

Г.И.Симонова  
 (Московский университет)

ОЦЕНИВАНИЕ ПАРАМЕТРА СДВИГА С ПОМОЩЬЮ  
 ПОРЯДКОВЫХ СТАТИСТИК

В задачах оценивания неизвестного параметра интересны методы, которые позволяют оценить параметр по результатам наблюдений из распределения неизвестного вида, принадлежащего некоторому классу распределений. В настоящей работе рассматривается такой подход применительно к оцениванию параметра сдвига симметричных распределений, удовлетворяющих некоторым условиям регулярности.

I. Некоторые методы оценивания центра распределения линейными комбинациями порядковых статистик

Рассмотрим последовательность независимых, одинаково распределенных случайных величин  $X_1, X_2, \dots, X_n$ , имеющих функцию распределения  $F(X-\theta)$ , где  $F(X)$  - симметричная, т.е.  $F(X) = 1 - F(-X)$  и  $\theta$  - неизвестный параметр.

Пусть  $X_{(1)} < X_{(2)} < \dots < X_{(n)}$  - вариационный ряд.

В качестве оценки неизвестного  $\theta$  рассмотрим линейную функцию порядковых статистик  $T_n(X_n) = \sum_{i=1}^n \lambda_i X_{(i)}$ . Будем искать лишь оценки, удовлетворяющие условию  $\sum_{i=1}^n \lambda_i = 1$ . Известно (см., например, [6]), что наилучшей линейной, несмещенной оценкой  $\theta$  при известной  $F$  является  $\theta_n^* = \sum_{i=1}^n \lambda_i^* X_{(i)}$ , где  $\lambda_i^* = \frac{(B^{-1}e)_i}{e^T B^{-1}e}$ ,  $e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ ,  $B$  - ковариационная матрица  $X = \|X_{(i)}\|$ , зависящая от функции распределения  $F$ . Для того чтобы использовать эту оценку, нужно знать матрицу  $B$ . Интересные оценки в случае неизвестной  $F$ , удовлетворяющей некоторым условиям регулярности, предлагает Такеучи [2]. Он оценивает ковариационную матрицу порядковых статистик порядка  $k$  по выборке объема "n" ( $n \gg k$ ). В качестве оценки  $\theta$  он берет  $\sum_{\alpha=1}^k \hat{c}_\alpha M\{Y_{(\alpha)} | X_{(1)} < \dots < X_{(n)}\}$ , где  $Y_{(1)} < Y_{(2)} < \dots < Y_{(n)}$  случайная выборка размера "k" из  $X_{(1)} < \dots < X_{(n)}$ , а  $\hat{c}_\alpha = (\hat{B}^{-1}e_x)_\alpha / e_x^T \hat{B}^{-1}e_x$ ,  $e_x = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_k$ ,  $\hat{B} = \|\hat{b}_{\alpha\beta} = \text{cov}(Y_{(\alpha)}, Y_{(\beta)}) | X_{(1)} < \dots < X_{(n)}\|$  - состоятельные оценки

$\theta_{\alpha\beta}$ ;  $\alpha, \beta = 1, 2, \dots, k$  и доказывает, что дисперсия таких оценок подходящим выбором „ $x$ “ может быть как угодно приближена к границе Рао-Крамера для любой функции из рассмотренного им класса распределений.

Другой подход, основанный на идеи оценивания „ $x$ “ коэффициентов линейной комбинации порядковых статистик по „ $n$ “ наблюдениям  $X_1, X_2, \dots, X_n$  из неизвестного распределения  $F$ , удовлетворяющего некоторым условиям регулярности, осуществлен Джонсоном [1]. Джонсон рассматривает линейные комбинации  $\theta_{xn} = \frac{1}{n} \sum_{i=1}^n c_i S_{in}$ ,

где  $S_{in} = \sum_{X_j \leq x_i} (X_{(j)} + X_{(n-j+1)})$ ,  $\sum_{i=1}^k c_i p_i = \frac{1}{2}$ , а  $\Delta_i$  – интервалы, порожденные разбиением  $[0, 1/2]$  на „ $x+1$ “ отрезков с соответствующими длинами  $p_0, p_1, \dots, p_k$ , причем в эту статистику  $[p_0]$  крайних членов вариационного ряда не включаются. Решив задачу  $\min \mathcal{D}(\sqrt{n} \theta_{xn})$  и заменив найденные значения  $c_i$  состоятельными оценками  $\hat{c}_i$ , Джонсон доказывает теорему [1].

Для любого  $\epsilon > 0$  существуют  $x_\epsilon$  и  $p_\epsilon$ , такие, что при  $p_0 = p_\epsilon$  для любых  $x > x_\epsilon$  и  $F \in \mathcal{F}$   $\sqrt{n}(\hat{\theta}_{xn} - \theta) \sim N(0, \sigma_x^2(F))$ , где

$\sigma_x^2(F) \leq \frac{1}{I(F)} + \epsilon$ . Исследуем поведение коэффициентов наилучших оценок  $\lambda_i$  как функций  $\frac{i}{n}$ , а затем аппроксимируем весовую функцию  $\lambda(t)$  функциями некоторого класса. При таком подходе можно исследовать связь между свойствами функций распределения и скоростью сходимости (при „ $x \rightarrow \infty$ “) асимптотической дисперсии к границе Рао-Крамера.

Теперь предположим, что коэффициенты в линейной комбинации порядковых статистик определяются посредством функции  $\lambda(t), t \in [0, 1]$  и рассмотрим  $S_n(\lambda) = \frac{1}{n} \sum_{i=1}^n \lambda\left(\frac{i}{n+1}\right) X_{(i)}$ , где  $\int_0^1 \lambda(t) dt = 1$ . Условия, при которых такая линейная комбинация асимптотически нормальна, исследовались многими авторами [4], [5]. Воспользуемся теоремой 2 [5].

Пусть  $M X_i^2 < \infty$ ,  $\lambda(t)$  – ограниченная и непрерывная функция. Тогда  $\frac{S_n - M S_n}{\sigma(S_n)} \xrightarrow[n \rightarrow \infty]{d} N(0, \sigma^2(\lambda, F))$ ,

где  $\sigma^2(\lambda, F) = \int_0^1 \lambda(F(x)) \lambda(F(y)) [F(\min(x, y)) - F(x)F(y)] dx dy$ .

Условия этой теоремы достаточны для того, чтобы  $M S_n \xrightarrow[n \rightarrow \infty]{d} \int_0^1 \lambda(t) F^{-1}(t) dt$ .

Рассмотрим следующий класс распределений  $\mathcal{F}$ :  $F \in \mathcal{F}$ , если:  
 1. существует плотность  $f(x)$ ,  $f(-x) = f(x)$ ,  $f(x) > 0$ ,  $x \in R$ ;  
 2.  $f(x) \in C^3(R)$ ;

3.  $f$  и все её производные равномерно ограничены;  
 4.  $I(F) = \int \psi_f^2(x) f(x) dx < \infty$ , где  $\psi_f(x) = -f'(x)/f(x)$ ;  
 5.  $\int x^2 f(x) dx < \infty$ .

Для  $F \in \mathcal{F}$  положим  $\lambda(t) = \frac{\psi_f'(F^{-1}(t))}{I(F)}$ . При таком определении функции  $\lambda(t)$  и класса  $\mathcal{F}$  выполнены условия теоремы 2 [5], а значит  $\sqrt{n} [S_n(\lambda) - \int_0^1 \lambda(t) F^{-1}(t) dt] \xrightarrow[n \rightarrow \infty]{\text{ac}} N(0, \frac{1}{I(F)})$ , т.е. оценка асимптотически эффективна для известной  $F$ .

## 2. Оценивание функции $\lambda(t)$ для классов распределений $\mathcal{F}$ и $\mathcal{F}'$

В случае, если  $F$  неизвестна, но  $F \in \mathcal{F}$ , весовую функцию можно оценить. Обозначим  $F^{-1}(t) = G(t)$ . Тогда  $G'(t) = \frac{1}{f(F^{-1}(t))}$ ,  $G''(t) = -\frac{f'(F^{-1}(t))}{f^2(F^{-1}(t))}$ ,  $G'''(t) = 3 \frac{f'^2(F^{-1}(t))}{f^5(F^{-1}(t))} - \frac{f''(F^{-1}(t))}{f^4(F^{-1}(t))}$ . Поскольку  $\psi_f'(F^{-1}(t)) = \left[ \frac{f'(F^{-1}(t))}{f(F^{-1}(t))} \right]^2 - \frac{f''(F^{-1}(t))}{f(F^{-1}(t))}$ , то  $\psi_f'(F^{-1}(t)) = \frac{G'''(t)}{G'^3(t)} - 2 \frac{G''(t)}{G'^4(t)}$ . Обозначим  $\mu'(F^{-1}(t)) = \mu(t)$ . Тогда

$\lambda(t) = \mu(t) / \int_0^t \mu(s) ds$ . Предположение о существовании третьей непрерывной производной у функции плотности  $f$  и условие  $f(x) > 0$  обеспечивают существование непрерывной  $G''(t)$ ,  $t \in (0, 1)$ , а следовательно, и непрерывную дифференцируемость  $\mu(t)$ ,  $t \in (0, 1)$ . Апроксимируем функцию  $\mu(t)$  ступенчатыми  $\mu_A(t)$  на  $[\varepsilon, 1-\varepsilon]$ . Выбор  $\varepsilon$  будет указан ниже.

$$\mu_A(t) = \sum_{i=1}^{n-1} \mu(t_i) \chi_{[t_i, t_{i+1})}(t), \quad \text{где } \chi_A(t) = \begin{cases} 1, & t \in A \\ 0, & t \notin A \end{cases}, \quad \text{а } t_0 < t_1 < \dots < t_n -$$

точки разбиения  $[\varepsilon, 1-\varepsilon]$  на  $n$  равных частей,  $t_i = t_0 + ih$ ,  $h = \frac{1-2\varepsilon}{n}$ . Таким образом, для того чтобы построить оценку, нужно знать  $\mu(t_i)$ . Для распределений класса  $\mathcal{F}$  выразим  $\mu(t_i)$  с точностью до  $O(h)$  через величины, которые могут быть легко оценены по выборке  $X_1, \dots, X_n$ . Разложив функцию  $G(t)$  в ряд Тейлора в четырех соседних точках:

$t_i-h, t_i, t_i+h, t_i+2h$  ( $t_i \neq t_0, t_n$ ) (так как  $F(x)$  — симметричная, то все рассуждения достаточно проводить для  $[0, 1/2]$ ). Поскольку  $G(t) \in C^4(0, 1)$ , то справедливо разложение:

$$r_{i-1} = G(t_i) - G(t_i-h) = G'_i h - G''_i h^2/2 + G'''_i h^3/3! + O(h^4),$$

$$r_i = G(t_i+h) - G(t_i) = G'_i h + G''_i h^2/2 + G'''_i h^3/3! + O(h^4),$$

$$r_{i+1} = G(t_i + 2h) - G(t_i + h) = G'_i h + G''_i \frac{3}{2} h^2 + G'''_i \frac{7}{3!} h^3 + O(h),$$

где  $G_i^{(j)} = G(t_i)$ ,  $j=1,2,3$ , причем  $\frac{\partial f(h^4)}{\partial t^j}$  общая функция для любых  $F \in \mathcal{F}$ . Чтобы найти  $\mu(t_i) = \frac{G'_i}{G'^3} - 2 \frac{G''_i}{G'^4}$  с точностью до  $O(h)$ , достаточно найти  $G''_i$  из этой системы линейных уравнений,  $G'_i$  – из укороченной системы двух уравнений, а за  $G'_i$  взять любое из выражений:  $G'_i = \frac{r_{i+1}}{h} + O_1(h) = \frac{r_i}{h} + O_2(h) = \frac{r_{i+1}}{h} + O_3(h)$ .

Решая указанные системы линейных уравнений, получим

$$G''_i = \frac{r_{i+1} + r_{i-1} - 2r_i}{h^3} + O(h); \quad G'_i = \frac{r_i - r_{i-1}}{h^2} + O(h)$$

$$\text{и, следовательно, } \mu(t_i) = \frac{r_{i+1} + r_{i-1} - 2r_i}{r_i r_{i-1} r_{i+1}} - 2 \left( \frac{r_i - r_{i-1}}{r_i r_{i-1}} \right)^2 + O(h) = \\ = \frac{1}{r_i} \cdot \frac{1}{r_{i-1}} + \frac{1}{r_i} \cdot \frac{1}{r_{i+1}} - 2 \frac{1}{r_{i+1}} \cdot \frac{1}{r_{i-1}} - 2 \left( \frac{1}{r_{i-1}} - \frac{1}{r_i} \right)^2 + O(h).$$

Подставив в эти выражения вместо  $r_i$  их оценки  $\hat{r}_i = X_{(\lfloor nt_{i+1} \rfloor + 1)} - X_{(\lfloor nt_i \rfloor + 1)}$ , получим оценки  $\hat{\mu}(t_i)$ , которые с точностью до  $O(h)$  совпадают с оценками Джонса:  $\hat{\sigma}_i = \frac{1}{\hat{r}_i} \left( \frac{2}{\hat{r}_i} - \frac{1}{\hat{r}_{i-1}} - \frac{1}{\hat{r}_{i+1}} \right) + \varphi h$ . Таким образом, заменив функцию  $\mu(t)$  на  $[\varepsilon, 1 - \varepsilon]$  ступенчатыми  $\mu_k(t)$ , получим оценку Джонса, а именно:  $\hat{\theta}_k = \frac{1/n \sum_{i=1}^n \hat{\mu}(t_i) (\sum_{j=1}^{k-1} X_{(j)})}{\sum_{i=1}^n \hat{\mu}(t_i) h}$ , и так как  $\sum_{i=1}^n \mu(t_i) h$  оценивает  $I(F)$ , то скорость стремления дисперсии этой оценки к  $\frac{1}{I(F)}$  для любых  $F \in \mathcal{F}$  имеет порядок  $O(\frac{1}{k})$ , если выбирать  $\varepsilon = \frac{1}{k}$ .

Для более гладкой функции  $\lambda(t)$  можно улучшить скорость сходимости дисперсии к  $\frac{1}{I(F)}$  за счет того, что для таких функций  $\int f(S) ds$  может быть оценен точнее, и значения  $\mu(t_i)$  вычисляются с большей точностью.

- Найдем, например, наилучшие оценки для следующего класса распределений  $\mathcal{F}'$ :  $F \in \mathcal{F}'$ , если
1. существует плотность  $f(x) = F'(x)$ ,  $f(-x) = f(x)$ ,  $f(x) > 0$ ,  $x \in \mathbb{R}$ ;
  2.  $f(x) \in C^4(\mathbb{R})$ ;
  3.  $f'(x)$  и все её производные равномерно ограничены;
  4.  $I(F) < \infty$ ;
  5.  $\int x^2 f(x) dx < \infty$ .

В этом случае можно обеспечить скорость сходимости асимптотической дисперсии к границе Рао-Крамера порядка  $O(\frac{1}{k^2})$  для  $\forall F \in \mathcal{F}'$ .

В самом деле, разобьем  $[\varepsilon, 1-\varepsilon]$  на  $"k"$  равных частей и на каждом участке заменим  $\mu(t)$  линейной функцией.

Введем  $\mu_k^*(t) = \sum_{i=1}^{k-1} \left\{ \mu(t_i) + \frac{\mu(t_i+h)-\mu(t_i)}{h}(t-t_i) \right\} X_{[t_i, t_{i+1}]}(t)$ ,  $\lambda_k^*(t) = \frac{\mu_k^*(t)}{\int_{\varepsilon}^{1-\varepsilon} \mu_k^*(s) ds}$ ,

тогда  $\int_{\varepsilon}^{1-\varepsilon} \mu(s) ds = \sum_{i=1}^{k-1} \frac{\mu(t_i) + \mu(t_i+h)}{2} h + O(h^2)$  (формула трапеций для  $\mu(t) \in C^2$ ).

Выбрав  $\varepsilon = \frac{1}{k^2}$  и вычислив  $\mu(t_i)$  с точностью до  $O(h^2)$ , обеспечим нужную скорость сходимости. Вычислить  $\mu(t_i)$  с точностью до  $O(h^2)$  можно аналогично предыдущему случаю. Разложив  $G(t)$  в ряд Тейлора в пяти соседних точках и решая соответствующие системы линейных уравнений, получим  $G_i'' = \frac{r_{i-2} - r_{i-1} - r_i + r_{i+1}}{2h^3} + O(h^2)$ ;

$$G_i'' = \frac{r_i - r_{i-1}}{h^2} + O(h^2); \quad G_i' = \frac{r_i + r_{i+1}}{2h} + O(h^2)$$

и тогда

$$\mu(t_i) = 4 \frac{r_{i-2} - r_{i-1} - r_i + r_{i+1}}{(r_i + r_{i+1})^3} - 32 \frac{(r_i - r_{i-1})^2}{(r_i + r_{i+1})^4} + O(h^2) \quad (*)$$

и

$$\theta_{kn} = \left\{ \frac{1}{n} \sum_{i=1}^{k-1} \mu(t_i) \left( \sum_{j \neq i} X_{(j)} \right) + \frac{1}{n} \sum_{i=1}^{k-1} \frac{\mu(t_i+h) - \mu(t_i)}{h} \left[ \sum_{j \neq i} \left( \frac{j}{n+1} - t_i \right) X_{(j)} \right] \right\} / \left\{ \sum_{i=1}^{k-1} \frac{\mu(t_i) + \mu(t_i+h)}{2} \right\}$$

где  $\mu(t_i)$  находится из  $(*)$ .

Подставляя в  $(*)$  вместо  $r_i$  их оценки, которые являются состоятельными при фиксированном  $"k"$  и  $"n" \rightarrow \infty$ , получим состоятельные оценки  $\hat{\mu}(t_i)$ , так как  $\mu(t)$  - непрерывная, ограниченная функция.  $\theta_{kn}$ , в которой вместо  $\mu(t_i)$  стоят  $\hat{\mu}(t_i)$ , обозначим через  $\hat{\theta}_{kn}$ .

$$\text{Тогда } P\{\sqrt{n}|\hat{\theta}_{kn} - \theta_{kn}| > \delta\} = P\{\sqrt{n} \frac{1}{n} \left| \sum_{i=1}^n [\hat{\lambda}_k^*(\frac{i}{n+1}) - \lambda_k^*(\frac{i}{n+1})] X_{(i)} \right| > \delta\} \leq$$

$$\leq P\left\{ \max_{i \neq 0, k} |\hat{\lambda}(t_i) - \lambda(t_i)| \cdot \frac{\sum X_i}{\sqrt{n}} > \delta \right\} \xrightarrow[n \rightarrow \infty]{} 0, \text{ поскольку}$$

$\hat{\lambda}(t_i) \xrightarrow[n \rightarrow \infty]{P} \lambda(t_i)$ ,  $i \neq 0, k$ , а  $\frac{\sum X_i}{\sqrt{n}}$  асимптотически нормальна

и, следовательно, ограничена по вероятности. А значит,  $\theta_{kn}$  можно заменить оценкой  $\hat{\theta}_{kn}$ , которая имеет при  $n \rightarrow \infty$  тот же предельный закон распределения, что и  $\theta_{kn}$ , т.е.

$$\sqrt{n}(\hat{\theta}_{kn} - \theta) \xrightarrow[n \rightarrow \infty]{D} N\left(0, \frac{1}{I(F)} + O(\frac{1}{K^2})\right) \quad \text{для } F \in \mathcal{F}'.$$

Автор благодарит Ю.Н.Тюрина за обсуждение результатов и ценные замечания.

### Л и т е р а т у р а

- 1.Johns M.V.Jr. Nonparametric Estimation of Location. American Statistical Association (JASA), 1974, v.69, 453-460.
- 2.Takeuchi K. A Uniformly Asymptotically Efficient Estimator of a Location Parameter.(JASA), 1971, v.66, 292-302.
- 3.Huber P.J. Robust Statistics: a Review. Annals of Mathematical Statistics. 1972, v.43, 1041-1067.
- 4.Chernoff, Gastwirth and Johns. Annals of Mathematical Statistics. 1967, v.38, 52-72.
- 5.Stigler, Stephen M. Annals of Statistics. 1974, v.2, 676-693.
- 6.К е н д а л л М.Дж., С т ь в а р т А. Статистические вы-  
воды и связи. М., Наука, 1973.