The Annals of Statistics
1980, Vol. 8, No. 4, 931–934

AN EDGEWORTH CURIOSUM[1]

BY STEPHEN M. STIGLER

University of Chicago

Taking the sample mean of a set of measurements need not be uniformly advantageous, even when $n = 2$ and all moments are finite. The examples presented include one given by Edgeworth in 1883.

Let $X_1, X_2, \cdots$ be independent, identically distributed random variables each with density $f(x - \theta)$, where $f(x)$ is assumed continuous and symmetric about zero. It has been "known" for centuries that in such a situation the sample mean $\bar{X}_n = (X_1 + \cdots + X_n)/n$ will give a better estimate of θ than will a single measurement X_1; this was proved as long ago as 1755, and, barring heavy Cauchy-like tails, it is often taken for granted today. (Simpson, 1755; Plackett, 1958). This widely held belief is apparently based on the assumption that the relationship $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X_1)/n$ extends at least qualitatively to reasonable loss functions other than squared error, but it is not strictly true, even for distributions with finite variances. In 1883 Edgeworth showed that one reasonable loss function where it may fail is that which assigns unit loss to errors larger than a given $\varepsilon > 0$, zero loss to smaller errors. In this case the criterion by which an estimate $\hat{\theta}$ is to be judged is $\Pr\{|\hat{\theta} - \theta| \leqq \varepsilon\}$; the larger this probability is, the better.

Edgeworth's example concerned the case $n = 2$. The fact that a single observation from a regular unimodal symmetric continuous density with finite moments of any order, could be preferred to the mean of two, may come as a surprise to some modern statisticians; it would have seemed particularly paradoxical in 1883. The nineteenth century had seen many inductive "proofs" of the superiority of the sample mean that were entirely independent of probability considerations, and they all began with the assumption that the mean of two measurements was the best combination of the two. Other steps in these "proofs" had been questioned, but the starting point had gone unchallenged. "But in Chance, as in other provinces of speculation which have been invaded by mathematics, common sense must yield to symbol." (Edgeworth, 1883). The symbols we employ are only slightly more modern than Edgeworth's.

If two competing estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ have continuous symmetric densities $f_1(x - \theta)$ and $f_2(x - \theta)$, then a simple sufficient condition that $\Pr\{|\hat{\theta}_1 - \theta| \leqq \varepsilon\} > \Pr\{|\hat{\theta}_2 - \theta| \leqq \varepsilon\}$ for some $\varepsilon > 0$ is that $f_1(0) > f_2(0)$. The density of $\bar{X}_2 = (X_1 + X_2)/2$ at θ is $2\int_{-\infty}^{\infty} f^2(x)\,dx$, and the ratio of the density of $\bar{X}_2$ at θ to that of X_1 is

Received January 1979; revised March 1979.

[1]This research was supported by the National Science Foundation, Grant No. SOC 78-01668.

AMS 1970 *subject classifications.* Primary 62G35; secondary 62-02.

Key words and phrases. Divergence, Edgeworth, sample mean, median, Pareto distribution, Hodges-Lehmann estimator, robustness.

$D = 2(f(0))^{-1}\int f^2(x)\,dx$. After Edgeworth we call D a coefficient of "divergence": if $D < 1$, then $\Pr\{|X_1 - \theta| \leqq \varepsilon\} > \Pr\{|\bar{X}_2 - \theta| \leqq \varepsilon\}$ for some $\varepsilon > 0$. If $X_1 - \theta$ has characteristic function $\phi(u)$ we may also write $D = \int_{-\infty}^{\infty}(\phi(u/2))^2\,du / \int_{-\infty}^{\infty}\phi(u)\,du$.

Now if each X_i has a Cauchy distribution, then, as has been known since 1824 (see Stigler 1974), X_1 and $\bar{X}_2$ have the same distribution, and of course $D = 1$. Consideration of two families to which the Cauchy distribution belongs, the symmetric stable laws and the Student's t distributions, shows that "divergence" (ie., $D < 1$) cannot be commonly expected with regular densities with shorter tails than the Cauchy. For a Student's t density with m degrees of freedom, an easy calculation shows

$$D = \frac{2\Gamma\left(\frac{m+1}{2}\right)\Gamma\left(m + \frac{1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma(m+1)}.$$

For $m = 1$, $D = 1$; for $m = 2$, $D = 1.18$, and as $m \to \infty$, D increases toward $2^{\frac{1}{2}}$, its value for normal densities. Similarly, for symmetric stable distributions, $\phi(u) = \exp\{-|u|^\alpha\}$, and an easy calculation gives $D = 2^{-((1-\alpha)/\alpha)}$. Here for $\alpha = 1$, the Cauchy case, $D = 1$, and $D < 1$ only for $\alpha < 1$, distributions with tails heavier than the Cauchy. Other examples which do not exhibit divergence include the symmetric beta distributions or Pearson's Type II, $f(x) = [\beta(p,p)2^{2p-1}]^{-1}(1 - x^2)^{p-1}$, for which $D = 2^{2p-1}\Gamma(2p)[\Gamma(2p-1)]^2/\{\Gamma(4p-2)[\Gamma(p)]^2\}$ decreases toward $2^{\frac{1}{2}}$ as $p \to \infty$, the logistic distribution $f(x) = e^{-x}/(1 + e^{-x})^2$, for which $D = \frac{4}{3}$, and the Laplace or double exponential $f(x) = 2^{-1}\exp(-|x|)$, for which $D = 1$.

Edgeworth (1883) gave a simple example that emphasizes that divergence is not dependent upon heavy tails. Take $f(x) = (k-1)(1 + |x|)^{-k}/2$, a sort of two-sided Pareto distribution. For this density, $D = 1 - (2k-1)^{-1} < 1$ for any $k > 1$, yet all moments up to and including the $(k-2)$nd are finite. This distribution arises naturally as a gamma mixture of Laplace distributions (see Johnson and Kotz, 1970, page 32), and has played a role in Bayesian analyses (e.g., Box and Tiao, 1962). The one-sided version of this distribution was discussed by Pareto in 1896.

The potential difference between the accuracies of $\bar{X}_2$ and X_1 can be seen by considering the barely tractable case $k = 4$, where $f(x) = 1.5(1 + |x|)^{-4}$. Table 1 gives $A = P(|X_1 - \theta| \geqq \varepsilon)$, $B = P(|\bar{X}_2 - \theta| \geqq \varepsilon)$, and B/A for this case. $A \leqq B$ for $0 \leqq \varepsilon \leqq \varepsilon^* \cong .1315$ while $A > B$ for $\varepsilon > \varepsilon^*$. B/A achieves its maximum value of 1.012 at $\varepsilon = .06$. Since the quartiles of the distribution of X_1 are $\pm .26$ and those of the distribution of $\bar{X}_2$ are $\pm .24$, the dominance of X_1 over $\bar{X}_2$ is seen to extend to half the "probable error" of a single measurement. The efficiency of X_1 relative to $\bar{X}_2$ as measured by B/A remains above .95 up to $\varepsilon = .25$.

Another family of distributions which contains divergent members is the family of power densities

$$f(x) = \left[2^{\beta^{-1}+1}\Gamma(\beta^{-1}+1)\right]^{-1}\exp\left[-\tfrac{1}{2}|x|^\beta\right]$$

TABLE 1

The relative performance of X_1 and $\bar{X}_2$ when $f(x - \theta) = 1.5(1 + |x - \theta|)^{-4}$
$A = P(|X_1 - \theta| \geq \varepsilon)$, and $B = P(|\bar{X}_2 - \theta| \geq \varepsilon)$

ε	A	B	B/A
0.00	1.000	1.000	1.000
0.02	.942	.950	1.009
0.04	.889	.898	1.011
0.06	.840	.849	1.012
0.08	.794	.802	1.010
0.10	.751	.757	1.007
0.1315	.690	.690	1.000
0.20	.579	.564	.975
0.2416	.522	.500	.957
0.2599	.500	.474	.948
0.30	.455	.423	.929
0.40	.364	.321	.880
0.50	.296	.246	.832
1.00	.125	.081	.648
2.00	.037	.018	.473
4.00	.008	.003	.360
∞	.000	.000	.250

(Box and Tiao, 1962; Johnson and Kotz, 1970, page 33). This family contains the Laplace distribution ($\beta = 1$), the normal ($\beta = 2$), and as a limiting case as $\beta \to \infty$, the rectangular distribution. It is easy to show that for this family, $D = 2^{1-\beta^{-1}}$, so that $D < 1$ for $\beta < 1$. For these densities, all moments are finite. Here, as is also true with Edgeworth's examples, $\bar{X}_2$ is a relative minimum of the likelihood function. For these examples, Bayesian or conditional confidence (conditioning on $X_1 - X_2$) procedures can produce reasonable interval estimates of θ that consist of unions of two disjoint intervals. From the point of view of robustness, however, this serves to emphasize a dependence on distributional assumptions that cannot be checked with samples of size $n = 2$.

The behavior exhibited by the examples given may seem counterintuitive because of a vague feeling that averaging two measurements should improve the accuracy of the estimate over that available from one measurement alone, particularly if we recall that for $n = 2$ the mean, median, midrange, and almost any other "average" agree! It is, however, easy to show that the value at zero of the density of the median of a sample from a symmetric population is never less than the corresponding value of the population density, if a true median exists, as will be the case when n is odd or fractional order statistics (Stigler, 1977) are allowed. The behavior noted also seems to run counter to Chebychev's inequality, which tells us that if $\mathrm{Var}(X_1) < \infty$, $\Pr\{|\bar{X} - \theta| \leq \varepsilon\} \geq 1 - \mathrm{Var}(X_1)/(n\varepsilon^2)$. The fact that this lower bound increases monotonically is, as the examples show, not a guarantee that the probability bounded does also.

One final curiosity may be noted: if E denotes the Pitman efficiency of the Wilcoxon test relative to the sign test (see Lehmann, 1975, page 380, for example),

then $E = 3D^2/4$. At first glance this is totally unexpected; we have an *asymptotic* efficiency, E, expressed as a constant (with respect to f) multiple of a *small sample* measure of efficiency, D^2. In a personal communication R. R. Bahadur has suggested a way of viewing D that may remove some of the mystery, however. Let $X_1', X_2', X_3', \cdots$ be an independent copy of the sequence $X_1, X_2, \cdots$, let $\hat{\theta}_n =$ median $\{X_1, X_2, \cdots, X_n\}$ and let $\hat{\tau}_n =$ median $\{(X_1 + X_1')/2, (X_2 + X_2')/2, \cdots, (X_n + X_n')/2\}$. Now D^2 is just the asymptotic relative efficiency (ratio of reciprocal asymptotic variances) of $\hat{\tau}_n$ to $\hat{\theta}_n$. The bearing of this fact upon the puzzling relationship $E = 3D^2/4$ is that the Pitman efficiency of the Wilcoxon test relative to the sign test equals the asymptotic efficiency of the related estimates of location, the Hodges-Lehmann estimate $\hat{\tau}_n^* =$ median $\{(X_i + X_j)/2;\ i \leqq j\}$ (see Hodges and Lehmann, 1963) and the sample median $\hat{\theta}_n$. That is, E equals the asymptotic efficiency of $\hat{\tau}_n^*$ relative to $\hat{\theta}_n$. Thus we see that the relationship in question is an expression of the fact that the relative efficiency of $\hat{\tau}_n^*$ to $\hat{\tau}_n$ is $\frac{3}{4}$ for any continuous, symmetric f! Thus, at least in the sense of asymptotic variances, $\hat{\tau}_n^*$ (a median of $n(n+1)/2$ dependent pairwise averages) is asymptotically equivalent to $\hat{\tau}_{.75n}$ (a median of $3n/4$ independent pairwise averages). The full implications of this fact remain to be explored.

Acknowledgment. I thank R. R. Bahadur, Robert Bell, Gouri Bhattacharyya, John W. Tukey, and Chien-Fu Wu for comments. The reader interested in other aspects of Edgeworth's 1883 paper may consult Stigler (1978).

REFERENCES

BOX, G. E. P. and TIAO, G. C. (1962). A further look at robustness via Bayes's theorem. *Biometrika* **49** 419–432.

EDGEWORTH, F. Y. (1883). The law of error. *Philosophical Magazine* (Fifth Series) **16** 300–309.

HODGES, J. L. JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611.

JOHNSON, N. L. and KOTZ, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions – 2*. Houghton Mifflin, Boston.

LEHMANN, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

PLACKETT, R. L. (1958). The principle of the arithmetic mean. *Biometrika* **45** 130–155. (Reprinted in *Studies in the History of Statistics and Probability*. (1970). (E. S. Pearson and M. G. Kendall, Eds.) Griffin, London.

SIMPSON, T. (1755). On the advantage of taking the mean of a number of observations, in practical astronomy. *Phil. Trans.* **49** 82–93.

STIGLER, S. (1974). Cauchy and the witch of Agnesi: an historical note on the Cauchy distribution. *Biometrika* **61** 375–380.

STIGLER, S. (1977). Fractional order statistics, with applications. *J. Amer. Statist. Assoc.* **72** 544–550.

STIGLER, S. (1978). Francis Ysidro Edgeworth, Statistician. (with discussion). *J. Roy. Statist. Soc.* Ser. A **141** 287–322.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
1118 EAST 58TH STREET
CHICAGO, ILLINOIS 60637