ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2016 Управление, вычислительная техника и информатика № 4 (37)

УДК 519.24
DOI: 10.17223/19988605/37/8

В.П. Шуленин

АСИМТОТИЧЕСКИЕ СВОЙСТВА МОДИФИЦИРОВАННЫХ СРЕДНИХ РАЗНОСТЕЙ ДЖИНИ

Изучаются свойства робастных оценок масштабного параметра. Показано, что модифицированная оценка средних разностей Джини имеет асимтотически нормальное распределение, является B-робастной оценкой и имеет ограниченную функцию влияния. Приводятся результаты сравнения оценок масштабного параметра в рамках гауссовской модели с засорением. Рассматривается адаптивный вариант предложенной оценки масштабного параметра.

Ключевые слова: масштабный параметр; робастные оценки; функция влияния; средняя разность Джини; U-статистики; адаптивные оценки.

В класс *U*-статистик Хёфдинга [1] входят многие конкретные оценки параметров, представляющие практический интерес. Обобщение класса *U*-статистик, описанное в работе [2], связанное с построением робастных оценок с ограниченными функциями влияния, приводит к рассмотрению *U*-статистик, основанных на урезанных выборках, что позволяет изучать многие известные в теории робастности оценки параметров с единых позиций и открывает широкие возможности для построения новых оценок. Например, выборочное α-урезанное среднее $\overline{X}_\alpha$, $0 \le \alpha < 1/2$, обычно применяемое в качестве робастной оценки параметра положения случайной величины (с.в.) X с функцией распределения (ф.р.) F, традиционно рассматривалось и изучалось как представитель семейства L-оценок в виде линейных комбинаций порядковых статистик $X_{(1)},...,X_{(n)}$ исходной выборки $X_1,...,X_n$. Доказанная ранее асимптотическая нормальность $\overline{X}_\alpha$-оценок (см., например [3, 4]) также непосредственно следует и из результатов работы [2] (см. пример 8.6.15 в [6]). В качестве другого примера приведем выборочную оценку средних разностей Джини, которая используется в качестве оценки масштабного параметра, характеризующего степень разброса с.в. X, и записывается в виде $\hat{\Delta}_0 = [n(n-1)]^{-1}\Sigma |X_i - X_j|$. Эта оценка имеет асимптотически нормальное распределение [5, 6], её асимптотическая относительная эффективность по отношению к традиционно применяемой на практике оценке $\hat{S}_1(0) = \{n^{-1}\Sigma(X_i - \overline{X})^2\}^{1/2}$ стандартного отклонения очень высокая, в частности, при нормальном распределении Ф она равна $AO\Theta_\Phi(\hat{\Delta}_0 : \hat{S}_1(0)) = 0{,}98$. И хотя $\hat{\Delta}_0$-оценка подвержена меньшему влиянию выбросов в выборке, чем $\hat{S}_1(0)$-оценка, обе они имеют *неограниченные* функции влияния Хампеля [4]. В качестве ещё одного примера оценок масштабного параметра приведём семейство интер-α-квантильных размахов $\hat{S}_3(\alpha) = [X_{(n-[\alpha n])} - X_{([\alpha n])}]/2$, $0 < \alpha < 1/2$ [5, 6]. Эти оценки имеют ограниченные функции влияния, но их асимптотические эффективности по отношению к оценке стандартного отклонения при нормальном распределении Ф очень низкие. Например, для оценки интерквартильного размаха $\hat{S}_3(0{,}25)$ асимптотическая относительная эффективность равна $AO\Theta_\Phi(\hat{S}_3(0{,}25) : \hat{S}_1(0)) = 0{,}37$. По этой причине в литературе (см., например, [5, 6]) рассматривают различные модификации оценок масштабного параметра с целью обеспечить и ограниченность функции влияния, и высокую эффективность по отношению к $\hat{S}_1(0)$-оценке при нормальном распределении Ф. К таким оценкам относится рассматриваемая в данной работе модифицированная оценка средних разностей Джини $\hat{\Delta}_\alpha$, $0 \le \alpha < 1/2$, определённая в (24). Эта оценка является

U-статистикой, основанной на урезанной выборке и её асимптотические свойства изучаются с использованием результатов работы [2]. В данном исследовании приводятся результаты сравнения $\hat{\Delta}_\alpha$-оценок с другими оценками масштабного параметра в рамках гауссовской модели с засорением и предлагается адаптивный вариант $\hat{\Delta}_\alpha$-оценок, для которых параметр α выбирается на основе информации, содержащейся в исходной выборке с использованием выборочной оценки функционала, характеризующего степень «затянутости хвостов» функции распределения *F* изучаемой случайной величины *X*.

1. U-статистики, основанные на урезанных выборках ($U_{\alpha\beta}$-оценки)

Пусть $X_1,...,X_n$ – последовательность нормальных оценок распределения (н.о.р.) случайных величин с ф.р. $F(x)$ и плотностью $f(x)$, $x \in R^1$, $X_{(1)},...,X_{(n)}$ – упорядоченная статистика исходной выборки $X_1,...,X_n$ и $X_{([\alpha n]+1)},...,X_{(n-[\beta n])}$, обозначает αβ-урезанную упорядоченную статистику выборки, где α и β – заданные пропорции урезания выборки, причем $0 \le \alpha, \beta \le 1/2$. Обозначим $n_{\alpha\beta} = n - [\alpha n] - [\beta n]$. Пусть задано «ядро» $h(X_1,...,X_m)$, $m < n$, которое является симметричной функцией своих аргументов. Множество *m*-наборов индексов $(i_i,...,i_m)$, удовлетворяющих условию $\{[\alpha n]+1 \le i_1 <,...,< i_m \le n-[\beta n]\}$, обозначим через $C_{\alpha\beta}$, т.е.

$$C_{\alpha\beta} = \{(i_1,...,i_m) : [\alpha n]+1 \le i_1 <,...,< i_m \le n-[\beta n]\}. \quad (1)$$

Следуя работе [2], рассмотрим *U*-статистики, основанные на αβ-урезанных выборках, которые определяются в виде

$$U_{n,\alpha\beta} = \binom{n_{\alpha\beta}}{m}^{-1} \sum_{C_{\alpha\beta}} h(X_{(i_1)},...,X_{(i_m)}). \quad (2)$$

Ниже рассмотрим случай $\alpha = \beta$ и переобозначим множество индексов $C_{\alpha\beta}$ из (1) через

$$C_\alpha = \{(i_1,...,i_m) : [\alpha n]+1 \le i_1 <,...,< i_m \le n-[\alpha n]\}. \quad (3)$$

Далее введем следующие обозначения. Пусть $n_\alpha = n - 2[\alpha n]$, и определим ограниченную функцию $g_\alpha(x;F)$, $0 \le \alpha < 1/2$, $x \in R^1$, в виде

$$g_\alpha(x;F) = \frac{I[F^{-1}(\alpha) \le x \le F^{-1}(1-\alpha)]}{(1-2\alpha)^m} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \cdots \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} h(x_1,...,x_{m-1},x) \prod_{i=1}^{m-1} dF(x_i), \quad (4)$$

где $I[A]$ – индикатор события *A*. Далее пусть $\xi_\alpha = F^{-1}(\alpha)$, $\overline{\xi}_\alpha = F^{-1}(1-\alpha)$ и значения функции $g_\alpha(x;F)$ в точках ξ_α и $\overline{\xi}_\alpha$ обозначим соответственно в виде

$$A_\alpha = -g_\alpha(\xi_\alpha, F),\ B_\alpha = g_\alpha(\overline{\xi}_\alpha, F). \quad (5)$$

Среднее значение и дисперсию $g_\alpha(X;F)$ соответственно обозначим в виде

$$U_\alpha(F) = M_F\{g_\alpha(X,F)\} = \frac{1}{(1-2\alpha)^m} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \cdots \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} h(x_1,...,x_m) \prod_{i=1}^{m} dF(x_i) \quad (6)$$

и

$$\Delta_\alpha(F) = D_F\{g_\alpha(X,F)\} = \int_{\xi_\alpha}^{\overline{\xi}_\alpha} g_\alpha^2(x,F) dF(x) - U_\alpha^2(F). \quad (7)$$

Теорема 1. Предположим, что ф.р. F имеет плотность f, которая непрерывна и ограничена в точках ξ_α и $\overline{\xi}_\alpha$. Пусть функция $g_\alpha(x;F)$ непрерывна в точках ξ_α, $\overline{\xi}_\alpha$, и предположим, что для некоторых $a < \xi_\alpha$ и $b > \overline{\xi}_\alpha$ выполняется условие

$$\sup_{a\le x_1,\ldots,x_m\le b} |\, h(x_1,\ldots,x_m)\,| = M_0 < \infty\,. \tag{8}$$

Тогда $\sqrt{n}U_{n,\alpha}$-статистика имеет асимптотически нормальное распределение, т.е. выполняется выражение, согласно которому *закон* (L – Law) распределения вероятностей случайной величины в фигурных скобках является стандартным нормальным

$$L\{\sqrt{n}\,[U_{n\alpha} - U_\alpha(F)]\,/\,\sigma(U_\alpha, F)\} = N(0,1) \text{ при } n\to\infty, \tag{9}$$

где $\sigma^2(U_\alpha, F)$ – асимптотическая дисперсия $\sqrt{n}U_{n,\alpha}$-статистики, вычисляемая по формуле

$$\sigma^2(U_\alpha, F) = m^2\{\Delta_\alpha(F) + \alpha(1-\alpha)(A_\alpha^2 + B_\alpha^2) + 2\alpha U_\alpha(F)(A_\alpha - B_\alpha) + 2\alpha^2 A_\alpha B_\alpha\}\,. \tag{10}$$

Доказательство. Схема доказательства этой теоремы основана на использовании результатов работы [2], в которой отмечается, что при выполнении предположений для плотности *f*, функции $g_\alpha(x;F)$ и ядра $h(x_1,\ldots,x_m)$ условия (8) применимы леммы (2.1–2.4) этой работы и при этом выполняется выражение вида

$$U_{n,\alpha} = U_\alpha(F) + n^{-1}\sum_{i=1}^{n} IF(X_i; F, U_\alpha) + o_p(n^{-1/2})\,, \tag{11}$$

в котором $IF(x;F,U_\alpha)$ обозначает функцию влияния Хампеля [4] оценки $U_\alpha(F_n)$ функционала $U_\alpha(F)$, определенного формулой (6). Из выражения (11) следует асимптотическая нормальность $\sqrt{n}U_{n,\alpha}$-статистик с использованием теоремы Слуцкого и центральной предельной теоремы с учетом, что элементы выборки $X_1,\ldots,X_n$ являются последовательностью н.о.р. случайных величин [5, 6]. Справедливость формулы (10) проверяется непосредственно, путем вычисления функции влияния $IF(x;F,U_\alpha)$ с учетом, что

$$\int IF(x;F,U_\alpha)\,dF(x) = 0 \text{ и } \sigma^2(U_\alpha, F) = \int IF^2(x;F,U_\alpha)\,dF(x)\,. \tag{12}$$

Стандартным способом убеждаемся (см.: [6. С. 195]), что функция влияния $IF(x;F,U_\alpha)$ имеет вид

$$IF(x;F,U_\alpha) = m\{g_\alpha(x,F) - U_\alpha(F) + A_\alpha(\alpha - I[x\le\xi_\alpha]) + B_\alpha(1-\alpha-I[x\le\overline{\xi}_\alpha])\}\,. \tag{13}$$

Отметим важное обстоятельство. Учитывая, что функция $g_\alpha(x;F)$ ограничена, функция влияния $IF(x;F,U_\alpha)$ вида (13) также является ограниченной функцией и, следовательно, $U_{n,\alpha}$-статистики являются B-робастными [4] и подвержены лишь ограниченному влиянию выбросов в выборке. Используя приведённое выражение (13) и вторую формулу в (12), после несложных преобразований получаем (10). Теорема доказана.

2. Средняя разность Джини и её модифицированный вариант

Обсуждение средней разности Джини как меры разброса случайных величин и её связь с кривой Лоренца приводятся в [7. С. 75]. Средняя разность Джини, как и медиана абсолютных разностей, относится ко второй группе функционалов, определяющих масштабный параметр (см. формулу (10.1.3) в [6]). В работах [5–7] функционал $T(F)$, определяющий среднюю разность Джини, записывают в разных вариантах. Обычно [7] его записывают в виде

$$\Delta_F = T(F) = \int\int |\, x - y\,|\, dF(x)\,dF(y)\,. \tag{14}$$

Другая форма записи функционала $T(F)$ [5, 6] имеет вид

$$\Delta_F = T(F) = \int_0^1 F^{-1}(t)(4t-2)\,dt = \int_0^1 F^{-1}(t)J(t)\,dt\,,\ J(t) = 4t-2\,,\ 0\le t\le 1\,. \tag{15}$$

Выборочная оценка $\tilde{\Delta}_0 = T(F_n)$ средней разности Джини Δ_F вида (14), построенная по выборке $X_1,\ldots,X_n$ методом подстановки, записывается в виде

$$\tilde{\Delta}_0 = T(F_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |\, X_i - X_j\,|\,. \tag{16}$$

Если же при построении оценки не используются совпадающие индексы (i, j), то оценку записывают в асимптотически эквивалентном варианте вида

$$\hat{\Delta}_0 = \frac{1}{n(n-1)} \sum_{i \neq j=1}^{n} | X_i - X_j | . \tag{17}$$

Отметим, что представление функционала $T(F)$ в виде (15) позволяет записать оценку в виде линейной комбинации порядковых статистик $X_{(1)},...,X_{(n)}$ исходной выборки $X_1,...,X_n$, т.е. в виде

$$\hat{\Delta} = \sum_{i=1}^{n} a_{ni} X_{(i)} , \tag{18}$$

где весовые коэффициенты a_{ni} вычисляются по формуле

$$a_{ni} = \int_{(i-1)/n}^{i/n} J(t)dt = \int_{(i-1)/n}^{i/n} (4t-2)dt = \frac{2i-1-n}{n^2}, \tag{19}$$

и $J(t) = 4t - 2$, $0 \le t \le 1$, – функция, определяющая *L*-оценку [5, 6]. Отметим, что оценка $\hat{\Delta}_0 = T(F_n)$ средней разности Джини $\Delta_F = T(F)$, согласно теореме (7.1.25) в [6], асимптотически нормальна, т.е. выполняется выражение

$$L\{\sqrt{n}(\hat{\Delta}_0 - \Delta_F) / \sigma_F(\Delta_n)\} = N(0,1) \text{ при } n \to \infty , \tag{20}$$

где асимптотическая дисперсия $\sqrt{n}\hat{\Delta}_0$-оценки вычисляется по формуле

$$\sigma_F^2(\hat{\Delta}_0) = \int IF^2(x; F, \Delta_F) dF(x) = \int \varphi_F^2(x) dF(x) - \left(\int \varphi_F(x) dF(x) \right)^2 , \tag{21}$$

в которой функция $\varphi_F(x)$ определяется в виде

$$\varphi_F(x) = 2x(2F(x)-1) - 4\mu_F(x), \ \mu_F(x) = \int_{-\infty}^{x} y dF(y) . \tag{22}$$

Можно убедиться (см., например, [5, 6]), что функция влияния $IF(x; F, \Delta_F)$ оценки $\hat{\Delta}_0 = T(F_n)$ средней разности Джини $\Delta_F = T(F)$ записывается в виде

$$IF(x; F, \Delta_F) = \varphi_F(x) - \int \varphi_F(x) dF(x) = \varphi_F(x) - 2(\Delta_F - \mu_F), \ x \in R^1 . \tag{23}$$

Отметим, что функция влияния $IF(x; F, \Delta_F)$ не является ограниченной функцией и, следовательно, выборочные оценки средней разности Джини, так же как и оценка $\hat{S}_1(0)$ стандартного отклонения, подвержены сильному влиянию выбросов в выборке.

Рассмотрим теперь модифицированную оценку средней разности Джини, которая была предложена в [8]. Эта оценка является *U*-статистикой, основанной на урезанной выборке, и имеет вид

$$\hat{\Delta}_\alpha = \{(n - 2[\alpha n])(n - 2[\alpha n] - 1)\}^{-1} \sum_{C_\alpha} \left| X_{(i)} - X_{(j)} \right|, \ 0 \le \alpha < 1/2 . \tag{24}$$

Функционал $U_\alpha(F)$ из (6), соответствующий этой оценке при $m = 2$ и ядре $h(x_1, x_2) = | x_1 - x_2 |$, записывается в виде

$$\Delta_\alpha(F) = \frac{1}{(1-2\alpha)^2} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \int |x - y| dF(x) dF(y), \ 0 \le \alpha < 1/2 . \tag{25}$$

Далее функция $g_\alpha(x; F)$ из (4) для данного случая при $m = 2$ и $h(x_1, x_2) = | x_1 - x_2 |$ запишется в виде

$$g_\alpha(x; F) = \frac{I[\xi_\alpha \le x \le \xi_{1-\alpha}]}{(1-2\alpha)^2} \{\varphi_\alpha(x, F)/2 + \mu_\alpha(F)\}, \tag{26}$$

где функция $\varphi_\alpha(x, F)$ и $\mu_\alpha(F)$ вычисляются по формулам

$$\varphi_\alpha(x, F) = 4xF(x) - 2x - \int_{F^{-1}(\alpha)}^{x} y dF(y), \ \mu_\alpha(F) = \int_{\alpha}^{1-\alpha} F^{-1}(t) dt . \tag{27}$$

Величины A_α и B_α из (5) в данном случае, соответственно, равны

$$A_\alpha = \left[(1-2\alpha)\xi_\alpha - \mu_\alpha(F)\right]/(1-2\alpha)^2,\ B_\alpha = \left[(1-2\alpha)\xi_{1-\alpha} - \mu_\alpha(F)\right]/(1-2\alpha)^2. \qquad (28)$$

Приведем выражение для функции влияния $IF(x;F,\hat{\Delta}_\alpha)$ оценки $\hat{\Delta}_\alpha$ вида (24). Для простоты рассмотрим симметричные распределения, т.е. $F \in \Im_{S|0}$, тогда $\mu_\alpha(F)=0$ и $\xi_\alpha = -\overline{\xi}_\alpha$. Введём дополнительные обозначения:

$$J_1 = J_1(F,\alpha) = \int_{\xi_\alpha}^{\overline{\xi}_\alpha} \varphi_\alpha(x,F)dF(x) = 2(1-2\alpha)^2\Delta_\alpha(F),\ J_2 = J_2(F,\alpha) = \int_{\xi_\alpha}^{\overline{\xi}_\alpha} \varphi_\alpha^2(x,F)dF(x). \qquad (29)$$

С использованием формул (26)–(29), формула (13) для *ограниченной* функции влияния $IF(x;F,\hat{\Delta}_\alpha)$ оценки $\hat{\Delta}_\alpha$ перепишется в виде

$$IF(x;F,\hat{\Delta}_\alpha) = \frac{1}{(1-2\alpha)^2}\begin{cases} [\varphi_\alpha(x,F) - J_1(F,\alpha) - 4\alpha(1-2\alpha)\overline{\xi}_\alpha], & |x| \le \overline{\xi}_\alpha, \\ [2(1-2\alpha)^2\overline{\xi}_\alpha - J_1(F,\alpha)], & |x| > \overline{\xi}_\alpha. \end{cases} \qquad (30)$$

Согласно (10) асимптотическая дисперсия $\sqrt{n}\hat{\Delta}_\alpha$-оценок вычисляется по формуле

$$\sigma^2(F,\hat{\Delta}_\alpha) = \int_{-\infty}^{\infty} IF^2(x;F,\hat{\Delta}_\alpha)dF(x) =$$

$$= \frac{J_2(F,\alpha) - 2J_1^2(F,\alpha) + 8J_1(F,\alpha)\alpha(1-2\alpha)F^{-1}(\alpha) + 8\alpha(1-2\alpha)^3[F^{-1}(\alpha)]^2}{(1-2\alpha)^4},\ F \in \Im_{S|0}. \qquad (31)$$

Используя формулы (29) и (31), получаем выражение для асимптотической стандартизованной дисперсии $\hat{\Delta}_\alpha$-оценок в виде

$$\tilde{\sigma}^2(F,\hat{\Delta}_\alpha) = \frac{\sigma^2(F,\hat{\Delta}_\alpha)}{\Delta_\alpha^2(F)} = 4\{J_2 - J_1^2 + 8\alpha(1-\alpha)J_1\xi_\alpha + 8\alpha(1-2\alpha)^3\xi_\alpha^2\}/J_1^2,\ F \in \Im_{S|0}. \qquad (32)$$

Пример 1. Приведём результаты вычислений асимптотической дисперсии оценки $\hat{\Delta}_0$ средней разности Джини при нормальном распределении, т.е. предполагаем, что $F(x) = \Phi(x) = (1/\sqrt{2\pi})\int_{-\infty}^{x} \exp\{-x^2/2\}dx$. В данном случае функция $\varphi_F(x)$ из (22) запишется в виде $\varphi_\Phi(x) = 2x(2\Phi(x)-1) + (4/\sqrt{2\pi})\exp\{-x^2/2\}$. Используя формулу (21), получаем (детали см. в примере (10.6.15) в [6]) значение асимптотической дисперсии $\sqrt{n}\hat{\Delta}_0$-оценки средней разности Джини для нормального распределения, равное

$$\sigma^2(\Phi,\hat{\Delta}_0) = \int \varphi_\Phi^2(x)d\Phi(x) - \left(\int \varphi_\Phi(x)d\Phi(x)\right)^2 = \frac{4(\pi+6\sqrt{3})}{3\pi} - \frac{16}{\pi} = \frac{4(\pi+6\sqrt{3}-12)}{3\pi} \approx 0{,}651.$$

Стандартизованная асимптотическая дисперсия $\sqrt{n}\hat{\Delta}_0$-оценки равна

$$\tilde{\sigma}^2(\Phi,\hat{\Delta}_0) = 4\cdot\frac{\int \varphi_\Phi^2(x)d\Phi(x) - \left(\int \varphi_\Phi(x)d\Phi(x)\right)^2}{\left(\int \varphi_\Phi(x)d\Phi(x)\right)^2} = 4\,\frac{4(\pi+6\sqrt{3}-12)}{3\pi\cdot(16/\pi)} = \frac{\pi+6\sqrt{3}-12}{3} \approx 0{,}511.$$

Пример 2. Рассмотрим супермодель с засорением $\Im_{\varepsilon,\tau}(\Phi) = \{F : F(x) = \Phi_{\varepsilon,\tau}(x)\}$, где

$$\Phi_{\varepsilon,\tau}(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/\tau),\ 0 \le \varepsilon \le 1,\ \tau \ge 1.$$

Численные значения асимптотической стандартизованной дисперсии $\tilde{\sigma}^2(F,\hat{\Delta}_0) = \sigma^2(F;\hat{\Delta}_0)/\Delta_F^2$ для $\sqrt{n}\hat{\Delta}_0$-оценки средней разности Джини для супермодели $\Im_{\varepsilon,\tau}(\Phi)$, вычисленные по формуле (10.6.32) в [6], приведены в табл. 1.

Для сравнения оценки $\hat{\Delta}_0$ средней разности Джини с оценкой $\hat{S}_1(0)$ стандартного отклонения в рамках супермодели $\Im_{\varepsilon,\tau}(\Phi)$ в табл. 2 приведены значения их асимптотических относительных эффективностей $AO\Im_F(\hat{\Delta}_0 : \hat{S}_1(0)) = \tilde{\sigma}^2(F;\hat{S}_1(0))/\tilde{\sigma}^2(F;\hat{\Delta}_0)$ для $F \in \Im_{\varepsilon,\tau}(\Phi)$.

Таблица 1

Асимптотическая стандартизованная дисперсия $\tilde{\sigma}^2(F,\hat{\Delta}_0)$ для $F \in \Im_{\varepsilon,\tau}(\Phi)$

τ \ ε	0,00	0,001	0,005	0,01	0,05	0,10	0,20	0,30
τ = 3	0,511	0,523	0,566	0,618	0,928	1,137	1,255	1,204
τ = 5	0,511	0,558	0,735	0,933	1,887	2,256	2,159	1,831
τ = 10	0,511	0,753	1,601	2,443	4,901	4,781	3,542	2,592

Таблица 2

Асимптотическая относительная эффективность $AO\Im_F(\hat{\Delta}_0 : \hat{S}_1(0))$ для $F \in \Im_{\varepsilon,\tau}(\Phi)$

τ \ ε	0,00	0,001	0,005	0,01	0,05	0,10	0,20	0.30
τ = 3	0,978	1,046	1,274	1,468	1,679	1,612	1,304	1,140
τ = 5	0,978	1,634	3,011	3,518	2,512	1,712	1,183	1,010
τ = 10	0.978	8,740	10,53	7,730	2,114	1,269	0,907	0,825

Из данных табл. 2 следует, что оценка $\hat{\Delta}_0$ средней разности Джини, проигрывая лишь 2% в эффективности оценке $\hat{S}_1(0)$ стандартного отклонения при нормальном распределении, становится предпочтительнее уже при небольших отклонениях от нормального распределения в рамках супермодели $\Im_{\varepsilon,\tau}(\Phi)$. Отметим, что при увеличении параметра τ относительная эффективность $AO\Im_{\Phi_{\varepsilon,\tau}}(\hat{\Delta}_0 : \hat{S}_1(0))$ принимает неожиданно большие значения. По всей вероятности, это является следствием того факта, что оценка $\hat{S}_1(0)$ стандартного отклонения имеет неограниченную (*квадратично возрастающую*) функцию влияния, и при увеличении τ и малых ε её дисперсия резко возрастает (см. также теорему (2) в [10]).

Пример 3. В табл. 3 приведены результаты сравнения различных оценок масштабного параметра, используя понятие дефекта оценки [12]. Напомним, что дефект $DE(F,\hat{\theta}_i)$ оценки $\hat{\theta}_i$, $i = 1,...,k$, среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра θ при распределении F определяют в виде

$$DE(F,\hat{\theta}_i) = 1 - \min\{\sigma^2(F,\hat{\theta}_1),...,\sigma^2(F,\hat{\theta}_k)\} / \sigma^2(F,\hat{\theta}_i),\ i = 1,...,k.$$

Отметим, что при сравнении оценок масштабного параметра используют вместо асимптотических дисперсий $\sigma^2(F,\hat{\theta}_i)$ их стандартизованные дисперсии $\tilde{\sigma}^2(F,\hat{\theta}_i)$, $i = 1,...,k$. Обычно дефекты оценок откладывают для наглядности на плоскости двух распределений [12]. В табл. 3 приведены дефекты оценок масштабного параметра с использованием их стандартизованных дисперсий в «плоскости двух распределений: $F_{(1)}$ – Гаусс, $F_{(3)}$ – Лаплас».

Таблица 3

Дефекты оценок масштабного параметра для распределений $F_{(1)} = \Phi$ и $F_{(3)}$

Дефекты оценок	$\hat{S}_1(0)$	$\hat{S}_2(0)$	$\hat{\Delta}_0$	$\hat{S}_4$	$\hat{S}_3(0,10)$	$\hat{S}_3(0,25)$
$DE(F_{(1)},\hat{S})$	0,00	0,12	**0,02**	0,14	0,38	0,63
$DE(F_{(3)},\hat{S})$	0,20	0,00	**0,04**	0,21	0,35	0,52

Итак, среди сравниваемых оценок предпочтение следует отдать $\hat{\Delta}_0$-оценке средней разности Джини. Отметим, что для симметричных распределений ограниченные функции влияния и асимптотические дисперсии $\hat{S}_3(0,25)$-оценки интерквартильного размаха и оценки $\hat{S}_3^* = \text{med}\{|X_i - \text{med}(X)|,\ 1 \le i \le n\}$, совпадают. Отметим также, что оценка $\hat{S}_4 = \text{med}\{|X_i - X_j|,\ 1 \le i < j \le n\}$, в отличие от оценок $\hat{S}_1(0)$, $\hat{S}_2(0)$ и $\hat{\Delta}_0$, также имеет ограниченную функцию влияния.

Пример 4. Вычисления асимптотической стандартизованной дисперсии $\tilde{\sigma}^2(F,\hat{\Delta}_\alpha)$ для $\hat{\Delta}_\alpha$-оценок вида (24) в рамках модели с засорением, т.е. для $F \in \Im_{\varepsilon,\tau}(\Phi)$, показывают, что она существенно зависит от параметра α. Например, для нормального распределения минимальное значение асимптотической стандартизованной дисперсия $\tilde{\sigma}^2(F,\hat{\Delta}_\alpha)$ для $\hat{\Delta}_\alpha$-оценок достигается при $\alpha = 0$, а при $\tau = 3$ и $\varepsilon = 0,05$ минимум достигается при $\alpha = 0,05$. Для распределения Коши минимум достигается при $\alpha = 0,20$. Отмеченный факт приводит к необходимости адаптации параметра α к меняющемуся распределению в рамках заданной супермодели при практическом использовании $\hat{\Delta}_\alpha$-оценок. Следуя работам [13, 14], определим адаптивный параметр $\hat{\alpha}(X_1,...,X_n)$ для $\hat{\Delta}_\alpha$-оценок вида (24), вычисляемый по формуле

$$\hat{\alpha}(X_1,...,X_n) = \begin{cases} \alpha_1,\ Q(F_n) \le Q_1, \\ \dfrac{\alpha_{i+1}-\alpha_i}{Q_2-Q_1}\{Q(F_n)-Q_i\}+\alpha_i,\ Q_i < Q(F_n) < Q_{i+1}, i=1,2, \\ \alpha_3,\ Q(F_n) \ge Q_3, \end{cases} \qquad (33)$$

где параметры α_1, α_2, α_3, Q_1, Q_2 и Q_3 задаются в соответствии с рассматриваемым типом супермодели. В формуле (33) $Q(F_n)$ – выборочная оценка, построенная по выборке $X_1,...,X_n$ методом подстановки, для функционала $Q(F;\nu,\mu)$, характеризующего степень затянутости хвостов распределения наблюдений [13]. Эта оценка записывается в виде

$$Q(F_n;\nu,\mu) = \frac{m}{k}\left(\sum_{i=n-k+1}^{n} X_{(i)} - \sum_{i=1}^{k} X_{(i)}\right) \Big/ \left(\sum_{i=n-m+1}^{n} X_{(i)} - \sum_{i=1}^{m} X_{(i)}\right),\ k=[\nu n],\ m=[\mu n], \qquad (34)$$

где $0 < \nu < \mu \le 0,5$ и $X_{(1)},...,X_{(n)}$ – порядковые статистики выборки $X_1,...,X_n$. Следуя работе [13], ниже полагаем $\nu = 0,2$ и $\mu = 0,5$. В качестве примера рассмотрим достаточно широкую супермодель Тьюки в виде λ-аппроксимаций квантильных функций заданных распределений [11]. Эта супермодель определяется в виде

$$\Im_\lambda = \{F : F_\lambda^{-1}(t) = \lambda_1 + [t^{\lambda_3} - (1-t)^{\lambda_3}]/\lambda_2,\ 0 \le t \le 1\}. \qquad (35)$$

При определении адаптивного параметра $\hat{\alpha}(X_1,...,X_n)$ по формуле (33) в рамках супермодели Тьюки $\Im_\lambda$ примем следующие значения параметров: $\alpha_1 = 0$, $\alpha_2 = 0,2$, $\alpha_3 = 0,3$, $Q_1 = 1,76$, $Q_2 = 2,50$, $Q_3 = 4,30$. Результаты сравнения адаптивной $\hat{\Delta}_{\hat{\alpha}}$-оценки средней разности Джини с семейством $\hat{\Delta}_\alpha$-оценок, для которых параметр α, $0 \le \alpha < 1/2$, *фиксирован*, приведены в табл. 4 в виде отношения асимптотической стандартизованной дисперсии оценки к минимальной стандартизованной дисперсии среди сравниваемых оценок при заданном распределении (т.е. при заданном значении параметра λ_3).

Таблица 4

Отношения асимптотических стандартизованных дисперсий $\hat{\Delta}_\alpha$-оценок для $F \in \Im_\lambda$

Ф.р.	λ_3	$Q(F_\lambda)$	$\hat{\Delta}_{0,0}$	$\hat{\Delta}_{0,01}$	$\hat{\Delta}_{0,05}$	$\hat{\Delta}_{0,10}$	$\hat{\Delta}_{0,20}$	$\hat{\Delta}_{0,30}$	$\hat{\Delta}_{\hat{\alpha}}$
Равномерная	1,000	1,60	**1,00**	1,14	1,67	2,50	5,00	10,0	1,00
	0,200	1,74	**1,00**	1,07	1,36	1,71	2,74	4,83	1,00
Гаусса–	0,1349	1,76	**1,00**	1,06	1,29	1,59	2,47	4,28	1,00
Лапласа	–0,0802	1,84	1,06	**1,00**	1,04	1,17	1,65	2,71	1,00
	–0,2450	1,91	1,51	1,11	**1,00**	1,05	1,35	2,12	1,00
	–0,5500	2,10	∞	1,67	1,11	**1,00**	1,10	1,57	1,00
Коши	–1,000	2,50	∞	3,87	1,59	1,16	**1,00**	1,22	1,00
	–2,000	4,38	∞	18,2	3,93	2,05	1,14	**1,00**	1,00

Приведенные в табл. 4 данные показывают, что адаптивная $\hat{\Delta}_{\hat{\alpha}}$-оценка средней разности Джини обладает преимуществом перед урезанными $\hat{\Delta}_\alpha$-оценками средней разности Джини, для которых параметр α, $0 \le \alpha < 1/2$, *фиксирован*, при изменении распределения выборки в достаточно широком семей-

стве распределений $\Im_\lambda$. Отметим, что приведенные результаты сравнения являются асимптотическими. При конечных объемах выборки требуются дополнительные исследования. Предварительные результаты моделирования показывают, что отмеченное преимущество адаптивных $\hat{\Delta}_{\hat{\alpha}}$-оценок начинает устойчиво проявляться уже при объемах выборки $n \geq 40$ и для других супермоделей.

Пример 5. Рассмотрим супермодель $\Im\{\Phi(c)\}$ в виде семейства обобщенных гауссовских распределений с плотностью $f(x,c)$, которая зависит от параметра c и определяется в виде

$$f(x,c) = \frac{c}{2A(c)\Gamma(1/c)} \exp\{-[|x|/A(c)]^c\},\ x \in R^1,\ 0{,}5 \leq c \leq 3{,}0,\ A(c) = \sqrt{\Gamma(1/c)/\Gamma(3/c)}.$$

Отметим, что супермодель $\Im\{\Phi(c)\}$ вида

$$\Im\{\Phi(c)\} = \{F : F(x,c) = [c/\Gamma(1/c)A(c)] \int_{-\infty}^{x} \exp\{-[|x|/A(c)]^c\} dx\},\ x \in R^1,$$

включает при $c = 1$ распределение Лапласа, и $c = 2$ соответствует нормальному распределению.

Изменение эффективности для $0{,}5 < C < 1{,}75$ приведены в табл. 5.

Таблица 5

Относительные эффективности $AO\Im_F(\hat{\Delta}_0 : \hat{S}_1(0))$ и $AO\Im_F(\hat{S}_2(0) : \hat{S}_1(0))$ для $F \in \Im\{\Phi(c)\}$

c	0,50	0,75	1,00	1,25	1,50	1,75	2,00	2,50
$AO\Im_F(\hat{\Delta}_0 : \hat{S}_1(0))$	1,74	1,41	1,21	1,09	1,03	1,00	0,98	0,96
$AO\Im_F(\hat{S}_2(0) : \hat{S}_1(0))$	1,78	1,52	1,25	1,09	1,00	0,93	0,88	0,81

Итак, при изменении параметра c в интервале $0{,}5 \leq c < 1{,}75$ оценка $\hat{S}_2(0)$ среднего абсолютных отклонений эффективнее оценки $\hat{S}_1(0)$ стандартного отклонения. Кроме того, оценка $\hat{\Delta}_0$ средней разности Джини также предпочтительнее оценки $\hat{S}_1(0)$.

Заключение

Традиционно применяемые на практике оценки масштабного параметра, характеризующего степень разброса случайной величины, имеют неограниченные функции влияния Хампеля и обладают повышенной чувствительностью к наличию выбросов в выборке, что приводит к существенным искажениям статистических выводов. В данной работе предложена модифицированная оценка средних разностей Джини $\hat{\Delta}_\alpha$, $0 \leq \alpha < 1/2$, которая входит в класс *U*-статистик, основанных на урезанных выборках. Показано, что эта оценка асимптотически нормальна, имеет ограниченную функцию влияния, и, следовательно, защищена от влияния выбросов. При этом она обладает высокой эффективностью при нормальном распределении наблюдений. В работе приведены результаты сравнения предложенной оценки с другими оценками масштабного параметра в рамках гауссовской модели с засорением и предложен адаптивный вариант $\hat{\Delta}_\alpha$-оценок, для которых параметр α, характеризующий пропорцию «урезания» исходной выборки, выбирается на основе информации, содержащейся в исходной выборке с использованием выборочной оценки функционала, характеризующего степень «затянутости хвостов» функции распределения F изучаемой случайной величины *X*.

ЛИТЕРАТУРА

1. Hoefding W. A class of statistics with asymptotically normal distribution //Ann. Math. Statist. 1948. V. 19. P. 292–325.
2. Janssen P., Serfling R., Veraverbeke M. Asymptotic normality of U-statistics based on trimmed samples // J. Statist. Planning and Inference. 1987. V. 16. P. 63–74.
3. Serfling R.J. Approximation Theorems of Mathematical Statistics. N. Y. : Wiley, 1980. 371 p.

4. Хампель Ф., Рончетти Э., Рауссей П., Штаэль В. Робастность в статистике. Подход на основе функций влияния. М. : Мир, 1989. 512 с.
5. Шуленин В.П. Введение в робастную статистику. Томск : Изд-во Том. ун-та, 1993. 227 с.
6. Шуленин В.П. Математическая статистика. Ч. 3: Робастная статистика : учеб. Томск : Изд-во НТЛ, 2012. 520 с.
7. Кендэлл М., Стьюарт А. Теория распределений. М. : Наука, 1966. 587 с.
8. Шуленин В.П. Исследование устойчивости и асимптотических свойств урезанной средней разности Джини // Тр. IV Международной конференции по теории вероятности и математической статистике. Вильнюс, 1985. С. 330–332.
9. Шуленин В.П. Асимптотические свойства GL и U-статистик // Вестник Томского государственного университета. Приложение. 2004. № 9 (11). С. 184–190.
10. Bickel P.J., Lehmann E.L. Descriptive statistics for nonparametric models. III. Dispersion // Ann. Statist. 1976. V. 4, No. 6. P.1139–1158.
11. Ramberg J.S., Schmeiser B.W. An approximative method for generating symmetric random variables // Commun ACM. 1972. V. 15. P. 987–990.
12. Andrews D.F., Bickel P.J., Hampel F.R., Huber P. J., Rogers W.H., Tukey J.W. Robust estimation of location: survey and advances. N. Y. : Princeton Univ. Press, 1972. 375 p.
13. Hogg R.V. Adaptive robust procedures: A partial review and some suggestions for future applications and theory // J. Amer. Statist. Assoc. 1974. V. 69. P. 909–923.
14. Шуленин В.П. Адаптивная оценка урезанной средней разности Джини // Методы и программное обеспечение обработки информации и прикладного статистического анализа данных на ЭВМ. Минск, 1985. С. 113–114.

Шуленин Валерий Петрович, канд. техн. наук. E-mail: shvp@fpmk.tsu.ru
Томский государственный университет

Поступила в редакцию 21 апреля 2016 г.

Shulenin Valery P. (Tomsk State University, Russian Federation).
The asymptotic properties of the modified Gini's mean difference.
Keywords: scale parameter; robust estimation; influence function; asymptotic relative efficiency; adaptive estimators.

DOI: 10.17223/19988605/37/8

The paper proposes a modified estimator Gini's mean difference, which is part of a class U-statistics based on the trimmed samples. It is shown that this estimate is asymptotically normal, has a limited influence function, it has a high efficiency for a normal distribution of observations.

We assume that $X_1,...,X_n$ a random sample from a distribution function $F(x)$ and we assume that has a density $f(x)$, $x \in R^1$, $X_{(1)},...,X_{(n)}$ - ordered statistics of the original sample $X_1,...,X_n$. Let $T(F)$, $F \in \Im$ common functional that characterizes the scale parameter, which describes the degree of dispersion of the study of a random variable X. We consider the functional which is defined as

$$\Delta_\alpha(F) = \frac{1}{(1-2\alpha)^2} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \int |x-y| dF(x) dF(y), \ 0 \le \alpha < 1/2 .$$

Sample estimator of this functional, called a modified Gini's mean difference, written in the form

$$\hat{\Delta}_\alpha = \{(n-2[\alpha n])(n-2[\alpha n]-1)\}^{-1} \sum_{C_\alpha} |X_{(i)} - X_{(j)}|, \ 0 \le \alpha < 1/2, \ C_\alpha = \{(i,j) : [\alpha n]+1 \le i < j \le n-[\alpha n]\} .$$

The results of the comparison of the proposed $\hat{\Delta}_\alpha$-estimators with other estimates of the scale parameter for Gaussian model with ε-fixed proportion of contamination, and proposed an adaptive version $\hat{\Delta}_\alpha$-estimators for which the parameter α characterizing the proportion of "trimmed" of the original sample, selected on the basis of information contained in the original sample using a sample estimate functional, characterizing the degree of "heavy tails" of the distribution function of the random variable X under study.

REFERENCES

1. Hoeffding, W. (1948) A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*. 19. pp. 292-325. DOI: 10.1214/aoms/1177730196
2. Janssen, P., Serfling, R. & Veraverbeke, M. (1987) Asymptotic normality of U-statistics based on trimmed samples. *J. Statist. Planning and Inference.* 16. pp. 63-74. DOI: 10.1016/0378-3758(87)90056-5
3. Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics*. New York: Wiley.
4. Hampel, F., Ronchetti, E., Rousseeuw, P.J. & Stahel, W. (1989) *Robastnost' v statistike. Podkhod na osnove funktsiy vliyaniya* [Robust Statistics. The Approach Based on Influence Functions]. Translated from English. Moscow: Mir.
5. Shulenin, V.P. (1993) *Vvedenie v robastnuyu statistiku* [Introduction to robust statistics]. Tomsk: Tomsk State University.
6. Shulenin, V.P. (2012) *Matematicheskaya statistika. Ch. 3: Robastnaya statistika* [Math statistics. Part 3. Robust statistics]. Tomsk: NTL.
7. Kendall, M. & Stewart, A. (1966) *Theory of distributions* [Theory of distributions]. Translated from English. Moscow: Nauka.

8. Shulenin, V.P. (1985) [Investigation of the stability and asymptotic properties of the Gini truncated mean difference]. *Proc. of the Fouth International Conference on the Theory of Probability and Mathematical Statistics.* Vilnius. pp. 330–332. (In Russian).
9. Shulenin, V.P. (2004) The asymptotic properties of GL and U- statistics. *Vestnik Tomskogo gosudarstvennogo universiteta. Prilozhenie – Tomsk State University Journal. Appendix.* 9(11). pp. 184-190. (In Russian).
10. Bickel, P.J. & Lehmann, E.L. (1976) Descriptive statistics for nonparametric models. III. Dispersion. *Ann. Statist.* 4(6). pp. 1139-1158. DOI: 10.1214/aos/1176343648
11. Ramberg, J.S. & Schmeiser, B.W. (1972) An approximative method for generating symmetric random variables. *Commun ACM*. 15. pp. 987-990. DOI: 10.1145/355606.361888
12. Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P. J., Rogers, W.H. & Tukey, J.W. (1972) *Robust estimation of location: survey and advances*. Princeton, New York: Princeton Univ. Press.
13. Hogg, R.V. (1974) Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *J. Amer. Statist. Assoc*. 69. pp. 909-923. DOI: 10.1080/01621459.1974.10480225
14. Shulenin, V.P. (1985) Adaptive estimation trimmed Gini mean difference. *Methods and software provision of information processing and application of statistical analysis of data on a computer.* Minsk. pp. 113-114. (In Russian).