Информационные системы и технологии: ИСТ`2000:
Сб. научн. ст. – Новосибирск. 2001. – С.26-41.

КОМПЬЮТЕРНЫЕ МЕТОДЫ ИССЛЕДОВАНИЯ СТАТИСТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ[1]

Лемешко Б.Ю.

Новосибирский государственный технический университет,
Новосибирск, Россия, тел. (383-2) 46-37-54, e-mail: headrd@fpm.ami.nstu.ru

1. Введение

Основным методом настоящих исследований, наряду с математическим аппаратом, является развиваемая нами методика компьютерного моделирования и исследования статистических закономерностей.

В последние годы при исследовании некоторых задач математической и прикладной статистики нами получен ряд результатов, связанных как с реализацией вычислительных алгоритмов, обеспечивающих построение оценок параметров с наилучшими свойствами при различной форме регистрации наблюдений, так и с исследованием статистических закономерностей. В результате этого построены достаточно простые модели законов распределений статистик различных критериев для целого множества проверяемых сложных гипотез. Удалось получить достаточно точные решения для задач, которые десятилетиями не могли быть получены аналитическими средствами. А самое главное, появилась уверенность, что с использованием данного подхода можно закрыть многие существующие в статистике “белые пятна”, применяя относительно простой вычислительный и математический аппарат.

2. Робастное оценивание

Разрабатываемые нами методы оценивания параметров предусматривают нахождение оценок при любой форме представления исходных данных: при точечных, группированных, частично группированных и цензурированных выборках [1].

При исследовании различных методов оценивания и свойств оценок было показано, что высокой устойчивостью к различным отклонениям от предположений и к наличию аномальных наблюдений обладают оценки максимального правдоподобия (ОМП) по группированным данным. Экспериментальные исследования свойств оценок и анализ вида функций влияния Хампеля показало, что ОМП по точечным выборкам за редким исключением не являются робастными. Из трёх десятков моделей законов распределений непрерывных случайных величин, наиболее часто используемых в приложениях, робастными оказались ОМП параметров сдвига и масштаба распределения Коши и параметра сдвига логистического распределения. Напротив, ОМП по группированным наблюдениям всегда оказываются робастными. Группирование наблюдений при оценивании позволяет получать устойчивые оценки [2-4]. Повышению качества таких оценок способствует применение асимптотически оптимального группирования, минимизирующего потери в информации Фишера [5].

Были предложены простые оптимальные L-оценки параметров сдвига и масштаба, построенные на оценках квантилей, соответствующих асимптотически оптимальному группированию [6]. L-оценки получаются в виде линейной комбинации выбороч-

[1] Выполнена при поддержке Российского фонда фундаментальных исследований (№ 00-01-00913).

ных асимптотически оптимальных квантилей. Для ряда параметрических моделей законов распределений, наиболее часто используемых в приложениях, построены таблицы коэффициентов для таких оценок (для 15 моделей законов при различном числе используемых квантилей), исследованы их свойства. Исследованы способы определения выборочных квантилей, при которых минимизируется величина смещения L-оценок. В настоящее время исследуется, как при проверке сложных гипотез отражается применение L-оценок на распределениях статистик критериев согласия (непараметрических критериев и типа χ^2).

Было показано, что с использованием предложенных робастных методов оценивания процедура параметрической отбраковки аномальных наблюдений становится очень эффективной [4].

3. Исследование вопросов применения критериев типа χ^2

При использовании критериев согласия типа χ^2 неоднозначность при построении и вычислении статистик бывает связана с выбором числа интервалов и тем, каким образом область определения случайной величины разбивается на интервалы. Естественно, что такой произвол отражается на статистических свойствах применяемых критериев, в частности, на их мощности при различении близких конкурирующих гипотез. Очевидно, что выбор числа интервалов и способа разбиения на интервалы следует осуществлять с позиций обеспечения максимальной мощности критерия.

С использованием критериев согласия могут проверяться простые гипотезы вида $H_0:\ F(x) = F_0(x,\theta)$, где $F_0(x,\theta)$ – функция распределения вероятностей, с которой проверяется согласие наблюдаемой выборки независимых одинаково распределенных величин $X_1, X_2, \ldots, X_n$, а θ – известное значение параметра (скалярного или векторного), и сложные гипотезы $H_0:\ F(x) \in \{F_0(x,\theta), \theta \in \Theta\}$, где Θ – пространство параметров. В процессе проверки сложной гипотезы оценка параметра $\hat{\theta}$ вычисляется по этой же самой выборке. Если оценка $\hat{\theta}$ вычислена по другой выборке, то гипотеза простая.

При использовании критериев согласия типа χ^2 область определения случайной величины разбивается на k интервалов граничными точками

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

Статистика X_n^2 Пирсона вычисляется в соответствии с соотношением

$$X_n^2 = n\sum_{i=1}^{k} \frac{(n_i / n - P_i(\theta))^2}{P_i(\theta)}, \qquad (1)$$

где n_i – количество наблюдений, попавших в интервал, $P_i(\theta) = \int_{x_{i-1}}^{x_i} f_0(x,\theta)dx$ – вероятность попадания наблюдения в i-й интервал, $n = \sum_{i=1}^{k} n_i$, $\sum_{i=1}^{k} P_i(\theta) = 1$. При справедливой простой гипотезе H_0 предельное распределение статистики $G(X_n^2 | H_0)$ есть χ^2-распределение с числом степеней свободы $k-1$. Если по выборке оценивалось m параметров закона в результате минимизации статистики X_n^2, статистика подчиняется χ^2-распределению с $k-m-1$ степенями свободы. При справедливой альтернативной

гипотезе H_1 предельное распределение $G(X_n^2|H_1)$ представляет собой нецентральное χ^2-распределение с тем же числом степеней свободы и параметром нецентральности

$$s(\theta)=\sum_{i=1}^{k}\frac{c_i^2(\theta)}{P_i}, \qquad (2)$$

где $c_i(\theta)=\sqrt{n}\int\limits_{x_{i-1}(\theta)}^{x_i(\theta)}\big(f_1(x,\theta)-f_0(x,\theta)\big)dx$ и $f_1(x,\theta)$ соответствует альтернативе.

В случае проверки сложных гипотез и оценивании по выборке параметров наблюдаемого закона использование в качестве предельных χ^2_{k-m-1}-распределений справедливо лишь при определении оценок минимизацией статистики X_n^2 или при вычислении по сгруппированным данным ОМП. Распределения статистик χ^2 Пирсона и отношения правдоподобия при оценивании параметров наблюдаемого закона по точечным наблюдениям и различных способах построения интервалов нами были исследованы в [7] методами компьютерного моделирования. Было показано, что распределения $G(X_n^2|H_0)$ статистики оказываются зависящими от того, как строятся интервалы, и хорошо аппроксимируются семейством гамма-распределений.

Предложенная в работах С.М. Никулина статистика[2] типа χ^2 отличается от X_n^2 только при сложных гипотезах. Предельное распределение этой статистики есть 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 $\mathbf{P}=(P_1,\ldots,P_k)^{\mathrm{T}}$ предполагается заданным, и граничные точки интервалов определяются соотношениями $x_i(\theta)=F^{-1}(P_1+\ldots+P_i)$, $i=\overline{1,(k-1)}$.

Данная статистика имеет вид

$$Y_n^2(\theta)=X_n^2+n^{-1}\mathbf{a}^{\mathrm{T}}(\theta)\mathbf{\Lambda}(\theta)\mathbf{a}(\theta), \qquad (3)$$

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

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

определяются оцениваемыми компонентами вектора параметров θ, $J(\theta_l,\theta_j)$ - элементы информационной матрицы $\mathbf{J}(\theta)$

$$\mathbf{J}(\theta)=\left[\int\left(\frac{\partial\ln f_0(x,\theta)}{\partial\theta_l}\frac{\partial\ln f_0(x,\theta)}{\partial\theta_j}\right)f_0(x,\theta)dx\right]_{m\times m},$$

$a(\theta_l)=w_{\theta_l 1}n_1/P_1+\ldots+w_{\theta_l k}n_k/P_k$ - элементы вектора $\mathbf{a}(\theta)$, величины $w_{\theta_l i}$ определяются соотношением

$$w_{\theta_l i}=-f_0[x_i(\theta),\theta]\frac{\partial x_i(\theta)}{\partial\theta_l}+f_0[x_{i-1}(\theta),\theta]\frac{\partial x_{i-1}(\theta)}{\partial\theta_l}.$$

[2] Никулин М.С. // Теория вероятностей и ее применение. 1973. Т. XVIII. № 3. – С.675-676.

При верной конкурирующей гипотезе статистика Y_n^2 имеет в качестве предельного $G(Y_n^2 \mid H_1)$ нецентральное χ^2_{k-1}-распределение с параметром нецентральности

$$s(\theta) = \sum_{i=1}^{k} \frac{c_i^2(\theta)}{P_i} + \mathbf{d}^T(\theta)\mathbf{\Lambda}(\theta)\mathbf{d}(\theta), \qquad (4)$$

где элементы вектора $\mathbf{d}(\theta)$ определяются как

$$d(\theta_l) = w_{\theta l 1} c_1(\theta)/P_1 + \ldots + w_{\theta l k} c_k(\theta)/P_k .$$

Зависимость мощности от способа группирования. Способ группирования особенно сильное влияние оказывает на предельное распределение $G(X_n^2|H_1)$. В работах [7-10] показано, что критерии согласия χ^2 Пирсона и отношения правдоподобия при проверке как простых, так и сложных гипотез имеют максимальную мощность против близких альтернатив, если использовать такое разбиение области определения случайной величины на интервалы, при котором потери в информации Фишера о параметрах закона, соответствующего гипотезе H_0, минимальны (асимптотически оптимальное группирование). Чем меньше потери в информации Фишера, связанные с группированием данных, тем больше параметр нецентральности, определяемый соотношением (2). В [5,10] для конкретных законов распределения представлен достаточно широкий состав (около 50 для, примерно, 20 законов распределений) построенных таблиц асимптотически оптимального группирования (АОГ-группирования), минимизирующего потери в информации Фишера. Использование АОГ-группирования при заданном числе интервалов обеспечивает максимальную мощность при близких гипотезах. Причем выигрыш в мощности даже по отношению к разбиению на интервалы равной вероятности очень существенный. Для тех параметрических моделей законов распределения, для которых решение задачи асимптотически оптимального группирования не может быть получено в виде, инвариантном относительно параметров закона, например, в [1] оптимальные граничные точки находятся в процессе проверки согласия.

Исследование распределений статистики Y_n^2 Никулина, которая отличается от X_n^2 только при сложных гипотезах, показало, что как $G(Y_n^2 \mid H_0)$, так и $G(Y_n^2 \mid H_1)$ несущественно зависят от способа группирования. Более того, наши исследования методами статистического моделирования показали, что с позиций наибольшей мощности разбиение на интервалы равной вероятности (РВГ-группирование) оказывается наиболее предпочтительным. Подчеркнем, что критерий типа χ^2 Никулина мощнее, чем критерии χ^2 Пирсона и отношения правдоподобия.

Зависимость мощности от числа интервалов k. За всю историю применения критериев типа χ^2 была предложена не одна формула для выбора числа интервалов[3], но ни одна из представленных в различных рекомендациях не выводилась с позиций максимальной мощности применяемого критерия, а, в основном, исходя из близости плотности к ее непараметрической оценке, гистограмме. Зная предельные распределения $G(S \mid H_0)$ и $G(S \mid H_1)$ статистики S, для любого заданного уровня значимости α можно оценить мощность соответствующего критерия, рассматривая её как функцию от числа интервалов k при заданном объеме выборки n. Исследование мощности критериев Пирсона и Никулина как функции от n и k проводилось аналитически и мето-

[3] Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. - Л.: Энергоатомизд., 1991. - 303 с.

дами статистического моделирования [11]. Результаты аналитических вычислений полностью подтверждаются оценками, полученными на основании моделирования.

Мощность критериев типа χ^2 определяется выражением[4]:

$$1-\beta = \mathrm{P}(\mathrm{s} \mid r, \alpha) = e^{-\mathrm{s}/2} \sum_{j=0}^{\infty} \frac{\mathrm{s}^j}{j! 2^{2j-1+r/2} \Gamma(j + r/2)} \times \int_{\sqrt{x(\alpha,r)}}^{\infty} y^{2j-1+r} e^{-y^2/2} dy, \qquad (5)$$

где s - параметр нецентральности, определяемый соотношениями (2) и (4), $x(\alpha, r)$ представляет собой $(1-\alpha)$-процентную точку χ^2_r-распределения с r степенями свободы (α - заданная вероятность ошибки первого рода, β - вероятность ошибки второго рода). Приводимые ниже функции мощности строились при уровне значимости $\alpha = 0.1$.

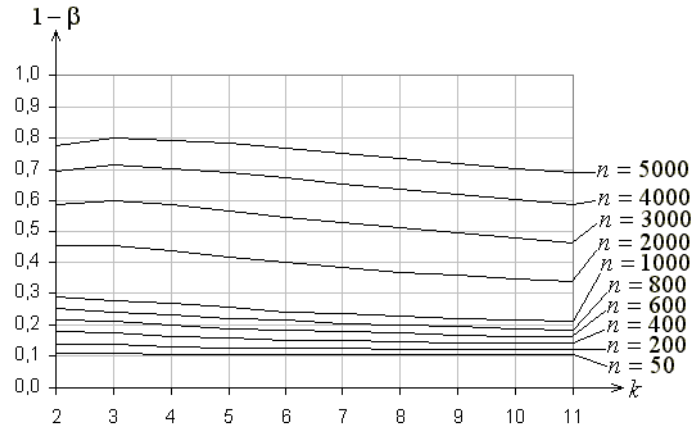

Рис. 1. Функции мощности критерия χ^2 Пирсона при проверке простой гипотезы о согласии с экспоненциальным законом при РВГ-группировании

На рис. 1 в зависимости от числа k равновероятных интервалов (РВГ-группирование) при различных n представлены функции мощности критерия χ^2 Пирсона при проверке простой гипотезы о согласии с экспоненциальным законом (H_0: $f_0(x) = \theta \exp\{-\theta x\}$ при $\theta = 1$; H_1: $f_1(x) = \theta \exp\{-\theta x\}$ при $\theta = 1.05$). На рис. 2 приведены аналогичные функции при использовании АОГ-группирования [5, 10]. И в том, и в другом случае с ростом k мощность падает, но в случае АОГ-группирования она выше, чем при РВГ-группировании. Падение мощности с k ростом вполне согласуется с результатами Чибисова[5] и Боровкова[6].

На рис. 3 представлены функции мощности критерия типа χ^2 Никулина при проверке сложной гипотезы о согласии с нормальным законом

$$H_0: \; f_0(x) = \frac{1}{\theta_1 \sqrt{2\pi}} \exp\left\{ -\frac{(x-\theta_0)^2}{2\theta_1^2} \right\},$$

[4] Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1983. - 416 с.
[5] Чибисов Д.М., Гванцеладзе Л.Г. // III сов.-яп. симп. по теор. Вер.. Ташкент: изд-во “Фан”, 1975. – С. 183-185.
[6] Боровков А.А. // Теория вероятностей и ее применение. 1977. Т. XXII. № 2. – С.375-378.

когда в качестве альтернативы рассматривается близкий ему логистический закон

$$H_1:\ f_1(x)=\frac{\pi}{\theta_1\sqrt{3}}\exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\}\Bigg/\left[1+\exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\}\right]^2$$

при значениях параметров $\theta_0=0$, $\theta_1=1$. Отметим, что функции мощности критерия χ^2 Пирсона в данной ситуации являются строго убывающими по k функциями и принимают максимальное значение при минимально возможном значении числа интервалов $k=4$.

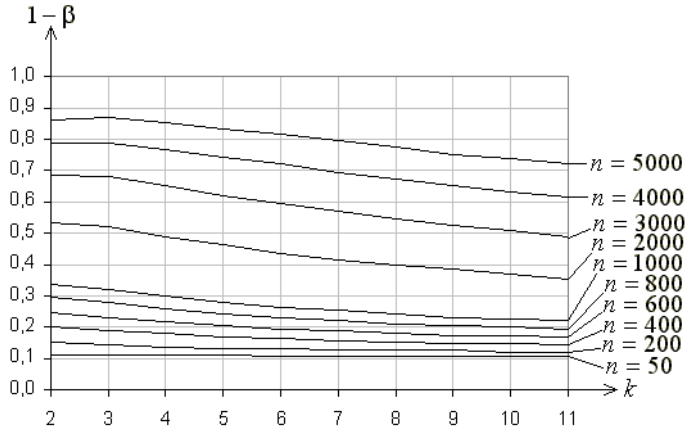

Рис. 2. Функции мощности критерия χ^2 Пирсона при проверке простой гипотезы о согласии с экспоненциальным законом при АОГ-группировании

Результаты расчета функций мощности по соотношению (5) контролировались статистическим моделированием функций мощности, при котором строились эмпирические функции распределений $G(S_n \mid H_0)$ и $G(S_n \mid H_1)$ для статистик S рассматриваемых критериев, и находились оценки мощности. Результаты моделирования оказываются очень близкими к расчетным.

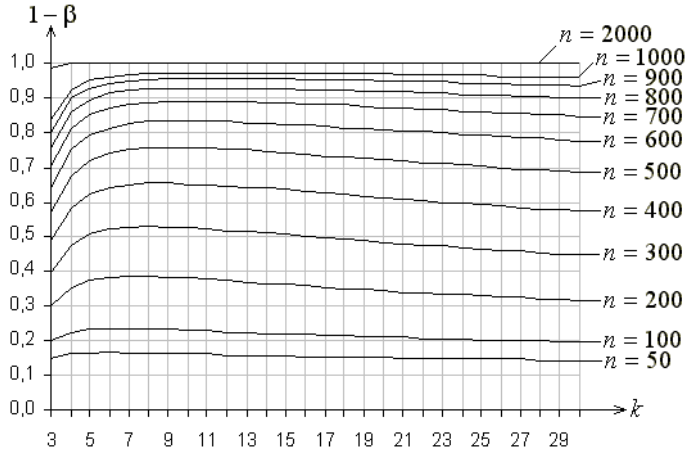

Рис. 3. Функции мощности критерия типа χ^2 Никулина при проверке сложной гипотезы о согласии с нормальным законом при РВГ-группировании и альтернативе, соответствующей логистическому закону

Таким образом, увеличение мощности критериев χ^2 Пирсона и отношения правдоподобия возможно за счет двух факторов: за счет выбора АОГ-группирования в качестве способа разбиения области определения случайной величины и за счет подбора оптимального числа интервалов k при заданном объеме выборки n. Увеличение мощности критерия типа χ^2 Никулина возможно только за счет выбора оптимального числа интервалов.

Оптимальное число интервалов k зависит от объема выборки n и от конкретной пары конкурирующих гипотез H_0 и H_1. Чаще всего оптимальное k оказывается существенно меньше значений, рекомендуемых различными регламентирующими документами и задаваемых множеством эмпирических формул.

Рассматривая пару альтернатив, всегда можно выбрать оптимальное число интервалов и подобрать оптимальное разбиение на интервалы, в результате чего будем иметь критерий максимальной мощности, наилучшим образом различающий данные конкурирующие гипотезы.

4. Исследование вопросов применения непараметрических критериев согласия при проверке сложных гипотез

К наиболее используемым критериям согласия относятся непараметрические критерии типа Колмогорова, типа ω^2 и Ω^2 Мизеса. В критерии Колмогорова в качестве расстояния между эмпирическим и теоретическим законом используется величина

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

где $F_n(x)$ – эмпирическая функция распределения, $F(x,\theta)$ – теоретическая функция распределения, n – объём выборки. При проверке гипотез обычно используется статистика вида (см. Большев Л.Н., Смирнов Н.В.)

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

где

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

n - объем выборки, $x_1, x_2, \ldots, x_n$ - упорядоченные по возрастанию выборочные значения, $F(x,\theta)$ - функция закона распределения, согласие с которым проверяется. Распределение величины S_K при простой гипотезе в пределе подчиняется закону Колмогорова $K(S)$ (см. Большев Л.Н., Смирнов Н.В.).

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

$$\int_{-\infty}^{\infty} \left\{E\left[F_n(x)\right] - F(x)\right\}^2 \psi\left(F(x)\right) dF(x),$$

где $E[\cdot]$ - оператор математического ожидания.

При выборе $\psi(t) \equiv 1$ в критериях типа ω^2 Мизеса пользуются статистикой (статистика Крамера-Мизеса-Смирнова) вида

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

при простой гипотезе подчиняющаяся распределению $a1(S)$ (см. Большев Л.Н., Смирнов Н.В.).

При выборе $\psi(t) \equiv 1/t(1-t)$ в критериях типа Ω^2 Мизеса применяемая статистика (статистика Андерсона-Дарлинга) имеет вид

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

В пределе эта статистика подчиняется распределению $a2(S)$ (см. Большев Л.Н., Смирнов Н.В.).

В случае простых гипотез предельные распределения статистик непараметрических критериев типа Колмогорова, ω^2 и Ω^2 Мизеса известны давно и не зависят от вида наблюдаемого закона распределения и значений его параметров. Говорят, что эти критерии являются "свободными от распределения". Это достоинство предопределяет широкое использование данных критериев в приложениях.

При проверке сложных гипотез, когда по той же самой выборке оцениваются параметры наблюдаемого закона $F(x,\theta)$, непараметрические критерии согласия теряют свойство "свободы от распределения". Однако, мощность непараметрических критериев при проверке сложных гипотез при тех же объемах выборок n всегда существенно выше, чем при проверке простых. И если при проверке простых гипотез непараметрические критерии типа Колмогорова, ω^2 и Ω^2 Мизеса уступают по мощности критериям типа χ^2, при условии, что в последних используется асимптотически оптимальное группирование [5,7-10], то при проверке сложных гипотез непараметрические критерии оказываются более мощными. Для того чтобы воспользоваться их преимуществами, надо только знать распределение $G(S|H_0)$ для проверяемой сложной гипотезы.

Различия в предельных распределениях тех же самых статистик при проверке простых и сложных гипотез очень существенны. Поэтому предостережения против неаккуратного применения критериев согласия при проверке сложных гипотез неоднократно поднимались на страницах печати[7]. При проверке сложных гипотез на условный закон распределения статистики $G(S|H_0)$ влияет целый ряд факторов, определяющих "сложность" гипотезы: вид наблюдаемого закона $F(x,\theta)$, соответствующего истинной гипотезе H_0; тип оцениваемого параметра и количество оцениваемых параметров; в некоторых ситуациях конкретное значение параметра (например, в случае гамма-распределения); используемый метод оценивания параметров и точность вычисления оценок.

Исходной точкой для исследований предельных распределений статистик непараметрических критериев согласия при сложных гипотезах послужила работа Каса-Кифера-Вольфовица[8]. В литературных источниках изложен ряд подходов к использованию непараметрических критериев согласия в случае проверки сложных гипотез. При достаточно большом объеме выборки ее можно разбить на две части и по одной из них оценивать параметры, а по другой проверять согласие[9]. В некоторых частных случаях предельные распределения статистик исследовались аналитическими методами[10], процентные точки распределений строились методами статистического моделирования[11].

[7] Орлов А.И. // Заводская лаборатория. – 1985. – Т. 51. – №1. – С. 60-62.
[8] Kac M., Kiefer J., Wolfowitz J. // Ann. Math. Stat. – 1955. – V.26. – P.189-211.
[9] Durbin J. // Lect. Notes Math. – 1976. – V. 566. – P. 33–44.
[10] Мартынов Г.В. Критерии омега–квадрат. – М.: Наука, 1978. – 80 с.
[11] Pearson E.S., Hartley H.O. Biometrica tables for Statistics. V.2. – Cambridge: University Press, 1972. – 634 p.

Для приближенного вычисления вероятностей “согласия” вида $P\{S > S^*\}$ (достигаемого уровня значимости) строились формулы, дающие достаточно хорошие приближения при малых значениях соответствующих вероятностей[12]. В наших работах [12-15] исследование распределений статистик непараметрических критериев согласия и построение моделей этих распределений осуществлялось с использованием методики компьютерного анализа статистических закономерностей.

Наши исследования показали, что распределения статистик критериев согласия существенно зависят от метода оценивания параметров. Вообще говоря, каждому типу оценок при конкретной сложной проверяемой гипотезе соответствует своё предельное распределение $G(S|H_0)$ статистики. В случае метода максимального правдоподобия распределения статистик $G(S|H_0)$ очень сильно зависят от закона, соответствующего гипотезе H_0. В то же время, разброс распределений $G(S|H_0)$ при использовании *MD*-оценок, минимизирующих статистику критерия, в существенно меньшей степени зависит от вида закона $F(x,\theta)$, соответствующего гипотезе H_0. Это позволяет говорить об определенной “свободе от распределения” для рассматриваемых критериев. Если опираться только на этот факт, то, казалось бы, что только такие методы оценивания и следует применять при проверке сложных гипотез. Однако исследование мощности рассматриваемых критериев при различных методах оценивания показало, что наибольшую мощность данные критерии при близких альтернативах имеют в случае использования ОМП.

Чтобы показать, насколько сильно меняются предельные распределения статистик при различных гипотезах, на рис. 4 иллюстрируется изменение распределений $G(S_K|H_0)$ статистики типа Колмогорова S_K в зависимости от числа оцениваемых параметров закона распределения *Su*-Джонсона с плотностью

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

На данном рисунке “1” отмечена функция распределения Колмогорова, которому подчиняется статистика при проверке простой гипотезы, “2” – распределение статистики при проверке сложной гипотезы и оценивании по данной выборке методом максимального правдоподобия только параметра θ_0, “3” – при оценивании только параметров θ_0 и θ_1, “4” – при оценивании параметров θ_0, θ_1 и θ_2, “5” – при оценивании всех четырех параметров $\theta_0, \theta_1, \theta_2, \theta_3$.

Распределения статистик непараметрических критериев согласия существенно зависят от вида и числа оцениваемых параметров. И даже при оценивании единственного параметра предельное распределение статистики резко отличается от предельного распределения той же самой статистики в случае проверки простой гипотезы. Различие возрастает с увеличением числа оцениваемых параметров. Пренебрежение этим фактом в практике применения критериев согласия приводит к большим ошибкам в вычислении вероятности вида $P\{S > S^*\}$ и неоправданному принятию проверяемой гипотезы.

Stephens M.A. // J. R. Stat. Soc. – 1970. – B. 32. – P. 115-122.
Stephens M.A. // J. Am. Statist. Assoc. – 1974. – V.69. – P. 730-737.
Chandra M., Singpurwalla N.D., Stephens M.A. // J. Am. Statist. Assoc. – 1981. – V.76. – P. 375.

[12] Тюрин Ю.Н. // Изв. АН СССР. Сер. Матем. – 1984. – Т. 48. – № 6. – С. 1314-1343.
Тюрин Ю.Н., Саввушкина Н.Е. // Изв. АН СССР. Сер. Техн. Кибернетика. – 1984. – № 3. – С. 109-112.
Саввушкина Н.Е. // Сб. тр. ВНИИ систем. исслед. – 1990, № 8.

При малых объемах выборки n распределения $G(S_n|H_0)$ зависят от n. Однако, существенная зависимость распределения статистик от n наблюдается только при небольших объемах выборки. Как показали наши исследования, при $n \geq 15 \div 20$ распределения $G(S_n|H_0)$ достаточно близки к предельным $G(S|H_0)$ и зависимостью от n можно пренебречь.

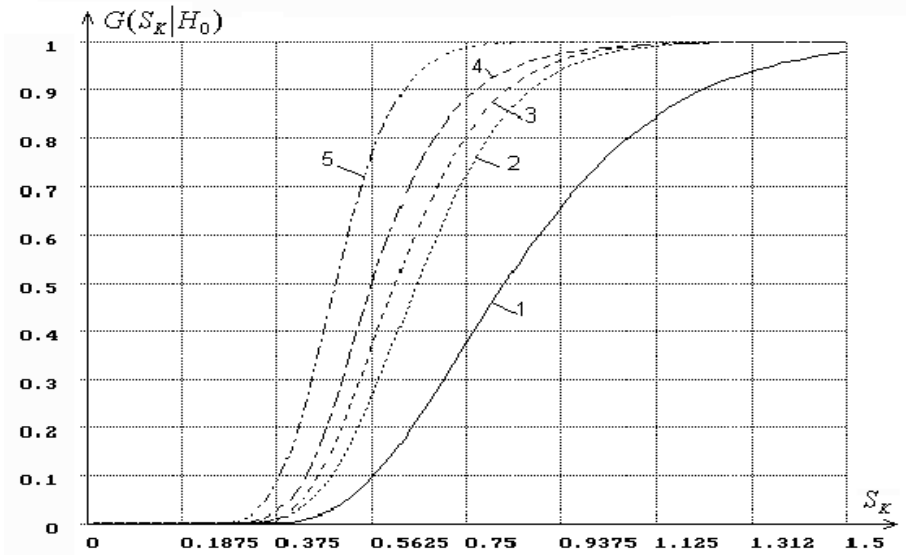

Рис. 4. Изменение распределений статистики типа Колмогорова в зависимости от числа оцениваемых параметров наблюдаемого распределения *Su*-Джонсона

Построенные на настоящий момент другими исследователями таблицы процентных точек и предельные распределения статистик непараметрических критериев ограничены относительно узким кругом сложных гипотез. Это объясняется тем, что построение предельного распределения аналитическими методами выливается в чрезвычайно непростую задачу. В то же время методика компьютерного анализа статистических закономерностей, хорошо зарекомендовавшая себя при моделировании распределений статистик критериев [11-14], позволяет при необходимости без существенных трудностей расширить этот круг.

В работах [13-14] нами были построены модели предельных распределений статистик рассматриваемых критериев при проверке сложных гипотез относительно 13 различных законов наблюдаемых случайных величин при использовании ОМП и *MD*-оценок. Это в общей сложности 280 моделей предельных распределений для такого же количества вариантов сложных гипотез. С учетом того, что к настоящему времени нами получены модели предельных распределений статистик при проверке различных сложных гипотез о согласии с распределениями *Sb*-, *Sl*- и *Su*-Джонсона и использовании ОМП, общее число моделей, построенных в результате применения методики превысило 350(!).

В соответствии с этой методикой следует по закону $F(x,\hat{\theta})$ смоделировать N выборок того же объема n, что и выборка, для которой необходимо проверить гипотезу $H_0:\ F(x) \in \{F(x,\theta), \theta \in \Theta\}$. Далее для каждой из N выборок вычислить оценки тех же параметров закона, а затем значение статистики S соответствующего критерия. В результате будет получена выборка значений статистики $S_1, S_2, .., S_N$ с законом распределения $G(S_n|H_0)$ для проверяемой гипотезы H_0. По этой выборке при значительных N можно построить достаточно гладкую эмпирическую функцию распределения $G_N(S_n|H_0)$, которой можно непосредственно воспользоваться для вывода о том, следу-

ет ли принимать гипотезу H_0. При необходимости можно по $G_N(S_n|H_0)$ построить приближенную аналитическую модель, аппроксимирующую $G_N(S_n|H_0)$, и тогда уже, опираясь на эту модель, принимать решение относительно проверяемой гипотезы.

Как показали наши исследования, хорошей аналитической моделью для $G_N(S_n|H_0)$ часто оказывается один из следующих четырех законов: логарифмически нормальный, гамма-распредление, распределение *Su*-Джонсона или распределение *Sl*-Джонсона [12-13]. В крайнем случае, всегда можно, опираясь на ограниченное множество законов распределения, построить модель в виде смеси законов [15,16].

Построенные нами аппроксимации предельных распределений статистик непараметрических критериев согласия расширяют область корректного применения этих критериев и могут быть рекомендованы широкому кругу исследователей. Расширение множества моделей распределений статистик, то есть расширение множества корректно проверяемых сложных гипотез, не является самоцелью. Это имеет смысл только для тех законов распределения случайных величин, которые наиболее часто используются в приложениях. Именно для этих законов следует продолжить, в том числе и наши, исследования. В целом же мы имеем дело с бесконечными множествами существующих случайных величин, их законов, возможных формулировок сложных гипотез. Следовательно, речь должна идти о постепенном *создании информационных технологий* исследования статистических закономерностей, технологий обеспечивающих статистический анализ и обработку экспериментальных наблюдений в различных областях знаний и человеческой жизнедеятельности.

5. Исследование вопросов оценивания параметров и проверки гипотез по сильно цензурированным выборкам

С задачей обработки цензурированных выборок, когда наблюдению оказывается доступной только часть области определения случайной величины, а для выборочных значений, попавших левее и/или правее этой области, фиксируется лишь сам факт этого попадания, приходится сталкиваться в различных приложениях. Особенно часто с цензурированными выборками встречаются в задачах надежности при оценивании продолжительности жизни. В такой неполной (цензурированной) выборке содержится меньше информации, чем в полной. Потеря части информации отражается на точности оценивания параметров аппроксимирующего закона распределения. При цензурировании наблюдений снижается способность критериев согласия различать близкие законы распределения. В среде специалистов по надежности, которым наиболее часто приходится сталкиваться с проблемами обработки сильно цензурированных выборок, сложилось даже мнение о бесперспективности различения моделей законов распределений, используемых в задачах надежности и контроля качества, с помощью критериев согласия [13].

В работе [17] и в дальнейшем нами были исследованы потери в информации Фишера в зависимости от степени цензурирования для различных законов распределения. Было показано, что в некоторых случаях, когда доступной наблюдению оказывается лишь незначительная область определения случайной величины, в цензурированной выборке сохраняется достаточно много информации. Например, в случае экспоненциального закона распределения при цензурировании слева, когда доступной наблюдению оказывается область справа, вероятность попадания в которую равна 0.05, сохраняется более 52% (!) от информации, имеющейся в полной выборке. Наличие подобных фактов позволяет надеяться, что по таким выборкам можно как находить хорошие

[13] Демидович О.Н. // Методы менеджмента качества. – 1999. – № 11. – С. 29-33

оценки параметров, так и достаточно уверенно проверять гипотезы о согласии эмпирического распределения с теоретическим. В таблице 1, приводимой ниже, представлены значения относительной информации Фишера о параметрах некоторых законов, сохраняющейся в цензурированной выборке, в зависимости от величины наблюдаемой области, вероятность попадания в которую показана в таблице в процентах.

Таблица 1

Отношение количества информации Фишера в наблюдении цензурированной выборки к количеству информации в нецензурированной $J_c(\theta)/J(\theta)$ $\left(\det \mathbf{J}_c(\theta)/\det \mathbf{J}(\theta)\right)$

Наблюдаемая часть %	О масштабном параметре распределений экспоненциального и Вейбулла, о параметрах сдвига распределений минимального и максимального[*)] значения		О параметре формы распределения Вейбулла, о параметрах масштаба распределений минимального и максимального[*)] значения		О двух параметрах распределения Вейбулла, распределений минимального и максимального[*)] значения	
	Цензурирование слева	Цензурирование справа	Цензурирование слева	Цензурирование справа	Цензурирование слева	Цензурирование справа
100	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
60	0.9914	0.6000	0.7091	0.4343	0.6389	0.2658
50	0.9805	0.5000	0.6343	0.4011	0.5256	0.1771
40	0.9597	0.4000	0.5680	0.3878	0.4076	0.1093
30	0.9212	0.3000	0.5168	0.3859	0.2878	0.0595
20	0.8476	0.2000	0.4883	0.3814	0.1707	0.0257
10	0.6891	0.1000	0.4830	0.3405	0.0654	0.0063
5	0.5223	0.0500	0.4654	0.2718	0.0234	0.0015
Наблюдаемая часть %	О параметре сдвига нормального распределения	О параметре масштаба нормального распределения	О двух параметрах нормального распределения	О параметре масштаба распределения Лапласа	О параметре распределения Рэлея	
	Цензурирование слева[**)]	Цензурирование слева[**)]	Цензурирование слева[**)]	Цензурирование слева[**)]	Цензурирование слева	Цензурирование справа
100	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
60	0.8753	0.5599	0.4399	0.6131	0.9914	0.6000
50	0.8183	0.5000	0.3296	0.6103	0.9805	0.5000
40	0.7467	0.4601	0.2311	0.5918	0.9597	0.4000
30	0.6550	0.4399	0.1457	0.5538	0.9212	0.3000
20	0.5336	0.4309	0.0754	0.4885	0.8476	0.2000
10	0.3591	0.4252	0.0239	0.3740	0.6891	0.1000
5	0.2318	0.3795	0.0073	0.2730	0.5223	0.0500

[*)] – для распределения максимального значения левое цензурирование меняется на правое;

[**)] – при левом и правом цензурировании ситуация идентична.

ОМП параметров распределений по цензурированным наблюдениям являются асимптотически эффективными. Однако при ограниченных объемах выборок и значительной степени цензурирования законы распределения ОМП весьма далеки от асимптотически нормального и, более того, оказываются *асимметричными*, а сами оценки *смещенными* [18]. С уменьшением n и увеличением степени цензурирования увеличивается асимметрия закона распределения оценок (см. рис. 5-6).

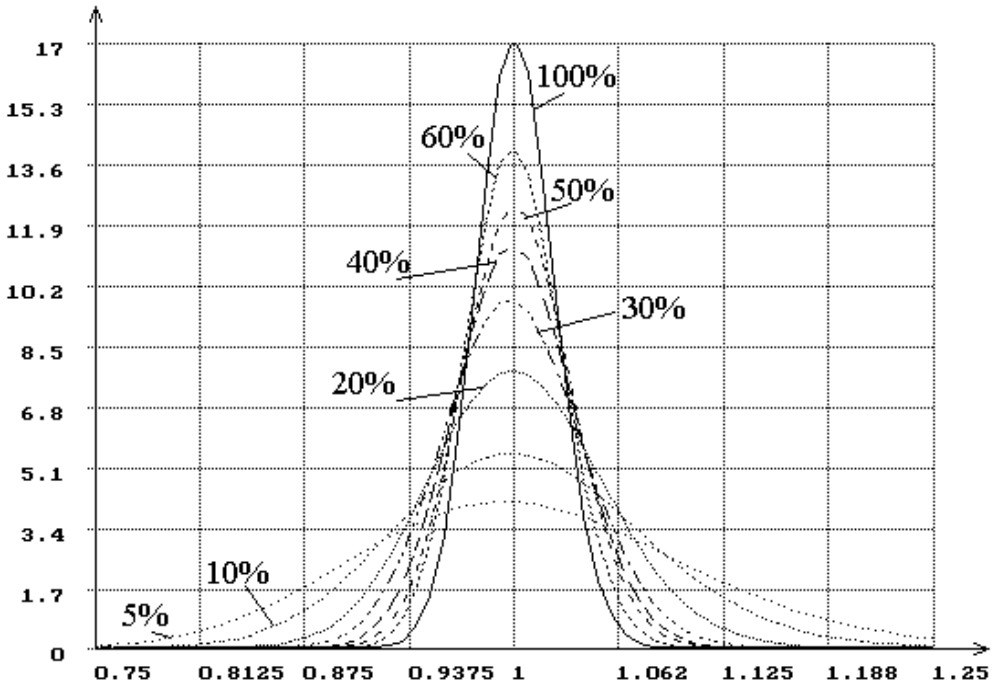

Рис. 5. Плотности распределения оценок масштабного параметра экспоненциального распределения при цензурировании справа при различной величине (%) наблюдаемой области определения случайной величины и полном объеме выборки n =2000.

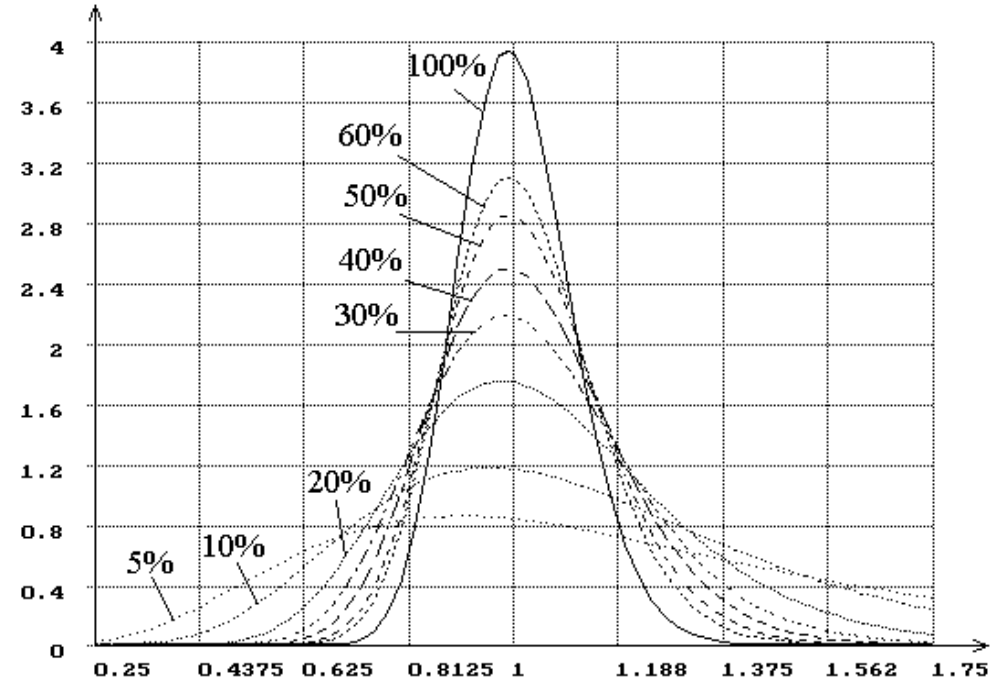

Рис. 6. Плотности распределения оценок масштабного параметра экспоненциального распределения при цензурировании справа при различной величине (%) наблюдаемой области определения случайной величины и полном объеме выборки n =100

С использованием методики компьютерного моделирования и анализа статистических закономерностей были исследованы величины смещения ОМП параметров некоторых законов в зависимости от объема всей выборки n и величины наблюдаемой её части. В результате исследования были получены оценки математических ожиданий смещений ОМП как эмпирические функции от объема выборки и степени цензурирования. Использование полученных величин в качестве поправок, ликвидирующих смещение при оценивании параметров экспоненциального и нормального законов, показало, что такие оценки с поправками оказываются несмещенными.

Для проверки согласия при цензурированных наблюдениях и простых гипотезах могут использоваться критерии типа Реньи (см. Большев Л.Н., Смирнов Н.В.), которые в этой ситуации являются "свободными от распределения". Очевидно, что при проверке сложных гипотез они теряют это свойство. Поэтому необходимы исследования распределений этих статистик при проверке различных сложных гипотез с целью построения моделей предельных распределений. Проверка сложных гипотез тесно взаимосвязана с проблемой оценивания параметров. В данном случае наибольшая неприятность связана со смещенностью ОМП параметров, вычисляемых по цензурированным выборкам. Показанная возможность построения несмещенных оценок параметров по цензурированным выборкам, позволяет исследовать распределения статистик критериев согласия при проверке соответствующих сложных гипотез, что, в свою очередь, позволит корректно проверять сложные гипотезы относительно законов распределения по цензурированным наблюдениям.

6. Исследование распределений статистик корреляционного анализа при отклонении многомерного закона от нормального

В различных приложениях статистического анализа многомерных случайных величин одну из ключевых позиций занимают задачи корреляционного анализа. В процессе решения этих задач вычисляются оценки коэффициентов и матриц парной, частной и множественной корреляции, проверяются различные статистические гипотезы относительно параметров многомерного распределения и коэффициентов корреляции. На основании результатов корреляционного анализа может делаться вывод или о наличии и характере функциональной зависимости, или о предпочтительности для описания исследуемого объекта регрессионной модели того или иного вида.

В основе существующего аппарата классического корреляционного анализа лежит предположение о принадлежности наблюдаемого случайного вектора *многомерному нормальному* закону. Базируясь на этом, получены предельные распределения статистик, используемых в классическом корреляционном анализе[14].

Большинство задач классического корреляционного анализа реализовано в нашей системе [19]. Целью исследований, проводимых в последнее время, явилось стремление проанализировать, что будет происходить с распределениями используемых в корреляционном анализе статистик, если наблюдаемый закон отличается от многомерного нормального.

Одной из возникающих проблем на пути исследования методами статистического моделирования предельных распределений статистик, является задача моделирования псевдослучайных векторов, "заданным образом" отличающихся от многомерного нормального. В [20] для моделирования многомерных распределений таких, как логи-

[14] Андерсон Т. Введение в многомерный статистический анализ. - М.: Физматгиз, 1963. - 500 с.
Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. – М.: Наука, 1976. – 736 с.

стическое и Лапласа, нами использовался подход принятый для нормального закона[15]. Статистический анализ моделируемых псевдослучайных векторов показал, что маргинальные функции распределения получаемых псевдослучайных величин имеют хорошее согласие с одноименными одномерными законами. В то же время, пока нельзя считать, что нами полностью решена задача моделирования псевдослучайных векторов с законом распределения, "заданным образом" отличающимся от многомерного нормального.

В [20] была исследована сходимость распределений ряда статистик корреляционного анализа к соответствующим предельным в зависимости от объемов выборок многомерной случайной величины. Было отмечено, что предельными распределениями различных статистик можно пользоваться, начиная с различных объемов выборок n. Было показано, что при "не слишком большом" отклонении многомерного закона от нормального предельные распределения большинства исследуемых статистик не претерпевают значимых изменений, за исключением некоторых. Была подтверждена эффективность применения методики компьютерного анализа статистических закономерностей для исследования распределений статистик задач корреляционного анализа многомерных случайных величин и показана возможность построения методами компьютерного анализа моделей предельных распределений статистик при любом виде наблюдаемого закона многомерной случайной величины.

7. Заключение

Авторским коллективом проводились и проводятся исследования, связанные с проверкой статистических гипотез по интервальным наблюдениям [21-23], с решением задач регрессионного анализа при точечных и группированных наблюдениях. Исследовались вопросы асимптотически оптимального группирования в задачах регрессионного анализа. Исследуются методы непараметрического оценивания плотностей распределений [24], начаты исследования вопросов применения критериев согласия при использовании непараметрических оценок.

Можно констатировать, что за последние 5 лет сформировалось научное направление, в основе которого лежат методы компьютерного анализа и исследования вероятностных закономерностей на базе аппарата математической статистики, вычислительных методов и статистического моделирования. Эффективность применяемых подходов и методов обуславливает постоянное расширение тематики исследований авторского коллектива, связанное ограничениями лишь материального характера и физическими возможностями.

Результаты и методика исследований активно используются в учебном процессе факультета прикладной математики и информатики Новосибирского государственного технического университета. Техническим комитетом ТК 125 "Стандартизация статистических методов управления качеством" готовятся на утверждение Госстандартом России 2 части методических рекомендаций, подготовленных на базе [10,14] и последних результатов.

Литература

1. Лемешко Б.Ю. Статистический анализ одномерных наблюдений случайных величин: Программная система. – Новосибирск: Изд–во НГТУ, 1995. – 125 с.

[15] Ермаков С.М., Михайлов Г.А. Статистическое моделирование. – Москва: Издательство «Наука», 1982. – 296 с.

2. Лемешко Б.Ю., Постовалов С.Н. // Сб. научных трудов НГТУ. – Новосибирск: Изд–во НГТУ. – 1996. – № 2(4). – С. 9–18.
3. Лемешко Б.Ю. // Надежность и контроль качества. – 1997. – № 5. – С. 26–35.
4. Лемешко Б.Ю. // Заводская лаборатория. – 1997. – Т.63. – № 5. – С. 43–49.
5. Денисов В.И., Лемешко Б.Ю., Цой Е.Б. Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов: В 2 ч. / Новосиб. гос. техн. ун–т. – Новосибирск, 1993. – 346 с.
6. Лемешко Б.Ю. // Труды III межд. конференции "Актуальные проблемы электронного приборостроения АПЭП–96". – Новосибирск, 1996. – Т. 6. – Ч.1. – С.37–44.
7. Лемешко Б.Ю., Постовалов С.Н. // Заводская лаборатория. 1998. Т. 64. – № 5. – С.56–63.
8. Лемешко Б.Ю. // Надежность и контроль качества. – 1997. – № 8. – С. 3–14.
9. Лемешко Б.Ю. // Заводская лаборатория, 1998. Т. 64. – №1. – С.56–64.
10. 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.
11. Лемешко Б.Ю., Чимитова Е.В. // Труды V межд. конференции "Актуальные проблемы электронного приборостроения АПЭП–2000". – Новосибирск, 2000. – Т. 6. – С. 21–23.
12. Лемешко Б.Ю., Постовалов С.Н. // Надежность и контроль качества. – 1997. – № 11. – С. 3–17.
13. Лемешко Б.Ю., Постовалов С.Н. // Заводская лаборатория. – 1998. – Т. 64. – № 3. – С. 61–72.
14. Лемешко Б.Ю., Постовалов С.Н. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Часть II. Непараметрические критерии. – Новосибирск: Изд–во НГТУ. – 1999. – 86 с.
15. Лемешко Б.Ю., Постовалов С.Н. // Изв. вузов. Физика. – Томск, 1995. – № 9. – С. 39–45.
16. Лемешко Б.Ю., Постовалов С.Н. // Сб. научных трудов НГТУ. – Новосибирск: Изд–во НГТУ, 1995. – № 1. – С. 25–31.
17. Лемешко Б.Ю., Гильдебрант С.Я., Постовалов С.Н. // Тр. IV межд. конференции "Актуальные проблемы электронного приборостроения АПЭП–98". Т. 3, Новосибирск, 1998. – С. 17–23.
18. Лемешко Б.Ю., Постовалов С.Н., Чимитова Е.В. // Тр. V межд. конференции "Актуальные проблемы электронного приборостроения АПЭП–2000". – Новосибирск, 2000. – Т. 7. – С. 188–191.
19. Лемешко Б.Ю. Корреляционный анализ многомерных наблюдений случайных величин: Программная система. – Новосибирск: Изд–во НГТУ, 1995. – 39 с.
20. Лемешко Б.Ю., Помадин С.С. // Тр. V межд. конференции "Актуальные проблемы электронного приборостроения АПЭП–2000". – Новосибирск, 2000. – Т. 7. – С. 184–187.
21. Лемешко Б.Ю., Постовалов С.Н. // Сб. научных трудов НГТУ. – Новосибирск: Изд–во НГТУ, 1995. – № 2. – С. 21–30.
22. Лемешко Б.Ю., Постовалов С.Н. // Вычислительные технологии. – 1997. – Т.2. – № 1. – С. 28–36.
23. Лемешко Б.Ю., Постовалов С.Н. // Вычислительные технологии. 1998. Т.3. – № 2. – С. 31–38.
24. Лемешко Б.Ю., Постовалов С.Н., Французов А.В. // Тр. V межд. конференции "Актуальные проблемы электронного приборостроения АПЭП–2000". – Новосибирск, 2000. – Т. 6. – С. 17–20.