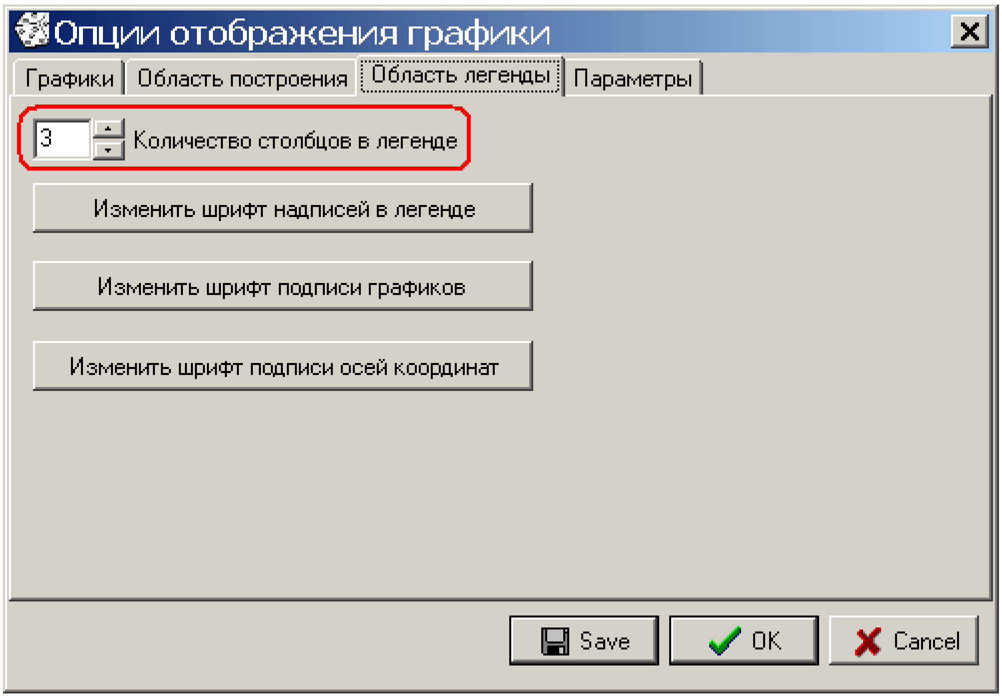


Рис. 13.11. Настройка области легенды

Пример 1.2. *Построить графики эмпирических функций распределения по выборкам «Выборка плацебо.dat» и «Выборка 6-MP.dat».*

Чтобы просмотреть в одном окне графики эмпирических функций распределения по выборкам, приведенным в *примере 1.1*, нужно в главном меню **Графики** выбрать пункт **Все выборки** (рис. 1.12).

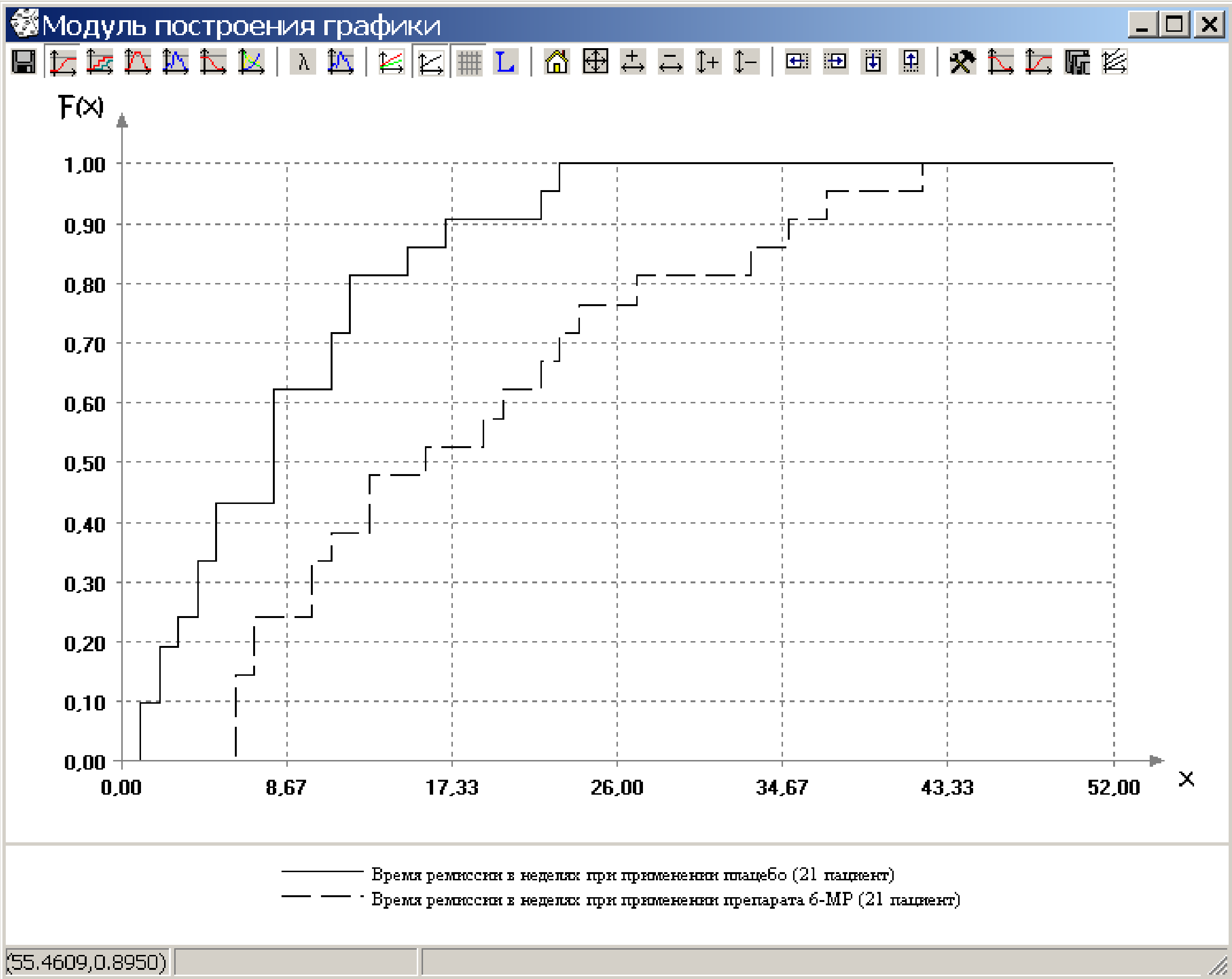


Рис. 1.12. Эмпирическая функция распределения

1.4.4. Проведение статистического анализа

1.4.4.1. Оценивание параметров и проверка согласия

Статистический анализ выборки производится в форме "Оценивание параметров и проверка согласия" (кнопка F→F̂ на панели инструментов), как показано на рис. 1.13. Необходимо выбрать выборку, закон распределения, метод оценивания и критерии согласия.

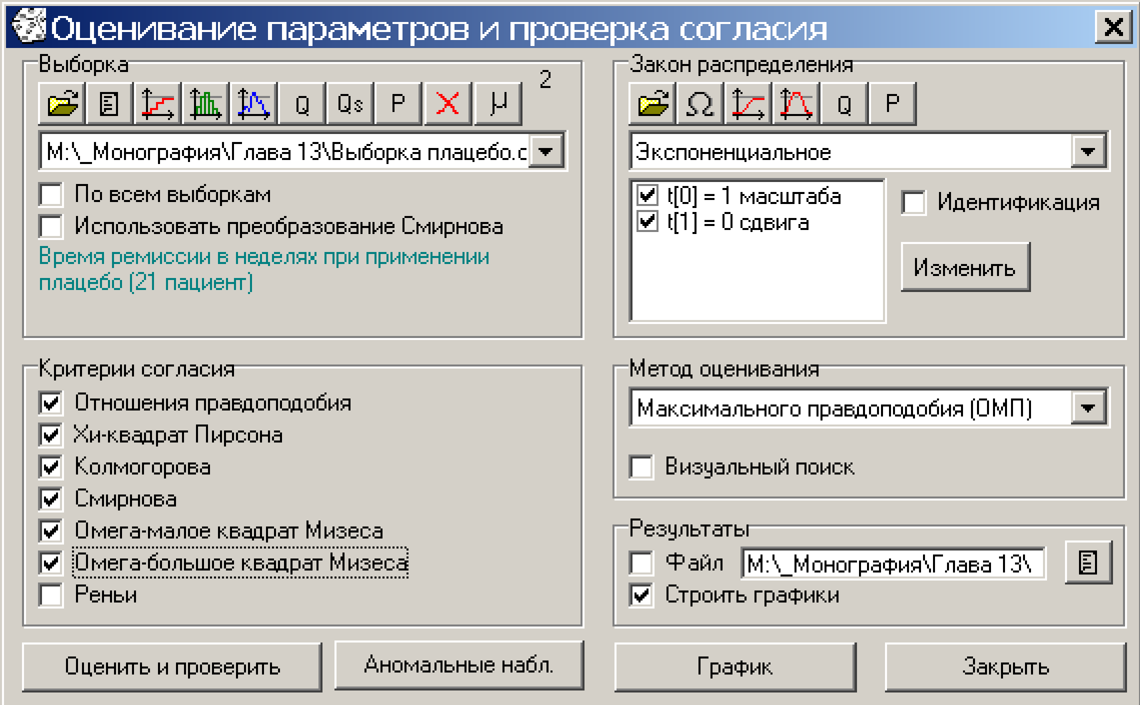


Рис. 1.13. Оценивание параметров и проверка согласия

- 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.

Чтобы отметить, какие параметры выбранного закона распределения требуется оценить, нужно поставить флажки рядом с оцениваемыми параметрами (рис. 1.14). Если, наоборот, параметр оценивать не надо, то флажок надо снять и с помощью кнопки **Изменить** задать значение параметра вручную.

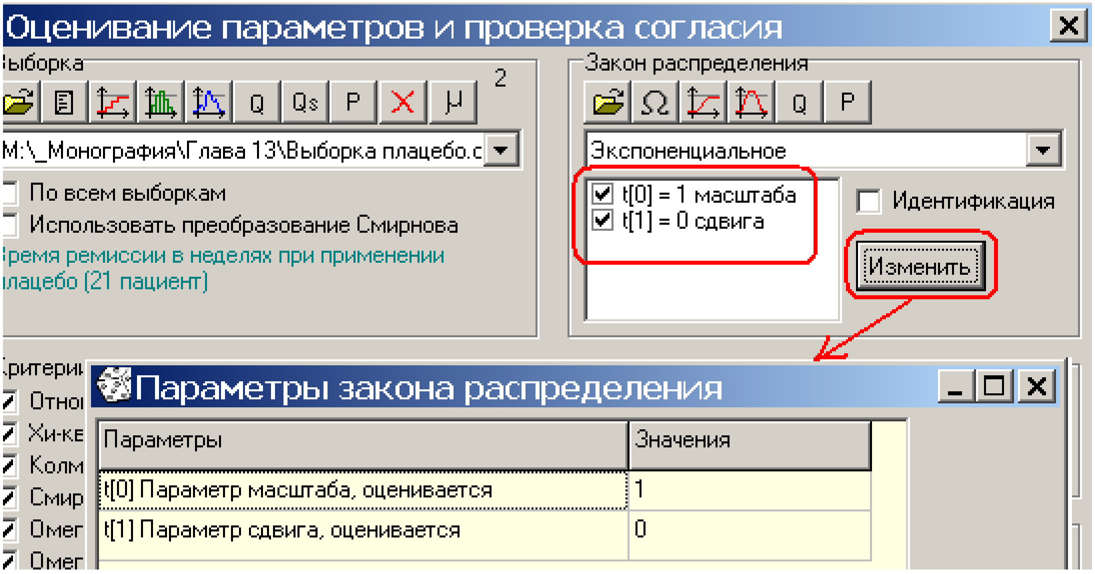


Рис. 13.14. Установка признака оценивания параметров

- 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.

Пример 1.3. *Оценить параметры экспоненциального распределения по выборкам «Выборка плацебо.dat» и «Выборка 6-MP.dat» и проверить их согласие с законами распределения:*

- *экспоненциальный;*
- *нормальный;*
- *Вейбулла-Гнеденко.*

Зададим параметры оценивания и проверки гипотез, как показано на рис. 1.13. В результате оценивания параметров по методу максимального правдоподобия мы получаем, что оценка максимального правдоподобия параметра масштаба равна 9.3601.

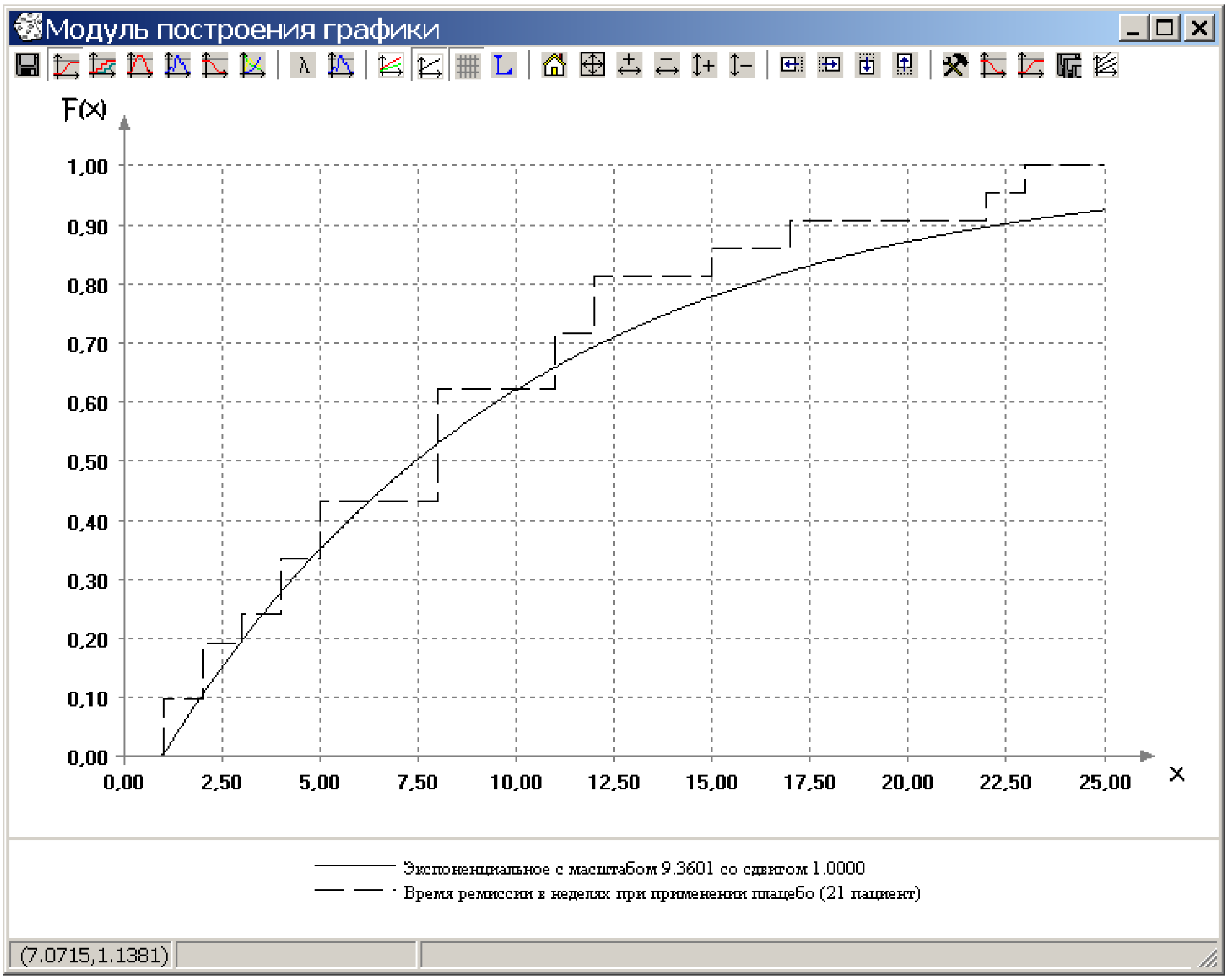


Рис. 1.15. Оценивание параметров экспоненциального распределения

Результаты проверки гипотезы о согласии выводятся в окне сообщений (рис. 1.16). Так мы выбрали использование нескольких критериев согласия, то система вычисляет достигаемый уровень значимости для каждого критерия, а затем вычисляет средний достигаемый уровень значимости. В нашем примере он получился равным 0.4 и при заданном уровне значимости (вероятности ошибки первого рода) 0.1 гипотеза о согласии не отвергается.

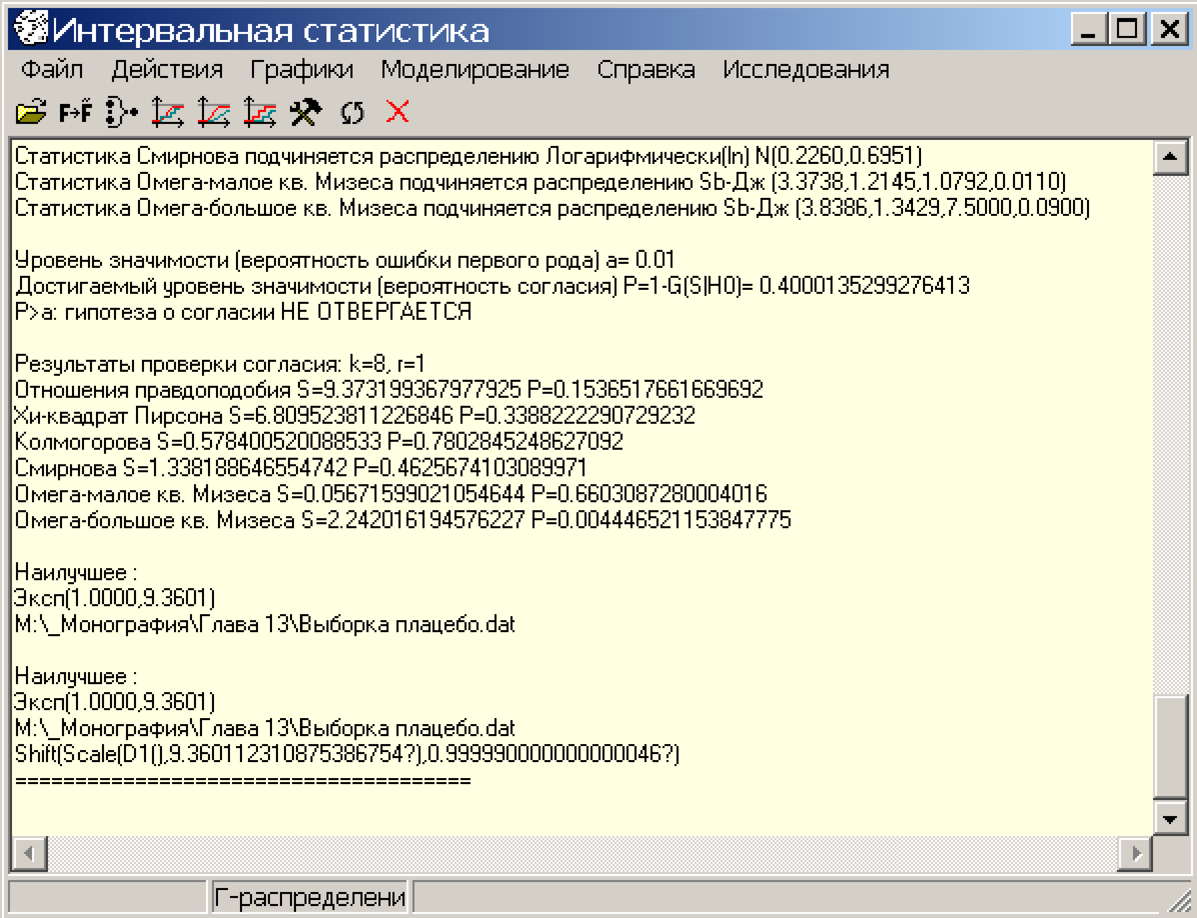


Рис. 1.16. Результаты оценивания и проверки гипотез о согласии

1.4.4.2. Проверка на нормальность

Проверка выборки на принадлежность семейству нормальных распределений является достаточно частой задачей на практике. Для проверки нормальности можно использовать и универсальные критерии согласия, доступные в форме **Оценивание параметров и проверка согласия**. Вместе с тем известно множество критериев, проверяющих именно гипотезу о нормальности, часть из которых реализована в системе.

Чтобы вызвать форму для проверки нормальности нужно в главном меню **Действия** выбрать пункт **Проверка на нормальность** (рис. 1.17). Далее флажками нужно выбрать критерии, которые будут использоваться, после чего нужно нажать на кнопку **Проверить!**

Результаты проверки гипотезы нормальности выводятся в окно сообщений (рис. 1.18).

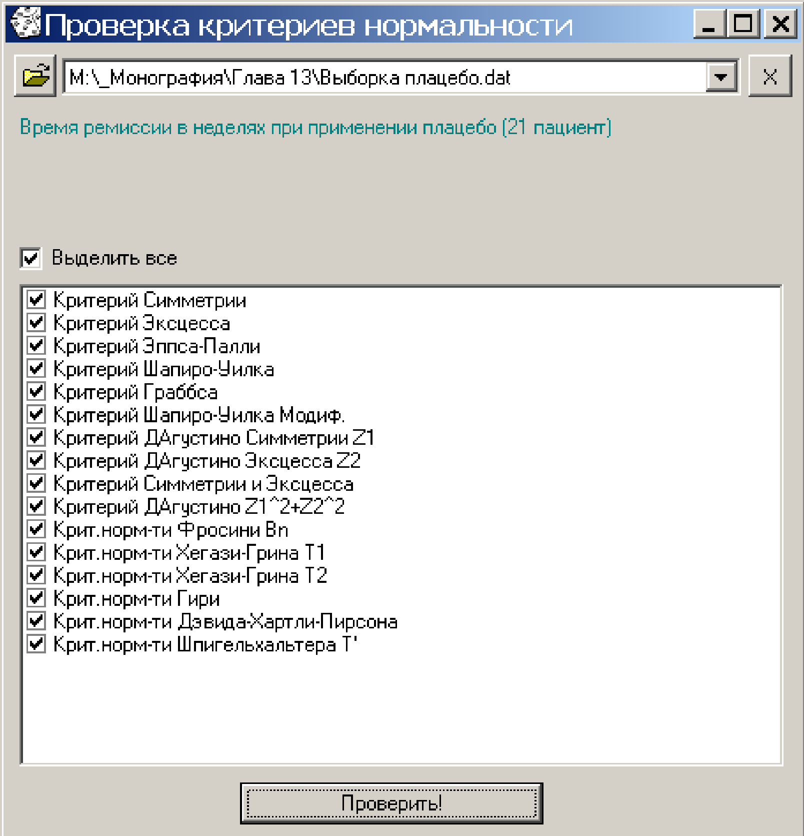


Рис. 1.17. Форма для проверки гипотезы нормальности

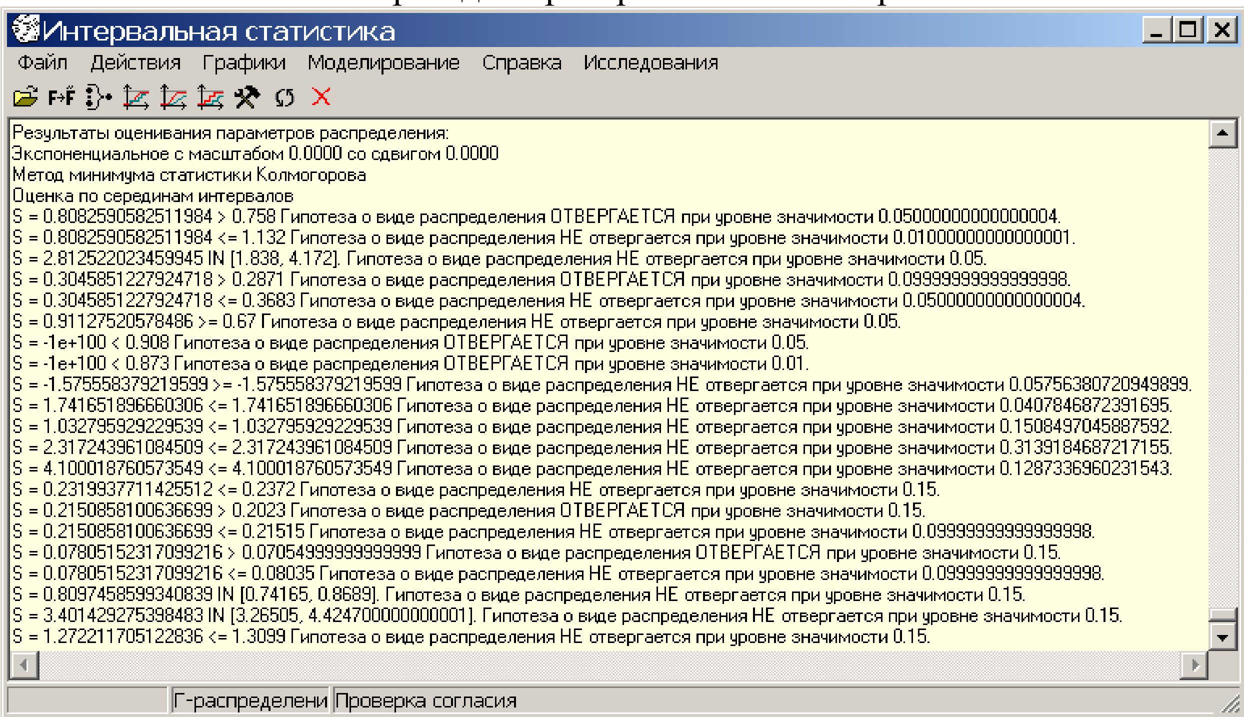


Рис. 1.18. Результаты проверки гипотезы о нормальности

1.5. Исследование свойств критериев согласия

Для исследования свойств критериев согласия также будем использовать программу ISW.

1.5.1. Моделирование распределений статистик критериев согласия

Форма "Моделирование распределений статистик критериев согласия" (для запуска в главном меню **Моделирование** выбрать пункт **Распределения статистик критериев**) позволяет сгенерировать распределения статистик критериев согласия (рис. 1.19). 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.

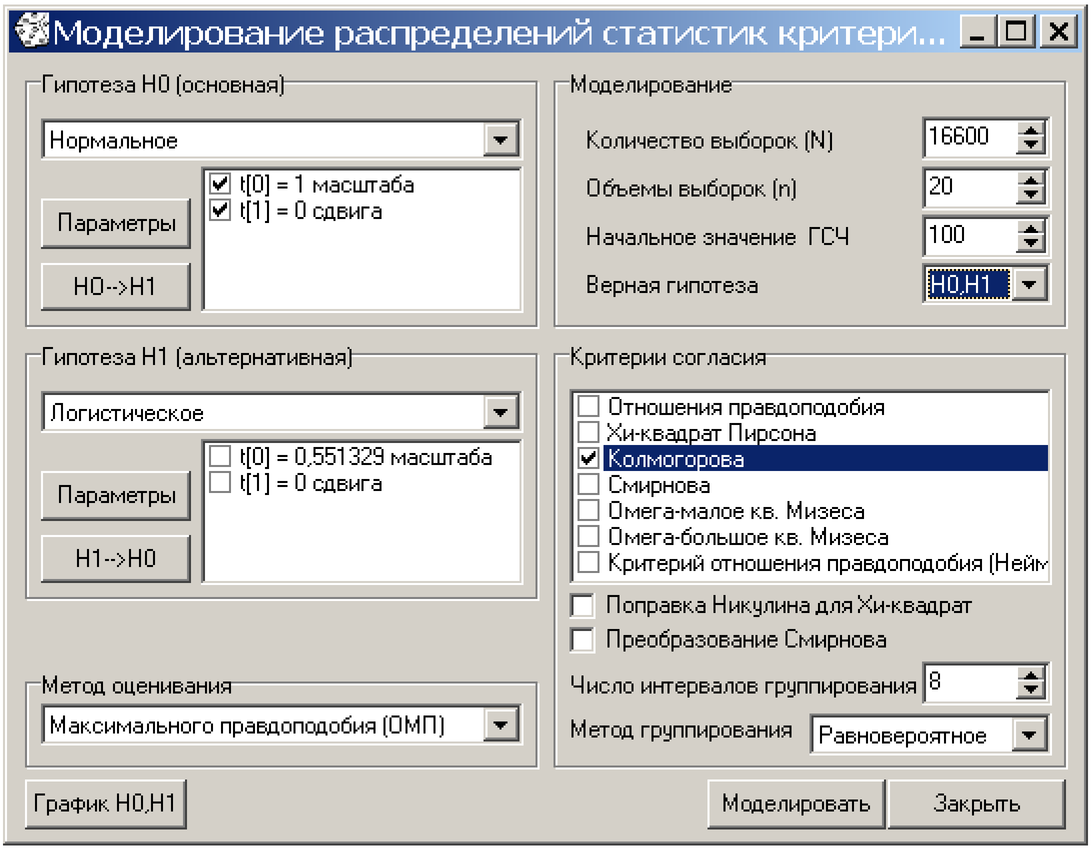


Рис. 1.19. Моделирование распределений статистик критериев согласия

Пример 1.4. *Определить мощность критерия согласия Колмогорова при проверке сложной гипотезы о нормальном распределении против простой гипотезы о логистическом распределении.*

Для решения данной задачи открываем форму **Моделирование распределений статистик критериев согласия** и заполняем параметры, как задано на рис. 1.19. Далее нажимаем на кнопку **Моделировать**.

В результате моделирования создаются два файла: распределение статистики при верной гипотезе H0, и при верной гипотезе H1 (рис. 1.20).
Эти выборки нужно открыть.

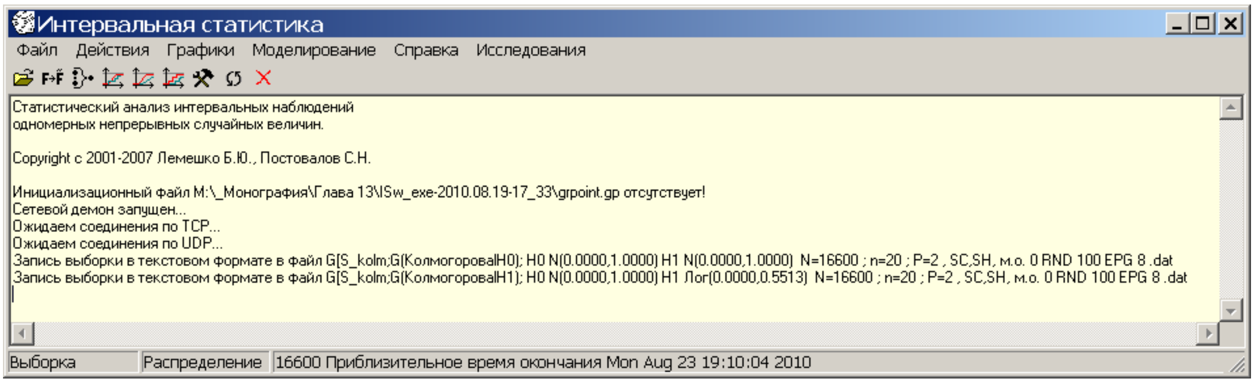


Рис. 1.20. Сообщение системы о создании файлов с выборами статистики Колмогорова

Для вычисления мощности критерия необходимо знать распределения статистики критерия при верной основной гипотезе, а также распределение статистики при верной альтернативной гипотезе.
Если распределения статистик получены с помощью моделирования и сохранены в виде выборок, то можно воспользоваться функцией **Вычисление мощности** (рис. 1.21).
Возможно два варианта задания критической области:

- **односторонний**, когда основная гипотеза отвергается при *больших* положительных значениях статистики;
- **двусторонний**, когда основная гипотеза отвергается при больших значениях *модуля* статистики.

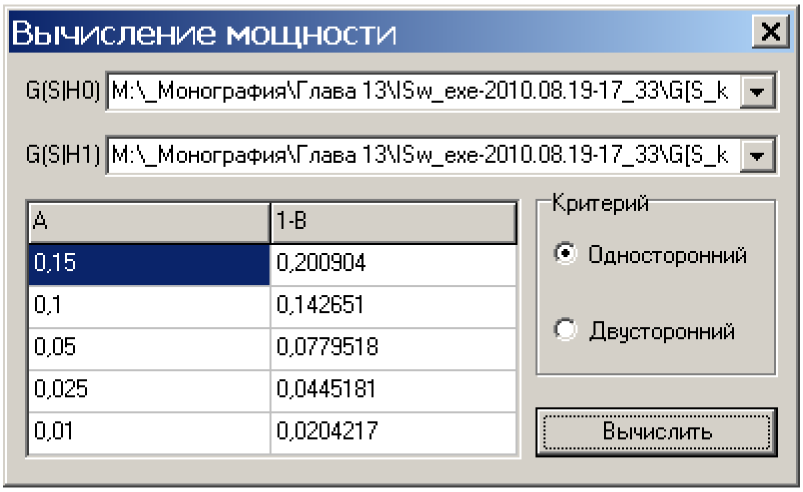


Рис. 1.21. Вычисление мощности

Результаты вычисления мощности можно проверить графически. Для этого нужно открыть графики эмпирических функций распределений на одном рисунке и визуально определить критическую область и вероятность попадания в критическую область при верной альтернативной гипотезе (рис. 1.22). Визуально мощность критерия при вероятности ошибки первого рода равна 0.15, что примерно совпадает с результатами на рис. 1.21.

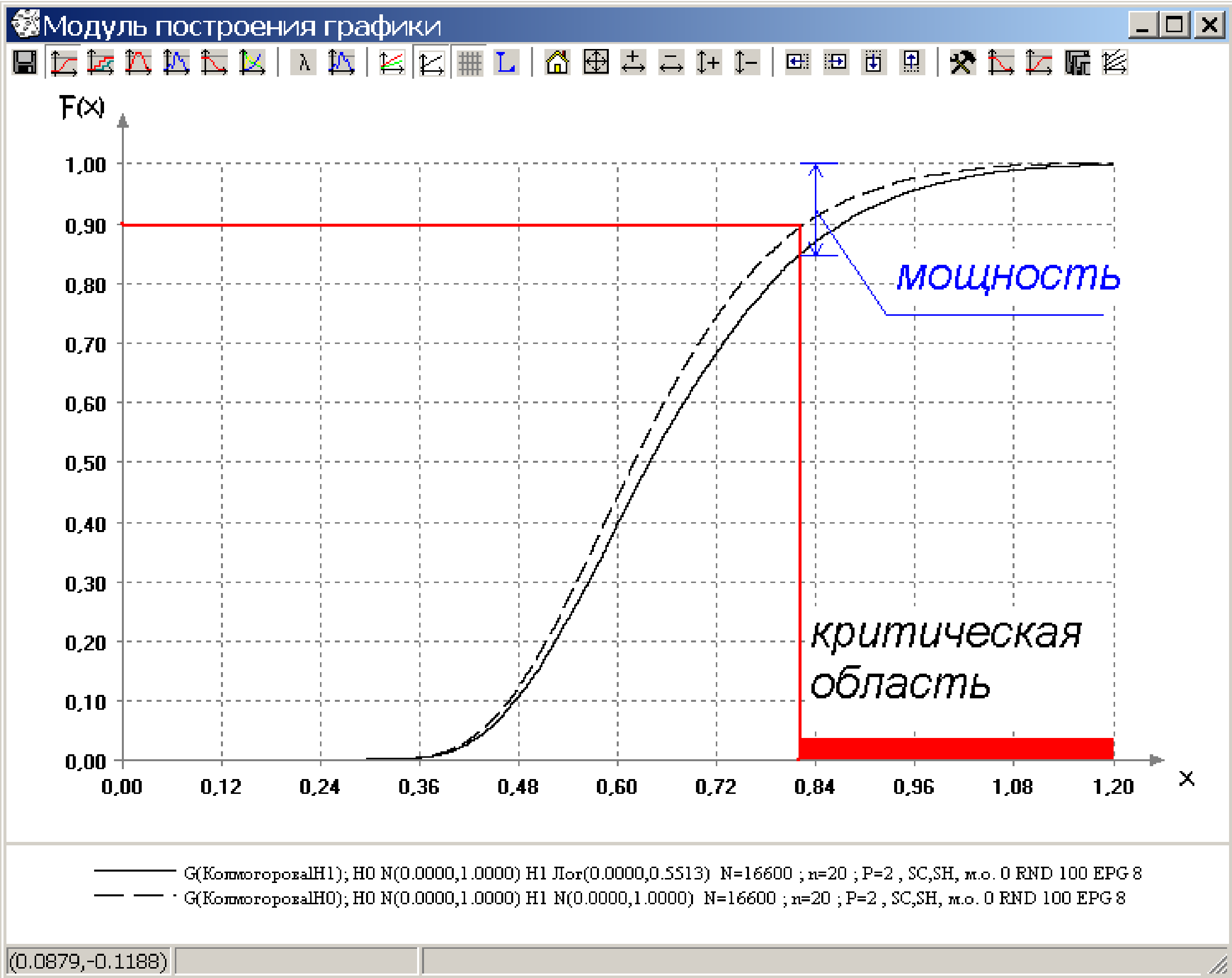


Рис. 1.22. Вычисление мощности графически критерия Колмогорова

2. Метод Монте-Карло

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].

2.1. 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 2.1. 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.

2.2. 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 2.2. 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.

2.3. 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

$$\left|\overline{X} - \mu\right| < \varepsilon, \qquad (1)$$

называют *точностью* оценки $\overline{X}$, а вероятность γ того, что неравенство (1) выполняется, ее *достоверностью*

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

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

2.3.1. 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

$$\varepsilon=t_\gamma\frac{\sqrt{p(1-p)}}{\sqrt{N}}. \qquad (2)$$

Отсюда количество реализаций, необходимое для достижения заданной точности ε равно

$$N=t_\gamma^2\frac{p(1-p)}{\varepsilon^2}. \qquad (3)$$

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 2.3.

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.

2.3.2. 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:

$$F_N(x) = \begin{cases} 0, \; x < x_{(1)} \\ m/N, \; x_{(m)} \le x < x_{(m+1)}, \; m = 1,2,\ldots, N-1, \\ 1, \; x \ge x_{(N)} \end{cases} \tag{4}$$

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:

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

Тогда формулы (2) и (3) для вычисления значения $p = F(x)$ с помощью случайной величины $m/N = F_N(x)$ примут вид

$$\varepsilon = t_\gamma \frac{\sqrt{F(x)(1 - F(x))}}{\sqrt{N}}, \; N = t_\gamma^2 \frac{F(x)(1 - F(x))}{\varepsilon^2}. \tag{5}$$

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. Как и следовало ожидать, точность моделирования вероятности в этих точках достаточно высокая.

Таблица 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

3. Порядок выполнения

1. Выполнить постановку задачи компьютерного моделирования закона распределения по варианту задания.
2. Написать программу для моделирования закона распределения (или некоторой его числовой характеристики).
3. Вычислить необходимый объем выборки N для заданной точности $\varepsilon = 0.01$.
4. Выполнить моделирование распределения статистики, провести исследования согласно варианту задания.
5. Идентифицировать закон распределения с использованием программной системы ISW. Сравнить полученные результаты с теоретическими.

Варианты заданий

Задание 1.

1. Проверка гипотезы о виде распределения

В следующей таблице приведены результаты измерений прочности провода на разрыв в деканьютонах.

235	235	230	232	226	230	231	229	237	235
238	234	229	231	240	237	239	231	233	240
235	239	234	230	236	231	240	232	231	228
234	233	235	227	226	231	230	232	237	238
238	236	230	235	231	230	235	228	233	240

Требуется проверить гипотезу о согласии полученной выборки с
а) нормальным распределением;
б) распределением Лапласа;
в) логистическим распределением.
Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Электричка в метро состоит из трех вагонов по k мест каждый. Три студента равновероятно садятся на любое место в любом вагоне. Найти распределение случайной величины ξ – «количество вагонов, в которых окажутся студенты».

Задание:

1. Написать программу для моделирования закона распределения ξ.
2. Вычислить необходимый объем выборки N для заданной точности $\varepsilon = 0.01$.
3. Выполнить моделирование распределения ξ. Исследовать это распределение при увеличении числа k от 1 до бесконечности.
4. Решить задачу аналитически. Сравнить аналитические результаты с результатами моделирования.*

Задание 2.

1. Проверка гипотезы о виде распределения

В таблице приведено распределение толщины 12 000 бобов.

Толщина, мм	До 7.00	7.00-7.25	7.25-7.5	7.5-7.75	7.75-8.00	8.00-8.25	8.25-8.5	8.5-8.75
Количество бобов	32	103	239	624	1187	1650	1883	1930
Толщина, мм	8.75-9.00	9.00-9.25	9.25-9.5	9.25-9.75	9.75-10.00	10.00-10.25	10.25-10.5	Свыше 10.5
Количество бобов	1638	1130	737	427	221	110	57	32

Проверить гипотезу о том, что толщина бобов подчиняется нормальному распределению. Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

В городе проживает $n+1$ человек. Один из них, узнав новость, сообщает ее другому, тот – третьему, и т.д., причем каждый человек передает новость наугад выбранному жителю, за исключением того от, которого он ее услышал. Пусть ξ – случайная величина, равная числу передач новости от одного человека к другому до момента возвращения к тому человеку, который узнал ее первым.

1. Написать программу для моделирования закона распределения ξ.
2. Вычислить необходимый объем выборки N для заданной точности $\varepsilon = 0.01$.
3. Выполнить моделирование распределения ξ статистики, вычислить среднее значение и дисперсию, исследовать зависимость от n.
4. Решить задачу аналитически. Сравнить аналитические результаты с результатами моделирования.*

Задание 3.

1. Проверка гипотезы о виде распределения

В следующей таблице представлены результаты измерений CO_2 *в граммах на литр в партии газированных напитков.*

7.30	7.00	7.20	6.50	7.00	7.00	7.20	7.20	6.80	6.80
6.40	6.80	6.80	6.60	6.90	7.20	6.60	7.30	7.00	6.80
6.70	6.70	6.40	6.80	7.00	6.40	6.80	6.80	7.20	7.20
6.90	7.10	7.40	7.00	7.20	6.80	7.00	7.40	6.60	7.00
6.30	6.60	7.20	6.60	7.20	6.20	7.00	7.20	6.60	6.80
6.50	7.00	6.80	7.00	7.00	6.40	7.20	7.40	7.10	7.00
7.10	7.10	6.90	7.10	6.80	7.40	7.00	6.80	6.60	6.80

Требуется проверить гипотезу о согласии полученной выборки с нормальным распределением. Определить достигаемый уровень значимости критериев методом Монте-Карло.

2*. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Для оплаты проезда пассажир пользуется двумя электронными карточками, оплачивая проезд наугад выбранной карточкой. Через некоторое время он обнаружил, что на одной из карточек осталось ноль поездок. Найти распределение случайной величины ξ *– «количество поездок, оставшихся на второй карточке», если считать, что на каждой карточке изначально было* n *поездок.*

1. Написать программу для моделирования закона распределения ξ*.*
2. Вычислить необходимый объем выборки N для заданной точности $\varepsilon = 0.01$.
3. Выполнить моделирование распределения ξ *статистики, вычислить среднее значение и дисперсию, исследовать зависимость от* n*.*
4. Решить задачу аналитически. Сравнить аналитические результаты с результатами моделирования.*

Задание 4.

1. Проверка гипотезы о виде распределения

В таблице приведено время реакции людей на звук в миллисекундах

223	104	209	183	180	168	215	172	200	191	197	183	174	176	155	115	163

Проверить гипотезу о нормальном распределении времени реакции на звук. Определить достигаемый уровень значимости критериев методом Монте-Карло.

2*. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Окружность единичного радиуса с центром в нуле имеет северный полюс на положительной полуоси абсцисс. Из полюса случайным образом направлен луч, причем его угол с осью абсцисс распределен равномерно на отрезке $[-\pi/2, \pi/2]$.

Найти распределение случайной величины ξ *- длины хорды внутри окружности. 1. Написать программу для моделирования закона распределения* ξ.

2. Вычислить необходимый объем выборки N для заданной точности $\varepsilon = 0.01$.

3. Выполнить моделирование распределения ξ *статистики, вычислить среднее значение и дисперсию.*

4. Решить задачу аналитически. Сравнить аналитические результаты с результатами моделирования.*

Задание 5.

1. Проверка гипотезы о виде распределения

В 1889 – 1890 гг. был измерен рост в сантиметрах 999 взрослых мужчин (рабочих московских фабрик). Результаты измерений представлены в следующей таблице.

Рост	143-146	147-149	150-152	153-155	156-158
число мужчин	1	2	8	26	65
Рост	159-161	162-164	165-167	168-170	171-173
число мужчин	120	180	201	170	120
Рост	174-176	177-179	180-182	183-185	186-188
число мужчин	64	28	10	3	1

Требуется проверить гипотезу о том, что рост взрослого мужчины имеет нормальное распределение. Определить достигаемый уровень значимости критериев методом Монте-Карло.

2*. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Равнобедренный треугольник образован единичным вектором в направлении оси абсцисс и единичным вектором в случайном направлении. Найти функцию распределения ξ - длины третьей стороны в $\mathbb{R}^2$.

1. Написать программу для моделирования закона распределения ξ.
2. Вычислить необходимый объем выборки N для заданной точности $\varepsilon = 0.01$.
3. Выполнить моделирование распределения ξ статистики, вычислить среднее значение и дисперсию.
4. Решить задачу аналитически. Сравнить аналитические результаты с результатами моделирования.*

Задание 6.

1. Проверка гипотезы о виде распределения

В следующей таблице приведен рост 65 студентов.

180	158	190	182	168	166	183	190	177	164
167	170	163	165	176	174	190	186	176	166
173	173	185	168	170	160	164	159	170	182
168	185	167	173	180	182	172	180	172	163
195	174	162	177	182	176	183	163	168	170
182	152	173	167	164	175	186	169	176	160
177	180	186	180	164					

Проверить гипотезу о нормальном распределении роста.
В следующей таблице приведен рост 42 студентов (мужчин).

180	190	182	168	166	183	190	177	170	176
174	190	186	173	173	185	168	160	170	182
185	173	180	182	172	180	172	195	174	177
182	176	183	170	182	175	186	176	177	180
186	180								

Проверить гипотезу о нормальном распределении роста.
В следующей таблице приведен рост 23 студентов (женщин).

158	164	167	163	165	176	166	170	164	159
168	167	163	162	163	168	152	173	167	164
169	160	164							

Проверить гипотезу о нормальном распределении роста.
Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Равнобедренный треугольник образован единичным вектором в направлении оси абсцисс и единичным вектором в случайном направлении. Найти функцию распределения ξ - длины третьей стороны в $\mathbb{R}^3$.

1. Написать программу для моделирования закона распределения ξ.
2. Вычислить необходимый объем выборки N для заданной точности $\varepsilon = 0.01$.
3. Выполнить моделирование распределения ξ статистики, вычислить среднее значение и дисперсию.
4. Решить задачу аналитически. Сравнить аналитические результаты с результатами моделирования.*

Задание 7.

1. Проверка гипотезы о виде распределения

В следующей таблице приведен вес 65 студентов.

65	55	78	70	55	56	70	75	57	56
58	56	54	52	54	73	67	73	53	60
62	60	63	63	62	59	55	53	55	78
53	78	50	61	75	60	68	80	62	52
75	66	50	73	67	65	60	44	61	60
70	45	55	53	47	65	80	55	64	45
65	75	78	80	70					

Проверить гипотезу о нормальном распределении веса.
В следующей таблице приведен вес 42 студентов (мужчин).

65	78	70	55	56	70	75	57	56	54
73	67	73	62	60	63	63	59	55	78
78	61	75	60	68	80	62	75	66	73
67	65	60	60	70	65	80	64	65	75
78	80								

Проверить гипотезу о нормальном распределении веса.
В следующей таблице приведен вес 23 студентов (женщин).

55	56	58	54	52	53	60	62	55	53
53	50	52	50	44	61	45	55	53	47
55	45	70							

Проверить гипотезу о нормальном распределении веса. Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Найти распределение случайной величины ξ - выборочного квантиля порядка α по выборке объема n *наблюдений на примере экспоненциального распределения.*

Написать программу для моделирования закона распределения ξ.
2. Вычислить необходимый объем выборки N для заданной точности $\varepsilon = 0.01$.
3. Выполнить моделирование распределения ξ статистики, вычислить среднее значение и дисперсию. исследовать их изменение с ростом объема выборки и величины α.
4. Решить задачу аналитически. Сравнить аналитические результаты с результатами моделирования.*

Задание 8.

1. Проверка гипотезы о виде распределения

В следующей таблице представлены результаты измерений CO_2 в граммах на литр в партии газированных напитков.

7.30	7.00	7.20	6.50	7.00	7.00	7.20	7.20	6.80	6.80
6.40	6.80	6.80	6.60	6.90	7.20	6.60	7.30	7.00	6.80
6.70	6.70	6.40	6.80	7.00	6.40	6.80	6.80	7.20	7.20
6.90	7.10	7.40	7.00	7.20	6.80	7.00	7.40	6.60	7.00
6.30	6.60	7.20	6.60	7.20	6.20	7.00	7.20	6.60	6.80
6.50	7.00	6.80	7.00	7.00	6.40	7.20	7.40	7.10	7.00
7.10	7.10	6.90	7.10	6.80	7.40	7.00	6.80	6.60	6.80

Требуется проверить гипотезу о согласии полученной выборки с нормальным распределением. Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Пусть $\{X_1, X_2, \ldots, X_n\}$ *–выборка случайной величины* ξ*, подчиненной закону распределения с функцией распределения* $F(x)$*. Требуется найти распределение статистики* $\sqrt{n}\dfrac{\bar{X} - M\xi}{\sqrt{D\xi}}$ *и исследовать его изменение с ростом объема выборки.*

Проверить зависит ли полученное распределение статистики от закона распределения случайной величины, а также от значений параметров распределения.

Задание 9.

1. Проверка гипотезы о виде распределения

В следующей таблице приведены результаты испытаний 200 ламп на продолжительность работы в часах.

Продолжительность	0 – 300	300 – 600	600 – 900	900 – 1200	1200 – 1500	1500 – 1800
Число ламп, вышедших из строя	53	41	30	22	16	12
Продолжительность	1800 – 2100	2100 – 2400	2400 – 2700	2700 – 3000	3000 – 3300	> 3300
Число ламп, вышедших из строя	9	7	5	3	2	0

Требуется проверить гипотезу о том, что продолжительность работы лампы подчиняется экспоненциальному закону распределения. Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Пусть $\{X_1, X_2, \ldots, X_n\}$ *–выборка случайной величины* ξ*, подчиненной закону распределения с функцией распределения* $F(x)$*. Требуется найти распределение статистики* $\sqrt{n}\frac{\left|\bar{X} - M\xi\right|}{\sqrt{D\xi}}$ *и исследовать его изменение с ростом объема выборки. Проверить зависит ли полученное распределение статистики от закона распределения случайной величины, а также от значений параметров распределения.*

Задание 10.

1. Проверка гипотезы о виде распределения

Диаметры головок заклепок в миллиметрах

Выборка 1

13,39	13,43	13,54	13,64	13,4	13,55	13,4	13,26	13,42	13,5
13,32	13,31	13,28	13,52	13,46	13,63	13,38	13,44	13,52	13,53
13,37	13,33	13,24	13,13	13,53	13,53	13,39	13,57	13,51	13,34
13,39	13,47	13,51	13,48	13,62	13,58	13,57	13,33	13,51	13,4
13,3	13,48	13,4	13,57	13,51	13,4	13,52	14,56	13,4	13,34

Выборка 2

13,23	13,37	13,48	13,48	13,62	13,35	13,4	13,36	13,45	13,48
13,29	13,58	13,44	13,56	13,28	13,59	13,47	13,46	13,62	13,54
13,2	13,38	13,43	13,36	13,56	13,51	13,47	13,4	13,29	13,2
13,46	13,44	13,42	13,29	13,41	13,39	13,5	13,48	13,53	13,34
13,45	13,42	13,29	13,38	13,45	13,5	13,55	13,33	13,32	13,69

Требуется проверить гипотезу о том, что диаметры головок заклепок подчиняются нормальному закону распределения. Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Пусть $\{X_1, X_2, \ldots, X_n\}$ *–выборка случайной величины* ξ*, подчиненной закону распределения с функцией распределения* $F(x)$*. Требуется найти распределение статистики* $\frac{1}{n\sqrt{D\xi}}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ *и исследовать его изменение с ростом объема выборки. Проверить зависит ли полученное распределение статистики от закона распределения случайной величины, а также от значений параметров распределения.*

Задание 11.

1. Проверка гипотезы о виде распределения

Имеются данные о численности 20 малых предприятий.

№	Число работников (чел)	№	число работников (чел)	№	Число работников (чел)	№	Число работников (чел)
1	12	6	40	11	42	16	63
2	15	7	30	12	50	17	28
3	36	8	19	13	45	18	34
4	17	9	27	14	10	19	42
5	70	10	29	15	21	20	27

Требуется проверить гипотезу о том, что численность малых предприятий подчиняется нормальному закону распределения. Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Пусть $\{X_1, X_2, \ldots, X_n\}$ *–выборка случайной величины* ξ, *подчиненной закону распределения с функцией распределения* $F(x)$. *Требуется найти распределение статистики* $\sqrt{n-1}\frac{\overline{X}}{S^2}$ *и исследовать его изменение с ростом объема выборки. Проверить зависит ли полученное распределение статистики от закона распределения случайной величины, а также от значений параметров распределения.*

Задание 12.

1. Проверка гипотезы о виде распределения

В следующей таблице приведены результаты испытаний 200 ламп на продолжительность работы в часах.

Продолжительность	0 – 300	300 – 600	600 – 900	900 – 1200	1200 – 1500	1500 – 1800
Число ламп, вышедших из строя	53	41	30	22	16	12
Продолжительность	1800 – 2100	2100 – 2400	2400 – 2700	2700 – 3000	3000 – 3300	> 3300
Число ламп, вышедших из строя	9	7	5	3	2	0

Требуется проверить гипотезу о том, что продолжительность работы лампы подчиняется экспоненциальному закону распределения. Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Пусть $\{X_1, X_2, \ldots, X_n\}$ *–выборка случайной величины* ξ, *подчиненной закону распределения с функцией распределения* $F(x)$. *Требуется найти распределение статистики* $\sqrt{n}D_n$, *где* $D_n = \max(D_n^+, D_n^-)$, $D_n^+ = \max\limits_{1\le i\le n}\left\{\frac{i}{n} - F(X_{(i)}, \theta)\right\}$,

$D_n^- = \max\limits_{1\le i\le n}\left\{F(X_{(i)}, \theta) - \frac{i-1}{n}\right\}$ *и исследовать его изменение с ростом объема выборки.*

Проверить зависит ли полученное распределение статистики от закона распределения случайной величины, а также от значений параметров распределения.

Задание 13.

1. Проверка гипотезы о виде распределения

В таблице приведено время реакции людей на свет в миллисекундах

181	194	173	153	168	176	163	152	155	156	178	160	164	169	155	122	144

Требуется проверить гипотезу о согласии полученной выборки с
а) нормальным распределением;
б) распределением Лапласа;
в) логистическим распределением.
Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Пусть $\{X_1, X_2, \ldots, X_n\}$ *–выборка случайной величины* ξ, *подчиненной закону распределения с функцией распределения* $F(x)$. *Требуется найти распределение статистики* $\frac{6nD_n+1}{6\sqrt{n}}$ *,где* $D_n = \max(D_n^+, D_n^-)$, $D_n^+ = \max\limits_{1\le i\le n}\left\{\frac{i}{n} - F(X_{(i)},\theta)\right\}$

$D_n^- = \max\limits_{1\le i\le n}\left\{F(X_{(i)},\theta) - \frac{i-1}{n}\right\}$ *и исследовать его изменение с ростом объема выборки.*

Проверить зависит ли полученное распределение статистики от закона распределения случайной величины, а также от значений параметров распределения.

.

Задание 14.

1. Проверка гипотезы о виде распределения

С целью совершенствования системы тарифов городская АТС проводит исследование продолжительности телефонных разговоров в рабочее время, распределение которых по данным выборки представлено в таблице:

Продолжительность телефонных разговоров, мин.	Количество разговоров
менее 4	10
4 – 5	20
5 - 6	30
6 - 7	35
7 - 8	25
8 - 9	15
9 - 10	10
10 и более	5
Итого	150

Проверьте, соответствует ли распределение продолжительности телефонных разговоров нормальному закону. Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Пусть $\{X_1, X_2, \ldots, X_n\}$ *–выборка случайной величины* ξ, *подчиненной закону распределения с функцией распределения* $F(x)$. *Требуется найти распределение статистики* $\dfrac{(6nD_n^+ + 1)^2}{9n}$ *,где* $D_n = \max(D_n^+, D_n^-)$, $D_n^+ = \max\limits_{1 \le i \le n} \left\{ \dfrac{i}{n} - F(X_{(i)}, \theta) \right\}$

$D_n^- = \max\limits_{1 \le i \le n} \left\{ F(X_{(i)}, \theta) - \dfrac{i-1}{n} \right\}$ *и исследовать его изменение с ростом объема выборки.*

Проверить зависит ли полученное распределение статистики от закона распределения случайной величины, а также от значений параметров распределения.

Задание 15.

1. Проверка гипотезы о виде распределения

Имеется распределение партий деталей по длительности производственного цикла:

Длительность, час	менее 28	28-113	113-198	198-283	283-368	368-453	453-538	538-623	623-708	708-и более	Итого
Количество партий	-	5	12	12	15	9	9	7	2	-	71

Требуется проверить гипотезу о согласии полученной выборки с
а) нормальным распределением;
б) распределением Лапласа;
в) логистическим распределением.
Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Пусть $\{X_1, X_2, \ldots, X_n\}$ *–выборка случайной величины* ξ*, подчиненной закону распределения с функцией распределения* $F(x)$*. Требуется найти распределение статистики* $\frac{(6nD_n^- + 1)^2}{9n}$ *,где* $D_n = \max(D_n^+, D_n^-)$, $D_n^+ = \max\limits_{1 \le i \le n}\left\{\frac{i}{n} - F(X_{(i)}, \theta)\right\}$

$D_n^- = \max\limits_{1 \le i \le n}\left\{F(X_{(i)}, \theta) - \frac{i-1}{n}\right\}$ *и исследовать его изменение с ростом объема выборки.*

Проверить зависит ли полученное распределение статистики от закона распределения случайной величины, а также от значений параметров распределения.

Задание 16.

1. Проверка гипотезы о виде распределения

В таблице приведено потребление вина в год (в литрах)

115	102	98	65	63	45	82	90	180	178
160	175	161	177	187	208	78	99	91	89
86	92	134	124	162	245	120	136	146	129
196	198								

Требуется проверить гипотезу о согласии полученной выборки с
а) нормальным распределением;
б) распределением Лапласа;
в) логистическим распределением.
Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Пусть $\{X_1, X_2, \ldots, X_n\}$ *–выборка случайной величины* ξ, *подчиненной закону распределения с функцией распределения* $F(x)$. *Требуется найти распределение статистики* $\frac{1}{12n} + \sum_{i=1}^{n} \left\{ F(X_{(i)}, \theta) - \frac{2i-1}{2n} \right\}^2$ *и исследовать его изменение с ростом объема выборки. Проверить зависит ли полученное распределение статистики от закона распределения случайной величины, а также от значений параметров распределения.*

Задание 17.

1. Проверка гипотезы о виде распределения

В таблице приведено потребление пива в год в литрах

420	350	320	398	388	345	355	362	295	299
209	236	198	195	185	203	270	312	308	250
260	235	255	265	170	150	120	143	133	121
116	118								

Требуется проверить гипотезу о согласии полученной выборки с
а) нормальным распределением;
б) распределением Лапласа;
в) логистическим распределением.
Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Пусть $\{X_1, X_2, \ldots, X_n\}$ *–выборка случайной величины* ξ*, подчиненной закону распределения с функцией распределения* $F(x)$*. Требуется найти распределение статистики* $-n-2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(X_{(i)},\theta)+\left(1-\frac{2i-1}{2n}\right)\ln(1-F(X_{(i)},\theta))\right\}$ *и исследовать его изменение с ростом объема выборки. Проверить зависит ли полученное распределение статистики от закона распределения случайной величины, а также от значений параметров распределения.*

Задание 18.

1. Проверка гипотезы о виде распределения

Имеются данные о ежедневной сумме выручки 15 магазинов (тыс. руб.)

№ магазинов	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Сумма выручки	20	15	30	19	27	12	13	29	32	17	18	21	23	36	25

Требуется проверить гипотезу о согласии полученной выборки с
а) нормальным распределением;
б) распределением Лапласа;
в) логистическим распределением.
Определить достигаемый уровень значимости критериев методом Монте-Карло.

2. Применение метода Монте-Карло в задачах теории вероятностей и математической статистики

Пусть $\{X_1, X_2, \ldots, X_n\}$ *–выборка случайной величины* ξ*, подчиненной закону распределения с функцией распределения* $F(x)$*. Требуется найти распределение статистики* $\frac{1}{n\sqrt{D\xi}}\sum_{i=1}^{n}(X_i - M\xi)^2$ *и исследовать его изменение с ростом объема выборки. Проверить зависит ли полученное распределение статистики от закона распределения случайной величины, а также от значений параметров распределения.*

Литература

1. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1983. - 416 с.
2. 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.
3. Мирвалиев М., Никулин М.С. Критерии согласия типа хи-квадрат / Заводская лаборатория. 1992. Т. 58. № 3. С.52-58.
4. Lemeshko B.Yu. Lemeshko S.B. and Postovalov S.N. Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses // Communications in Statistics - Theory and Methods, 2010. Vol. 39, No. 3. – P. 460-471.
5. Freireich, E. J., Gehan, E. A., Frei, E. The Effect of 6-Mercaptopurine on the Duration of Steroid-Induced Remissions in Acute Leukemia: A Model for Evaluation of Other Potential Useful Therapy. // Blood, 1963. V. 21 No. 6. — P. 699-716.