Report 99-007
Does Increasing the Sample Size
Always Increase the Accuracy
of a Consistent Estimator
Paul van der Laan
Constance van Eden
ISSN: 1389–2355
DOES INCREASING THE SAMPLE SIZE ALWAYS INCREASE THE ACCURACY OF A CONSISTENT ESTIMATOR?

Paul van der Laan and Constance van Eeden 1

Abstract

Birnbaum (1948) introduced the notion of peakedness about $\theta$ of a random variable $T$, defined by $P(|T - \theta| < \varepsilon), \varepsilon > 0$. What seems to be not well-known is that, for a consistent estimator $T_n$ of $\theta$, its peakedness does not necessarily converge to 1 monotonically in $n$. In this article some known results on how the peakedness of the sample mean behaves as a function of $n$ are recalled. Also, new results concerning the peakedness of the median and the interquartile range are presented.

1 Introduction

Suppose $X_1, \ldots, X_n$ are a sample from a distribution with finite variance and one wants to estimate $\mu = \mathbb{E}X_1$ based on $(X_1, \ldots, X_n)$. Then it is, of course, well-known that $\bar{X}_n = (\sum_{i=1}^n X_i)/n$ is a consistent estimator of $\mu$, i.e., for all $\varepsilon > 0$,

$$p_{\bar{X}_n}(\varepsilon) = P(|\bar{X}_n - \mu| < \varepsilon) \to 1 \quad \text{as } n \to \infty.$$  (1.1)

What seems to be less well-known and is seldom, if ever, mentioned when the subject of consistency is discussed in a course, is that $p_{\bar{X}_n}(\varepsilon)$ does not necessarily converge to one monotonically in $n$. Thus, judging the accuracy of $\bar{X}_n$ by $p_{\bar{X}_n}(\varepsilon)$, $\varepsilon > 0$, a larger $n$ might give a worse estimator.

In this article we first recall in Section 2 some known results on how $p_{\bar{X}_n}(\varepsilon)$ behaves as a function of $n$. Then, in Section 3, we present new results on this question for the case where the median or the midrange are used to estimate the median or the mean of $X_1$.

1Paul van der Laan is Professor, Department of Mathematics and Computing Science, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands (E-mail: PvdLaan@win.tue.nl). Constance van Eeden is Honorary Professor, Department of Statistics, The University of British Columbia, Vancouver, B.C., Canada, V6T 1Z2 (E-mail: vaneeden@stat.ubc.ca).
2 Results for $\bar{X}_n$ and some generalizations

Birnbaum (1948) calls
\[ p_T(\epsilon) = P(|T - \theta| < \epsilon) \quad \epsilon > 0 \]
the peakedness (with respect to $\theta$) of $T$ and calls $T$ more peaked than $S$ when $p_T(\epsilon) \geq p_S(\epsilon)$ for all $\epsilon > 0$. He proves several properties of the peakedness and gives, e.g., conditions under which, for the same $\theta$ and the same sample size, one of two sample means is more peaked than the other.

Proschan (1965) gives several results on the behaviour of $p_{T_n}(\epsilon)$ as a function of $n$ where $T_n$ is a convex combination of $X_1, \ldots, X_n$, a sample from a distribution $F$. He supposes that $F$ has a density which is symmetric with respect to $\theta$ and is logconcave on the support of $F$. In particular, Proschan shows that for such a distribution $p_{X_n}(\epsilon)$ is, for each $\epsilon > 0$, strictly increasing in $n$ (i.e., of course, for those $\epsilon > 0$ which are in the interior of the support of $X_1 - \theta$).

Proschan also gives an example where $p_{X_n}(\epsilon)$ is not increasing in $n$. In fact, he gives a distribution for which $X_1$ is more peaked about 0 than $(X_1 + X_2)/2$. This distribution is the convolution of a distribution with a symmetric (about zero) logconcave density and a Cauchy distribution with median zero. Then, for $\phi$ strictly increasing and convex on $(0, \infty)$ with $\phi(x) = \phi(-x)$ for all $x$, $\phi(X_1)$ is more peaked with respect to zero than $(\phi(X_1) + \phi(X_2))/2$. Of course, for this case $\bar{X}_n$ does not converge to zero in probability, so the result might not be too surprising. However, Dharmadhikari and Joag-Dev (1988, p. 171-172) show that, e.g., for the density
\[ f(x) = \frac{1}{3} I(|x| \leq 1) + \frac{1}{18} (1 \leq |x| \leq 4), \]
$X_1$ is more peaked with respect to zero than $(X_1 + X_2)/2$. And for this distribution (1.1) clearly holds.

The results of Proschan (1965) have been extended to the multivariate case by Olkin and Tong (1987) (see also Dharmadhikari and Joag-Dev (1988, Theorem 7.11)).

3 The case of the median and the midrange

Assume that $X_1, \ldots, X_n$ is a sample from a distribution function with a density and that $n$ is odd. Let $M_n$ be the median of $X_1, \ldots, X_n$, let $\mathcal{M} = [m_1, m_2]$ be the set of medians of the distribution of $X_1$ and let $F$ be the distribution function of $X_1$. Then the following theorem holds.

**Theorem 3.1** Under the above conditions, the peakedness of $M_n - m$ is, for $m \in \mathcal{M}$ and $\epsilon > 0$ such that $\frac{1}{2} < F(m + \epsilon) < 1$, strictly increasing in $n$. 

2
Proof. Assume without loss of generality that \( m = 0 \). First note that, for \( x \in (-\infty, \infty) \),
\[
P(M_n > x) = \sum_{i=0}^{(n-1)/2} \binom{n}{i} F(x)^i (1 - F(x))^{n-i} = 1 - \frac{1}{B\left(\frac{n+1}{2}, \frac{n+1}{2}\right)} \int_0^{F(x)} t^{\frac{n-1}{2}} (1-t)^{\frac{n-1}{2}} dt.
\]
So, as a function of \( y = F(x), 0 < y < 1 \),
\[
\frac{d}{dy} P(M_n > x) = -\frac{y^{\frac{n-1}{2}} (1-y)^{\frac{n-1}{2}}}{B\left(\frac{n+1}{2}, \frac{n+1}{2}\right)}.
\]
Putting \( Q_n(y) = P(M_n > x) - P(M_{n+2} > x) \), this gives
\[
\frac{d}{dy} Q_n(x) = \frac{(n+2)!}{\left(\left(\frac{n+1}{2}\right)!ight)^2} y^{\frac{n+1}{2}} (1-y)^{\frac{n+1}{2}} - \frac{n!}{\left(\left(\frac{n-1}{2}\right)!ight)^2} y^{\frac{n-1}{2}} (1-y)^{\frac{n-1}{2}}
\]
\[
= y^{\frac{n-1}{2}} (1-y)^{\frac{n-1}{2}} \frac{n!}{\left(\left(\frac{n+1}{2}\right)!ight)^2} \left( (n+1)(n+2)y(1-y) - \left(\frac{n+1}{2}\right)^2 \right).
\]
This last expression is, for \( 0 < y < 1, 0 = 0, < 0 \) if and only if
\[
G(y) = -y^2 + y - \frac{n+1}{4(n+2)} = \frac{1}{4(n+2)} - (y - \frac{1}{2})^2 \left\{ \begin{array}{ll} > 0 \quad & (\quad) \\ = 0 \quad & (\quad) \\ < 0 \quad & (\quad) \end{array} \right.,
\]
which is equivalent to
\[
\left| y - \frac{1}{2} \right| \left\{ \begin{array}{ll} < \quad & (\quad) \\ = \quad & (\quad) \\ > \quad & (\quad) \end{array} \right. \iff c = \frac{1}{2} \sqrt{(n+2)^{-1}}.
\]
So, \( Q_n(y) \) is increasing on \( (\frac{1}{2} - c, \frac{1}{2} + c) \) and decreasing on \( (0, \frac{1}{2} - c) \) and on \( (\frac{1}{2} + c, 1) \). Combining this with the fact that, for all \( n \),
\[
P(M_n > x) = \left\{ \begin{array}{ll} 1 \quad & (\quad) \quad \text{for } y = 0 \\ \frac{1}{2} \quad & (\quad) \quad \text{for } y = \frac{1}{2} \\ 0 \quad & (\quad) \quad \text{for } y = 1,
\end{array} \right.
\]
shows that
\[
P(M_n > x) - P(M_{n+2} > x) \left\{ \begin{array}{ll} > 0 \quad & (\quad) \quad \text{for } x \text{ such that } \frac{1}{2} < F(x) < 1 \\ < 0 \quad & (\quad) \quad \text{for } x \text{ such that } 0 < F(x) < \frac{1}{2},
\end{array} \right.
\]
3
which proves the result. □

Note, from Theorem 3.1, that the conditions on $F$ for the median to have increasing peakedness in $n$ are much weaker than those for the mean. All one needs for the median is a density, while for the mean a logconcave symmetric density is needed in the proofs. But in order for the median to be a consistent estimator of the population median, the condition $f\left(F^{-1}\left(\frac{1}{2}\right)\right) > 0$ is needed.

Now take the case of a sample $X_1, \ldots, X_n$ from a uniform distribution on the interval $[\theta - 1, \theta + 1]$ and let $S_n$ be the midrange of this sample, i.e.

$$S_n = \frac{1}{2}\left(\min_{1 \leq i \leq n} X_i + \max_{1 \leq i \leq n} X_i\right).$$

Then the following theorem holds.

**Theorem 3.2** The peakedness of $S_n$ with respect to $\theta$ is strictly increasing in $n$ for $n \geq 2$ and each $\epsilon \in (0, 1)$.

**Proof.** Suppose, without loss of generality, that $\theta = 0$. Then the joint density of $\min_{1 \leq i \leq n} Y_i$ and $\max_{1 \leq i \leq n} Y_i$ at $(x, y)$ is, for $n \geq 2$, given by

$$\frac{n(n-1)}{2^n} (y-x)^{n-2} \quad \left(1 \leq x < y \leq 1\right).$$

So, for $-1 \leq t \leq 0$,

$$P\left(\min_{1 \leq i \leq n} Y_i + \max_{1 \leq i \leq n} Y_i \leq 2t\right) = \frac{n(n-1)}{2^n} \int_{-1}^{t} dx \int_{x}^{2t-x} (y-x)^{n-2} dy = \frac{(1+t)^n}{2}$$

and, for $0 < t \leq 1$,

$$P\left(\min_{1 \leq i \leq n} Y_i + \max_{1 \leq i \leq n} Y_i \leq 2t\right) = 1 - P\left(\min_{1 \leq i \leq n} Y_i + \max_{1 \leq i \leq n} Y_i \leq -2t\right) = 1 - \frac{(1-t)^n}{2},$$

which gives, for $|t| < 1$,

$$P(|S_n| < t) = 1 - (1-t)^n,$$

from which the results follows immediately. □

**Remark**

Note that, in quoting Proschan’s (1965) results, we ask for the distribution function $F$ to have a density $f$ which is logconcave on the support of $F$, while Proschan asks for this density to be a Pólya frequency function of order 2 (PF$_2$). However, it was shown by Schoenberg (1951) that

$$f \text{ is PF}_2 \iff f \text{ is logconcave on the support of } F,$$
so the two conditions are equivalent.

Further note that Ibragimov (1956) showed that, for a distribution function $F$ with a density $f$,

$$f \text{ is strongly unimodal } \iff f \text{ is logconcave on the support of } F,$$

where a density is strictly unimodal if its convolution with all unimodal densities is unimodal. So, the condition of logconcavity of $f$ can also be replaced by the condition of its strict unimodality. For more results on Pólya frequency functions see e.g. Marshall and Olkin (1979, Chapter 18) and Karlin (1968).

4 References


