

УДК 519.2+004.4

М.А.Тынкевич, О.С.Болотова, Е.И.Латышева

ИНФОРМАЦИОННАЯ СИСТЕМА СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ЭКОНОМИЧЕСКИХ ДАННЫХ (СТЭК)

Человек познавал окружающий мир и самого себя в этом мире на основе статистики, которая позволяла ему находить связи между явлениями. Человеческий разум при посредстве статистики и неотделимо сопровождающего ее математического аппарата породил как современную науку – астрономию и геологию, физику твердого тела и генетику, так человеческую этику понятий «хорошо» или «плохо». Экспоненциальный рост объемов информации во всех сферах и усложнения связей в обществе за последние три столетия привел к примату математики, для которой присущ лаконизм и обоснованность выводов. «В каждой естественной науке заключено столько истины, сколько в ней математики» (И. Кант).

Как и при решении любой *обратной задачи – воспроизведения закона, которому следовала или следует некоторая система* (техническая, экономическая и пр.), по результатам ее действий (часто неполным и извращенным внесистемными факторами, особенно в сфере денежных интересов) *невозможно гарантировать надежность и однозначность получаемых выводов*. Тем не менее такой поиск остается жизненной потребностью.

Наряду с многообразием других подходов, для анализа статистических данных используются методы математической статистики (*регрессионный и дисперсионный анализ, анализ главных компонент, связанные с ними методы проверки гипотез*), математическим аппаратом которой является теория вероятностей. Там, где удается из моря «дурной вероятности» выловить статистические закономерности (законы распределения и тенденции), с целью последующей оптимизации вступают в работу *методы массового обслуживания* (без них не могли бы проектироваться, например, современные средства массовых коммуникаций) и *методы статистического моделирования* (Монте-Карло).

В последние 70 лет появилось великое множество алгоритмов *анализа временных рядов*, основанных на сочетании регрессионного анализа, спектрального анализа и эвристических подходов, вокруг которых не утихают споры. Предлагается многообразие специфических подходов к информации, предлагаемой статистикой рынка ценных бумаг, страхового дела. Хотя никаких революционных идей в обработке данных здесь не возникло, существует много фирм, дающих вполне правдоподобные прогнозы.

Для облегчения применения методов статистического анализа и прогнозирования на рынке программного обеспечения распространяются

сотни разных программных пакетов. Специализированные пакеты (Эвриста, МЕЗОЗАР, ОЛИМП, СтатЭксперт, Forecast Expert) ориентированы на использование в специфической предметной области - страховое дело, статистика ценных бумаг и другие области анализа временных рядов

Пакеты общего назначения более популярны. Они обладают продуманным, дружественным к пользователю интерфейсом и относительно подробной документацией, широким спектром статистических функций, что привлекает как специалистов, так и новичков в статистике (SPSS, STATISTICA, STATGRAPHICS, S-plus, SAS, STADIA).

Тем не менее, наряду с весьма богатыми возможностями, есть в них и определенные минусы. Большинство пакетов имеет свою оригинальную систему подготовки данных, ограниченно русифицировано (многие понятия нестандартны и не согласуются с традиционными для классической литературы); богатство возможностей, обеспеченное большим коллективом разработчиков в естественном стремлении «объять необъятное», перерастает в трудность восприятия рядовым пользователем, не отличающимся глубокимзнакомством с подводной частью айсберга математической статистики.

Соответственно, экономист, инженер или студент вынужден для выполнения небольшого и подчас элементарного анализа обращаться к объемистому дорогостоящему руководству, в котором, вместо хотя бы описания методов достижения целей и пояснения используемых понятий, он обнаруживает кнопочную технологию и последовательность экранов (если в описаниях библиотек программ 40-летней давности как зарубежные, так и отечественные авторы приводили методы и даже исходные модули, что делает их публикации вечными, то современные пособия, за малым исключением, живут не более 5- 10 лет, до усиленной рекламы следующего пакета).

Мы попытались построить компьютерную систему, предназначенную для пользователя, знакомого лишь с азами статистического анализа, который без длительного обучения мог бы «занести в компьютер» свою информацию и удовлетворить свое любопытство на традиционном уровне познания основных статистических величин (среднее, дисперсия, вариация и пр.), установления факта *нормальности* или специфики данных, без чего последующий анализ может оказаться неправомерным, наличия корреляции, характера связей и тенденций (в рамках «джентльменского набора» математических кривых, популярного в

среде эконометристов). Мы не добивались повышенной точности получаемых оценок там, где это связано с излишествами расхода времени и памяти, поскольку точность порядка 0.1% в экономических оценках более чем приемлема.

Естественно, разработке пакета предшествовал определенный анализ математических соотношений, которые легли в его основу.

Основу всякого статистического анализа одномерной выборки $X = \{x_i, i=1, N\}$ составляют т.н. моменты, их производные и некоторые другие параметры эмпирического распределения.

Моменты первых порядков определяются традиционно, например, в несмещенных оценках

$$\mu_x = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i - \text{среднее значение};$$

$$Dx = \frac{1}{N-1} \sum_{i=1}^N [x_i - \bar{x}]^2 - \text{дисперсия};$$

$$M_3x = \frac{1}{(N-1)(N-2)} \sum_{i=1}^N [x_i - \bar{x}]^3,$$

$$M_4x =$$

$$M_4x = \frac{N(N^2-2N+3)m_4x - 3 \cdot N(2N-3)D_x^2}{(N-1)(N-2)(N-3)}$$

- моменты третьего и четвертого порядков, где

$$m_4x = \frac{1}{N} \sum_{i=1}^N [x_i - \bar{x}]^4.$$

Среди других традиционных базовых характеристик:

- $s_x = \sqrt{Dx}$ - стандартное отклонение;

- $Ax = \frac{M_3x}{s_x^3}$ - коэффициент асимметрии, определяющий смещение распределения вправо ($Ax > 0$) и влево ($Ax < 0$);

- $\mathcal{E}x = \frac{M_4x}{s_x^4} - 3$ - коэффициент эксцесса, характеризующий остроту пика распределения (для нормального распределения эксцесс нулевой);

- x_{min}, x_{max} - предельные значения массива данных;

- $Rx = x_{max} - x_{min}$ - размах;

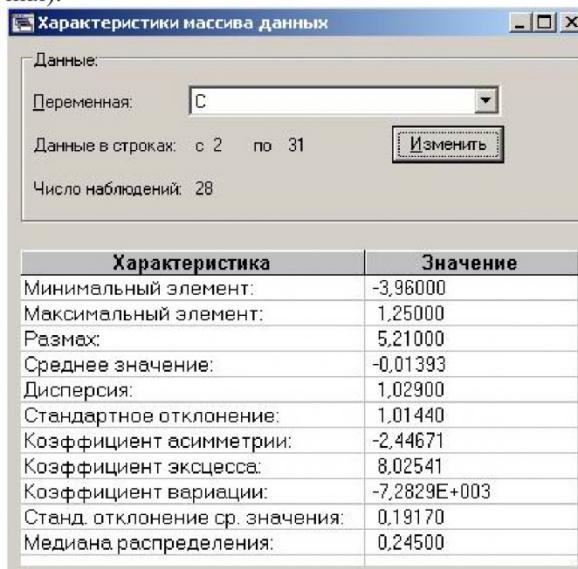
- $Vx = \frac{s_x}{Mx}$ - коэффициент вариации по Пирсону;

- $S_x = \frac{s_x}{\sqrt{N}}$ - стандартное отклонение среднего значения;

- Me - медиана распределения с плотностью $p(x)$ (срединное значение упорядоченной выборки - $X_{[N/2]+1}$ при нечетном N или $0.5 [X_{[N/2]} + X_{[N/2]+1}]$ при четном N):

$$\int_{-\infty}^{Me} p(x)dx = \int_{Me}^{\infty} p(x)dx;$$

- Mo - мода распределения (значение x , соответствующее максимуму плотности распределения).



Характеристика	Значение
Минимальный элемент	-3.96000
Максимальный элемент	1.25000
Размах	5.21000
Среднее значение	-0.01393
Дисперсия	1.02900
Стандартное отклонение	1.01440
Коэффициент асимметрии	-2.44671
Коэффициент эксцесса	8.02541
Коэффициент вариации	-7.2829E+003
Станд. отклонение ср. значения	0.19170
Медиана распределения	0.24500

Рис. 1. Вывод результатов расчета статистических характеристик

К базовой статистике следует отнести и коэффициент корреляции ($|r_{xy}| \leq 1$)

$$r_{xy} = \frac{1}{N} \sum_{i=1}^N \frac{x_i - \mu_x}{s_x} \cdot \frac{y_i - \mu_y}{s_y} \equiv \frac{\overline{xy} - \mu_x \mu_y}{s_x s_y}.$$

На основе такой базовой статистики строится весь последующий анализ эмпирических данных.

Часто практики не уделяют должного внимания распределениям вероятностей случайных величин (это присуще во многом и известным пакетам). Знание характера эмпирического распределения существенно, например, для правомерности применения методов регрессионного анализа (в их основе близость распределения к нормальному), методов теории массового обслуживания (экспоненциальное и Пуассона) и др. Отсутствие унимодальности приводит подчас к абсурдным выводам.

Из многообразия типов распределений непрерывных случайных величин, нуждающихся в компьютерном анализе, мы выбрали лишь те, которые наиболее часто встречаются в приложениях (рис.2), и установили представление плотности $p(x)$, функции $F(x)$ и параметров распределения через моменты.

1. Равномерное распределение

$$p(x) = \begin{cases} \frac{1}{B-A} & A < x < B \\ 0 & x \notin [A, B] \end{cases}; \quad F(x < A) = 0;$$

$$F(x) = \frac{x-A}{B-A} \quad (A < x < B), \quad F(x > B) = 1;$$

$$A = \mu_x - s_x \sqrt{3}, \quad B = \mu_x + s_x \sqrt{3}.$$

2. Нормальное распределение

$$p(x) = \frac{1}{B\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-A}{B}\right)^2}; \quad F(x) = \Phi\left(\frac{x-A}{B}\right);$$

$$A=\mu_x, \quad B=s_x.$$

$$\text{Здесь } \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{1}{2}t^2} dt.$$

3. Распределение Лапласа-Шарлье

$$p(x) = \frac{1}{B\sqrt{2\pi}} e^{-\frac{R^2}{2}} \times$$

$$\left[1 - \frac{A_x}{6}(3R - R^3) + \frac{E_x}{24}(R^4 - 6R^2 + 3) \right],$$

$$R = \frac{x-A}{B};$$

$A=\mu_x$; $B=s_x$; A_x и E_x – коэффициенты асимметрии и эксцесса. Это оригинальное распределение, определяемое как отрезок ряда Грама-Шарлье первого рода, сохраняет унимодальность и неотрицательность $p(x)$, как нами установлено аналитическим и численным анализом, в ограниченном диапазоне $|R| < \sqrt{3}$ при ограниченных значениях асимметрии $|A_x| < 3$ и эксцесса $\approx -\pi/2 < E_x < 4$.

4. Распределение Лапласа (двустороннее показательное)

$$p(x) = \frac{1}{2B} e^{-\left|\frac{x-A}{B}\right|}; \quad F(x) = \begin{cases} \frac{1}{2} e^{\frac{x-A}{B}}, & x < A \\ 1 - \frac{1}{2} e^{\frac{A-x}{B}}, & x > A \end{cases}$$

$$A=\mu_x, \quad B=s_x/\sqrt{2}.$$

5. Логарифмически нормальное распределение

$$p(x) = \frac{1}{x \cdot B\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(x)-A}{B}\right)^2\right];$$

$$F(x) = \Phi\left(\frac{\ln(x)-A}{B}\right), \quad x > 0;$$

$$A = \ln \frac{Mx^2}{\sqrt{Mx^2 + s_x^2}}, \quad B = \sqrt{\ln \frac{Mx^2 + s_x^2}{Mx^2}};$$

$$Mo = \exp(A-B^2), \quad Me = \exp(A).$$

6. Распределение Вейбулла

$$p(x) = \frac{A}{B} x^{A-1} e^{-\frac{x^A}{B}}; \quad F(x) = 1 - e^{-\frac{x^A}{B}}, \quad x > 0;$$

$$A \geq 1, \quad B > 0;$$

$$Me = (B \cdot \ln(2))^{1/A}; \quad Mo = [B \cdot (A-1)/A]^{1/A};$$

A – корень уравнения (существует при $V_x < 1$)

$$L(A) = V_x^2 + 1 - \Gamma(1 + \frac{2}{A}) / \Gamma^2(1 + \frac{1}{A}) = 0;$$

$$B = \left[\frac{Mx}{\Gamma(1 + \frac{1}{A})} \right]^A,$$

$$\text{где } \Gamma(\lambda) = \int_0^\infty t^{\lambda-1} e^{-t} dt \text{ -гамма-функция, } V_x -$$

коэффициент вариации; распределение часто используется при оценках надежности и риска.

7. Экспоненциальное распределение

$$p(x) = \lambda e^{-\lambda x}, \quad x > 0; \quad F(x) = 1 - e^{-\lambda x};$$

$\lambda = 1/Mx$; $Me = \ln(2)/\lambda$; $Mo=0$; $A_x = 2$; $E_x = 6$. Одно из основных распределений в теории массового обслуживания.

8. Распределение Рэлея

$$p(x) = \frac{x}{\beta^2} e^{-\frac{1}{2}\left(\frac{x}{\beta}\right)^2}, \quad x > 0;$$

$$F(x) = 1 - e^{-\frac{x^2}{2\beta^2}}.$$

Параметр масштаба $\beta = \sqrt{\frac{2}{\pi}} Mx$; $Dx = \frac{4-\pi}{2} \beta^2$, асимметрия положительна.

9. Распределение Максвелла

$$p(x) = \sqrt{\frac{2}{\pi}} \frac{x^2}{\beta^3} e^{-\frac{1}{2}\left(\frac{x}{\beta}\right)^2}, \quad x > 0;$$

$$F(x) = -\sqrt{\frac{2}{\pi}} \cdot \frac{x}{\beta} e^{-\frac{1}{2}\left(\frac{x}{\beta}\right)^2} - 1 + 2\Phi(x/\beta),$$

$$x > 0.$$

$$Mo = \beta\sqrt{2}, \quad Mx = \beta\sqrt{\frac{8}{\pi}}, \quad Dx = \frac{3\pi-8}{\pi} \beta^2,$$

$$\text{откуда } \beta = s_x \sqrt{\frac{\pi}{3\pi-8}}.$$

10. Распределение χ^2 (ХИ-квадрат)

$$p(x) = \frac{x^{\frac{A}{2}-1} e^{-\frac{x}{2}}}{\frac{A}{2} \Gamma(\frac{A}{2})}, \quad x > 0, A \geq 1;$$

$$F(x) = \frac{1}{\Gamma(A/2)} \int_0^{x/2} t^{\frac{A}{2}-1} e^{-t} dt;$$

$A=\mu_x$; $s_x=\sqrt{2A}$; при $A>2$ мода= $A-2$; $A_x=2\sqrt{\frac{2}{A}}$, $E_x=12/A$; составляет базу критерия Пирсона.

11. ХИ-распределение

$$p(x) = \frac{x^{A-1} e^{-\frac{x^2}{2}}}{2^{\frac{A-1}{2}} \Gamma(\frac{A}{2})}, x > 0,$$

$$F(x) = \frac{1}{\Gamma(A/2)} \int_0^{x^2/2} t^{A/2-1} e^{-t} dt$$

$$A = s_x^2 + \mu_x^2 ; \text{ мода} = \sqrt{A - 1}.$$

12. Гамма-распределение

$$p(x) = \frac{x^{A-1} e^{-\frac{x}{B}}}{B^A \Gamma(A)}; x > 0;$$

$$A = \left[\frac{Mx}{s_x} \right]^2 = \frac{1}{V_x^2}, B = \frac{s_x^2}{Mx} = s_x V_x.$$

Мода распределения равна $x = B(A-1)$.13. Распределение Фишера (распределение отношения дисперсий, v^2 -распределение)

$$p(x) = \frac{\Gamma(\frac{A+B}{2})}{\Gamma(\frac{A}{2}) \cdot \Gamma(\frac{B}{2})} \times$$

$$\left(\frac{A}{B} \right)^{A/2} x^{\frac{A}{2}-1} \left(1 + \frac{A}{B} x \right)^{-\frac{A+B}{2}},$$

$$A, B > 0; x > 0;$$

$$B = \frac{2 \cdot Mx}{Mx - 1}; A = \frac{2B^2(B-2)}{(B-2)^2(B-4) \cdot s_x^2 - 2B^2};$$

$$\text{при } A > 2 \quad Mo = \frac{B}{A} \cdot \frac{A-2}{B+2}.$$

14. Распределение Парето

$$p(x) = \frac{c}{x^{c+1}}, 1 \leq x < \infty; c > 0;$$

$$F(x) = c \int_1^x \frac{dx}{x^{c+1}} = 1 - \frac{1}{x^c},$$

где параметр (формы) распределения

$$c = \frac{Mx}{Mx - 1} > 1.$$

15. Распределение Стьюдента

$$p(x) = \frac{1}{\sqrt{\pi A}} \frac{\Gamma(\frac{A+1}{2})}{\Gamma(\frac{A}{2})} \left(1 + \frac{x^2}{A} \right)^{-\frac{A+1}{2}};$$

$$A = \frac{2 \cdot S_x^2}{S_x^2 - 1} > 2; \text{ симметричное распределение с}$$

экспессом $= 3(A-2)/(A-4)$ (при $A > 4$). Одно из популярнейших распределений для оценки доверительных интервалов.

16. T^2 -распределение Хотеллинга

$$p(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n-k+1}{2}) \Gamma(\frac{k}{2})} \cdot \frac{x^{\frac{k}{2}-1} (1 + \frac{x}{n})^{-\frac{n+1}{2}}}{n^{\frac{k}{2}}}$$

 $0 < x < \infty, n > k > 1.$

$$k = \frac{M_x(n-1)}{M_x+n}; n - \text{корень уравнения}$$

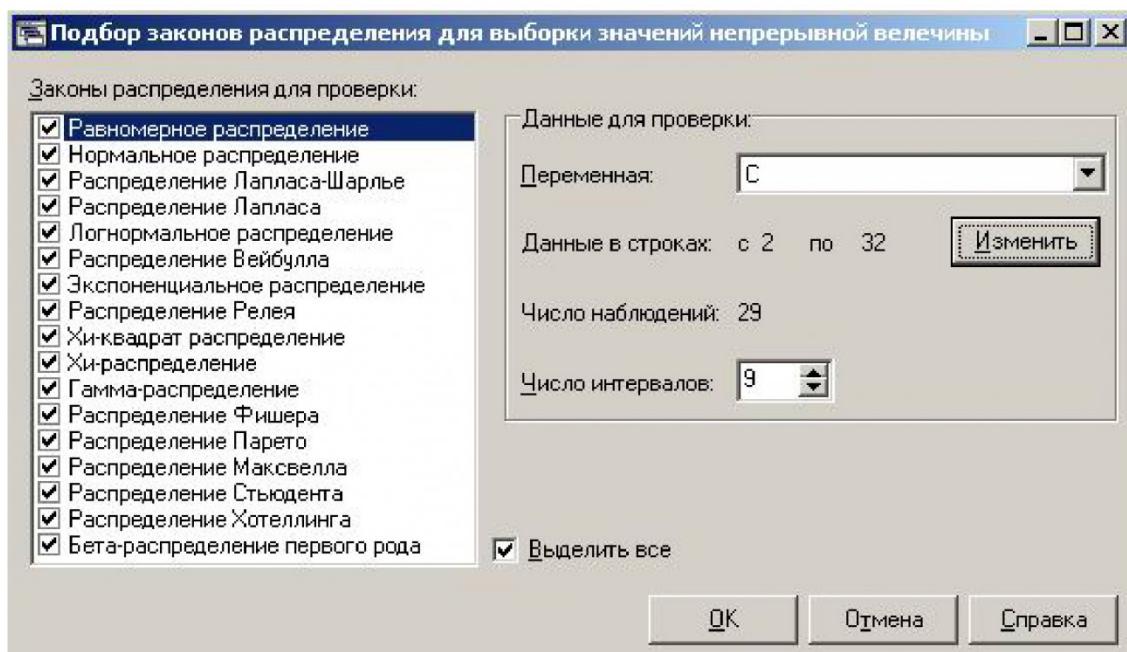


Рис. 2. Выбор данных для проверки эмпирического распределения на соответствие гипотетическим распределениям

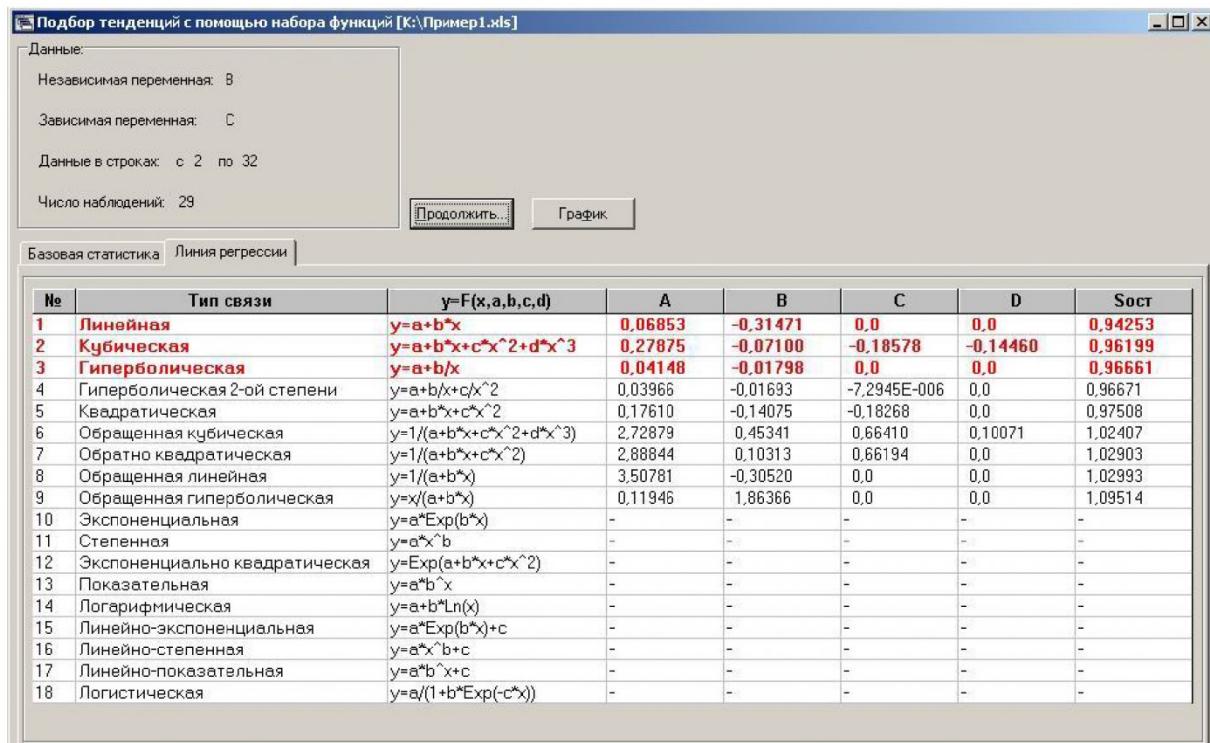


Рис. 3. Подбор тенденций для динамических рядов

$$n^2(s_x^2 - 2M_x) - n(4M_x^2 + 3s_x^2) - \\ - 2M_x(M_x^2 + s_x^2) = 0$$

(в зависимости от знака при n^2 в его решении выбираем минус или плюс).

17. Бета -распределение первого рода

$$p(x) = \frac{\Gamma(A+B)}{\Gamma(A)\cdot\Gamma(B)} x^{A-1} (1-x)^{B-1}, \quad 0 < x < 1;$$

$$F(x) = \begin{cases} 0, x \leq 0 \\ \frac{B_x(A,B)}{B(A,B)}, 0 < x < 1, \\ 1, x \geq 1 \end{cases}$$

где $B_x(\alpha, \beta) = \int_0^x t^{\alpha-1} (1-t)^{\beta-1} dt$ – неполная бета-функция;

$$A = \frac{(Mx)^2}{s_x^2} (1-Mx) - Mx; \quad B = A \frac{1-Mx}{Mx};$$

при $A, B > 1$ существует мода = $(A-1)/(A+B-2)$.

Предлагаемый нами пакет СТЭК позволяет найти выполнить базовую статистику эмпирического распределения объема N и осуществить проверку его на близость к вышеперечисленным по критерию

$$\chi^2 = \sum_{i=1}^k \frac{(p_i^{meop} - p_i^{mn})^2}{Np_i^{meop}}, \quad f=k-1,$$

где k – число подинтервалов области данных, определяемое по формуле Стерджеса

$$k=[1,446 \ln(N)+3,5]$$

или принудительно с учетом “порога”.

Естественно, пакет обеспечивает графику плотности и функции распределения при любых допустимых значениях параметров, генерацию последовательностей псевдослучайных чисел с указанным распределением

$$\int_{-\infty}^z p(z) dz = x, \quad x \in [0,1] \text{ р.р.,}$$

поиск квантилей по заданному значению $F(x)=\alpha$ (эта возможность позволяет избавиться от традиционного использования таблиц при поиске уровня значимости получаемых статистических оценок).

Другая составная часть пакета связана с традиционным регрессионным анализом. Обычная парная корреляция и регрессия предлагает оценки по 5 традиционным для эконометрики типам связей:

$$y=a+bx; \quad y=a+b/x; \quad y=a+b \cdot \ln(x); \quad y=a \cdot \exp(bx); \quad y=a+bx+cx^2$$

(коэффициенты уравнений регрессии, корреляционные отношения, остаточная дисперсия, средняя невязка, их оценки по Стьюденту и уровень значимости).

Блок множественной регрессии

$$y \equiv x_1 = a_0 + \sum_{i=2}^m a_j f(x_j).$$

также допускает линейное, гиперболическое, ло-

гарифмическое и параболическое включения всех факторов (в том числе и x_i) с последующим расчетом матрицы парных регрессионных отношений, признаков, упомянутых выше, и коэффициентов уравнения регрессии в стандартизованном масштабе. Кстати, проблема «пропущенных значений» решается по усмотрению пользователя пакета (удалением «наблюдения» или вставкой среднего значения).

Пакет обеспечивает и ранговую корреляцию по Спирмену и Кендаллу с учетом возможной связности рангов [4]

С ориентацией на анализ трендов динамических рядов подобрана система популярных в эконометрике функций (рис.2). С той же целью создана и возможность подбора аппроксимирующего алгебраического полинома заданной степени или с автоматическим выбором таковой по критерию 100-кратного уменьшения остаточной дисперсии. Здесь нами предусмотрена возможность построения полинома как по обычным значениям факторов, так и по их стандартизованным (нормированным) значениям.

Для создания системы СТЭК по соображениям разработки с минимальными затратами эффективного пользовательского интерфейса, соответствующего стандартам операционной системы Windows, и высокого быстродействия создаваемых приложений использована интегрированная среда разработки Borland Delphi 6.0.

Ввод данных в системе организован в виде знакомой всем пользователям MS Excel электронной таблицы, состоящей из столбцов (переменных, факторов) и строк (наблюдений). Максимально число переменных – 26, число наблюдений – 512.

Сохраняются операции с использование буфера обмена Windows, операции с выделенными

блоками (аналогично MS Excel) и пр.

Ввод данных в электронную таблицу можно осуществить непосредственно с клавиатуры, на основе уже введенных при помощи формул, копированием через буфер обмена или открытием готовых файлов (расширения файлов *.xls, *.txt). В качестве дополнительных функций система позволяет сортировать и нормировать данные.

Вывод численных и текстовых результатов анализа производится в отдельные окна, где расположены таблицы, подобные таблицам ввода; можно копировать в буфер обмена или сохранять в файлах указанных выше типов. Имеется возможность вывода результатов на печать. Вывод графических результатов производится также в отдельные окна. Пользователь может настроить различные компоненты графика (цвет линий, фона, вид легенда, заголовок, тип линии и др.), сохранить его в графическом растровом формате (расширение *.bmp), копировать в буфер обмена, а также вывести на печать.

Доступ к статистическим процедурам осуществляется через пункты меню и через кнопки панели инструментов.

Как мы уже указывали выше, система СТЭК обеспечивает определение базовых числовых характеристик совокупностей, подбор законов распределения, графику законов распределения, генерацию псевдослучайных чисел заданного типа, корреляционный и регрессионный анализ, подбор тенденций для динамических рядов.

Предусмотрено получение справки не только о работе системы, но и теоретической информации об описательных статистиках, рассчитываемых системой, обо всех распределениях, представленных для анализа, а также теоретические выкладки по корреляционному и регрессионному анализу.

СПИСОК ЛИТЕРАТУРЫ

1. Поллард Дж. Справочник по вычислительным методам статистики. –М.: Финансы и статистика. 1982. -344 с.
2. Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. –М.: Наука. 1984. -832 с.
3. Программное обеспечение ЭВМ Мир-1 и Мир-2. Том 2. –Киев: Наукова думка. 1976. -371 с.
4. Математический энциклопедический словарь. – М.: Советская энциклопедия. 1988.
5. Янке Е., Эмде Ф., Леш Ф. Специальные функции. – М.: Наука. 1964. -344с.

Авторы статьи:

Тынкевич
Моисей Аронович
- канд. физ.-мат. наук, проф. каф.
вычислительной инженерии и информационных технологий

Болотова
Ольга Сергеевна
- дипломант каф. вычислительной
инженерии и информационных техно-
логий

Латышева
Евгения Ильинична
- дипломант каф. вычислительной
инженерии и информационных техно-
логий