**УДК 004.588**

**ББК 74.5**

**Ч-751**

**В. З. Чокой**

**Иркутск, Россия**

# КОМПЬЮТЕРНЫЕ СРЕДСТВА ОПИСАТЕЛЬНОЙ СТАТИСТИКИ – ПЕРВИЧНОГО АНАЛИЗА И ПОДБОРА ТЕОРЕТИЧЕСКИХ РАСПРЕДЕЛЕНИЙ ДЛЯ НАБЛЮДАЕМЫХ ПРОЦЕССОВ

Рассмотрены математические подходы к первичному анализу статистических данных по наблюдаемым процессам и подбора для них адекватных теоретических распределений, а также оценке параметров подобранных распределений. Представлен обзор авторских компьютерных средств (инструментов) решения перечисленных задач описательной статистики. Рассмотрены основные интерфейсные решения по инструментам и показания по их применению в образовательной, инженерной и исследовательской практиках.

**Ключевые слова**: наблюдаемый процесс, описательная статистика, статистические показатели, эмпирическое распределение, закон теоретического распределения, критерии проверки гипотез.

**V. Z. Chokoj**

**Irkutsk, Russia**

# COMPUTER AIDS OF DESCRIPTIVE STATISTICS AS INITIAL ANALYSIS AND SELECTION OF PREDICTIVE DISTRIBUTIONS FOR OBSERVED PROCESSES

The article considers mathematical approaches to the initial analysis of statistic data and selection of adequate predictive distributions as well as estimation of parameters of selected distributions. The author’s computer aids (tools) for solving the

problems of descriptive statistics are reviewed. The article describes the main tools interface solutions and proposes their use in educational and engineering practices and in research scientific training.

**Key words**: observed process, descriptive statistics, statistical values, empirical distribution, predictive distribution law, hypothesis tests.

Аппарат описательной статистики, в частности, первичный анализ статистических данных актуален при решении различных практических задач, базирующихся на использовании накопленных числовых массивов. При этом у пользователя появляются возможности для количественной характеристики эмпирических массивов стандартизованными единичными показателями и функциями, для подбора эмпирическим массивам адекватных теоретических распределений, для настройки подобранных распределений путем оценки численных значений параметров распределений.

Главное, подобрав для наблюдаемого процесса соответствующее теоретическое распределение, становится доступным мощный аппарат этого распределения, позволяющий выполнять параметрические оценки гаммы показателей, характеризующих, например, надежность наблюдаемых объектов: безотказность (вероятность отказа и безотказной работы, среднюю наработку на отказ, интенсивность отказов), долговечность (ресурс и строк службы), сохраняемость (средний срок хранения, периодичность профилактических работ), ремонтопригодность (средний срок восстановления, параметр потока восстановлений), а также комплексных показателей в виде различных коэффициентов готовности.

*Оценка статистических показателей и функций эмпирического распределения* в целом для пользователей не составляет особых затруднений, однако при большом размере исходного ряда вычисления становятся громоздкими и затруднительными для ручного счета. К основным статистическим показателям и функциям следует отнести: размах, среднее, верхнюю и нижнюю доверительные границы среднего, усеченное среднее, моду, медиану, дисперсию и среднее

квадратическое отклонение, ассиметрию, эксцесс, коэффициент вариации, предельную относительную ошибку, интегральную функцию распределения, дифференциальную функцию (плотность) распределения. Расчетные формулы и правила их определения достаточно широко освещены в литературе, например, в [Айвазян, 1998; Андерсон, 1976; Тюрин, 1995].

Найденные значения показателей и функции позволяют решать широкий круг частных задач, например, выполнять первичный анализ процесса, сравнивать несколько процессов, оценивать достаточность накопленных данных, выполнять непараметрическую (статистическую) оценку надежности или иных свойств объектов, выдвигать гипотезы о подходящем теоретическом законе распределения наблюдаемых параметров и т. д.

*Подбор теоретических распределений*. Часто проверку гипотезы о соответствии выбранного теоретического распределения эмпирическому ряду выполняют с использованием одного из критериев – Пирсона (критерий $\chi^2$) или Фишера (*F*-критерий).

Подбор распределения по критерию Пирсона:

1. Выдвигается гипотеза о соответствии эмпирического ряда одному из конкретных теоретических распределений.

2. Для исходного эмпирического ряда рассчитывают потребный единичный интервал $\Delta_t = \frac{t_{max}-t_{min}}{1+3{,}21\cdot \lg(n)}$ и потребное число интервалов разбиения $k = \frac{t_{max}-t_{min}}{\Delta_t}$, где $t_{max} - t_{min}$ – размах вариации аргумента *t*.

3. Рассчитывают фактическое значение критерия Пирсона

$$\chi^2_{\Phi} = \sum_{i=1}^{k} \frac{(m_i - n\cdot p_i)^2}{n\cdot p_i},$$

где $m_i$ – эмпирическое количество значений аргумента *t*, попадающих в *i*-й интервал;

*n* – количество значений в эмпирическом ряде;

*k* – количество интервалов разбиения в эмпирическом ряде;

$p_i$ – теоретическая вероятность попадания аргумента *t* в *i*-й интервал.

4. Рассчитав число степеней свободы $r = k - s$ (где $s$ – число параметров, характеризующих выбранное теоретическое распределение) и задавшись уровнем значимости $\alpha$ по таблице квантилей распределения $\chi^2$ (например, в [Айвазян, 1998]) определяют пороговое значение критерия Пирсона $\chi^2_{r,\alpha}$.

5. Сравнивают фактическое и пороговое значения. Если $\chi^2_{\text{ф}} \leq \chi^2_{r,\alpha}$, то гипотеза о соответствии эмпирического ряда теоретическому распределению подтверждается, а при невыполнении неравенства – отвергается (что требует перехода к пункту 1).

Подбор распределения по критерию Фишера:

1. Выдвигается гипотеза о соответствии эмпирического ряда одному из конкретных теоретических распределений. Для значений аргумента из эмпирического ряда по формулам теоретического распределения рассчитывают теоретические значения функции $y_{i\,\text{т}}$.

2. Рассчитывают фактическое значение критерия Фишера

$$F_{\text{ф}} = \frac{D \cdot (n-m-1)}{m \cdot (1-D)},$$

где: $n$ – число значений функции в эмпирическом ряде;

$m$ – количество параметров, характеризующих теоретическое распределение;

$D$ – коэффициент множественной детерминации, определяемый как

$$D = 1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_{\text{общ}}},$$

здесь $\sigma^2_{\text{ост}}$ – остаточная дисперсия, определяемая как

$$\sigma^2_{\text{ост}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_{i\,\text{ф}} - y_{i\,\text{т}}\right)^2;$$

$\sigma^2_{\text{общ}}$ – общая дисперсия, определяемая как $\sigma^2_{\text{общ}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_{i\,\text{ф}} - \bar{y}_{\text{ф}}\right)^2$;

$\bar{y}_{\text{ф}}$ – среднее функции по эмпирическому ряду, $\bar{y}_{\text{ф}} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_{i\,\text{ф}}$;

$y_{i\,\text{ф}}$ – $i$-е значение функции по эмпирическому ряду;

$y_{i\,\text{т}}$ – $i$-е значение функции по теоретическому распределению.

3. Определив два числа степеней свободы ($f_1 = m$; $f_2 = n - m - 1$) и задавшись уровнем значимости $\alpha$ по таблице квантилей $F$-распределения (например, в [1])

определяют пороговое значение критерия Фишера $F_{f1,f2,\alpha}$.

4. Сравнивают фактическое и пороговое значения. Если $F_{\text{ф}} \leq F_{f1,f2,\alpha}$, то гипотеза о соответствии эмпирического ряда теоретическому распределению подтверждается, а при невыполнении неравенства – отвергается (что требует перехода к пункту 1).

*Оценка параметров теоретических распределений* для таких широко используемых теоретических распределений как: равномерное, биномиальное, пуассоновское, экспоненциальное, нормальное, логнормальное− оценка параметров обычно не вызывает вычислительных трудностей. Порядок оценки этих параметров представлен, например, в [Айвазян, 1998; Андерсон, 1976; Тюрин, 1995]. Ниже детально рассмотрен достаточно громоздкий порядок оценки параметров распределения Вейбулла, нашедшего самое широкое распространение на практике в силу определенной универсальности и развитого математического аппарата.

*Оценка параметров распределения Вейбулла*. Точечную оценку параметров места $a$ и параметра формы $b$ (для числа наблюдаемых объектов $N \geq 10$) вычисляют в зависимости от плана наблюдений в следующей последовательности.

Для планов [NUr], [NUT] значение $a$ рассчитывается по формуле

$$a = \frac{\left[\sum_{i=1}^{m} t_i^b + (N-m)\cdot t_m^b\right]^{1/b}}{m},$$

а значение $b$ определяется из уравнения

$$\left(\frac{m}{b} + \sum_{i=1}^{m} \ln t_i\right) \cdot \left[\sum_{i=1}^{m} t_i^b + (N-m)\cdot t_m^b\right] - m \cdot \left[\sum_{i=1}^{m} t_i^b \cdot \ln t_i + (N-m)\cdot t_m^b \cdot \ln t_m\right] = 0.$$

Для планов [NRr], [NRT] значение $a$ рассчитывается по формуле

$$a = \frac{\left[\sum_{i=1}^{m} t_i^b + \sum_{j=1}^{N}\left(t_m - \sum_{k=1}^{m_j} t_{jk}\right)^b\right]^{1/b}}{m},$$

а значение $b$ определяется из уравнения

$$\left(\frac{m}{b} + \sum_{i=1}^{m} \ln t_i\right) \cdot \left[\sum_{i=1}^{m} t_i^b + \sum_{j=1}^{N}\left(t_m - \sum_{k=1}^{m_j} t_{jk}\right)^b\right] - m \cdot \left[\sum_{i=1}^{m} t_i^b \cdot \ln t_i + \sum_{j=1}^{N}\left(t_m - \sum_{k=1}^{m_j} t_{jk}\right)\cdot \ln\left(t_m - \sum_{k=1}^{m_j} t_{jk}\right)\right] = 0.$$

В обоих случаях решение уравнений для $b$ осуществляют методом последовательных приближений, суть которого состоит в следующем. Вначале решается уравнение правдоподобия относительно $b$ по рекуррентной формуле

$$b_{k+1} = b_k + \frac{\frac{1}{b_k} + \frac{S_1}{m} - \frac{S_3}{S_2}}{\frac{1}{b_k} + \frac{S_2 \cdot S_4 - S_3^2}{S_2^2}},$$

где $S_1 = \sum_{i=1}^{m} \ln t_i$;

$S_2 = \sum_{i=1}^{m} t_i^{b_k} + (N-m) \cdot t_m^{b_k}$ – для планов наблюдений [NUr], [NUT];

$S_2 = \sum_{i=1}^{m} t_i^{b_k} + \sum_{j=1}^{N} \left(t_m - \sum_{k=1}^{m_j} t_{jk}\right)^{b_k}$ – для планов наблюдений [NRr], [NUT];

$S_3 = \sum_{i=1}^{m} t_i^{b_k} \cdot \ln t_i + (N-m) \cdot t_m^{b_k} \cdot \ln t_m$ – для планов наблюдений [NUr], [NUT];

$S_3 = \sum_{i=1}^{m} t_i^{b_k} \cdot \ln t_i + \sum_{j=1}^{N} \left(t_m - \sum_{k=1}^{m_j} t_{jk}\right)^{b_k} \cdot \ln\left(t_m - \sum_{k=1}^{m_j} t_{jk}\right)$ – для планов наблюдений [NRr], [NRT];

$S_4 = \sum_{i=1}^{m} t_i^{b_k} \cdot \ln t_i + (N-m) \cdot t_m^{b_k} \cdot ln^2 t_m$ – для планов наблюдений [NUr], [NUT];

$S_4 = \sum_{i=1}^{m} t_i^{b_k} \cdot ln^2 t_i + \sum_{j=1}^{N} \left(t_m - \sum_{k=1}^{m_j} t_{jk}\right)^{b_k} \cdot ln^2\left(t_m - \sum_{k=1}^{m_j} t_{jk}\right)$ – для планов наблюдений [NRT], [NRr];

$b_k$ – это $k$-е приближение к искомому корню $b$;

$t_i$ – $i$-е значение в исходном ряде (например, наработка до $i$-го отказа);

$t_m$ – $m$-е (последнее) значение в исходном ряде;

$t_{jk}$ – наработка (например) $j$-го объекта между ($k$ –1)-м и $k$-м отказами ($j = \overline{1, N}$);

$m$ – число значений в исходном ряде наблюдений;

$m_j$ – число регистрируемых событий (например, отказов) у $j$-го изделия;

[N..] – планы наблюдений при формировании исходного ряда – в соответствии с [4].

Процесс нахождения $b$ прекращают, как только разность между очередными смежными приближениями $b_k$ не станет меньше задаваемой точности решения

ε (обычно ε = 0.05…0.1). Начальное приближение $b_0$ с целью уменьшения числа вычислительных итераций табулировано в [3] в зависимости от величины коэффициента вариации υ, определяемого по формуле

$$\upsilon = \frac{\sqrt{\frac{1}{m-1}\cdot\sum_{i=1}^{m}\left(t_i-\frac{1}{m}\cdot\sum_{i=1}^{m}t_i\right)^2}}{\frac{1}{m}\cdot\sum_{i=1}^{m}t_i}.$$

Для малого числа наблюдаемых объектов ($N < 10$) корректная оценка параметров теоретических распределения проблематична.

*Функциональность и интерфейсные решения по инструментам описательной статистики*. Исходя из очевидной актуальности, и в соответствии с рассмотренными выше подходами, на факультете Эксплуатации летательных аппаратов Иркутского филиала МГТУ ГА сформирован пакет инструментов описательной статистики. Эти инструменты включены в многофункциональный пакет Модельер 2.0 и для пользователей доступны через группу «Процессы в системах» головного меню (*рис. 1*).

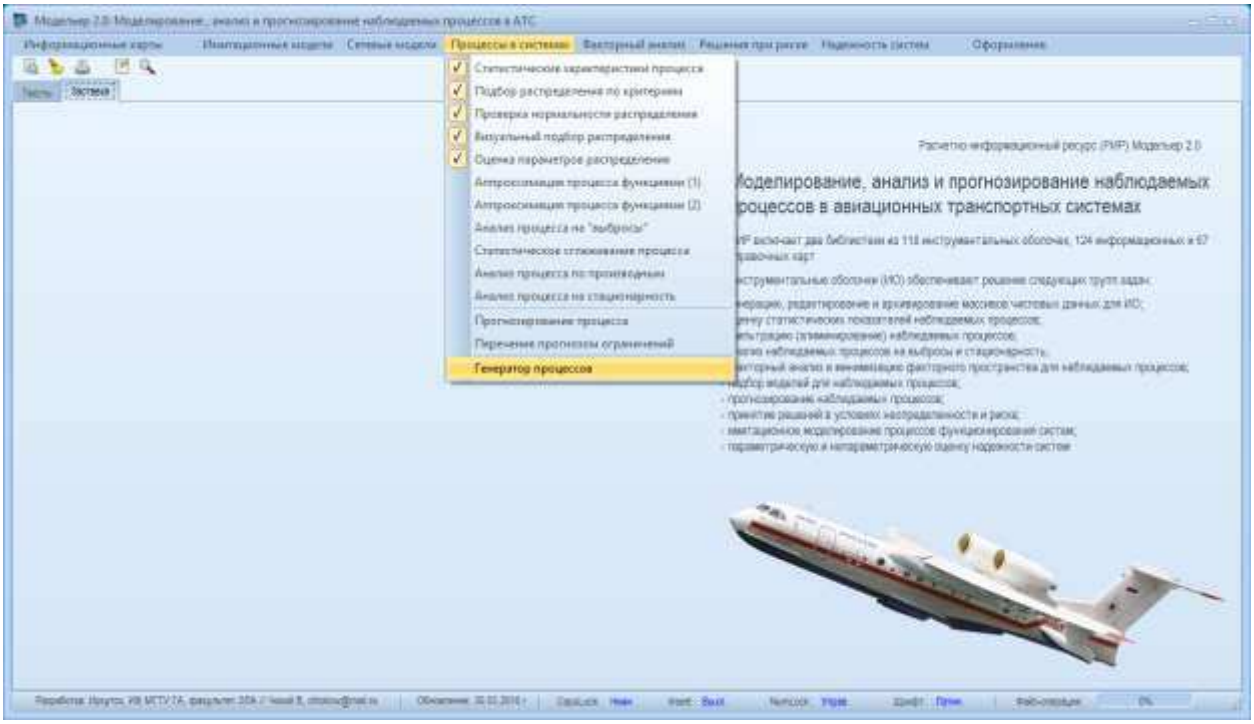


*Рис. 1.* Головная панель пакета Модельер 2.0 (раскрыта группа «Процессы в системах» головного меню)

Из данной группы меню в пакет Модельер 2.0 включены инструменты:

- статистические характеристики процессов (*рис. 2*);

- доверительная вероятность среднего значения процесса;

- проверка нормального распределения по критерию $\chi^2$ (*рис. 3*);

- подбор закона распределения процесса по критерию Романовского, коэффициенту вариации, эксцессу, интенсивности событий (*рис. 4*);

- визуальный подбор адекватного закона распределения (*рис. 5*);

- оценка параметров теоретических распределений (*рис. 6*).

*Рис. 2.* Панель инструментов «Статистические характеристики процессов» и «Доверительная вероятность среднего значения процесса»

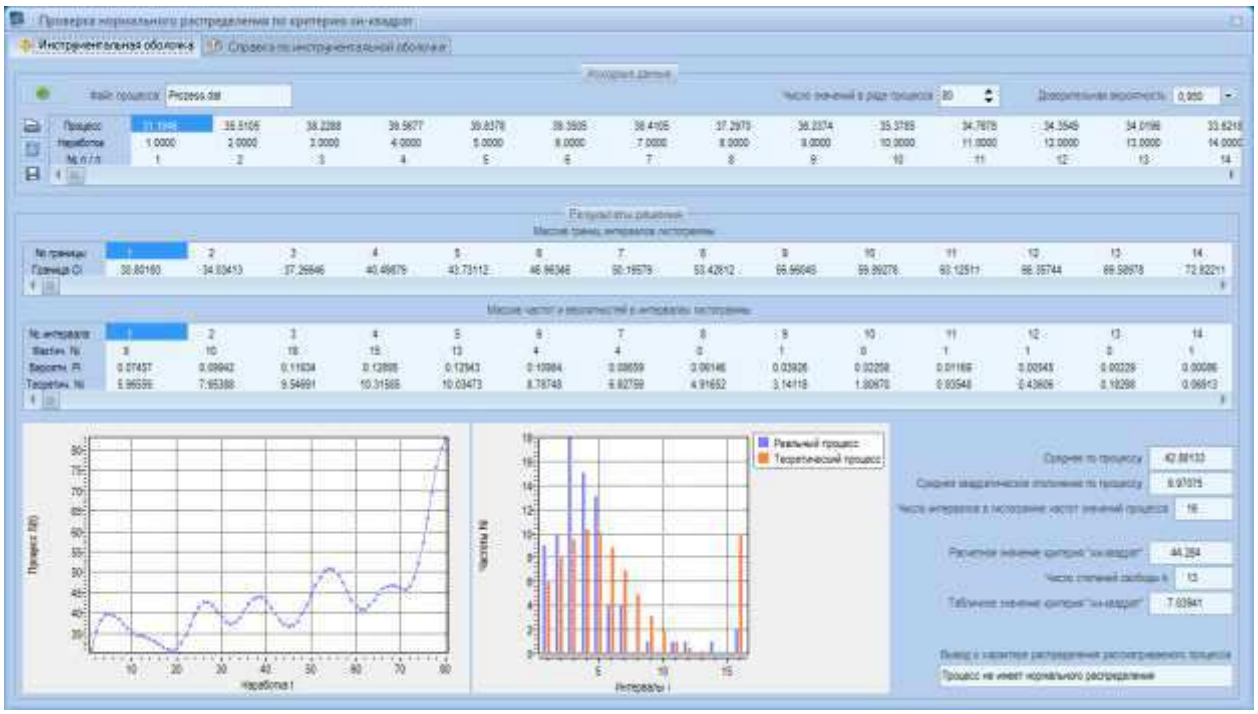

*Рис. 3.* Панель инструмента «Подбор распределения по критерию $\chi^2$»

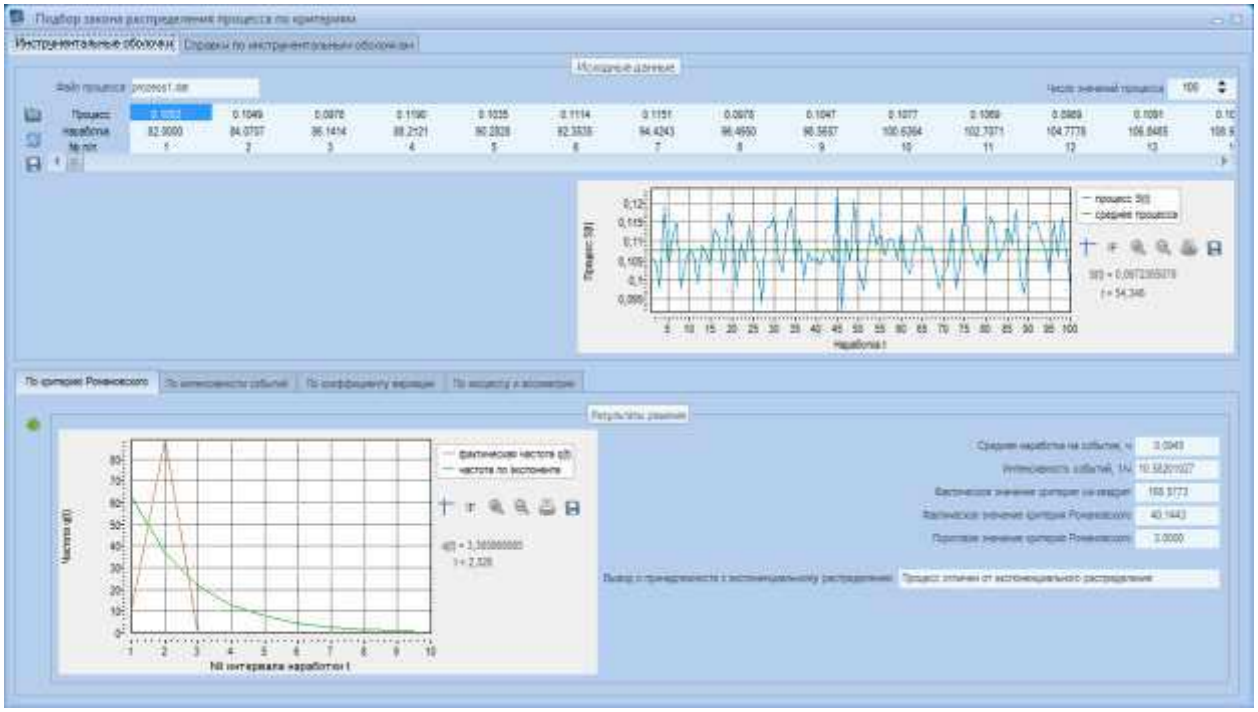

*Рис. 4.* Панель инструмента «Подбор закона распределения по критериям»

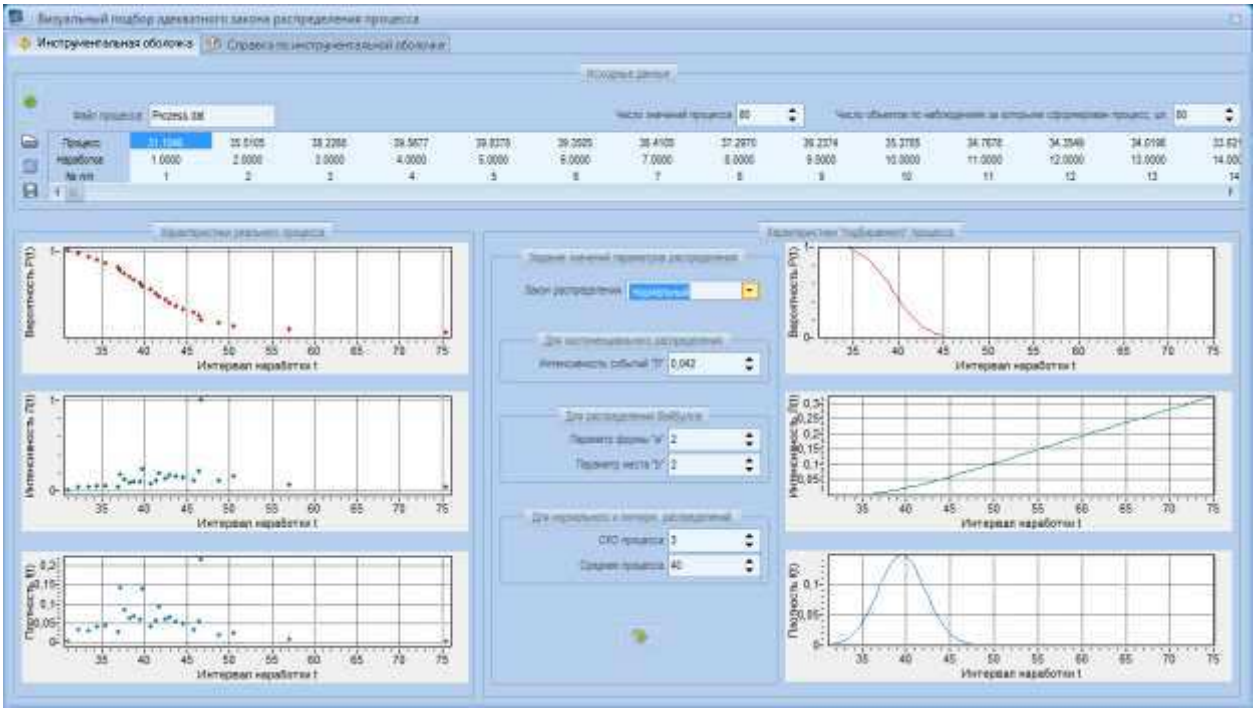

*Рис. 5.* Панель инструмента «Визуальный подбор адекватного распределения»

На панелях всех перечисленных инструментов предусмотрена закладка «Справка по инструментальной оболочке», в которую загружается справочная информация, необходимая пользователям с недостаточным опытом работы (*рис. 7*). Справка содержит следующую информацию: проверка работоспособности инструмента, показания к применению инструмента, примеры решаемых задач, особенности подготовки исходных данных и интерпретации получаемых результатов.

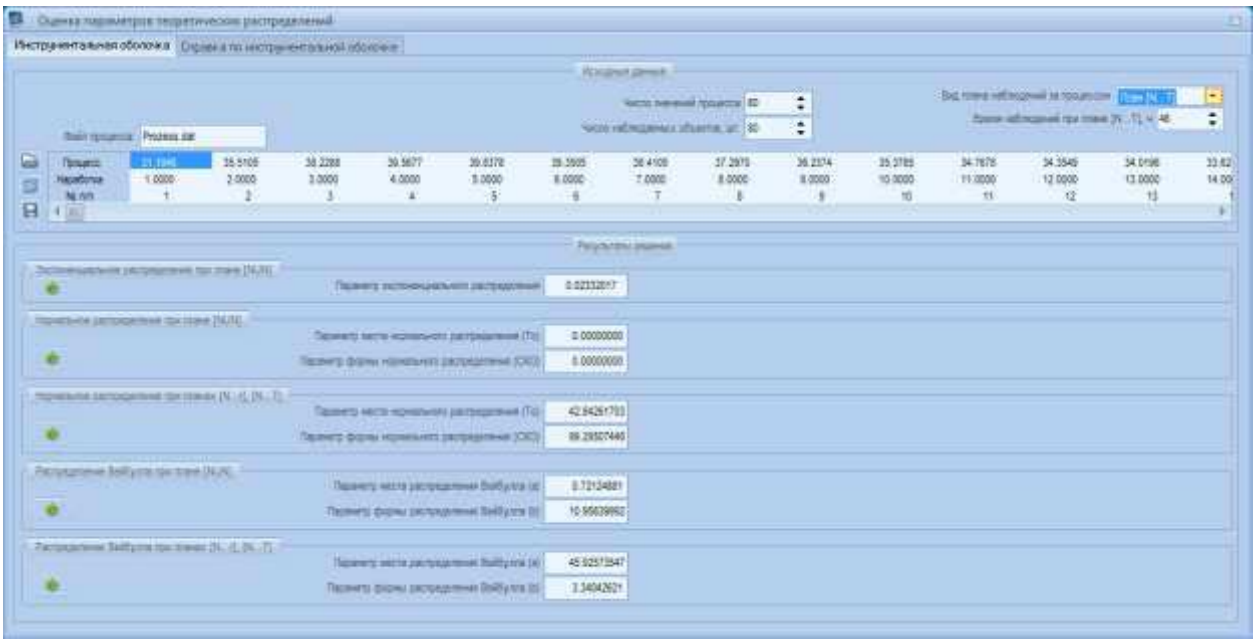

*Рис. 6.* Панель инструмента «Оценка параметров теоретических распределений»

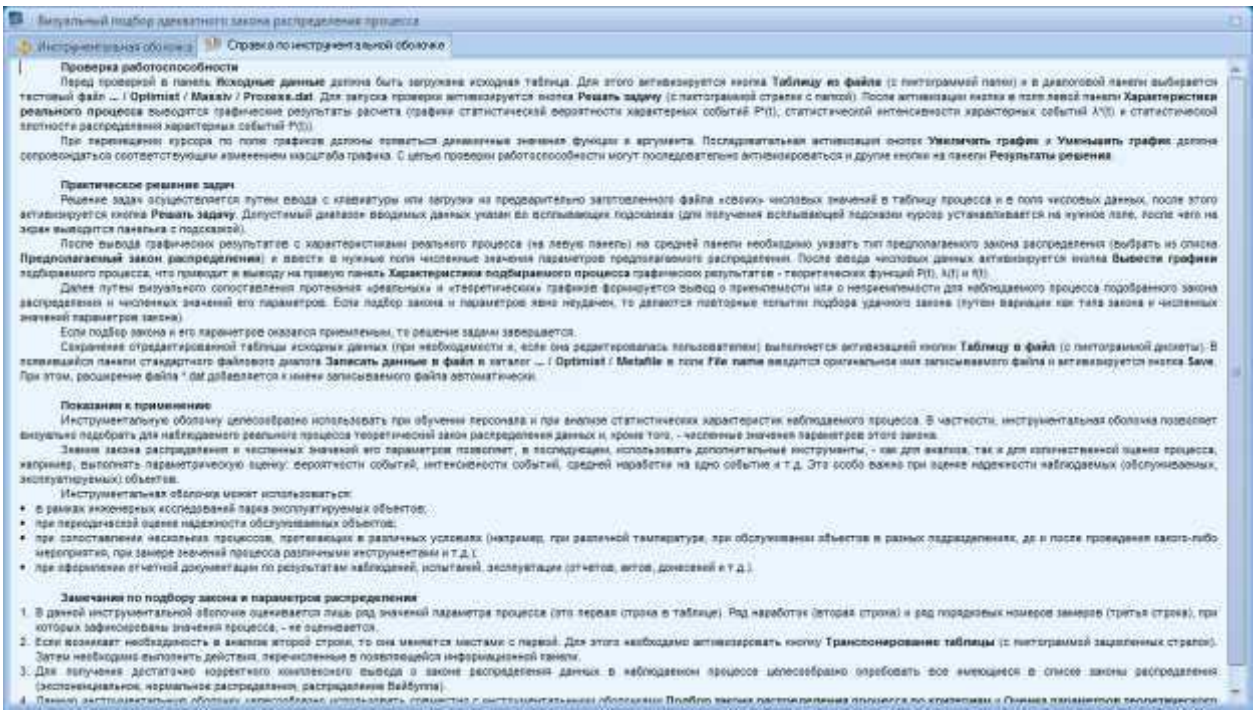

*Рис. 7.* Пример закладки «Справка по инструментальной оболочке»

Основные интерфейсные решения по инструментам унифицированы в соответствии с требованиями к пакету Модельер 2.0 и рассмотрены в [Чокой, 2016]. Для информационного обеспечения работы пользователей в пакет включены информационные карты, доступ к которым возможен через группу «Информационные карты» головного меню. Эти карты содержат иллюстрированные ма-

териалы как теоретического, так и практического характера и предназначены для самостоятельной проработки актуальной предметной области перед использованием инструментальных оболочек. Данные материалы также могут быть использованы и при проведении лекций.

Пакет Модельер 2.0 представляет собой автономное полнофункциональное windows-приложение, функционирующее на типовых IBM-подобных ЭВМ с операционной системой Windows-xx. Для инсталляции пакета на жестком диске достаточно 1,8 Гб памяти.

Особенностями пакета являются:

- наличие двух встроенных справочных систем – по общим вопросам моделирования и статистического анализа (информационные карты, доступные из головного меню) и справок по каждому инструменту (представлены на отдельной вкладке каждой инструментальной панели);

- наличие возможности одномоментного изменения дизайна панелей путем загрузки соответствующего скина (через группу «Оформление» головного меню);

- наличие всплывающих подсказок по назначению интерфейсных элементов, а также по формату и допустимому диапазону вводимых числовых данных;

- ограниченное использование в справочных текстах и в наименованиях интерфейсных элементов специфических терминов, требующих углубленной математической подготовки.

**Библиографический список**

1. *Айвазян С. А*. Прикладная статистика и основы эконометрики / С. А. Айвазян, В. С. Мхитарян. М.: Издательское объединение ЮНИТИ, 1998. 1022 с.

2. *Андерсон Т.* Статистический анализ временных рядов. М.: МИР, 1976. 755 с.

3. Надежность в технике. Система сбора и обработки информации. Методы оценки показателей надежности. ГОСТ 27.503–81. М.: Издательство стандартов, 1982. 55 с.

4. Надежность в технике. Система сбора и обработки информации. Планирование наблюдений. ГОСТ 27.502–83. М.: Издательство стандартов, 1984. 23 с.

5. *Тюрин Ю. Н.* Анализ данных на компьютере / Ю. Н. Тюрин, А. А. Макаров. М.: Финансы и статистика, 1995. 384 с.

6. *Чокой В. З.* Электронный тренажер Speller-TSM по локализации отказов оборудования самолетов Airbus A320 // «Crede Experto: транспорт, общество, образование, язык». 2016. № 1. URL: http://ce.if-mstuca.ru/index.php/2016-1 (Дата обращения: 07.07.2016).

## References

1. *Ajvazjan S. A.* (1998). Applied statistics and econometrics fundamentals / S. A. Ajvazjan, V. S. Mhitarjan. M.: Publishing union UNITY, 1998. 1022 p. (In Russian).

2. *Anderson T.* (1976). The statistical analysis of temporal series. M.: MIR, 1976. 755 p. (In Russian).

3. Industrial product dependability. The system of information collection and processing. The methods of estimate of reliability ratio. GOST 27.503–81. M.: Standard publisher, 1982. 55 p. (In Russian).

4. Industrial product dependability. The system of information collection and processing. Observations planning. GOST 27.502–83. M.: Standard publisher, 1984. 23 p. (In Russian).

5. *Tjurin Ju. N. (1995).* Data computer-aided analysis / Ju. N. Tjurin, A. A. Makarov. M.: Finances and statistics, 1995. 384 p. (In Russian).

6. *Chokoj V. Z.* (2016). SPELLER-TSM electronic simulator for troubleshooting of A320 equipment // «Crede Experto: transport, society, education, language». 2016. № 1. URL: http://ce.if-mstuca.ru/index.php/2016-1 (accessed date: 07.07.2016). (In Russian).