СБОРНИК НАУЧНЫХ ТРУДОВ НГТУ. - 1996. - № 1. - 3 - 12
УДК 519.2 (075.8)

СТАТИСТИЧЕСКИЙ АНАЛИЗ НАБЛЮДЕНИЙ, ИМЕЮЩИХ ИНТЕРВАЛЬНОЕ ПРЕДСТАВЛЕНИЕ

Б.Ю. ЛЕМЕШКО♣, С.Н. ПОСТОВАЛОВ♥

В статье рассмотрены вопросы проверки гипотез о согласии и вопросы построения точечных оценок по исходным данным, имеющим интервальное представление. Предложена модификация критериев согласия χ^2 Пирсона, отношения правдоподобия, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса и метода максимального правдоподобия вычисления точечных оценок по интервальной выборке. Приведены результаты программной реализации.

1. ВВЕДЕНИЕ

Наиболее общим представлением исходных данных в классической статистике будет частично группированная выборка, частными случаями которой являются негруппированные, группированные и цензурированные данные. Дальнейшим обобщением частично группированной выборки является *интервальная* выборка, в которой каждое наблюдение представлено интервалом $[a_i, b_i]$, которому принадлежит неизвестное точно значение x_i. Классификация одномерных выборок показана на рис.1. Группированная выборка задана непересекающимися интервалами, негруппированная - вырожденными интервалами, у которых $a_i = b_i$. Интервальное представление наблюдения можно интерпретировать как неточное измерение случайной величины, связанное либо с заведомо известной погрешностью измерительного прибора, либо с особенностями измеряемой величины. Например, уровень воды в реке невозможно измерить точно, так как он непрерывно колеблется, но можно представить в виде интервала задающего его верхнюю и нижнюю границы [1]. Интервальное представление можно получить также в результате обработки точных наблюдений, сопровождающихся усечением части данных (например, при *группировании* или *цензурировании*).

Подробная библиография работ, связанных с вопросами интервальной статистики, приведена в [2].

Для статистического анализа интервальных наблюдений большинство классических методов непосредственно неприменимо, требуется их модификация, либо разработка новых. В настоящей статье рассмотрена модификация непараметрических критериев согласия Колмогорова, Смир-

---

♣ Доцент кафедры прикладной математики, канд. техн. наук
♥ Аспирант кафедры прикладной математики

нова, ω^2 и Ω^2 Мизеса и предложен способ построения точечных оценок, минимизирующих статистики соответствующих критериев согласия.

По сравнению с частным случаем поставленной задачи, рассмотренным в [3], нам удалось уточнить и упростить полученные выражения.

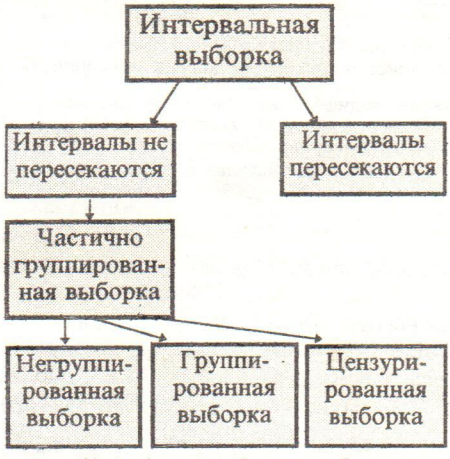

Рис. 1 Классификация выборочных наблюдений

1. ЭМПИРИЧЕСКАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

Пусть задана интервальная выборка:

$$\mathbf{X_k}=\left\{\left\langle a_i,b_i,n_i\right\rangle \,\middle|\, a_i\le x_{ij}\le b_i,\ j=1,\ldots,n_i;\ a_i\in R,\ b_i\in R,\ i=1,\ldots,k\right\},$$

где k-число интервалов; $n=\sum_{i=1}^{k} n_i$ - объём выборки; n_i - число наблюдений в i-м интервале; a_i и b_i - границы i-го интервала; x_{ij} - точные значения наблюдений. Упорядочим граничные точки интервалов:

$$a_{(1)} < a_{(2)} < \ldots < a_{(k)};$$
$$b_{(1)} < b_{(2)} < \ldots < b_{(k)}.$$

Обозначим через $n_{a_{(i)}}$ - число наблюдений, принадлежащих интервалу с левой границей $a_{(i)}$, и $n_{b_{(i)}}$ - число наблюдений, принадлежащих интервалу с правой границей $b_{(i)}$. Тогда

$$N_{a_{(i)}} = \sum_{j=1}^{i} n_{a_{(j)}}; \quad N_{b_{(i)}} = \sum_{j=1}^{i} n_{b_{(j)}}, \quad i = 1, \ldots, k.$$

Эмпирическая функция распределения $F_n(x)$, построенная по точным наблюдениям x_{ij}, будет ограничена снизу и сверху функциями $\underline{F_n}(x)$ и $\overline{F_n}(x)$, имеющими следующий вид:

$$\overline{F_n}(x) = \begin{cases} 0, & x < a_{(1)} \\ \dfrac{N_{a_{(i)}}}{n}, & a_{(i)} \le x < a_{(i+1)}, \quad i = 1, \ldots, k-1, \\ 1, & x \ge a_{(k)} \end{cases}$$

$$\underline{F_n}(x) = \begin{cases} 0, & x < b_{(1)} \\ \dfrac{N_{b_{(i)}}}{n}, & b_{(i)} \le x < b_{(i+1)}, \quad i = 1, \ldots, k-1. \\ 1, & x \ge b_{(k)} \end{cases}$$

Таким образом,

$$\underline{F_n}(x) \le F_n(x) \le \overline{F_n}(x), \quad \forall\, x \in R. \tag{1}$$

На рис. 2 показана эмпирическая функция распределения ($\underline{F_n}(x)$ и $\overline{F_n}(x)$) для различных типов исходных данных.

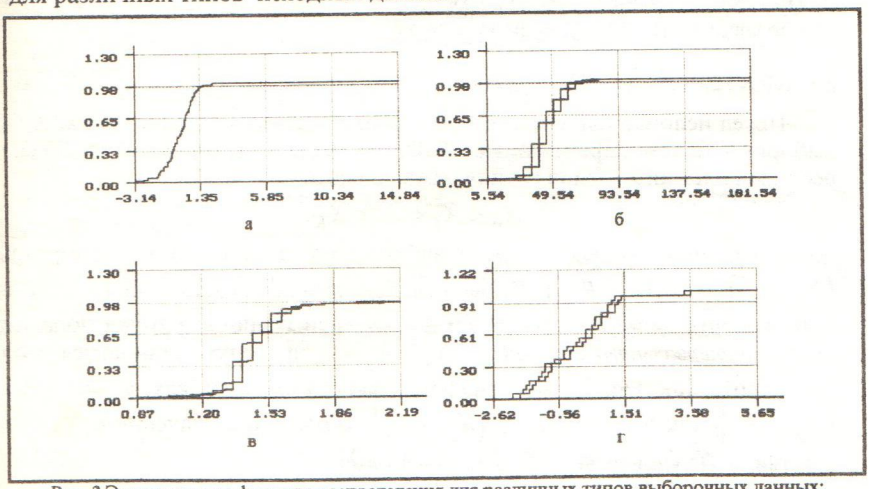

Рис. 2 Эмпирическая функция распределения для различных типов выборочных данных: а - негруппированные; б - группированные; в - частично группированные; г - интервальные

2. ПРОВЕРКА ГИПОТЕЗ О СОГЛАСИИ

При проверке гипотез о согласии для найденного значения соответствующей статистики S^* вычисляется вероятность

$$p = P\left\{S > S^*\right\} = \int_{S^*}^{\infty} g(s)ds,$$

где $g(s)$ - плотность распределения статистики при условии истинности нулевой гипотезы. При заданном уровне значимости α гипотеза о согласии не отвергается, если $p > \alpha$. Когда выборка задана неточно, то статистика принадлежит интервалу $\left[\underline{S^*}, \overline{S^*}\right]$, где на основании (1) границы определяются следующим неравенством:

$$\underline{S^*} = \inf_{\underline{F_n} \le F_n \le \overline{F_n}} S^*(F_n, F) \le S^*(F_n, F) \le \sup_{\underline{F_n} \le F_n \le \overline{F_n}} S^*(F_n, F) = \overline{S^*}$$

Вероятность $P\left\{S > S^*\right\}$ будет принадлежать интервалу $\left[p_{\min}, p_{\max}\right]$, где $p_{\min} = \int_{\overline{S^*}}^{\infty} g(s)ds$, $p_{\max} = \int_{\underline{S^*}}^{\infty} g(s)ds$. Тогда, при заданном уровне значимости α, гипотезу о согласии следует отклонить, если $p_{\max} \le \alpha$; гипотезу о согласии не следует отвергать, если $p_{\min} > \alpha$.

2.1. КРИТЕРИЙ χ^2-ПИРСОНА

Перед использованием критерия необходимо сгруппировать исходную выборку. Область определения случайной величины разбивается на K непересекающихся интервалов граничными точками

$$X_0 < X_1 < \ldots < X_K,$$

после чего подсчитывается число наблюдений m_i, попавших в интервалы $\left(X_i, X_{i+1}\right]$, $i = 0, 1, \ldots, K-1$. Процедура подсчета не представляет трудности, если выборка задана точно. В случае же, если граничная точка попадает внутрь *интервального* наблюдения $a_i < X_j < b_i$ процедура становится неоднозначной, так как точное значение можно отнести как к интервалу $\left(X_{j-1}, X_j\right]$, так и к интервалу $\left(X_j, X_{j+1}\right]$. Возможные значения m_i в соответствии с (1) удовлетворяют ограничениям:

$$n \cdot \underline{F_n}(X_i) \le \sum_{j=0}^{i} m_j \le n \cdot \overline{F_n}(X_i), \;\; m_i \ge 0, \;\; i = 0,1,\ldots,K-1 \; ; \qquad (2)$$

$$\sum_{i=0}^{K-1} m_i = n. \qquad (3)$$

Статистика критерия имеет вид

$$\chi^2 = \sum_{i=0}^{K-1} \frac{(m_i - np_i)^2}{p_i}.$$

Если m_i определяются неоднозначно, то можно найти максимум и минимум статистики χ^2 на области, заданной формулами (2) и (3):

$$\underline{\chi^2} = \min_{m_i} \chi^2; \qquad \overline{\chi^2} = \max_{m_i} \chi^2.$$

Аналогично ограничивается статистика критерия отношения правдоподобия.

2.2. КРИТЕРИЙ СОГЛАСИЯ КОЛМОГОРОВА

Статистика критерия имеет вид

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

где $F_n(x)$ - эмпирическая функция распределения, $F(x)$ - теоретическая, согласие с которой проверяется, n - объем выборки. Преобразуем неравенство (1) к виду:

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

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

Эти неравенства выполняются для всех x, поэтому они сохраняются при взятии супремума:

$$\sup_x \left(\underline{F_n}(x) - F(x) \right) \le \sup_x \left(F_n(x) - F(x) \right) \le \sup_x \left(\overline{F_n}(x) - F(x) \right); \qquad (4)$$

$$\sup_x \left(F(x) - \overline{F_n}(x) \right) \le \sup_x \left(F(x) - F_n(x) \right) \le \sup_x \left(F(x) - \underline{F_n}(x) \right). \qquad (5)$$

Объединим эти неравенства в одно:

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

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

$$\le \overline{D_n} = \max\left\{\sup_x\left(\overline{F_n}(x) - F(x)\right),\ \sup_x\left(F(x) - \underline{F_n}(x)\right)\right\}. \qquad (6)$$

2.3. КРИТЕРИЙ СОГЛАСИЯ СМИРНОВА

Статистика критерия имеет вид $D_n^+ = \sup_x\left(F_n(x) - F(x)\right)$. Из неравенства (2) следует:

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

2.4. КРИТЕРИЙ СОГЛАСИЯ ω^2 МИЗЕСА

Построим вариационный ряд для точных значений наблюдений:

$$x_{(1)} < x_{(2)} < \ldots < x_{(n)}.$$

Если наблюдения заданы неточно, то каждый член вариационного ряда известен с точностью до интервала

$$\underline{x_{(i)}} < x_{(i)} < \overline{x_{(i)}}\ , \qquad (7)$$

где $\underline{x_{(i)}}$ и $\overline{x_{(i)}}$ можно определить из неравенства (1), так как между вариационным рядом и эмпирической функцией распределения имеется взаимно-однозначное соответствие (см. рис. 3) и $F_n(x_{(i)}) = i/n$:

$$\underline{x_{(i)}} = \overline{F_n^{-1}}(i/n),\quad \overline{x_{(i)}} = \underline{F_n^{-1}}(i/n).$$

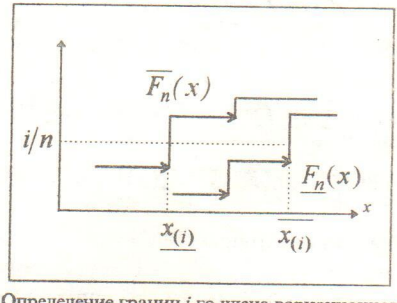

Рис. 3 Определение границ i-го члена вариационного ряда

Статистика критерия имеет вид

$$n\omega_n^2 = n\int_{-\infty}^{+\infty}(F_n(x)-F(x))^2 dF(x) = \frac{1}{12n} + \sum_{i=1}^{n}\left[F(x_{(i)}) - \frac{2i-1}{2n}\right]^2.$$

Пусть $s_i = \left[F(x_{(i)}) - \frac{2i-1}{2n}\right]^2$. Тогда из монотонности функции распределения $F(x)$ и условия (7) следует, что $\underline{s_i} \le s_i \le \overline{s_i}$, где $\underline{s_i}$ и $\overline{s_i}$ имеют вид:

$$\underline{s_i} = \begin{cases} \left[F(\overline{x_{(i)}}) - \frac{2i-1}{2n}\right]^2, \text{ при } F(\underline{x_{(i)}}) \le F(\overline{x_{(i)}}) \le \frac{2i-1}{2n}, \\ 0, \qquad \text{при } F(\underline{x_{(i)}}) \le \frac{2i-1}{2n} \le F(\overline{x_{(i)}}), \\ \left[F(\underline{x_{(i)}}) - \frac{2i-1}{2n}\right]^2, \text{ при } \frac{2i-1}{2n} \le F(\underline{x_{(i)}}) \le F(\overline{x_{(i)}}), \end{cases}$$

$$\overline{s_i} = \max\left\{\left[F(\overline{x_{(i)}}) - \frac{2i-1}{2n}\right]^2, \left[F(\underline{x_{(i)}}) - \frac{2i-1}{2n}\right]^2\right\}.$$

Тогда

$$\underline{n\omega_n}^2 = \frac{1}{12n} + \sum_{i=1}^{n}\underline{s_i}, \quad \overline{n\omega_n}^2 = \frac{1}{12n} + \sum_{i=1}^{n}\overline{s_i}.$$

2.5. КРИТЕРИЙ СОГЛАСИЯ Ω^2 МИЗЕСА

Статистика имеет вид

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

Из неравенства (7) и монотонности функции распределения следует:

$$\left(F(x_{(i)})\right)^{\frac{2i-1}{2n}}\left(1-F(x_{(i)})\right)^{1-\frac{2i-1}{2n}} \le \left(F(\overline{x_{(i)}})\right)^{\frac{2i-1}{2n}}\left(1-F(\underline{x_{(i)}})\right)^{1-\frac{2i-1}{2n}};$$

$$\left(F(x_{(i)})\right)^{\frac{2i-1}{2n}}\left(1-F(x_{(i)})\right)^{1-\frac{2i-1}{2n}} \ge \left(F(\underline{x_{(i)}})\right)^{\frac{2i-1}{2n}}\left(1-F(\overline{x_{(i)}})\right)^{1-\frac{2i-1}{2n}}.$$

Отсюда

$$\underline{\Omega_n^2} = -n - 2\ln\prod_{i=1}^{n}\left(F(\underline{x_{(i)}})\right)^{\frac{2i-1}{2n}}\left(1-F(\overline{x_{(i)}})\right)^{1-\frac{2i-1}{2n}};$$

$$\overline{\Omega_n^2} = -n - 2\ln\prod_{i=1}^{n}\left(F(\overline{x_{(i)}})\right)^{\frac{2i-1}{2n}}\left(1-F(\underline{x_{(i)}})\right)^{1-\frac{2i-1}{2n}}.$$

2.6.ЧИСЛЕННЫЙ ПРИМЕР

Рассмотрим следующий пример. Была смоделирована негруппированная выборка объемом 100 наблюдений по нормальному закону с параметрами (0,1). Оценки максимального правдоподобия параметров сдвига и масштаба по негруппированной выборке равны соответственно $\hat{\mu}=0.048574$ и $\hat{\sigma}=1.07144$ (рис. 4). Далее, предположили, что исходные наблюдения получены с некоторой ошибкой, не превосходящей по модулю $\Delta=0.05$. Тогда каждое i-е наблюдение является интервалом вида $[x_i-\Delta, x_i+\Delta]$. Зафиксировав значение параметра сдвига равным оценке максимального правдоподобия по негруппированной выборке, было проверено согласие по всем рассмотренным критериям в зависимости от параметра масштаба (рис. 5). На рисунке изображены функции $p_{\max}(\sigma)$ и $p_{\min}(\sigma)$.

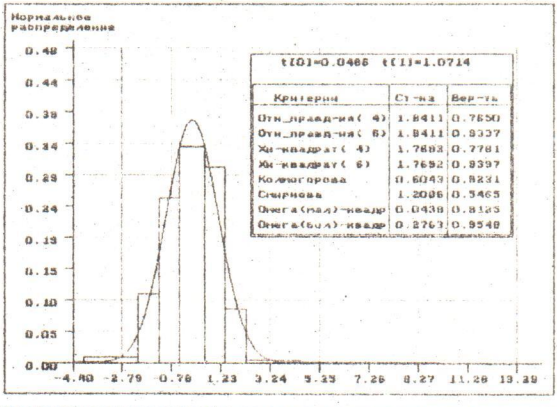

Рис. 4. Оценивание параметров нормального распределения по негруппированной выборке

На основании той же негруппированной выборки, была исследована зависимость длины интервала $p_{\max}-p_{\min}$ от Δ. Проверялось согласие с нормальным распределением с параметрами, оцененными по методу максимального правдоподобия. Очевидно, что с ростом неопределенности ис-

ходных данных (Δ), происходит увеличение неопределенности вероятности согласия ($\Delta p = p_{max} - p_{min}$). При этом критерии можно упорядочить по возрастанию Δp следующим образом: Колмогорова, Смирнова, ω^2 Мизеса, отношения правдоподобия, χ^2 Пирсона, Ω^2 Мизеса. Поэтому, при *интервальном* задании исходной выборки наиболее предпочтительным оказывается использование *критериев Колмогорова, Смирнова, ω^2 Мизеса.*

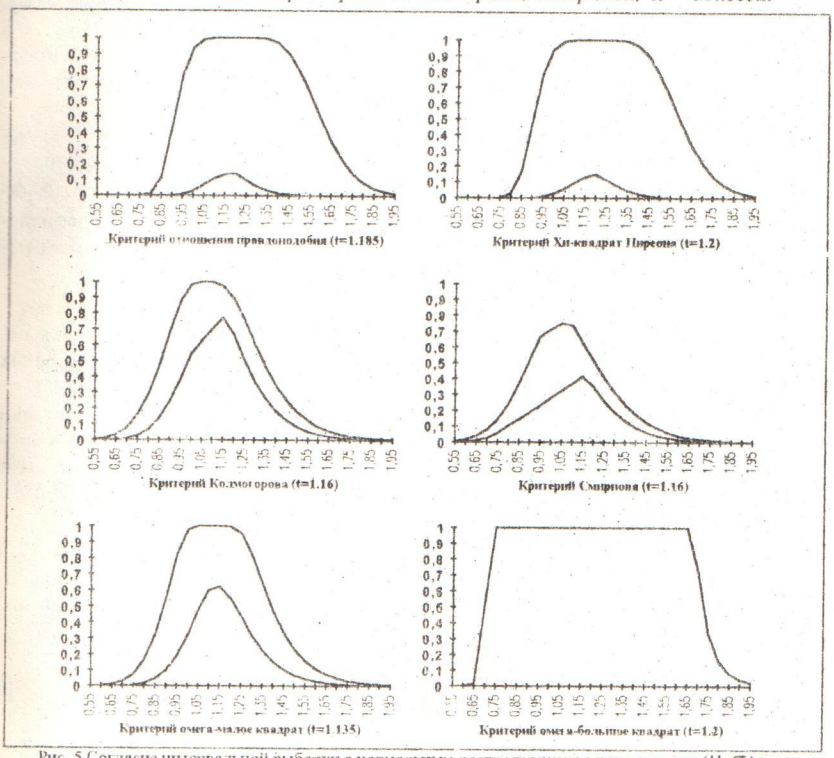

Рис. 5 Согласие интервальной выборки с нормальным распределением с параметрами (μ, σ), где μ =0.048574, а σ изменяется от 0.55 до 1.95. Верхняя линия соответствует p_{max}, нижняя - p_{min}

3. ОЦЕНИВАНИЕ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ

Для получения оценок параметров распределения ни один из известных методов не применим в явном виде. Однако, используя полученные выше выражения оценок для статистик критериев согласия, не представляет труда построить оценки параметров распределений, аналогичные оценкам

типа минимума χ^2. Очевидно, что чем меньше статистика критерия согласия, тем выше будет вероятность не отклонения гипотезы. Поскольку точное значение статистики неизвестно, то можно минимизировать ее верхнюю границу:

$$\hat{\theta} = \arg\min_{\theta} \overline{S^*}(\underline{F_n}, \overline{F_n}, F(\theta)).$$

Оценку такого типа рассматриваем в терминах теории принятия решений в условиях неопределенности: как оценку, максимизирующую вероятность согласия при наихудшем расположении точных значений. На рис. 5 показаны оценки параметра σ в виде значений t, полученные таким способом. По всем критериям оценка масштабного параметра оказалась больше, чем оценка максимального правдоподобия по исходной точной выборке.

Более точные оценки удалось получить, используя модификацию *метода максимального правдоподобия*, когда при достаточно малых интервалах $[a_i, b_i]$ использовали разностную аппроксимацию производной функции распределения. Оценка масштабного параметра нормального распределения, полученная таким образом, оказалась равной σ=1,0657, совсем незначительно отличаясь от точного значения.

4. ЗАКЛЮЧЕНИЕ

Получены общие выражения оценок снизу и сверху для статистик непараметрических критериев согласия в случае частично группированных данных или интервального (нечеткого) их представления.

Показано, что в случае интервального представления исходной выборки достаточно определенные выводы о согласии можно делать на основании полученных оценок для статистик критериев согласия. При этом наиболее эффективными оказываются непараметрические критерии Колмогорова, ω^2 Мизеса и Смирнова.

При оценивании параметров распределения было отмечено, что при малых длинах интервалов наилучшим является использование предложенной модификации метода максимального правдоподобия с разностной аппроксимацией производной функции распределения.

[1] FRUHWIRTH-SCHNATTER S. *On statistical inference for fuzzy data with applications to discriptive statistics* // Fuzzy Sets and Systems, 1992. - № 50. - С. 143 - 165.

[2] ОРЛОВ А.И. *Интервальная статистика* //Интервальные вычисления.- 1992. -№ 1. - С. 44 - 52.

[3] ЛЕМЕШКО Б.Ю., ПОСТОВАЛОВ С.Н. *К использованию непараметрических критериев по частично группированным данным* // Сб. науч. тр. НГТУ. - Новосибирск, 1995. - Вып. 2. - С. 21 - 30.