The Annals of Statistics
1985, Vol. 13, No. 4, 1418–1431

THE RANDOM AVERAGE MODE ESTIMATOR[1]

BY GARY W. OEHLERT

University of Minnesota

This paper proposes an estimator for the mean of a positive random variable given a sample from that random variable. The positive mean estimation problem is common, but there are few results for this problem in the literature. This paper shows the proposed estimator to be less sensitive to outlying values than the sample mean and determines its asymptotic mean squared error.

1. Introduction. There are many times when a statistician wishes to nonparametrically estimate the mean of a positive random variable; for example, estimating total income tax collected from a sample of returns, or deadwood biomass of a forest from a sample of quadrats. Unfortunately, the mean is a difficult parameter to estimate well, and the sample mean, which is the natural estimate of the population mean, is very nonrobust. Standard robust estimates of location are not designed for this situation and can give poor results. This paper studies the large sample error properties of a new estimator of the population mean, the random average mode estimator, which can be used as an alternative to the sample mean in the positive mean estimation problem, and which is robust in a weak sense.

It is worthwhile to stress that there are situations when the mean is the parameter of choice and other location parameters will not do. This is because these situations use the mean not only as a measure of location but also as a measure of total in the population. Measures of location other than the mean, such as quantiles or averages of quantiles, do not possess this relationship with total and cannot be used in place of some estimate of the mean. Efron (1981) and Collins [in the discussion of Efron (1981)] discuss the point. This is considerably different from the usual robustness setup, where, for example, we want to estimate the center of symmetry of a distribution regardless of the shape of any contamination present. In the present situation, we have positive observations which come from a distribution which is typically asymmetric and heavily skewed to the right. For example, in the case of observations on deadwood in a forest, most observations correspond to some twigs and branches, but very few observations correspond to large fallen trees. These large observations are not outliers in the sense of being "rogue" or suspect, and thus we do not want to disregard them entirely when estimating the mean. In fact, overly down-weighting these values can lead to severe underestimates of the mean.

Received July 1983; revised October 1984.

[1]This paper is taken from the author's 1981 Yale University Ph.D. dissertation, which was supported in part by the Air Force Office of Scientific Research, Contract No. F49620-79-C-0164.

AMS 1980 *subject classifications*. Primary 62G05; secondary 62F35.

Key words and phrases. Nonparametric mean estimation, robust estimation.

When the mean is not needed for its special interpretation, some other measure of location should nearly always be used. In the case of a symmetric distribution with possible contamination, the center of symmetry is the natural parameter. Bickel and Lehmann (1975) discuss alternative location parameters for nonparametric, possibly asymmetric, situations. See also Collins (1976).

Standard robust estimates of location do not perform well in the present case. This is not surprising, as they were not designed for estimating means. Unfortunately, the robustness literature gives little insight into what should be used to estimate the mean. Hampel (1968, 1973), Bickel (1976), and Huber (1981) all mention the problem of estimating the mean and give essentially the same advice: "... there is little choice except to pray and use the sample mean" (Huber 1981). However, the sample mean is unsatisfactory in the present situation for the same reason it is unsatisfactory in the usual estimation problem: The sample mean is not robust. We need an estimator of the mean which will be less influenced by the large values than the sample mean, yet still allow them to have some effect.

Let us formalize the model. We have $X_1, X_2, \dots, X_n$ independent identically distributed (iid) positive random variables with distribution F, mean μ, and finite variance σ^2. Denote the higher moments of X_1, if they exist, by $\mu_k = EX_1^k$. We wish to estimate μ using an estimator that is both robust and accurate. Our concept of robustness is informal and differs from the function analytic approach of Hampel (1968, 1971) and others. (We must, in fact, use some other concept of robustness than Hampel, since the mean is a discontinuous functional on the space of distributions, and hence cannot be estimated robustly in Hampel's sense.) We will say that an estimator is robust if it is less influenced by large values than the sample mean, that is, if the estimator has a sensitivity curve that is sublinear. Note that this definition of robustness still admits unbounded sensitivity curves, and thus is in contrast with the bounded and even redescending influence curves of standard robustness theory. We will measure accuracy of an estimator by its mean square error (MSE). Mean square error is a convenient combination of bias and variability; other measures of accuracy may of course be used.

Any consistent estimator of the mean for this model must be asymptotically equivalent to the sample mean. Within this class of estimates however, the robustness and accuracy properties vary considerably. The random average mode estimator that we propose performs well for small samples, has sublinear sensitivity, and will be shown to have good MSE properties asymptotically.

There are some philosophical difficulties involved in evaluating estimators of the mean. For example, the mean is useful because of its relation to total, but total only makes sense for finite populations. The finiteness of the population reduces the usefulness of asymptotic calculations and makes small sample results more important. Also, much of the difficulty in trying to estimate the mean in a nonparametric framework is inherent in the problem and cannot be removed by clever choice of estimator or criterion. Bahadur and Savage (1956) have derived some important limitations to the estimation of means; minor modifications of their results are appropriate for positive random variables.

The next section defines the random average mode (RAM) estimator and presents several of its properties. These include asymptotic equivalence with the sample mean, a logarithmic sensitivity to large values, and an asymptotic formula for the MSE of RAM. The last section summarizes the RAM estimator, discusses possible extensions, and reports the results of some sampling experiments comparing the RAM estimator with several competitors.

2. The random average mode. Let $e_1, e_2, \dots, e_n$ be iid mean 1 exponential random variables independent of the sample $X_1, X_2, \dots, X_n$. (We will denote the observed value of the random variable X_i by x_i.) Conditional on the sample, the random variable $u = \sum_{i=1}^n x_i e_i$ is unimodal. For $n > 1$, the RAM estimate is the mode of $u/(n-1)$; for $n = 1$, RAM is just x_1. RAM is not a randomized estimator; it is a function of the xs expressed in a random-looking fashion.

Consideration of the RAM estimator can be heuristically motivated by a number of arguments. If we take $y_0 = 0$, $y_{n+1} = \infty$, and $y_1 < y_2 < \cdots < y_n$ the ordered observations, then $\mu = \sum_{j=1}^{n+1} c_{j-1}[F(y_j) - F(y_{j-1})]$, where c_{j-1} is the conditional mean of X_1 on $[y_{j-1}, y_j)$. Neither the subinterval means nor probabilities are known, but the subinterval means may be approximated by the left endpoints, and for continuous F, the subinterval probabilities have a joint Dirichlet distribution. Making these approximations, the mean should be close to a known approximating random variable. For theoretical and practical simplicity, a further approximation is made, and the Dirichlet random variables are replaced by independent scaled exponentials. Now the mean is approximated in distribution by the random variable u defined above; to estimate the mean, take the mode of the approximating random variable.

There is also a Bayesian interpretation for RAM. Suppose there is a Dirichlet prior $D(\alpha(x))$ on the distribution F (Ferguson, 1973). If the prior is diffuse, $\alpha \equiv 0$, then μ given the observations is distributed as $\sum_{i=1}^n x_i s_i$, where the s_i are distributed uniformly on the positive simplex summing to 1. Since for large n the simplex random variables are close in distribution to independent scaled exponentials, the RAM estimate is a convenient approximation to the mode of the posterior distribution of the mean. Breth (1979) has used similar formulations to get Bayesian nonparametric lower confidence bounds for the mean.

From this Bayesian point of view, it is possible to see that RAM is a cousin of resampling schemes such as the jackknife, bootstrap, and random subsampling. Each of these resampling schemes takes the data and some function of the data and creates some new data called pseudovalues, bootstrapped values, or subsample values. The distribution of these new values can often be expressed in terms of a posterior distribution given the data (see Efron, 1981), and inference on the parameter of interest is based on this posterior distribution, either analytically or via Monte Carlo techniques. For example, inference on the mean for the bootstrap is based on the distribution $(1/n)\sum_{i=1}^n x_i m_i$, where $(m_1, m_2, \dots, m_n)$ have a joint symmetric multinomial distribution with marginal means equal to one. The RAM estimator, in essence, performs an analytic resampling, and inference is made using the mode of the distribution of the resampled values, rather than the mean.

Let the random variable $u = \sum_{i=1}^{n} x_i e_i$, conditional on the xs, have density $h(u; x_1, x_2, \ldots, x_n)$, abbreviated $h_n(u)$ when there is no confusion about the sample. In the case where all the xs are positive and distinct, the density may be expressed:

$$h(u; x_1, x_2, \ldots, x_n) = \sum_{j=1}^{n} x_j^{n-2} \prod_{k \neq j} (x_j - x_k)^{-1} e^{-u/x_j}$$

(Johnson and Kotz, 1970, page 222). There are a number of properties of the RAM estimator that can be deduced directly from expressions for the density.

PROPOSITION. (*a*) *The random variable u has a unique mode, and the RAM estimator is thus well defined.*

(*b*) *The mode of u is a strictly increasing function of each* x_i.

(*c*) *For ordered distinct observations* $0 < y_1 < y_2 < \cdots < y_n$ *and* $n > 1$,

$$\frac{(n-1)\mathrm{RAM}}{y_{n-1}\log(y_n)} \to 1 \quad \text{as } y_n \to \infty,$$

and the other y's *are held fixed.*

PROOF. (a) For $n = 1$ or 2, the mode of u is expressed in closed form as 0 or $x_1 x_2(\log x_2 - \log x_1)/(x_2 - x_1)$, respectively. For $n \geq 3$, express the density

$$h_n(u) = \int_0^u h_{n-1}(u - z) e^{-z/x_n} \frac{dz}{x_n}.$$

Note that the density is zero at zero for $n > 1$, and differentiate with respect to u to get

$$h_n'(u) = \frac{1}{x_n} e^{-u/x_n} \int_0^u h_{n-1}'(v) e^{v/x_n}\, dv.$$

Hence there is at most one sign change in h_n', and part (a) is true by induction on n.

(b) Suppose that $h_n'(u_0) = 0$ and that $h_{n-1}'(z_0) = 0$. Then

$$0 = \int_0^{z_0} h_{n-1}'(v) e^{v/x_n}\, dv + \int_{z_0}^{u_0} h_{n-1}'(v) e^{v/x_n}\, dv,$$

where the first term is positive and the second is negative. If $\tilde{x} > x_n$, then $h'(u_0; x_1, x_2, \ldots, x_{n-1}, \tilde{x})$ is $1/\tilde{x} \exp(-u_0/\tilde{x})$ times

$$\int_0^{z_0} h_{n-1}'(v) e^{v/x_n} p(v)\, dv + \int_{z_0}^{u_0} h_{n-1}'(v) e^{v/x_n} p(v)\, dv,$$

where $p(v) = \exp[v(1/\tilde{x} - 1/x_n)]$ is a strictly decreasing positive function. This implies that $h'(u_0; x_1, x_2, \ldots, x_{n-1}, \tilde{x}) > 0$, which implies that RAM is increasing in each x_i.

(c) We first show that RAM is unbounded as $y_n \to \infty$. By dominated convergence for $n > 2$ as $y_n \to \infty$,

$$y_n h'(u; y_1, \dots, y_n) \to \int_0^u h'(u - v; y_1, \dots, y_{n-1})\, dv$$
$$= h(u; y_1, \dots, y_{n-1}) > 0.$$

Thus for any u, RAM will eventually be greater than that u as $y_n \to \infty$. RAM is $1/n - 1$ times the u that solves the equation

$$\begin{aligned} 0 &= \sum_{j=1}^{n} \frac{y_j^{n-3}}{\prod_{k \neq j} (y_j - y_k)} e^{-u/y_j} \\ &= \left(1 + O\left(\frac{1}{y_n}\right)\right) + \sum_{j=1}^{n-1} y_n c_j \left(1 + O\left(\frac{1}{y_n}\right)\right) e^{-u[y_j^{-1} - y_n^{-1}]} \\ &= \left(1 + O\left(\frac{1}{y_n}\right)\right) + y_n c_{n-1} \left(1 + O\left(\frac{1}{y_n}\right)\right) e^{-u[y_{n-1}^{-1} - y_n^{-1}]} \left(1 + O\left(e^{-u[y_{n-2}^{-1} - y_{n-1}^{-1}]}\right)\right), \end{aligned}$$

for constants $c_j = -y_n^{n-3}/\prod_{k \neq j,\, k < n} (y_j - y_k)$, where the last step is valid because u is unbounded as $y_n \to \infty$. Taking logs, we get

$$O\left(\frac{1}{y_n}\right) = \log(y_n) + \log(c_{n-1}) - u\left[y_{n-1}^{-1} - y_n^{-1}\right] + O\left(e^{-u[y_{n-2}^{-1} - y_{n-1}^{-1}]}\right).$$

Solving for u, we get result (c) for $n > 2$. For $n = 2$, (c) follows immediately from the closed form given for RAM in (a). □

This logarithmic sensitivity to large values qualifies RAM as robust according to our criterion, though it is a weak form of robustness. The properties we have deduced so far imply that RAM behaves as we would like it to behave; RAM increases as each x value increases, but increases more and more slowly as an x value begins to separate itself from the other values. Unfortunately, it is not the case that as two x values increase they will have logarithmic influence. The RAM estimator is linear in the second largest observation. Sampling experiments show that this is not usually a serious problem in practice; we will return to this topic in Section 3 when we discuss a generalization of the RAM estimator.

We have stated that asymptotically any estimator of the mean must be equivalent with the sample mean $\bar{x}$. We now show that this is the case for RAM.

THEOREM 1. *For iid nonnegative finite variance observations $X_1, X_2, \dots, X_n$,*

$$n^{1/2}(RAM - \bar{x}) \to 0,$$

almost surely.

PROOF. Since $E|X_1|^{2/3}$ is finite, $\lim_{n\to\infty} n^{-3/2}\sum_{i=1}^n x_i^3$ must equal zero almost surely (Loeve 1977, page 255). Thus

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^n x_i^3}{\left(\sum_{i=1}^n x_i^2\right)^{3/2}} = \limsup_{n\to\infty} \frac{n^{-3/2}\sum_{i=1}^n x_i^3}{\left(\frac{1}{n}\sum_{i=1}^n x_i^2\right)^{3/2}} = 0, \quad \text{a.s.,}$$

and

$$\frac{\sum_{i=1}^n x_i e_i - \sum_{i=1}^n x_i}{\left(\sum_{i=1}^n x_i^2\right)^{1/2}} \to_D N(0,1),$$

conditional on the x sequence, for almost all sequences. Continuing,

$$n^{1/2}\left(\sum_{i=1}^n x_i \frac{e_i}{n-1} - \bar{x}\right) \to_D N(0, \mu_2),$$

conditional on the xs. Now both sides of the last expression are random variables with unique modes. Since they are converging in distribution to each other, their modes must converge also, and we get the result. □

We now turn our attention to the mean square error of the RAM estimator. We would like this MSE to be substantially less than the MSE of the sample mean, though this is possible only when the sample size is small. What we show now is that RAM has an MSE which is σ^2/n plus an order n^{-2} correction term which is a function of the moments of X_1 and is usually negative for distributions with long right tails, plus an $o(n^{-2})$ error term. We prove this under the condition that $Ee^{tX_1} < \infty$ for all t in some neighborhood of zero. In fact, $EX_1^k < \infty$ for a finite (though rather large) k is sufficient for the proof of our result, but the moment generating function assumption streamlines the proof and will be retained.

THEOREM 2. *Let $X_1, X_2, \dots, X_n$ be iid nonnegative random variables with positive mean. Assume $Ee^{tX_1} < \infty$ for all t in some neighborhood of zero. Then the MSE of the RAM estimator can be expressed*

$$MSE = \frac{\sigma^2}{n} + \left(n^2\mu_2^2\right)^{-1}\left[3\mu_3^2 - 2\mu_4\mu_2 - 2\mu_1\mu_2\mu_3 + 2\mu_2^3 - \mu_1^2\mu_2^2\right] + o(n^{-2}).$$

The details of the proof of the theorem are tedious, but the idea is simple. For "good" x sequences, the mode of h_n is very close to the mode of an approximating Edgeworth expansion. The MSE of the mode of the Edgeworth expansion as an estimator of the mean can be found by the delta method and is equal to the

MSE given in the statement of the theorem. Finally, "bad" x sequences occur with probability exponentially small in n, so that their contribution can be ignored.

Before proving the theorem, we state a few lemmas. Proofs of lemmas will be omitted when obvious. We will use the notation $S_k = \sum_{i=1}^n x_i^k$ and $c_{j,n} = S_j/(S_2)^{j/2}$.

LEMMA 1. *For positive numbers* $x_1, x_2, \ldots, x_n$,

(1) $c_{j,n} \geq c_{j+1,n}$,

(2) *for* $j > 1$, $1 \geq c_{j,n} \geq n^{1-j/2}$.

PROOF. The Cauchy–Schwartz and Jensen inequalities. □

The good sets where the approximations will work are defined as follows. For $\varepsilon > 0$, say that conditions A hold if both the following inequalities hold:

$$\left|\frac{1}{n}S_k - \mu_k\right| < \varepsilon \quad \text{for } k = 1, 2, \ldots, 7, \tag{A1}$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} I(x_i \leq \mu) - F(\mu)\right| < \varepsilon. \tag{A2}$$

LEMMA 2. *For given* ε *and* xs *as in Theorem* 2, *the probability that conditions* A *do not hold decreases exponentially with* n.

PROOF. The result follows from the first Bonferroni inequality and standard large deviations results (Feller, 1971, page 549). □

LEMMA 3. (*Johnson and Rodgers*, 1951). *For any unimodal distribution with unique mode* u, *mean* μ, *and variance* σ^2, $(u - \mu)^2 \leq 3\sigma^2$ *holds*.

We next derive the specific Edgeworth expansions we will use in the proof of Theorem 2. The proof of the uniform approximation is based on Feller (1971, page 533), but we show it in some detail because we need explicit bounds on the error. Let $H_j(u)$ be the jth Hermite polynomial defined by

$$D_j e^{-u^2/2} = (-1)^j H_j(u) e^{-u^2/2}.$$

LEMMA 4. *Let* $X_1, X_2, \ldots, X_n$ *be as in Theorem* 2, *and suppose* $e_1, e_2, \ldots, e_n$ *are iid exponential random variables independent of the* Xs. *Let* $f_n(u)$ *be the density of* $u = \sum_{i=1}^n x_i(e_i - 1)(S_2)^{-1/2}$, *conditional on the* xs. *If conditions* A *hold* *for* $\varepsilon > 0$, *then*

$$\left| f_n(u) - \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\left[1 + \frac{1}{3}c_{3,n}H_3(u) + \frac{1}{4}c_{4,n}H_4(u) + \frac{1}{5}c_{5,n}H_5(u) + \frac{1}{6}c_{6,n}H_6(u) + \frac{1}{18}c_{3,n}^2H_6(u) + \frac{1}{12}c_{3,n}c_{4,n}H_7(u) + \frac{1}{162}c_{3,n}^3H_9(u) + \frac{1}{15}c_{3,n}c_{5,n}H_8(u) + \frac{1}{32}c_{4,n}^2H_8(u) + \frac{1}{72}c_{3,n}^2c_{4,n}H_{10}(u) + \frac{1}{1944}c_{3,n}^4H_{12}(u)\right]\right| < B_1 n^{-5/2} + B_2 e^{-n^r},$$

for positive constants B_1, B_2, *and* r *depending on* ε *and the moments of* X, *but not the particular* x *sequence.*

PROOF. The difference between the densities is uniformly less than the integral of the absolute difference of their characteristic functions. Split the interval of integration into $|t| < \frac{1}{2}c_{3,n}^{-1/3}$ and the complement of this interval. Under conditions A, the end points of the interval go to infinity at least as fast as Rn^p for some $0 < p < \frac{1}{2}$, and real constant R. On the outer interval, we only need to bound the absolute integrals of each characteristic function. For f_n, this is

$$\int_{|t|\geq Rn^p} \prod_{j=1}^{n}\left(1 + \frac{t^2x_j^2}{S_2}\right)^{-1/2} dt \leq \int_{|t|\geq Rn^p} \prod_{j:\, x_j\geq\mu}\left(1 + \frac{t^2x_j^2}{S_2}\right)^{-1/2} dt$$
$$\leq \left(1 + \frac{R^2n^{2p}\mu^2}{S_2}\right)^{-1/2[\sum_{j=1}^n I(x_j\geq\mu)-2]} \times \int_{-\infty}^{\infty}\left(1 + \frac{t^2\mu^2}{S_2}\right)^{-1} dt.$$

Under conditions A, the last integral is bounded by a constant times $n^{1/2}$ and $\sum_{j=1}^n I(x_j \geq \mu)$ must be going to infinity as fast as n, so this term is bounded by $\frac{1}{2}B_1e^{-n^r}$. Since the $c_{j,n}$ are bounded, the integral of the Edgeworth characteristic function on the outer interval is also bounded by a term $\frac{1}{2}B_1e^{-n^r}$.

On the inner interval, $tx_j(S_2)^{-1/2}$ is less than $\frac{1}{2}$, so the log characteristic function of f_n may be expanded in a Taylor series, rearranged, and back exponentiated to be expressed

$$e^{-t^2/2}\left(1 + \left[\frac{1}{3}(it)^3c_{3,n} + \frac{1}{4}(it)^4c_{4,n} + \cdots\right] + \frac{1}{2!}\left[\frac{1}{3}(it)^3c_{3,n} + \frac{1}{4}(it)^4c_{4,n} + \cdots\right]^2 + \cdots\right).$$

We may bound certain sums appearing in the characteristic function. Since $|tx_j/(S_2)^{1/2}| < \frac{1}{2}$, we have

$$\begin{aligned}
&\left|\frac{1}{k}(it)^k c_{k,n} + \frac{1}{k+1}(it)^{k+1} c_{k+1,n} + \cdots\right| \\
&= \left|\sum_{j=1}^{n}\left[\frac{1}{k}(it)^k \frac{x_j^k}{(S_2)^{k/2}} + \frac{1}{k+1}(it)^{k+1}\frac{x_j^{k+1}}{S_2^{(k+1)/2}} + \cdots\right]\right| \\
&\le \frac{2}{k}|t|^k c_{k,n}.
\end{aligned}$$

Further, $|t|^3 c_{3,n} < 1$ on the inner interval, so geometric sums with this term are bounded. Using the these facts, we find (after some algebra) that the absolute difference of the two characteristic functions on the innter interval is bounded by

$$\begin{aligned}
e^{-t^2/2}\Big[&\tfrac{2}{7}|t|^7 c_{7,n} + \tfrac{1}{9}|t|^9 c_{3,n}c_{6,n} + \tfrac{1}{10}|t|^9 c_{4,n}c_{5,n} \\
&+ \tfrac{2}{25}|t|^{10} c_{5,n}^2 + \tfrac{1}{45}|t|^{11} c_{3,n}^2 c_{5,n} + \tfrac{1}{24}|t|^{11} c_{3,n} c_{4,n}^2 \\
&+ \tfrac{1}{48}|t|^{12} c_{4,n}^3 + \tfrac{1}{324}|t|^{13} c_{3,n}^3 c_{4,n} + \tfrac{1}{144}|t|^{14} c_{3,n}^2 c_{4,n}^2 \\
&+ \tfrac{1}{144}|t|^{15} c_{3,n} c_{4,n}^4 + \tfrac{1}{384}|t|^{16} c_{4,n}^4 + \tfrac{1}{40}\left[\tfrac{2}{3}|t|^3 c_{3,n}\right]^5\Big].
\end{aligned}$$

Since conditions A hold, the coefficient of each $|t|^k$ term in this expression is bounded by $B_2 n^{-5/2}$, where B_2 is independent of the xs. Integrating over the interval we conclude that the lemma is true. □

Using the same methods of proof, we get Lemma 5.

LEMMA 5. *Under the assumptions of Lemma* 4,

(a) $|f_n'' - \phi''| \le B_3 n^{-1/2} + B_4 e^{-n^r}$,

(b) $|f_n - \phi| \le B_5 n^{-1/2} + B_6 e^{-n^r}$,

where ϕ is the standard normal density, and the Bs are independent of the xs.

PROOF OF THEOREM 2. Let u_0 be the mode of the f_n density in Lemma 4. Then RAM $= (n/n-1)[\bar{x} + (1/n)(S_2)^{1/2}u_0]$. Let a_0 be the mode of the approximating Edgeworth density g_n, and let Ra be the estimator we get using a_0 instead of u_0. For this approximation,

$$\begin{aligned}
E(\text{RAM} - Ra)^2 &= O(n^{-2})E\left[S_2(u_0 - a_0)^2\right] \\
&\le O(n^{-1})\left[E\left(\frac{1}{n}S_2\right)^2 E(u_0 - a_0)^4\right]^{1/2} \\
&= O(n^{-1})\left[E(u_0 - a_0)^4\right]^{1/2}.
\end{aligned}$$

By Lemma 3, $|u_0| \le \sqrt{3}$. The solution a_0 is a root of a degree 13 polynomial. From the form of this polynomial and Lemma 1, we can infer that all roots are bounded by Bn, so that $|u_0 - a_0| \le Bn$. (We will use B as a generic positive constant that does not depend on the xs.) When conditions A hold, we may get a much tighter bound on $|u_0 - a_0|$. By Lemma 5, $|u_0| \le Bc_{3,n}^{1/2}$, and for some $p > 0$ and large n, $f_n''(u) < -p$ for $|u| < p$. By Lemma 4, $|g_n(u) - f_n(u)| \le Bn^{-5/2}$, which, combined with Taylor's theorem, gives

$$f_n(u_0) - Bn^{-5/2} \le g_n(a_0) \le f_n(u_0) - p(u_0 - a_0)^2 + Bn^{-5/2},$$

which implies that $(u_0 - a_0)^2 \le Bn^{-5/2}$ for large n. Thus $E(u_0 - a_0)^4 = O(n^{-5}) + O(n^4)\Pr$ (conditions A do not hold) $= O(n^{-5})$ by Lemma 2. Hence, $E(\mathrm{RAM} - Ra)^2 = O(n^{-7/2})$. This implies that

$$\begin{aligned}\mathrm{MSE(RAM)} &= E(\mathrm{RAM} - Ra + Ra - \mu)^2 \\ &= E(Ra - \mu)^2 + 2E[(Ra - \mu)(\mathrm{RAM} - Ra)] + E(Ra - \mathrm{RAM})^2 \\ &= E(Ra - \mu)^2 + o(n^{-2}),\end{aligned}$$

using Cauchy–Schwartz to bound the cross product term.

The mode a_0 is the root of a degree 13 polynomial. If we write $a_0 = -c_{3,n} + d$, then, when conditions A hold, we must have that

$$0 = d\left[1 + O\left(\frac{1}{n}\right)\right] + BO\left(\frac{1}{n^{3/2}}\right) + a_0^2 BO\left(\frac{1}{n^{1/2}}\right) + \cdots + a_0^{13} BO\left(\frac{1}{n^2}\right).$$

Since under conditions A, a_0 itself is $O(n^{-1/2})$, we must have that $|d| < Bn^{-3/2}$, for n sufficiently large.

Consider a third estimate Rb which uses $-c_{3,n}$ instead of a_0. Then

$$E(Rb - Ra)^2 \le O(n^{-1})[E(d^4)]^{1/2}.$$

Since $|d| < Bn^{-3/2}$ under conditions A, and $|d| < Bn$ otherwise, we have as before that

$$E(\mathrm{RAM} - \mu)^2 = E(Rb - \mu)^2 + o(n^{-2}).$$

The estimate Rb can be expressed as a rational function of the sample:

$$Rb = \frac{1}{n-1}S_1 - \frac{1}{n-1}S_3/S_2.$$

We need the expected value of $(Rb - \mu)^2$; this may be found by using the delta method (Oehlert, 1983). To use the delta method, we must find polynomials $P1$ and $P2$ in the sample moments such that $n^2|(Rb - \mu)^2 - P1| \to 0$, in probability, $n^2|(Rb - \mu)^2 - P1| < P2$, and $EP1^{1+\delta}$ and $EP2^{1+\delta}$ are finite for some positive δ. Under these conditions, $n^2E[(Rb - \mu)^2 - P1] \to 0$. Let $v_j = n^{-1/2}\sum_{i=1}^n (x_i^j - \mu_j)$. After some algebra, the approximating polynomial is found to be

$$P1 = \frac{v_1^2}{n} - \frac{2v_1}{n^{3/2}}\left\{\frac{\mu_3}{\mu_2} + \frac{v_3}{\mu_2\sqrt{n}} - \frac{\mu_3 v_2}{\mu_2^2\sqrt{n}}\right\} + \frac{1}{n^2}\frac{\mu_3^2}{\mu_2^2}$$

$$+\frac{2v_1^2}{n^2} + \frac{2\mu_1 v_1}{n^{3/2}} - \frac{2\mu_1\mu_3}{\mu_2 n^2} + \frac{\mu_1^2}{n^2}.$$

Rb lies between zero and $\bar{x}$, so that $\bar{x}^2 + \mu^2$ bounds the squared error and $P2$ may be constructed from this bounding polynomial and $P1$. All moments exist due to the moment generating function assumption. Following the chain of equalities back, we find that

$$\begin{aligned}\text{MSE(RAM)} &= E(Rb - \mu)^2 + o(n^{-2}) \\ &= \frac{\sigma^2}{n} + (n^2\mu_2^2)^{-1}\left[3\mu_3^2 - 2\mu_4\mu_2 - 2\mu_1\mu_2\mu_3 + 2\mu_2^3 - \mu_1^2\mu_2^2\right] \\ &\quad + o(n^{-2}),\end{aligned}$$

as required, and the theorem is proved. □

3. Summary and remarks. The mean is important due to its connection with total in a population or interpretation as a long-run sum of observations, but there is little guidance in the literature as to how to estimate a mean. The most obvious estimate of the mean is the sample mean, but the sample mean is not robust and will occasionally err drastically due to the influence of a few points. A second reasonable possibility is an asymmetric trimmed mean, with a small fraction of the large observations not included in the mean. Symmetric trimmed means work well in the standard robustness problem, and we might suggest (symmetric or asymmetric) trimmed means for the positive mean estimation problem. However, the bias introduced using trimmed means is typically much larger than the reduction in variance obtained (Oehlert, 1981), so trimmed means cannot be recommended wholeheartedly. One possible solution to the problem is to employ some form of stratified sampling scheme, rather than sampling from the whole population at once. This enables the researcher to concentrate on the tail, but this type of sampling is not always possible. The random average mode is an alternative estimator which can be used when the mean must be estimated and fancy sampling is not possible. The RAM estimator is influenced only logarithmically by a large value and has good MSE properties for distributions with long right tails.

One important criticism of the RAM estimator is that while it is reasonably robust against one large value, it is not robust against two large values. RAM is robust against one large value because the mode of an exponential random variable is at zero, and not robust against two large values because the mode of the sum of two exponentials is larger than zero. This suggests that the RAM estimator be generalized to be the mode of $\sum_{i=1}^{n} x_i d_i$, where the d random variables are iid and independent of the xs. Different choices of distribution for the ds will lead to different properties for the resulting modal estimator; some choices of distribution for d may lead to estimators which are robust against more than one large value. The only requirement on the ds is that the convolution be unimodal. For example, analogy suggests that letting the d_i be iid gamma random variables with shape parameter $n^{-\alpha}$, $0 < \alpha < 1$, would result in an estimator which was sublinearly sensitive to the $[n^{\alpha}]$ largest observations. A second method of increasing the robustness of the RAM estimator is to compute an estimate in stages. For example, for n points, compute the RAM estimate of the mean of the smallest $n - 1$; call this value $R1$. Then compute the RAM estimate of $R1$ repeated $n - 1$ times and the largest original value. This leads to a logarithmic sensitivity to the two largest observations.

The RAM estimator and the bootstrap distribution for the mean are more than superficially similar; both make a correction of the same order to the simple normal approximation to the distribution of the mean. Singh (1981) has shown that the bootstrap distribution for the mean corresponds to an Edgeworth approximation with a skewness correction. (See also Bickel and Freedman 1981.) Our proof of the MSE of the RAM estimator showed that asymptotically RAM was equivalent to the mode of the corresponding Edgeworth approximation with skewness correction.

Finally, we briefly report on some small sample experiments which compare the RAM estimator with three alternatives. [See Oehlert (1981) for a more complete discussion.] The estimators will be evaluated by empirically calculating their mean squared errors (based on 500 trials) for three sample sizes, $n = 10, 30, 100$, and three distributions: the exponential and two exponential mixtures $[(1 - p)\exp(1) + p\exp(5)$, for $p = 0.1, 0.05]$. The three alternative estimators are the sample mean $\bar{x}$, a trimmed mean formed by taking the mean of the $n - 1$

TABLE 1
Scaled mean squared errors

n	$\bar{x}$	shrunken mean	trimmed mean
	Distribution: $exp(1)$		
10	1.064	0.983	1.198*
30	1.009	0.984	1.176*
	Distribution: $0.9\,exp(1) + 0.1\exp(5)$		
10	1.650*	1.108	1.203*
30	1.283*	1.048	1.215*
100	1.085	1.011	1.214*
	Distribution: $0.95\,exp(1) + 0.05\exp(5)$		
10	1.573*	1.067	1.226*
30	1.414*	1.090*	1.209*
100	1.234*	1.095*	1.135*

$^*p < 0.01$.

smallest observations, and an adaptively shrunken mean $\bar{x}^3/(\bar{x}^2 + s^2/n)$. [The optimal shrinking factor of $\bar{x}$ toward zero is $\mu^2/(\mu^2 + \sigma^2/n)$.]

To facilitate comparison, for every sample size and distribution the MSE of each of the alternative estimators is divided by the MSE of the RAM estimator. The ratios are displayed in Table 1, and ratios which are significantly different from 1 (two-sided, pairwise t test on the squared errors, $\alpha = 0.01$) are marked with an asterisk.

Table 1 shows that the sample mean and the trimmed mean have significantly larger MSEs than RAM for these kinds of asymmetric distributions. The adaptive shrunken mean can compete with RAM in terms of MSE, but only on the shorter tailed distributions.

The asymptotic MSEs of the RAM estimator for these distributions are $1/n + O/n + o(n^{-2})$ for the exponential, $4.84/n - 49.6/n^2 + o(n^{-2})$ for the 0.1 mixture, and $2.96/n - 78.3/n^2 + o(n^{-2})$ for the 0.05 mixture. For $n = 100$, the observed MSEs for RAM on the mixtures were 0.0459 and 0.0252, respectively, about halfway between σ^2/n and the order n^{-2} correction.

Acknowledgments. This paper is taken from the author's 1981 Yale Ph.D. dissertation. The author would like to thank John Hartigan for his guidance and encouragement, and the referees for some helpful comments.

REFERENCES

BAHADUR, R. R. and SAVAGE, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.* **27** 1115–1122.

BICKEL, P. J. (1976). Another look at robustness: a review of reviews and some new developments. *Scand. J. Statist.* **3** 145–168.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BICKEL, P. J. and LEHMANN, E. L. (1975). Descriptive statistics for nonparametric models, I, II. *Ann. Statist.* **3** 1038–1069.

BRETH, M. (1979). Nonparametric Bayesian interval estimation. *Biometrika* **66** 641–644.

COLLINS, J. R. (1976). Robust estimation of a location parameter in the presence of asymmetry. *Ann. Statist.* **4** 68–85.

EFRON, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canad. J. Statist.* **9** 139–172.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**, 2nd ed. Wiley, New York.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. Ph.D. dissertation, Univ. of California, Berkeley.

HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1877–1896.

HAMPEL, F. R. (1973). Robust estimation: a condensed partial survey. *Z. Wahrsch. verw. Gebiete* **27** 87–104.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

JOHNSON, N. L. and KOTZ, S. (1970). *Continuous Univariate Distributions II*. Houghton-Mifflin, Boston.

JOHNSON, N. L. and RODGERS, C. A. (1951). The moment problem for unimodal distributions. *Ann. Math. Statist.* **22** 433–439.

LOEVE, M. (1977). *Probability Theory I*. Springer, New York.

OEHLERT, G. W. (1981). Estimating the mean of a positive random variable. Ph.D. dissertation, Yale Univ., New Haven.

OEHLERT, G. W. (1983). A note on the delta method. Unpublished manuscript.

SINGH, K. (1981). On the accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195.

DEPARTMENT OF APPLIED STATISTICS
SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
ST. PAUL, MINNESOTA 55108