A Sharp Inequality for Medians of L-Statistics in a Nonparametric Statistical Model

Ryszard Zieliński

Institute of Mathematics, Polish Academy of Science, Warszawa, Poland.
(R.Zieliński@impan.gov.pl)

Abstract. Sharp bounds for medians of L-statistics in the nonparametric statistical model with all continuous and strictly increasing distribution functions are given. As a corollary we conclude that L-statistics are very poor nonparametric quantile estimators.

1 Result

Let $X_1, \ldots, X_n$ be a sample from a distribution $F \in \mathcal{F}$, where $\mathcal{F}$ is the class of all continuous and strictly increasing distribution functions on their supports. Let $X_{1:n}, \ldots, X_{n:n}$ be the order statistics, let $T = \sum_{j=1}^{n} \lambda_j X_{j:n}; \lambda_j \geq 0, j = 1, 2, \ldots, n; \sum_{j=1}^{n} \lambda_j = 1$, be a nontrivial L-statistic (at least two $\lambda$'s are positive). Let $S = S(X_1, \ldots, X_n)$ be any function of observations $X_1, \ldots, X_n$ and let $Med(F, S)$ denote a median (of the distribution) of $S$ if the sample comes from the distribution $F$. Our primary interest are functions of the form $S(\cdot) = F(T(\cdot))$.

Key words and phrases: Harrell-Davis estimator, Kaigh-Cheng estimator, L-statistics, quantiles, quantile estimators.
Theorem 1.1. If $T = \sum_{j=k}^{m} \lambda_j X_{j:n}$ is an L-statistic such that $\lambda_k > 0$, $\lambda_m > 0$, $k < m$, and $\lambda_k + \lambda_{k+1} + \ldots + \lambda_m = 1$, then

\[(*) \quad m(U_{k:n}) \leq Med(F, F(T)) \leq m(U_{m:n}),\]

where $m(U_{k:n})$ and $m(U_{m:n})$ are medians of order statistics $U_{k:n}$ and $U_{m:n}$ from a sample of size $n$ from the uniform $U(0, 1)$ parent distribution. The bounds are sharp in the sense that for every $\varepsilon > 0$ there exists $F \in \mathcal{F}$ such that $\text{Med}(F, F(T)) > m(U_{m:n}) - \varepsilon$ and for every $\eta > 0$ there exists $G \in \mathcal{F}$ such that $\text{Med}(G, G(T)) < m(U_{k:n}) + \eta$.

Proof. The first statement follows easily from the fact that $X_{k:n} < T < X_{m:n}$ and hence for every $F \in \mathcal{F}$ we have $U_{k:n} = F(X_{k:n}) < F(T) < F(X_{m:n}) = U_{m:n}$. To prove the second part of the theorem it is enough to construct families of distributions $F_\alpha$, $\alpha > 0$, and $G_\alpha$, $\alpha > 0$, such that $\text{Med}(F_\alpha, F_\alpha(T)) \to m(U_{m:n})$ and $\text{Med}(G_\alpha, G_\alpha(T)) \to m(U_{k:n})$, as $\alpha \to 0$.

Consider the family of power distributions $F_\alpha(x) = x^\alpha$, $0 < x < 1$, $\alpha > 0$. Then $X_{j:n} = F_\alpha^{-1}(U_{j:n}) = U_{j:n}^{1/\alpha}$ and

\[F_\alpha(T) = \left(\lambda_k U_{k:n}^{1/\alpha} + \lambda_{k+1} U_{k+1:n}^{1/\alpha} + \ldots + \lambda_{m-1} U_{m-1:n}^{1/\alpha} + \lambda_m U_{m:n}^{1/\alpha}\right)^\alpha\]
\[= U_{m:n} \left[\lambda_k \left(U_{k:n}/U_{m:n}\right)^{1/\alpha} + \lambda_{k+1} \left(U_{k+1:n}/U_{m:n}\right)^{1/\alpha} + \ldots + \lambda_{m-1} \left(U_{m-1:n}/U_{m:n}\right)^{1/\alpha} + \lambda_m\right]^\alpha\]

If $\alpha \to 0$ then $F_\alpha(T) \to U_{m:n}$ and $\text{Med}(F_\alpha, F_\alpha(T)) \to m(U_{m:n})$.

Now consider the family $G_\alpha$ with $G_\alpha(x) = 1 - (1 - x)^\alpha$; in full analogy to the above we conclude that then $G_\alpha(T) \to U_{k:n}$ and $\text{Med}(G_\alpha, G_\alpha(T)) \to m(U_{k:n})$ as $\alpha \to 0$. \(\square\)

Corollary 1.1. If an L-statistic $T = \sum_{j=k}^{m} \lambda_j X_{j:n}$, $\lambda_k > 0$, $\lambda_m > 0$, $\lambda_k + \lambda_{k+1} + \ldots + \lambda_m = 1$, $k < m$, and $\lambda_j = \lambda_j(q)$, $j = k, \ldots, m$, is considered as a nonparametric estimator of the q-th quantile $x_q(F) = F^{-1}(q)$ of an unknown distribution $F \in \mathcal{F}$, then the error of estimation may be arbitrarily large in the sense that for every $C > 0$ there exists a distribution $F \in \mathcal{F}$ such that $|\text{Med}(F, T) - x_q(F)| > C$.

Proof. Suppose that $q < m(U_{m:n})$. The case that $q > m(U_{k:n})$ can be considered in full analogy.
Choose \( \varepsilon > 0 \) such that \( m(U_{m:n}) - \varepsilon > q \). By the Theorem there exists a distribution \( F \in \mathcal{F} \) such that \( \text{Med}(F,F(T)) > m(U_{m:n}) - \varepsilon > q \). By the obvious equality that states that \( \text{Med}(F,F(T)) = F(\text{Med}(F,T)) \) we obtain that \( \text{Med}(F,T) - x_q(F) > 0 \). For an \( \sigma > 0 \) consider the distribution \( F_{\sigma} \in \mathcal{F} \) defined by the formula \( F_{\sigma}(x) = F(x/\sigma) \). Then \( x_q(F_{\sigma}) = \sigma \cdot x_q(F) \) and, due to the fact that \( T \) is scale equivariant, \( \text{Med}(F_{\sigma},T) = \sigma \cdot \text{Med}(F,T) \). Hence \( \text{Med}(F_{\sigma},T) - x_q(F_{\sigma}) = \sigma \cdot (\text{Med}(F,T) - x_q(F)) \) which by a suitable choice of \( \sigma > 1 \) may be arbitrarily large.

\[ \square \]

2 Numerical illustrations (simulations)

To demonstrate that \( L \)-statistics may produce very large errors in estimating quantiles in the nonparametric model \( \mathcal{F} \) with all continuous and strictly increasing distribution functions we decided to present the problem of estimating the median of an unknown \( F \in \mathcal{F} \) with the following well known estimators:

- **Davis and Steinberg (1986)**
  \[ X_{(n+1)/2:n}, \quad \text{if } n \text{ is odd}; \quad \left( X_{n/2:n} + X_{n/2+1:n} \right)/2, \quad \text{if } n \text{ is even}, \]

- **Harrell and Davis (1982)**
  \[ HD = \frac{n!}{\left((n-1)!\right)^2} \sum_{j=1}^{n} \left[ \int_{(j-1)/n}^{j/n} [u(1-u)]^{(n-1)/2} du \right] X_{j:n}, \]

- **Kaigh and Cheng (1991) for \( n \) odd**
  \[ KC = \frac{1}{\binom{2n-1}{n}} \sum_{j=1}^{n} \left( \frac{n-3}{2} + j \right) \left( \frac{3n-1}{2} - j \right) X_{j:n}. \]

As the distributions for studying our problem we have chosen

- **Pareto with cdf**
  \[ 1 - \frac{1}{x^\alpha}, \quad x > 1, \quad \text{heavy tails, no moments of order } k \geq \alpha, \]

- **Power (special case of Beta) with cdf**
  \[ x^\alpha, \quad x \in (0,1), \quad \text{no tails, all moments}, \]
Exponential with cdf

\[ 1 - \exp\{-\alpha x\}, \quad x > 0, \quad \text{very regular}, \]

all distributions for \( \alpha = 1/2, 1/4, \) and \( 1/8. \)

Results of our numerical investigations for samples of size \( n = 9 \)
(Harrell-Davis and Kaigh-Cheng) or for samples of size \( n = 10 \)
(Davis-Steinberg statistic \( (X_{5:10} + X_{6:10})/2 \)) are presented in the Table below. The number of simulated samples, and consequently the number
of simulated values of the estimator under consideration, was
\( N = 9,999, \) and the median from the sample of size \( N = 9,999 \)
has been taken as an estimator of the median of the distribution of the
estimator under consideration. Observe that \( m(U_{n,n}) - m(U_{1:n}) \)
increases with \( n \) so that errors of estimators with \( k = 1 \) and \( m = n \)
(e.g. HD and KC) increase with \( n.\)

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Median</th>
<th>HD</th>
<th>KC</th>
<th>( X_{5:10} + X_{6:10} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pareto</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha = 1/2 )</td>
<td>4</td>
<td>7.72</td>
<td>13.71</td>
<td>4.13</td>
</tr>
<tr>
<td>( \alpha = 1/4 )</td>
<td>16</td>
<td>255</td>
<td>1107</td>
<td>18.45</td>
</tr>
<tr>
<td>( \alpha = 1/8 )</td>
<td>256</td>
<td>3.3 \times 10^6</td>
<td>2.8 \times 10^7</td>
<td>383</td>
</tr>
<tr>
<td>Power</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha = 1/2 )</td>
<td>0.25</td>
<td>0.2780</td>
<td>0.2919</td>
<td>0.2535</td>
</tr>
<tr>
<td>( \alpha = 1/4 )</td>
<td>0.0625</td>
<td>0.1055</td>
<td>0.1286</td>
<td>0.0692</td>
</tr>
<tr>
<td>( \alpha = 1/8 )</td>
<td>0.0039</td>
<td>0.0241</td>
<td>0.0432</td>
<td>0.0053</td>
</tr>
<tr>
<td>Exponential</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha = 1/2 )</td>
<td>1.3863</td>
<td>1.5138</td>
<td>1.6235</td>
<td>1.4079</td>
</tr>
<tr>
<td>( \alpha = 1/4 )</td>
<td>2.7726</td>
<td>3.0571</td>
<td>3.2731</td>
<td>2.8036</td>
</tr>
<tr>
<td>( \alpha = 1/8 )</td>
<td>5.5452</td>
<td>6.0595</td>
<td>6.4897</td>
<td>5.6143</td>
</tr>
</tbody>
</table>

3 A remark

A reason for the bad behavior of nontrivial \( L\)-statistics as quantile
estimators is that they are not equivariant under monotonic trans-
formation of data while the class $\mathcal{F}$ of all continuous and strictly increasing distribution functions allows such transformations. In some parametric families of distributions L-statistics may perform excellently. The problem is discussed thoroughly in a Technical Report (Zieliński 2005).

**Acknowledgment**

The author wish to express his gratitude to referees for their helpful comments which improved the presentation of the paper, especially the proof of Corollary 1.1.

**References**


