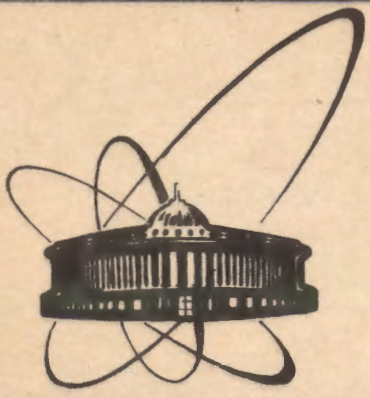

сообщения
объединенного
института
ядерных
исследований
дубна

Р5-91-468

Е.В.Белякова

УСТОЙЧИВАЯ РАВНОМЕТРИЗАЦИЯ ОДНОМЕРНОЙ ВЫБОРКИ

1991

ВВЕДЕНИЕ

В настоящей работе представлено математическое обоснование перехода от исходного семейства распределений к семейству распределений достаточной статистики*. Этот переход называется редукцией. Смысл его — уменьшение размерности пространства наблюдений при сохранении информации и устойчивости данных /1/.

Рассмотрено одномерное пространство наблюдений, следовательно, задача редукции сводится здесь к уменьшению количества наблюдений переходом к новой выборке, состоящей из достаточных статистик, которые, однако, будут все более представительными и устойчивыми при соответственно неограниченном росте объемов исходной и новой выборок.

1. ПОРЯДКОВЫЕ СТАТИСТИКИ ОДНОМЕРНОЙ ВЫБОРКИ

Пусть $\xi^{(n)} = \{\xi_i, i = \overline{1, n}\}$ — последовательность независимых случайных величин /рис. 1/ и $R_n(\xi_i)$ — ранг ξ_i в выборке $\xi^{(n)}$. Значение $R_n(\xi_i)$ складывается из числа ξ_j, строго меньших ξ_i, числа ξ_j, равных ξ_i плюс 1 — само ξ_i.

Рассмотрим двумерные точки $Q_i = \left(\xi_i, \frac{R_n(\xi_i)}{n+1}\right) = (x, y)$ в полосе: $x \in (-\infty, +\infty)$, $y \in (0,1)$. Точки Q_i можно упорядочить по второй компоненте: $Q_1^*, \ldots, Q_n^*$, таким образом, что $Q_j^* = \left(\xi_n(j), \frac{j}{n+1}\right)$, где $\xi_n(j)$ - порядковая статистика выборки $\xi^{(n)}$ /рис. 2/.

На основе выборки $\xi^{(n)}$ получим другую, удобную для нас выборку. Поэтому введем последовательность целых чисел s(n). Исключительно для упрощения работы с целыми числами можно считать, что $\frac{n+1}{s(n)+1} = m_n$ — целое число.

Пусть $\lambda_k = \frac{k}{s(n)+1} = \frac{k*m_n}{n+1}$, $k = 0, 1, \ldots, s(n)$.

Пример 1.

Положим $n = 7$. $\frac{7+1}{s(n)+1} = m_n$ - целое по условию. Пусть $m_7 = 2$. Тогда $s(7) = 8/2 - 1 = 3$.

** Для семейства распределений вероятностей $P_\vartheta, \vartheta \in \Theta$ или для параметра $\vartheta \in \Theta$ статистика X (векторная случайная величина) называется достаточной статистикой, если для любого события A существует вариант условной вероятности $P_\vartheta(A/X = x)$, не зависящий от ϑ.*

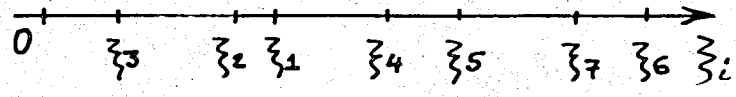

Рис. 1. Исходная выборка $\xi^{(n)}$. Для определенности все $\xi_i > 0$.

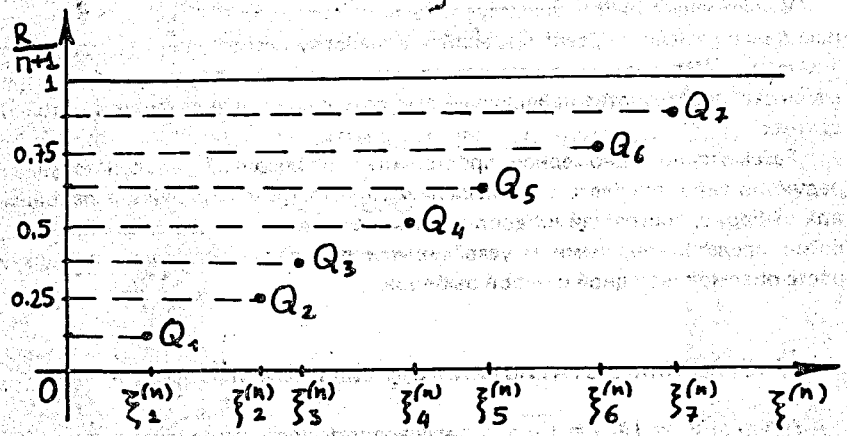

Рис. 2. Представление точек Q_i.

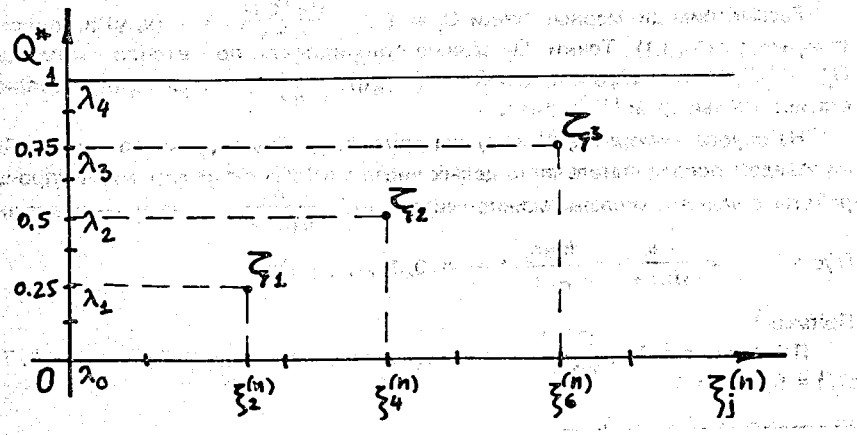

Рис. 3. Порядковые статистики из набора $\xi^{(n)}$, соответствующие числам λ_k, $k = \overline{0,4}$.

Определим числа λ_k, $k = \overline{0,4}$:

$\lambda_0 = 0*2/8 = 0$, $\lambda_1 = 1*2/8 = 0{,}25$, $\lambda_2 = 2*2/8 = 0{,}5$,

$\lambda_3 = 3*2/8 = 0{,}75$, $\lambda_4 = 4*2/8 = 1$ / рис. 3/.

2. ТЕОРЕМА О СХОДИМОСТИ

Введем величину $\tau_j^{(n)}$:

$$\tau_j^{(n)} = \sum_{k=1}^{s(n)+1} \lambda_{k-1} \; I\{ \frac{R_n(\xi_j)}{n+1} \in (\lambda_{k-1}, \lambda_k] \}. \qquad (1)$$

Пусть $F(x) = P\{\xi_j < x\}$ — функция распределения некоторой случайной величины ξ_j .

ТЕОРЕМА

Если $F(x)$ непрерывна и $s(n) \to \infty$, то

$$r_n(Z) = \frac{1}{n} \sum_{j=1}^{n} I\{\tau_j^{(n)} < Z\} \text{ при } Z \in (0,1) \qquad (2)$$

сходится к Z в среднем квадратичном.

ДОКАЗАТЕЛЬСТВО

Прежде всего отметим, что все $\tau_1, \ldots, \tau_n$ одинаково распределены совместно, и их распределение в классе непрерывных ф.р. $F(x)$ от самой $F(x)$ не зависит. Далее

$$r_n^2(Z) = \frac{1}{n} r_n(Z) + \frac{1}{n^2} \sum_{i \neq j} I \; \max(\tau_i^{(n)}, \tau_j^{(n)}) < Z \;, \qquad (3)$$

поскольку элементами суммы в $r_n(Z)$ являются индикаторы. Из (3) сразу получим, что

$$Er_n^2(Z) = \frac{1}{n} Er_n(Z) + \frac{1}{n^2} \sum_{i \neq j} P\{\max(\tau_1^{(n)}, \tau_2^{(n)}) < Z\},$$

но τ_i одинаково распределены, поэтому

$$Er_n(Z) = P\{\tau_1^{(n)} < Z\} \text{ и}$$

$$Er_n^2(Z) = \frac{1}{n} P\{\tau_1^{(n)} < Z\} + \frac{n-1}{n} P\{\max(\tau_1^{(n)}, \tau_2^{(n)}) < Z\}, \qquad (4)$$

Следовательно, осталось выяснить поведение двух величин:

$$\Delta_1^{(n)}(Z) = \Delta_1 = P\{\tau_1^{(n)}, \tau_2^{(n)}) < Z\},$$

$$\Delta_2^{(n)}(Z) = \Delta_2 = P\{\max(\tau_1^{(n)}, \tau_2^{(n)}) < Z\}. \quad (5)$$

Рассмотрим частный случай $Z = \lambda_k$, $1 \leqslant k \leqslant s(n)$. $\tau_1^{(n)}$ и $\tau_2^{(n)}$ могут при этом принимать значения лишь $\lambda_0, \lambda_1, \ldots, \lambda_{k-1}$, т.е. $\frac{R_n(\xi_1)}{n+1}$ и $\frac{R_n(\xi_2)}{n+1}$ должны принадлежать отрезку $(0, \lambda_k]$:

$$P\{\tau_1^{(n)} < \lambda_k\} = P\{\frac{R_n(\xi_1)}{n+1} \leqslant \lambda_k\} \text{ и,}$$

$$P\{\max(\tau_1^{(n)}, \tau_2^{(n)}) < \lambda_k\} = P\{\frac{\max(R_n(\xi_1), R_n(\xi_2))}{n+1} \leqslant \lambda_k\}, \quad (6)$$

но $\lambda_k(n+1) = \frac{k(n+1)}{s(n)+1} = k*m_n$, так что

$$\Delta_1^{(n)}(\lambda_k) = P\{R_n(\xi_1) \leqslant k*m_n\} = \frac{k*m_n}{n},$$
$$\Delta_2^{(n)}(\lambda_k) = P\{\max(R_n(\xi_1), R_1(\xi_2)) \leqslant k*m_n\} = \frac{k*m_n(k*m_n - 1)}{n(n-1)}, \quad (7)$$

так как $R_n(\xi_1)$ и $R_n(\xi_2)$ совместно принимают все возможные пары чисел i и j, $1 \leqslant i \neq j \leqslant n$, с вероятностью $\frac{1}{n(n-1)}$, ибо все перестановки для рангов $(R_n(\xi_1), \ldots, R_n(\xi_n))$ равновероятны.

$\Delta_1^{(n)}$ и $\Delta_2^{(n)}$ неубывающие функции по Z. Для $Z \in (\lambda_{k-1}, \lambda_k]$ приведем без вывода оценки для названных функций:

$$|\Delta_1^{(n)}(Z) - Z| \leqslant \frac{1}{n} + \frac{1}{s(n)+1}, \quad (8)$$

$$|\Delta_2^{(n)}(Z) - Z^2| \leqslant 3*(\frac{1}{n} + \frac{1}{s(n)+1}). \quad (9)$$

Неравенства (8) и (9) дают возможность оценить математическое ожидание и дисперсию величины $r_n(Z)$ (приведем без вывода):

$$|Er_n(Z) - Z| \leqslant \frac{2}{s(n)+1}, \quad (10)$$

$$Dr_n(Z) \leqslant \frac{const}{s(n)+1}. \quad (11)$$

при $n \to \infty$ верхние границы обоих неравенств стремятся к нулю, следовательно, $r_n(Z)$ сходится к Z на (0,1) в среднем квадратичном.

3. ВЫВОДЫ

Из теоремы в п. 2 следует, что:

1. Величины $\tau_j^{(n)}$ асимптотически распределены равномерно на [0,1].
2. Величины $\tau_j^{(n)}$ равномерно группируют выборку $\xi^{(n)}$ на шкале вероятностей.

3. Информация о F(x) остается в порядковых статистиках $\xi^{(n)}$, $j = \overline{1,n}$, среди которых конечное число с номерами $j_n = j_n(k) = k*m_n$ являются $\frac{1}{s(n)+1}$ – достаточной статистикой [2].
4. Выборка $\xi^{(n)}$ может быть устойчиво равнометризована, т.е. возможен переход к новой представительной выборке $\zeta^{(n)}$ гораздо меньшего объема.

Пример 2.

Получим величины $\tau_j^{(7)} \equiv \tau_j$:

$$\tau_j = \sum \lambda_{k-1} * I\{\frac{R_7(\xi_j)}{8} \in (\lambda_{k-1}, \lambda_k]\}.$$

По приведенной ранее выборке $\xi^{(7)}$ /рис. 2./:

$R_7(\xi_1) = 3$, $R_7(\xi_2) = 2$,

$R_7(\xi_3) = 1$, $R_7(\xi_4) = 4$,

$R_7(\xi_5) = 5$, $R_7(\xi_6) = 7$, $R_7(\xi_7) = 6$.

$\tau_1 = 0*I\{3/8 \in (0,1/4]\} + 1/4*I\{3/8 \in (1/4, 1/2]\} +$

$+1/2*I\{3/8 \in (1/2, 3/4]\} + 3/4*I\{3/8 \in (3/4, 1]\} = 1,4$, т.е. $\tau_1 = 1/4$.

Аналогично вычисляются:

$\tau_2 = 0$, $\tau_3 = 0$, $\tau_4 = 1/4$,

$\tau_5 = 1/2$, $\tau_6 = 3/4$, $\tau_7 = 1/2$.

Положим $\epsilon = \frac{1}{s(n)+1} = \frac{1}{3+1} = 1/4$.

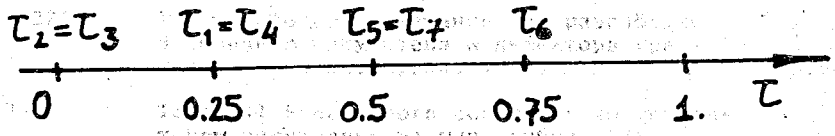

Рис. 4. Значения величин $\tau_j^{(n)}$, n = 7.

Тогда в наборе статистики $\xi^{(7)}$ $\xi_7(2)$, $\xi_7(4)$, $\xi_7(6)$ являются ϵ-достаточными статистиками, и эта новая выборка $\zeta^{(7)}$:

$$\zeta^{(7)} = \{\zeta_1 = \xi_n(k*m_n), i = \overline{1, s(n)}\}$$

может быть успешно использована в дальнейшем /рис. 4/.

Автор выражает искреннюю признательность Е.П.Жидкову за поддержку работы.

ЛИТЕРАТУРА

1. *Математическая энциклопедия. М.: Наука, 1979, Т.2, с.375.*
2. *Сираждинов С.Х. є-достаточность конечных наборов порядковых статистик. — Прага, Труды Международной школы-конференции по теории вероятности и математической статистике. 1982.*

Рукопись поступила в издательский отдел
30 октября 1991 года.