ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
№ 293 Декабрь 2006

ТЕОРИЯ ВЕРОЯТНОСТИ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

УДК 519.2

О.Н. Головчинер, Ю.Г. Дмитриев

СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ФУНКЦИОНАЛА С УЧЕТОМ СИММЕТРИИ РАСПРЕДЕЛЕНИЯ

Рассматривается задача статистического оценивания математического ожидания функции от случайной величины при наличии дополнительной информации о симметрии неизвестного распределения относительно центра α, принимающего одно из m заданных значений. Предлагаются две адаптивные оценки с одинаковыми асимптотическими свойствами. Путем имитационного моделирования проводится сравнение качества оценок при конечных выборках.

Постановка задачи

Пусть по результатам N независимых наблюдений $X_1, X_2, \ldots, X_N$ над случайной величиной X с неизвестной функцией распределения $F(x)$ требуется оценить функционал $\theta(F) = \mathrm{M}\,\varphi(X) = \int_{R^1} \varphi(x) dF(x)$ от заданной скалярной функции $\varphi(x)$ на R^1, если известно, что $F(x)$ симметрично относительно точки α, т.е. $F(x) = 1 - F(2\alpha - x) \quad \forall x \in R^1$, и центр симметрии α может принимать одно из m заданных значений: $\alpha \in \{\alpha_i\}_{i=\overline{1,m}}$.

Если центр симметрии известен ($m = 1$ и $\alpha = \alpha_1$), то для оценивания $\theta(F)$ с учетом информации о симметрии применяется оценка

$$\hat{\theta}_\alpha = \int_{R^1} \frac{\varphi(x) + \varphi(2\alpha - x)}{2} dF_N(x) = \\ = \frac{1}{2N} \sum_{i=1}^{N} \left[\varphi(X_i) + \varphi(2\alpha - X_i)\right], \qquad (1)$$

где $F_N(x)$ – эмпирическая функция распределения. Эта оценка является несмещенной, $\mathrm{M}\,\hat{\theta}_\alpha = \theta(F)$ и ее дисперсия равна

$$\mathrm{D}\,\hat{\theta}_\alpha = \frac{\mathrm{D}\,\varphi(X) - \mathrm{cov}\left(\varphi(X), \varphi(2\alpha - X)\right)}{2 \cdot N} = \frac{\sigma_0^2}{N}. \qquad (2)$$

Поскольку $\mathrm{D}\,\varphi(X) = \mathrm{D}\,\varphi(2\alpha - X) = \mathrm{D}\,\varphi$, то, согласно неравенству Коши–Буняковского, величина $\left[\mathrm{D}\,\varphi(X) - \mathrm{cov}\left(\varphi(X), \varphi(2\alpha - X)\right)\right]$ не отрицательна, следовательно, дисперсия оценки $\hat{\theta}_\alpha$ не превосходит дисперсию безусловной эмпирической оценки $\hat{\theta} = N^{-1} \sum_{i=1}^{N} \varphi(X_i)$, определяемую выражением $\mathrm{D}\,\hat{\theta} = N^{-1}\,\mathrm{D}\,\varphi$.

Для случая $m > 1$, т.е. если центр симметрии задан с точностью до конечного множества значений, для оценивания $\theta(F)$ предлагаются два типа оценок, основанных на методе коррелированных процессов В.Н. Пугачева:

1. Оценка $\hat{\theta}_N$, для построения которой определяются вспомогательные функции.

2. Оценка $\hat{\theta}_{\alpha N}$, получаемая подстановкой в (1) оценки центра симметрии α.

Оценка первого типа

Определим для каждого из m заданных значений центра симметрии вспомогательную функцию

$$\psi_i(x) = \frac{\varphi(x) - \varphi(2\alpha_i - x)}{2} \qquad (3)$$

и величину $\Delta_i = \int_{R^1} \psi_i(x) dF(x)$, принимающую нулевое значение, если $\alpha_i = \alpha$. Поскольку множество $\{\alpha_i\}_{i=\overline{1,m}}$ обязательно содержит истинное значение центра симметрии, то $\prod_{i=1}^{m} \Delta_i = 0$. Таким образом, имеющаяся дополнительная информация сводится к равенству

$$\prod_{i=1}^{m} \Delta_i = \int_{R^1} \ldots \int_{R^1} \psi_1(x_1) \ldots \psi_m(x_m)\, dF(x_1) \ldots dF(x_m) = 0.$$

Рассмотрим оценку $\theta_N^{(\lambda)} = \hat{\theta} - \lambda \cdot U_N$, в которой

$$U_N = \left(C_N^m\right)^{-1} \sum_{1 \le i_1 < \ldots < i_m \le N} g\left(X_{i_1}, \ldots, X_{i_m}\right)$$

есть U-статистика с невырожденным ядром

$$g\left(x_{i_1}, \ldots, x_{i_m}\right) = (m!)^{-1} \sum_{(p)} \prod_{t=1}^{m} \psi_t\left(x_{i_t}\right),$$

где обозначение (p) под знаком суммы означает суммирование по всем $m!$ перестановкам $(i_1, \ldots, i_m)$ чисел $(1, \ldots, m)$.

Если $\mathrm{M}\varphi^2(X) < \infty$ и коэффициент λ принимается равным $\lambda^* = \mathrm{cov}(\hat{\theta}, U_N)/\mathrm{D}U_N$, то оценка $\theta_U^{(\lambda^*)}$ обладает минимальной дисперсией

$$\mathrm{D}\theta_N^{(\lambda^*)} = \min_{\lambda} \mathrm{D}\theta_N^{(\lambda)} = \frac{1}{N}\left(\mathrm{D}\varphi - \frac{\mathrm{cov}(\hat{\theta}, U_N)}{\mathrm{D}U_N}\right).$$

Вычисления показывают, что

$$\mathrm{cov}(\hat{\theta}, U_N) = \frac{1}{N}\sum_{i=1}^{m} \Delta^{(i)}\,\mathrm{cov}(\varphi, \psi_i);$$

$$\mathrm{D}U_N = \frac{1}{N}\sum_{i=1}^{m}\left(\Delta^{(i)}\right)^2 \mathrm{D}\psi_i + \frac{1}{N^2}Q + O(N^{-3}),$$

где $\Delta^{(i)} = \prod_{k\neq i}^{m}\Delta_k$, $\Delta^{(ij)} = \prod_{k\neq i,j}^{m}\Delta_k$,

$$Q = \left(\sum_{k=1}^{m}\sum_{l=1,l>k}^{m}\Delta^{(kl)}\,\mathrm{cov}(\psi_k, \psi_l)\right)^2 +$$

$$+\sum_{k=1}^{m}\sum_{l=1,l>k}^{m}\left(\Delta^{(kl)}\right)^2 \mathrm{D}\psi_k \cdot \mathrm{D}\psi_l +$$

$$+2\sum_{k=1}^{m}\sum_{l=1,l\neq k}^{m}\sum_{p=1,p>l}^{m}\Delta^{(kl)}\Delta^{(kp)}\,\mathrm{D}\psi_k \cdot \mathrm{cov}(\psi_l, \psi_p).$$

По условию задачи одно из значений центра симметрии является истинным. Пусть, например, $\alpha_k = \alpha$, тогда $\Delta_k = 0$ и только $\Delta^{(k)}$ и $\Delta^{(ki)}$, $i \neq k$, не равны нулю. Поэтому приведенные выражения можно упростить и представить дисперсию рассматриваемой оценки в виде

$$N \cdot \mathrm{D}\theta_N^{(\lambda^*)} = \mathrm{D}\varphi - \frac{\mathrm{cov}^2(\varphi, \psi_k)}{\mathrm{D}\psi_k} + O(N^{-1}). \qquad (4)$$

Так как оптимальный коэффициент обычно неизвестен, то применим адаптивную оценку $\hat{\theta}_N = \hat{\theta} - \hat{\lambda}\cdot U_N$, в которой вместо λ^* используется оценка его главной части $\hat{\lambda}$, полученная заменой неизвестных величин их эмпирическими значениями $\hat{\Delta}_i$, $\hat{\mathrm{cov}}(\varphi, \psi_i)$, $\hat{\mathrm{D}}\psi_i$, $\hat{\mathrm{cov}}(\psi_i, \psi_j)$, $i, j = \overline{1,m}$, вычисленными по исходной выборке. Свойства адаптивной оценки определяет следующая теорема.

Теорема 1. Пусть $\mathrm{M}\varphi^2(X) < \infty$ и

$$\sigma_0^2 = \frac{1}{2}\mathrm{D}\varphi(X) - \mathrm{cov}\left(\varphi(X), \varphi(2\alpha - X)\right) > 0.$$

Тогда при $N \to \infty$:

а) $\hat{\theta}_N \xrightarrow{p} \theta(F)$,

б) $\mathcal{L}\left(\sqrt{N}\left(\hat{\theta}_N - \theta\right)\right) \to \mathcal{N}\left(0, \sigma_0^2\right)$, где $\mathcal{N}\left(0, \sigma_0^2\right)$ – нормальное распределение с нулевым математическим ожиданием и дисперсией σ_0^2.

Доказательство. Докажем утверждение «а». Так как $\mathrm{M}\varphi^2(X) < \infty$, то из закона больших чисел и первой теоремы непрерывности [1] следует, что

$$\hat{\Delta}_i \xrightarrow{p} \Delta_i,\ \hat{\mathrm{cov}}(\varphi, \psi_i) \xrightarrow{p} \mathrm{cov}(\varphi, \psi_i),$$

$$\hat{\mathrm{D}}\psi_i \xrightarrow{p} \mathrm{D}\psi_i,\ \hat{\mathrm{cov}}(\psi_i, \psi_j) \xrightarrow{p} \mathrm{cov}(\psi_i, \psi_j)$$

для всех $i, j = \overline{1,m}$, поэтому $\hat{\lambda} \xrightarrow{p} \lambda^*$. Отсюда, принимая во внимание несмещенность U-статистик [2. С. 16], получаем $\hat{\theta}_N \xrightarrow{p} \theta(F)$.

Для доказательства утверждения «б» рассмотрим разность $\sqrt{N}\left(\hat{\theta}_N - \theta_N^{(\lambda^*)}\right) = \sqrt{N}\left(\hat{\lambda} - \lambda^*\right)U_N = R_N$.

Здесь U_N – невырожденная U-статистика, так как

$$\mathrm{M}\left(g(X_1, \ldots, X_m)\middle| X_1 = x_1\right) = m^{-1}\sum_{i=1}^{m}\Delta^{(i)}\psi_i(x_1) \neq 0.$$

Согласно [2. С. 92], при $N \to \infty$ величина $\sqrt{N}U_N$ слабо сходится к нормально распределенной случайной величине с нулевым средним и конечной дисперсией. Учитывая, что $\hat{\lambda} \xrightarrow{p} \lambda^*$, получим, что R_N слабо сходится к нулю, т.е. асимптотические распределения величин $\sqrt{N}(\hat{\theta}_N - \theta)$ и $\sqrt{N}(\theta_N^{(\lambda^*)} - \theta)$ совпадают.

Представим оценку $\theta_N^{(\lambda^*)} = N^{-1}\sum_{i=1}^{N}\varphi(X_i) - \lambda^* U_N$ в виде U-статистики $\overline{U}_N$ с симметрическим невырожденным ядром $\bar{g}(x_1, \ldots, x_m) = (m)^{-1}\sum_{i=1}^{m}\varphi(x_i) - \lambda^*\cdot g(x_1, \ldots, x_m)$, для которого выполняются все условия одномерной теоремы об асимптотической нормальности U-статистик [2. С. 90]. Согласно указанной теореме, при $N \to \infty$ случайная величина $\sqrt{N}\left(\overline{U}_N - \theta\right)$ имеет нормальное распределение с нулевым средним и дисперсией, совпадающей с главной частью выражения (4). Принимая во внимание (3) и $\alpha_k = \alpha$, получаем

$$\mathrm{D}\varphi - \frac{\mathrm{cov}^2(\varphi, \psi_k)}{\mathrm{D}\psi_k} = \frac{\mathrm{D}\varphi - \mathrm{cov}\left(\varphi(X)\varphi(2\alpha - X)\right)}{2} = \sigma_0^2,$$

поэтому $\mathcal{L}\left(\sqrt{N}\left(\theta_N^{(\lambda^*)} - \theta\right)\right) \to \mathcal{N}\left(0, \sigma_0^2\right)$ и, следовательно, $\mathcal{L}\left(\sqrt{N}\left(\hat{\theta}_N - \theta\right)\right) \to \mathcal{N}\left(0, \sigma_0^2\right)$.

Оценка второго типа

Применим метод коррелированных процессов для оценивания центра симметрии. Оценим функционал $\alpha(F) = \mathrm{M}X_1 = \int_{R^1} x\,dF(x)$ с учетом дополнительной информации о том, что $\alpha \in \{\alpha_i\}_{i=\overline{1,m}}$. Обозначим $\delta_i = \alpha - \alpha_i$, $i = \overline{1,m}$, тогда дополнительная информация представляется в виде

$$\prod_{i=1}^{m}\delta_i = \int_{R^1}\dots\int_{R^1}(x_1-\alpha_1)\dots(x_m-\alpha_m)dF(x_1)\dots dF(x_m)=0.$$

Полученное равенство несмещенно оценивает U-статистика U_α с симметрическим ядром

$$g_\alpha\left(x_{i_1},\ \dots,\ x_{i_m}\right)=(m!)^{-1}\sum_{(p)}\prod_{t=1}^{m}\left(x_{i_t}-\alpha_t\right).$$

Рассмотрим оценку $\alpha_N=\hat{\alpha}-\hat{\gamma}\cdot U_\alpha$, где $\hat{\alpha}=N^{-1}\sum_{i=1}^{N}X_i$, а $\hat{\gamma}$ – адаптивная оценка главной части оптимального коэффициента

$$\gamma^*=\frac{\operatorname{cov}(\hat{\alpha},U_\alpha)}{\mathrm{D}U_\alpha}=\left[a_1+\frac{2\,a_2^2\,\mathrm{D}X}{(N-1)a_1}\right]^{-1}+O\left(N^{-2}\right),$$

в которой вместо неизвестных величин $\mathrm{D}X$, $a_1=\sum_{j=1}^{m}\prod_{i\neq j}^{m}(\alpha-\alpha_i)$ и $a_2=\sum_{j=1}^{m}\sum_{k=1}^{m}\prod_{i\neq j,k}^{m}(\alpha-\alpha_i)$ используются их эмпирические оценки.

Лемма. Если $\mathrm{D}X=\mathrm{M}(X-\alpha)^2<\infty$, то при $N\to\infty$ $\alpha_N\xrightarrow{\ p\ }\alpha$ и $\sqrt{N}(\alpha_N-\alpha)\xrightarrow{\ p\ }0$.

Доказательство. Пусть коэффициент γ^* известен, тогда $\alpha_N^*=\hat{\alpha}-\gamma^*\cdot U_\alpha$.

Рассмотрим разность $\sqrt{N}\left(\alpha_N-\alpha_N^*\right)=\sqrt{N}\left(\hat{\gamma}-\gamma^*\right)U_\alpha$. На основании закона больших чисел и первой теоремы непрерывности имеем $\hat{\gamma}\xrightarrow{\ p\ }\gamma^*$. Из выражения $\mathrm{M}\left(g_\alpha\left(X_1,\ \dots,\ X_m\right)\middle|X_1=x_1\right)=m^{-1}a_1\left(x_1-\alpha\right)\neq 0$ следует невырожденность U-статистики U_α. Поэтому, согласно [2. С. 92], при $N\to\infty$ величина $\sqrt{N}U_\alpha$ слабо сходится к нормально распределенной случайной величине с нулевым средним и конечной дисперсией $a_1^2\,\mathrm{D}X$. Следовательно, $\sqrt{N}\left(\hat{\gamma}-\gamma^*\right)\cdot U_\alpha\xrightarrow{\ p\ }0$ и асимптотические распределения величин $\sqrt{N}(\alpha_N-\alpha)$ и $\sqrt{N}\left(\alpha_N^*-\alpha\right)$ совпадают.

Представим α_N^* в виде U-статистики $\overline{U}_\alpha$ с ядром $\overline{g}_\alpha\left(x_1,\ \dots,\ x_m\right)=(m)^{-1}\sum_{i=1}^{m}x_i-\gamma^*g_\alpha\left(x_1,\ \dots,\ x_m\right)$. Очевидно, что $\mathrm{M}\,\overline{U}_\alpha=\mathrm{M}\,\alpha_N^*=\alpha$. Вычисления показывают, что

$$\mathrm{M}\left(\overline{g}_\alpha\left(X_1,\ \dots,\ X_m\right)\middle|X_1=x_1\right)=0$$

и

$$\mathrm{M}\left(\overline{g}_\alpha\left(X_1,\ \dots,\ X_m\right)\middle|X_1=x_1,X_2=x_2\right)\neq 0,$$

т.е. рассматриваемая U-статистика является вырожденной, ее ранг равен 2 и, согласно [2], при $N\to\infty$

$$\mathrm{D}\,\overline{U}_\alpha=N^{-2}\cdot 2\left(\frac{a_2\,\mathrm{D}X}{a_1}\right)^2+O\left(N^{-3}\right).$$

Отсюда, принимая во внимание неравенство Чебышева, получаем утверждения леммы. ■

Подставляя α_N вместо неизвестного центра симметрии α в (1), получим оценку

$$\hat{\theta}_{\alpha N}=\frac{1}{2N}\sum_{i=1}^{N}\left[\varphi(X_i)+\varphi(2\alpha_N-X_i)\right].$$

Теорема 2. Пусть функция $\varphi(x)$ непрерывна в окрестности α и имеет непрерывные производные до второго порядка включительно, $\mathrm{M}\,\varphi^2(X)<\infty$, $\mathrm{M}\,\varphi'(X)<\infty$, $\mathrm{M}\,\varphi''(X)<\infty$ и $\sigma_0^2=\frac{1}{2}\mathrm{D}\varphi(X)-\operatorname{cov}\left(\varphi(X),\varphi(2\alpha-X)\right)>0$. Тогда при $N\to\infty$ $\hat{\theta}_{\alpha N}\xrightarrow{\ p\ }\theta(F)$ и $\mathcal{L}\left(\sqrt{N}\left(\hat{\theta}_{\alpha N}-\theta\right)\right)\to\mathcal{N}\left(0,\sigma_0^2\right)$, где $\mathcal{N}\left(0,\sigma_0^2\right)$ – нормальное распределение с нулевым математическим ожиданием и дисперсией σ_0^2.

Доказательство. В окрестности точки α справедливо разложение

$$\hat{\theta}_{\alpha N}=\hat{\theta}_\alpha+(\alpha_N-\alpha)\cdot\frac{1}{N}\sum_{i=1}^{N}\varphi'(2\alpha-X_i)+$$
$$+(\alpha_N-\alpha)^2\cdot\frac{1}{N}\sum_{i=1}^{N}\varphi''(2z-X_i),$$

где $z\in(\alpha,\alpha_N)$. Отсюда

$$\sqrt{N}\left(\hat{\theta}_{\alpha N}-\theta\right)=\sqrt{N}(\alpha_N-\alpha)\cdot\frac{1}{N}\sum_{i=1}^{N}\varphi'(2\alpha-X_i)+$$
$$+(\alpha_N-\alpha)^2\cdot\frac{1}{N}\sum_{i=1}^{N}\varphi''(2z-X_i)+\sqrt{N}\left(\hat{\theta}_\alpha-\theta\right).\qquad(5)$$

На основании закона больших чисел

$$\frac{1}{N}\sum_{i=1}^{N}\varphi'(2\alpha-X_i)\xrightarrow{\ p\ }\mathrm{M}\,\varphi'(X)<\infty,$$

$$\frac{1}{N}\sum_{i=1}^{N}\varphi''(2z-X_i)\xrightarrow{\ p\ }\mathrm{M}\,\varphi''(X)<\infty,$$

а согласно лемме, $\sqrt{N}(\alpha_N-\alpha)\xrightarrow{\ p\ }0$. В силу этого два первых слагаемых в правой части (5) стремятся по вероятности к нулю. Следовательно, асимптотическое распределение величины $\sqrt{N}\left(\hat{\theta}_{\alpha N}-\theta\right)$ совпадает с предельным распределением величины $\sqrt{N}\left(\hat{\theta}_\alpha-\theta\right)$, которое является нормальным $\mathcal{N}\left(0,\sigma_0^2\right)$ в силу центральной предельной теоремы. ■

Имитационное моделирование

Согласно доказанным теоремам, предложенные оценки $\hat{\theta}_N$ и $\hat{\theta}_{\alpha N}$, учитывающие информацию о центре симметрии, заданном с точностью до конечного множества значений, при $N\to\infty$ распределены одинаково, причем их асимптотическое распределение совпа-

дает с предельным распределением оценки $\hat{\theta}_\alpha$ – с известным центром симметрии.

Для исследования свойств оценок при конечных объемах наблюдений было проведено численное моделирование, при котором по наборам стандартных нормальных величин, сгенерированных средствами пакета Statistica 6.0, вычислялись различные оценки функционала $\theta(F) = \int_{R^1} e^x dF(x)$ с учетом информации о симметрии распределения с центром $\alpha \in \{\alpha_1, \alpha_2\}$. Величина α_2 принимала различные значения в диапазоне от –1,5 до 1,5, а $\alpha_1 = \alpha = 0$. Критерием качества для сравнения оценок выбрана величина $N\text{СКО} = N/n \cdot \sum_{i=1}^{n} \left(\theta_k^{(i)} - \theta\right)^2$, где $\theta_k^{(i)}$ – значение оценки, вычисленное на *i*-м прогоне, а n=1000 – количество прогонов. Изменение величины *N*СКО в зависимости от значения α_2 для разных оценок и объемов наблюдений показано на приведенных графиках (см. рис. 2).

Результаты проведенного моделирования показали, что:

1. Привлечение дополнительной информации о симметрии исследуемого распределения позволяет повысить точность оценивания по сравнению с безусловной оценкой при конечных объемах наблюдений. Оценить величину уменьшения СКО оценок позволяют приведенные графики. Например, при N=50 и $\alpha_2 = \pm 0,75$ СКО оценок $\hat{\theta}_N$ и $\hat{\theta}_{\alpha_N}$ примерно вдвое меньше, чем СКО безусловной оценки (рис. 1, 3, 4).

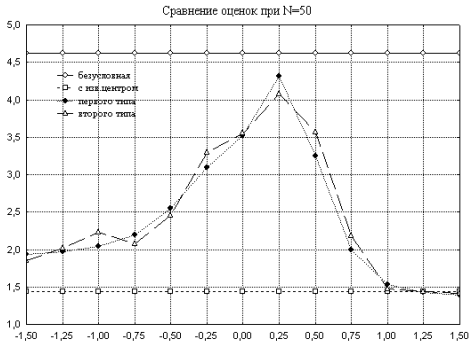


Рис. 1. *N*СКО оценок $\hat{\theta}$, $\hat{\theta}_\alpha$, $\hat{\theta}_N$ и $\hat{\theta}_{\alpha N}$ для $N = 50$

2. Величина выигрыша в точности оценивания зависит от «расстояния» между возможными значениями центра симметрии. Чем ближе друг к другу значения α_i, $i = \overline{1,m}$, тем меньше «ценность» дополнительной информации. Особенно ясно эта зависимость проявляется при оценивании центра симметрии (оценка α_N), когда не оказывает влияния функция $\varphi(x)$.

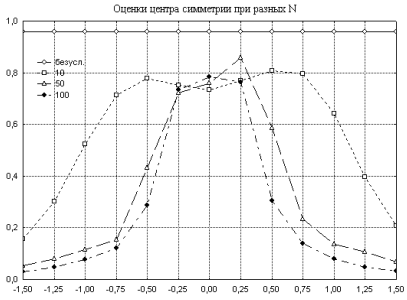


Рис. 2. *N*СКО оценок $\hat{\alpha}$ и α_n для разных *N*

3. Неопределенность в задании центра симметрии увеличивает СКО оценок по сравнению с оценкой $\hat{\theta}_\alpha$ (с известным центром). Кроме того, при маленьких *N* (до 50) очень заметно влияние адаптации. С ростом объемов наблюдений различия между оценками уменьшаются. На рис. 3 и 4 приведены графики *N*СКО оценок $\hat{\theta}_N$ и $\hat{\theta}_{\alpha N}$ для разных *N*.

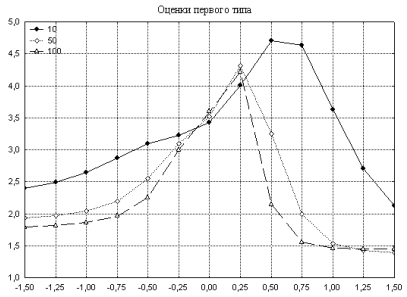


Рис. 3. *N*СКО оценки $\hat{\theta}_N$ для разных *N*

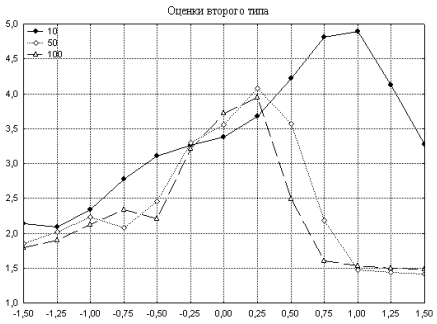


Рис. 4. *N*СКО оценки $\hat{\theta}_{\alpha N}$ для разных *N*

4. Свойства оценок зависят от вида подынтегральной функции. Для исследуемой функции $\varphi(x) = e^x$ при $N = 50$ и более свойства оценок $\hat{\theta}_N$ и $\hat{\theta}_{\alpha N}$ совпадают (рис. 1).

ЛИТЕРАТУРА

1. *Боровков А.А.* Математическая статистика. Проверка гипотез. Оценка параметров. М.: Наука, 1984. 472 с.
2. *Королюк В.С., Боровских Ю.В.* Теория U-статистик. Киев: Наукова думка, 1981. 384 с.

Статья представлена кафедрой теоретической кибернетики факультета прикладной математики и кибернетики Томского государственного университета, поступила в научную редакцию «Кибернетика» 30 мая 2006 г.