Biometrika (1985), **72**, 1, pp. 216–8
Printed in Great Britain

The effect of an outlier on L-estimators of location in symmetric distributions

By H. A. DAVID
Statistical Laboratory, Iowa State University, Ames, Iowa 50011, U.S.A.

and J. K. GHOSH
Indian Statistical Institute, Calcutta, India

Summary

The effect is studied of an outlier which has the same symmetric distribution as the other observations except for a change in location and a possible increase in scale. We show that the median is the most bias-resistant estimator, in the class of L-statistics with symmetric nonnegative coefficients that add up to one, for a class of distributions which includes the normal, double-exponential and logistic distributions.

Some key words: Bias; Convexity; Inequality; Order statistic; Robustness.

Let $Z_1, \ldots, Z_n$ be independent variates with finite expectations. Suppose that one of the Z's, we do not know which, represents an outlier. The outlier has distribution function $G(x)$, the other variates have distribution function $F(x)$ which is symmetric about μ and have density $f(x)$.

It is desired to estimate μ, the mean of the target population, in the presence of the unidentified outlier. Let the order statistics formed from the Z's be denoted by $Z_{1:n} \leqslant \ldots \leqslant Z_{n:n}$. As estimators we shall consider in this note the class $\{L_n\}$ of linear functions of order statistics, so-called L-estimators,

$$L_n = \sum_{i=1}^{n} a_i Z_{i:n}, \quad \sum_{i=1}^{n} a_i = 1, \quad a_i = a_{n+1-i} \geqslant 0 \quad (\text{for all } i). \tag{1}$$

We are concerned with the bias of L_n in finite samples. For a general small-sample study of L-estimators, see Rosenberger & Gasko (1983).

We begin by considering the subclass $\{M_{r,n}\}$ of basic estimators

$$M_{r,n} = \tfrac{1}{2}(Z_{n-r+1:n} + Z_{r:n}) \quad (r = [\tfrac{1}{2}n]+1, \ldots, n),$$

which includes the median. Clearly, L_n can be written as a linear function of the $M_{r,n}$, with nonnegative coefficients. It will be shown that under certain conditions $E(M_{r,n})$ is an increasing function of r. From this follows in particular a formal proof of the intuitively appealing result that the median is the most bias-resistant estimator in the class $\{L_n\}$ of (1).

Let $\delta_{r,n} = E(Z_{r+1:n} - Z_{r:n})$. Then $2E(M_{r+1,n} - M_{r,n}) = \delta_{r,n} - \delta_{n-r,n}$. We have, compare David & Groeneveld (1982), taking $\mu = 0$ without loss of generality,

$$\delta_{r,n} - \delta_{n-r,n} = \binom{n}{r} \int_{-\infty}^{\infty} \{1 - G(x) - G(-x)\} F^{r-1}(x) \{1 - F(x)\}^{n-1-r} \{F(x) - r/n\}\, dx. \tag{2}$$

This expression, without the factor $1 - G(x) - G(-x)$, has been studied by David & Groeneveld (1982). From the argument there given, Case 2, it follows that a sufficient condition ensuring $\delta_{r,n}-\delta_{n-r,n}>0$ is that the positive even function

$$R(x)=\{1-G(x)-G(-x)\}/f(x) \tag{3}$$

be increasing in x for $x>0$.

Of special interest is the situation $G(x)=F\{(x-\lambda)/\sigma\}$ for $\lambda>0$, $\sigma>0$ when

$$R(x)=\frac{1}{\sigma}\int_0^{\lambda} r(x,t)\,dt,$$

where

$$r(x,t)=\left\{f\left(\frac{x+t}{\sigma}\right)+f\left(\frac{x-t}{\sigma}\right)\right\}\Big/f(x). \tag{4}$$

Example 1. If $f(x)=(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}x^2}$, then

$$r(x,t)=2\exp\{-\tfrac{1}{2}t^2/\sigma^2+\tfrac{1}{2}x^2(1-1/\sigma^2)\}\cosh(tx\sigma^{-2}).$$

Thus $r(x,t)$, and hence $R(x)$, is an increasing function of x for $x>0$ and $\sigma\geqslant 1$.

The same result is easily shown to hold for the double exponential $f(x)=e^{-|x|}$. Note that, in the special case $\lambda=\infty$, stronger results are possible (David, 1985).

For $\sigma=1$, the location-shift case, there is an interesting connection with hypothesis testing. It is well known (Lehmann, 1959, p. 330) that $f(x-\theta)$ is a monotone likelihood ratio family if and only if $\psi=-\log f$ is convex. Assume this is so and also that ψ has a derivative, ψ'. We continue to take f to be symmetric, so that ψ is symmetric. It can be shown that, under these assumptions,

$$r(x,t)=\{f(x+t)+f(x-t)\}/f(x),$$

being an increasing function of x for $x\geqslant 0$ and $t>0$, is equivalent to the concavity of $\psi'(x)$ for $x\geqslant 0$. It is also easy to show that, for $\sigma>1$, the concavity of $\psi'(x)$ continues to imply the increasing character of

$$r(x,t)=\frac{1}{\sigma}\left\{f\left(\frac{x+t}{\sigma}\right)+f\left(\frac{x-t}{\sigma}\right)\right\}\Big/f(x).$$

Example 2. For the logistic distribution, with density $f(x)=e^{-x}(1+e^{-x})^{-2}$, $\psi(x)$ is convex and $\psi'(x)$ is concave for $x\geqslant 0$, so that (3) is increasing in x for $x>0$, with

$$G(x)=\left[1+\exp\left\{-\left(\frac{x-\lambda}{\sigma}\right)\right\}\right]^{-1}\quad(\sigma\geqslant 1).$$

For the uniform distribution, (3) is not applicable but some direct arguments are possible. The median is no more bias-robust than any other L_n in (1) not involving the extremes.

Acknowledgement

This work was supported by a grant from the U.S. Army Research Office.

References

David, H. A. (1985). Order statistics under non-standard conditions. In *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences*, Ed. P. K. Sen. Amsterdam: North-Holland. To appear.

David, H. A. & Groeneveld, R. A. (1982). Measures of local variation in a distribution: Expected length of spacings and variances of order statistics. *Biometrika* 69, 227–32.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. New York: Wiley.
ROSENBERGER, J. L. & GASKO, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In *Understanding Robust and Exploratory Data Analysis*, Ed. D. C. Hoaglin, F. Mosteller and J. W. Tukey, pp. 297–338. New York: Wiley.

[*Received August* 1984. *Revised October* 1984]