Statistica Sinica 1(1991), 137-158

# SIMPLE ESTIMATORS FOR THE MEAN OF SKEWED POPULATIONS

Wayne A. Fuller

*Iowa State University*

*Abstract:* Simple estimators of the mean are developed and investigated. The Weibull is used as the distributional model for the tail of the observed distribution. No assumption is made about the left portion of the distributions. It is proven that the once-Winsorized mean is superior to the sample mean for Weibull populations with shape parameter greater than one. Estimators for the mean based upon a simple preliminary test for the exponential distribution are illustrated.

*Key words and phrases:* Skewed distributions, robust estimation, outliers, Weibull, survey sampling.

## 1. Introduction

The problem of estimating the population mean from a sample containing a few "very large" observations has been faced by most sampling practitioners. The definition of "very large" must itself be part of a study of estimation for such samples. An individual practitioner may specify values as "outlier" or "unusual" on the basis of experience. However, the definition of "very large" that appears most useful to the statistician is a definition that separates cases wherein the sample mean performs well as an estimator from those cases wherein alternative estimators are markedly superior to the mean.

In the experimental situation, it is common to reject a large observation on the basis that the unusual observation is the result of contamination or of errors in procedure. Early studies of outliers include Dixon (1950), Anscombe (1960, 1961), Veale and Huntsberger (1969), and Kale and Sinha (1971). The books by Barnett and Lewis (1984) and Hawkins (1980) treat the general problem of outlier detection.

Tukey (1962) suggested the consideration of "longer-tailed" distributions as an explanation of outliers. Tukey and McLaughlin (1963) and Dixon and Tukey (1968) studied an estimator suggested by Charles Winsor and called by Tukey the Winsorized mean. In its general form this estimator replaces the largest $r$ observations by the $(r+1)$st largest observation and replaces the $s$ smallest observations by the $(s+1)$st smallest.

A large body of literature has appeared on robust estimation in the last twenty years. Robust techniques are discussed in the texts by Huber (1981) and Hampel, et al. (1986). Chambers (1986) studied robust estimators of the superpopulation mean.

Bershad (1961) investigated estimators for the mean of a finite population using a fixed number, $A$, to divide the population into two groups. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ be the ordered observations in a sample of size $n$. Bershad considered the estimator of the finite population total defined by

$$Nn^{-1}\sum_{i=1}^{n-r} X_{(i)} + C \sum_{i=n-r+1}^{n} X_{(i)},$$

where the first summation is over sample values less than $A$ and the second summation is over the sample values greater than $A$. The values of $A$ and $C$ which minimize the mean square error are a function of the population mean and variance for each of the two groups. Bershad demonstrated that gains can be made for a rather wide choice of $A$ and $C$. Hidiroglou and Srinath (1981) and Ernst (1980) also investigated estimators that give a different weight to the largest observations. These estimators are called reweighted estimators by some authors.

As a further illustration of estimators of the reweighting type, let $n-r$ of the units in a sample of size $n$ be less than a specified value, $A$, and let $r$ of the units be greater than $A$. Consider the estimator of the infinite population mean

$$\widetilde{\mu} = a(n-r)^{-1}\sum_{i=1}^{n-r} X_{(i)} + (1-a)r^{-1} \sum_{i=n-r+1}^{n} X_{(i)}, \tag{1.1}$$

where $X_{(i)}$ are the ordered observations and $a$ is to be determined. If $a$ depends only on $r$, it is possible to show that the $a(r)$ that minimizes the mean square error of $\widetilde{\mu}$ is

$$a(r) = \frac{(n-r)\sigma_2^2 + r(n-r)p_1(\mu_1-\mu_2)^2}{r\sigma_1^2 + (n-r)\sigma_2^2 + r(n-r)(\mu_1-\mu_2)^2}, \tag{1.2}$$

where $(\mu_1, \sigma_1^2)$ is the mean and variance of the population of elements less than $A$, $(\mu_2, \sigma_2^2)$ is the mean and variance of the population of elements greater than $A$, and $p_1$ is the proportion of the population less than $A$. In order to implement the estimator $\widetilde{\mu}$, it is necessary to use outside information to specify the value for $a(r)$. This requires the practitioner to have some prior knowledge of the parameters of the population. Hidiroglou and Srinath suggest that such information can be obtained from previous surveys or from a previous census.

Our investigation of the estimation of the mean is undertaken on the assumption that the right tail of the distribution is well approximated by the right tail of a Weibull distribution. The nature of the left portion of the distribution is of modest importance to the problem, and casual investigation of empirical distributions suggests that it is difficult to specify a form for the left portion that is widely applicable. On the other hand, many empirical distributions display "tail characteristics" of the Weibull.

We shall be interested in relatively simple estimators. In particular, we shall consider estimators that are linear in the order statistics and estimators that are a function of a simple preliminary test. Our investigation suggests that the loss in efficiency associated with the use of our estimators for populations where the mean performs well (such as the exponential) is very small relative to the gains made in using the estimators on heavily skewed populations. It is demonstrated that these estimators can easily have mean square errors on the order of one-half or one-third the variance of the mean for the type of populations encountered in practice. We begin by considering estimation for an infinite population mean. In the final section we show how the results can be extended to the estimation of a finite population mean.

## 2. Winsorization for the Weibull

The Weibull density is given by

$$\begin{aligned} f(y;\alpha,\lambda) &= \alpha\lambda^{-1}y^{\alpha-1}\exp\{-\lambda^{-1}y^{\alpha}\}, & y > 0 \\ &= 0, & \text{otherwise} \end{aligned} \tag{2.1}$$

where $\lambda > 0$ and $\alpha > 0$. If $X$ is defined by the one-to-one transformation $X = Y^{\alpha}$, then $X$ is distributed as an exponential random variable with parameter $\lambda$. Conversely, the Weibull variable is the power of an exponential variable, $X^{\gamma}$, where $\gamma = \alpha^{-1}$.

The once-Winsorized mean for a right skewed distribution is

$$\widehat{W}_1 = n^{-1}\left[\sum_{i=1}^{n-1} Y_{(i)} + Y_{(n-1)}\right]. \tag{2.2}$$

Using the mean square error as the criterion, we prove that the once-Winsorized mean is superior to the sample mean for the Weibull if the shape parameter $\gamma$ is greater than one, has the same efficiency as the mean if $\gamma = 1$, and is less efficient than the mean if $\gamma < 1$.

The behavior of the order statistics of a random sample selected from the exponential distribution plays a central role in our investigation. Let $X$ be an exponential random variable with parameter $\lambda$. Let the order statistics of a random sample of size $n$ be denoted by $X_{(i)}$ , where $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ and set $X_{(0)} = 0$. Then it is well known that the random variables

$$Z_k = (n-k+1)\big(X_{(k)} - X_{(k-1)}\big), \qquad k = 1,2,\ldots,n, \tag{2.3}$$

are independently, identically distributed exponential random variables with parameter $\lambda$. See David (1981, p.20).

The following two lemmas permit us to express the expected value of positive powers of the largest exponential order statistic as a function of expected values of that power of the second largest order statistic and of a smaller power of the largest. These results can also be obtained from Theorem 3 of Lin (1989).

**Lemma 1.** *Let $X_{(i)}$ be the* ith *order statistic for a random sample of* n *selected from the standard exponential. Then, for $\gamma > 0$,*

$$E\{X_{(n)}^{\gamma}\} = E\{X_{(n-1)}^{\gamma}\} + \gamma E\{X_{(n)}^{\gamma-1}\}.$$

**Proof.** We have

$$\begin{aligned} E\{X_{(n)}^{\gamma}\} &= \int_0^{\infty} x^{\gamma} n(1-e^{-x})^{n-1} e^{-x}\, dx \\ &= n\sum_{j=0}^{n-1} \binom{n-1}{j} (-1)^j (j+1)^{-(\gamma+1)}\Gamma(\gamma+1), \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function, and

$$E\{X_{(n-1)}^{\gamma}\} = n(n-1)\sum_{j=0}^{n-2} \binom{n-2}{j} (-1)^j (j+2)^{-(\gamma+1)}\Gamma(\gamma+1).$$

It follows that

$$\begin{aligned} E\{X_{(n)}^{\gamma}\} - \gamma E\{X_{(n)}^{\gamma-1}\} &= n\Gamma(\gamma+1)\sum_{j=0}^{n-1} \binom{n-1}{j} (-1)^j [(j+1)^{-(\gamma+1)} - (j+1)^{-\gamma}] \\ &= n\Gamma(\gamma+1)\sum_{j=0}^{n-1} \binom{n-1}{j} (-1)^j (-j)(j+1)^{-(\gamma+1)} \\ &= E\{X_{(n-1)}^{\gamma}\}. \end{aligned}$$

**Lemma 2.** *Let $X_{(i)}$ be the ith order statistic for a random sample of size n selected from the standard exponential. Then, for $\gamma > 0$,*

$$E\{X_{(n-1)}^{\gamma} X_{(n)}^{\gamma}\} = E\{X_{(n-1)}^{2\gamma}\} + \gamma E\{X_{(n-1)}^{\gamma} X_{(n)}^{\gamma-1}\}.$$

**Proof.** We may write $X_{(n)}^{\gamma} = [X_{(n-1)} + Z]^{\gamma}$, where $Z$ is an exponential random variable distributed independently of $X_{(n-1)}$. Letting $f_{(n-1)}(x)$ denote the density of $X_{(n-1)}$, we have

$$E\{X_{(n-1)}^{\gamma} X_{(n)}^{\gamma}\} = \int_0^{\infty}\int_0^{\infty} x^{\gamma} f_{(n-1)}(x)(x+z)^{\gamma} e^{-z}\,dz\,dx$$

and, using integrating by parts,

$$\begin{aligned} E\{X_{(n-1)}^{\gamma} X_{(n)}^{\gamma}\} &= \int_0^{\infty} x^{2\gamma} f_{(n-1)}(x)\,dx \\ &\quad + \int_0^{\infty}\int_0^{\infty} \gamma x^{\gamma}(x+z)^{\gamma-1} f_{(n-1)}(x) e^{-z}\,dz\,dx \\ &= E\{X_{(n-1)}^{2\gamma}\} + \gamma E\{X_{(n-1)}^{\gamma} X_{(n)}^{\gamma-1}\}. \end{aligned}$$

The covariance between any two exponential order statistics is equal to the variance of the smallest. The following lemma demonstrates that the exponential ($\gamma = 1$) represents a boundary between those $\gamma < 1$ where the covariance is less than variance of the smallest and those $\gamma > 1$ where the covariance is greater than the variance of the smallest.

**Lemma 3.** *Given a sample of size n from the Weibull distribution, the covariance of the kth ($k \geq 2$) order statistic, $Y_{(k)}$, with any smaller order statistic, $Y_{(j)}$, $0 < j \leq k-1$, is greater than, equal to, or less than $\mathrm{Cov}(Y_{(k-1)}, Y_{(j)})$ if the parameter $\gamma$ is greater than, equal to, or less than one, respectively.*

**Proof.** We express the Weibull order statistics as powers of the exponential order statistics and recall that

$$X_{(k)} = \sum_{i=1}^{k} (n-i+1)^{-1} Z_i,$$

where the $X_{(k)}$ are the exponential order statistics and the $Z_i$ are independent exponential random variables. We then express, for $0 < j < k-1$,

$$X_{(k)} = X_{(j)} + W_{jk} = X_{(j)} + W_{j,k-1} + (n-k+1)^{-1} Z_k,$$

where

$$W_{jk} = \sum_{i=1}^{k-j} (n-j+1-i)^{-1} Z_{j+i}, \qquad n \ge k > j.$$

We consider the conditional expected value of $Y_{(k)} - Y_{(k-1)}$ given $Y_{(j)}$;

$$\begin{aligned} &E\{X^\gamma_{(k)} - X^\gamma_{(k-1)} | X^\gamma_{(j)} = x^\gamma_{(j)}\} \\ &\quad = E\left\{[x_{(j)} + W_{j,k-1} + (n-k+1)^{-1} Z_k]^\gamma - (x_{(j)} + W_{j,k-1})^\gamma\right\}. \end{aligned} \tag{2.4}$$

If $\gamma = 1$, the expected value on the right of (2.4) is $E\{(n-k+1)^{-1}Z_k\}$. If $0 < \gamma < 1$, the quantity in (2.4) whose expectation is desired is decreasing in $x_{(j)}$ for every fixed $W_{j,k-1} \ge 0$ and fixed $Z_k > 0$. Hence, if $0 < \gamma < 1$, the expectation on the right of (2.4) is a decreasing function of $x_{(j)}$. By a similar argument, if $\gamma > 1$, the expectation is an increasing function of $x_{(j)}$. Now the covariance between $X$ and an increasing function of $X$ is positive. See, for example, Bickel (1967, p. 576). Hence, if $\gamma$ is less than one, equal to one, or greater than one, then

$$\text{Cov}\left(Y_{(k)} - Y_{(k-1)}, Y_{(j)}\right), \qquad 0 < j < k-1,$$

is negative, zero, or positive, respectively.

We now give the main result.

**Theorem.** *Let a random sample of n elements be selected from the Weibull distribution defined in* (2.1). *Let the once-Winsorized mean in* (2.2) *be used as an estimator of the population mean. Then the mean square error of* (2.2) *is less than, equal to, or greater than the variance of the sample mean if the parameter,* $\gamma = \alpha^{-1}$, *is greater than, equal to, or less than one, respectively.*

**Proof.** Expressing the Weibull order statistics as the $\gamma$th powers of the exponential order statistics, $X_{(i)}$, we consider two estimators of the expectation of the sum of the two largest order statistics; $X^\gamma_{(n)} + X^\gamma_{(n-1)}$ and $2X^\gamma_{(n-1)}$. The mean square error of these estimators is

$$\begin{aligned} \text{MSE}\{X^\gamma_{(n)} + X^\gamma_{(n-1)}\} &= E\{X^{2\gamma}_{(n)} + 2X^\gamma_{(n-1)}X^\gamma_{(n)} + X^{2\gamma}_{(n-1)}\} \\ &\quad - \left[E(X^\gamma_{(n)}) + E(X^\gamma_{(n-1)})\right]^2, \\ \text{MSE}\{2X^\gamma_{(n-1)}\} &= 4E(X^{2\gamma}_{(n-1)}) - 4E(X^\gamma_{(n-1)})E(X^\gamma_{(n)} + X^\gamma_{(n-1)}) \\ &\quad + \left[E(X^\gamma_{(n)}) + E(X^\gamma_{(n-1)})\right]^2. \end{aligned}$$

Using Lemmas 1 and 2, we have

$$\begin{aligned} &\text{MSE}\{X^\gamma_{(n)} + X^\gamma_{(n-1)}\} - \text{MSE}\{2X^\gamma_{(n-1)}\} \\ &\quad = 2\gamma\text{Cov}(X^\gamma_{(n)}, X^{\gamma-1}_{(n)}) + 2\gamma\text{Cov}(X^\gamma_{(n-1)}, X^{\gamma-1}_{(n)}). \end{aligned} \tag{2.5}$$

Now $X_{(n)}^{\gamma-1}$ is a monotone increasing, constant, or monotone decreasing function of $X_{(n)}$ if $\gamma > 1$, $\gamma = 1$, or $0 < \gamma < 1$, respectively. Furthermore,

$$E\{X_{(n)}^{\gamma-1} | X_{(n-1)} = x_{(n-1)}\} = E\{(X_{(n-1)} + Z)^{\gamma-1} | X_{(n-1)} = x_{(n-1)}\}$$

is monotone increasing, constant, or monotone decreasing function of $x_{(n-1)}$ if $\gamma > 1$, $\gamma = 1$, or $0 < \gamma < 1$, respectively. Hence, the two covariances on the right of (2.5) are positive, zero, or negative, as $\gamma > 1$, $\gamma = 1$, or $0 < \gamma < 1$, respectively. It follows that $2X_{(n-1)}^{\gamma}$ as an estimator of $E\{X_{(n-1)}^{\gamma} + X_{(n)}^{\gamma}\}$ has smaller, the same, or larger mean square error than $X_{(n)}^{\gamma} + X_{(n-1)}^{\gamma}$ as $\gamma > 1$, $\gamma = 1$, or $0 < \gamma < 1$, respectively. Now

$$\begin{aligned} &\mathrm{MSE}\left\{\sum_{i=1}^{n-1} Y_{(i)} + Y_{(n-1)}\right\} - \mathrm{MSE}\left\{\sum_{i=1}^{n} Y_{(i)}\right\} \\ &= \mathrm{Cov}\left\{\sum_{i=1}^{n-2} Y_{(i)}, 2Y_{(n-1)}\right\} + \mathrm{MSE}\left\{2Y_{(n-1)}\right\} \\ &\quad - \mathrm{Cov}\left\{\sum_{i=1}^{n-2} Y_{(i)}, [Y_{(n-1)} + Y_{(n)}]\right\} - \mathrm{Var}\left\{Y_{(n-1)} + Y_{(n)}\right\}. \end{aligned}$$

But, by Lemma 3,

$$\mathrm{Cov}\left\{\sum_{i=1}^{n-2} Y_{(i)}, 2Y_{(n-1)}\right\}$$

is less than, equal to, or greater than

$$\mathrm{Cov}\left\{\sum_{i=1}^{n-2} Y_{(i)}, [Y_{(n-1)} + Y_{(n)}]\right\}$$

if $\gamma > 1$, $\gamma = 1$, or $0 < \gamma < 1$, respectively. Using (2.5), the result follows.

## 3. Efficiency of the Winsorized Mean for the Weibull

The once-Winsorized mean $\widehat{W}_1$ is defined in (2.2). In general, the $r$th Winsorized mean for a right skewed distribution is

$$\widehat{W}_r = n^{-1}\left(\sum_{j=1}^{n-r} X_{(j)} + rX_{(n-r)}\right). \tag{3.1}$$

If the $X_{(j)}$ are distributed as exponential order statistics and if we consider only estimators employing the first $(n - r)$ order statistics, then it can be shown that the minimum mean square error estimator of the mean is

$$\widehat{M}_r = (n-r+1)^{-1}\left(\sum_{j=1}^{n-r} X_{(j)} + rX_{(n-r)}\right). \tag{3.2}$$

For the exponential distribution with parameter $\lambda$ we have

$$E\{\widehat{W}_r, \widehat{M}_r\} = \{n^{-1}, (n-r+1)^{-1}\}(n-r)\lambda,$$
$$\text{MSE}\{\widehat{W}_r, \widehat{M}_r\} = \{n^{-2}(n-r+r^2), (n-r+1)^{-1}\}\lambda^2.$$

For $r = 1$, $\widehat{M}_1 = \widehat{W}_1$ has a mean square error that is equal to the variance of the sample mean.

The best unbiased estimator for the parameter $\lambda$ of the exponential distribution using only the first $(n-r)$ observations is

$$\hat{\lambda}_r = (n-r)^{-1}\left(\sum_{j=1}^{n-r} X_{(j)} + rX_{(n-r)}\right), \tag{3.3}$$

which has a variance of $(n-r)^{-1}\lambda^2$. Discarding the largest $r$ observations has the effect of reducing the sample size by $r$.

The efficiencies of these three estimators relative to the sample mean for the Weibull distribution have been tabulated by McElhone (1970). A portion of McElhone's results are given in Table 1. The striking aspect of this table is the large gains in efficiency possible with the use of the once-Winsorized mean. For example, given a Weibull with shape parameter $\gamma = 2$ and a sample of size 25, the once-Winsorized mean is 24 percent more efficient than the mean. A population displaying this degree of skewness would not be unusual in practice. For $n = 25$, the once-Winsorized mean is twice as efficient as the sample mean for a population with $\gamma = 3$ and more than four times as efficient as the sample mean for a population with $\gamma = 4$.

For those configurations in the table and $r > 1$, $\widehat{M}_r$ is uniformly superior on the basis of mean square error to $\widehat{W}_r$. For reasonable sized samples, say $n > 4$, there is little difference among the mean square errors of the three estimators.

Relative to the gains made for $r = 1$, the gains from proceeding to $r = 2$ are modest, and in no case is $r = 3$ superior to $r = 2$. For $\gamma = 2$, $r = 2$ is inferior to $r = 1$. Of course, the sample sizes of the table are relatively small. As sample size increases and as $\gamma$ increases, $r$ greater than one becomes superior to $r$ equal to one.

Table 1. Efficiencies of estimators relative to the mean for the Weibull distribution

| Sample size $n$ | Censoring $r$ | $\gamma=2$: $\widehat{W}$ | $\gamma=2$: $\hat{\lambda}$ | $\gamma=2$: $\widehat{M}$ | $\gamma=3$: $\widehat{W}$ | $\gamma=3$: $\hat{\lambda}$ | $\gamma=3$: $\widehat{M}$ | $\gamma=4$: $\widehat{W}$ | $\gamma=4$: $\hat{\lambda}$ | $\gamma=4$: $\widehat{M}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 2.86 | 1.67 | 2.86 | 8.94 | 5.43 | 8.94 | 30.04 | 18.71 | 30.04 |
| 3 | 1 | 2.27 | 1.69 | 2.27 | 6.34 | 4.81 | 6.34 | 20.12 | 15.43 | 20.12 |
| 4 | 1 | 1.99 | 1.63 | 1.99 | 5.13 | 4.25 | 5.13 | 15.55 | 12.95 | 15.55 |
| 5 | 1 | 1.82 | 1.57 | 1.82 | 4.42 | 3.84 | 4.42 | 12.89 | 11.23 | 12.89 |
| | 2 | 1.58 | 1.61 | 1.64 | 4.50 | 4.54 | 4.56 | 14.67 | 14.67 | 14.73 |
| 7 | 1 | 1.62 | 1.48 | 1.62 | 3.61 | 3.30 | 3.61 | 9.91 | 9.05 | 9.91 |
| | 2 | 1.42 | 1.50 | 1.49 | 3.61 | 3.72 | 3.68 | 11.06 | 11.18 | 11.14 |
| 10 | 1 | 1.47 | 1.39 | 1.47 | 2.99 | 2.83 | 2.99 | 7.69 | 7.25 | 7.69 |
| | 2 | 1.31 | 1.40 | 1.37 | 2.95 | 3.07 | 3.02 | 8.38 | 8.54 | 8.47 |
| | 3 | 1.04 | 1.26 | 1.18 | 2.54 | 2.78 | 2.68 | 7.70 | 7.98 | 7.87 |
| 15 | 1 | 1.34 | 1.30 | 1.34 | 2.49 | 2.41 | 2.49 | 5.92 | 5.72 | 5.92 |
| | 2 | 1.22 | 1.30 | 1.27 | 2.43 | 2.54 | 2.48 | 6.30 | 6.44 | 6.37 |
| | 3 | 0.98 | 1.18 | 1.12 | 2.06 | 2.26 | 2.19 | 5.66 | 5.90 | 5.81 |
| 20 | 1 | 1.28 | 1.25 | 1.28 | 2.22 | 2.17 | 2.22 | 5.00 | 4.88 | 5.00 |
| | 2 | 1.17 | 1.25 | 1.22 | 2.15 | 2.25 | 2.20 | 5.24 | 5.37 | 5.30 |
| | 3 | 0.96 | 1.14 | 1.08 | 1.82 | 2.00 | 1.94 | 4.65 | 4.86 | 4.78 |
| 25 | 1 | 1.24 | 1.22 | 1.24 | 2.05 | 2.02 | 2.05 | 4.42 | 4.35 | 4.42 |
| | 2 | 1.14 | 1.21 | 1.18 | 1.98 | 2.07 | 2.03 | 4.59 | 4.70 | 4.64 |
| | 3 | 0.95 | 1.11 | 1.06 | 1.68 | 1.83 | 1.78 | 4.04 | 4.22 | 4.16 |

## 4. A Simple Test for the Weibull Shape Parameter

The theory and empirical evidence of the previous sections demonstrate that considerable gains in efficiency relative to the mean are possible if the parameter, $\gamma$, of the Weibull is greater than one. It is therefore desirable to be able to test the hypothesis that $\gamma = 1$ versus the hypothesis that $\gamma > 1$.

A number of tests have been suggested in the literature, resting upon the fact that $Z_k = (n-k+1)(X_{(k)} - X_{(k-1)})$ are distributed as independent exponential random variables when the original $X_{(k)}$ are the exponential order statistics. Several tests are discussed in McElhone (1970). Also see Jackson (1967) and Mann, et al. (1974).

We desire a test to use as a basis for choosing between competing estimators of the population mean. We also desire a test that is location invariant because we are concerned only with the tail of the empirical distribution. The location invariance requirement rules out tests based upon the log-Weibull distribution.

If the two competing estimators are the sample mean and the $j$th-Winsorized mean, a very simple test statistic that is a function of the difference between the two estimators is

$$F_{Tj} = \frac{j^{-1} \sum_{i=n-j+1}^{n} Z_{Yi}}{(T-j)^{-1} \sum_{i=n-T+1}^{n-j} Z_{Yi}}, \tag{4.1}$$

where $Z_{Yi} = (n-i+1)(Y_{(i)} - Y_{(i-1)})$ and $j+1 \leq T \leq n$ is the number of large observations used to construct the test. We consider the use of $T < n$ in constructing the test because we only postulate a model for the tail portion of the distribution.

If the $Y$'s are selected from an exponential distribution, the $Z_{Yi}$ are independent exponential random variables and $F_{Tj}$ is distributed as Snedecor's $F$ with $2j$ and $2(T-j)$ degrees of freedom. If $T = n$ and $j > 1$, the statistic is seen to be the ratio of a multiple of the difference between the sample mean and the once-Winsorized mean to the once-Winsorized mean. Thus it is an intuitively appealing criterion for choosing between the two estimators. The following lemma demonstrates that the test possesses power to discriminate against Weibull distributions with $\gamma > 1$.

**Lemma 4.** *Let $Y_{(i)}$ be the order statistics of the Weibull distribution. The test that accepts $\gamma = 1$ or accepts $\gamma > 1$ as $F_{Tj}$ is less than or greater than $F_{2(T-j)}^{2j}(\delta)$, where $F_{2(T-j)}^{2j}(\delta)$ is the $\delta$ percentile of Snedecor's $F$ with $2j$ and $2(T-j)$ degrees of freedom, has a power function that is monotone increasing in $\gamma$.*

**Proof.** The order statistics of the Weibull sample may be expressed as powers of the order statistics from the exponential distribution. Thus the test statistic may be written as

$$F_{Tj} = \frac{j^{-1}\left[\sum_{i=n-j+1}^{n} \left(X_{(n-j)}^{\gamma} - X_{(n-T)}^{\gamma}\right)^{-1} \left(X_{(i)}^{\gamma} - X_{(n-T)}^{\gamma}\right) - j\right]}{(T-j)^{-1}\left[\sum_{i=n-T+1}^{n-j-1} \left(X_{(n-j)}^{\gamma} - X_{(n-T)}^{\gamma}\right)^{-1} \left(X_{(i)}^{\gamma} - X_{(n-T)}^{\gamma}\right) + j + 1\right]},$$

where $X_{(i)}$ are the order statistics of the exponential. Now the ratio of $X_{(i)}^{\gamma} - X_{(n-T)}^{\gamma}$ to $X_{(n-j)}^{\gamma} - X_{(n-T)}^{\gamma}$ is monotone decreasing for $i \leq n-j-1$ and is monotone increasing for $i > n-j$. Therefore $F_{Tj}$ is monotone increasing in $\gamma$ and the result follows.

## 5. Estimation Following a Preliminary Test

In Section 3 we demonstrated that there exists a simple estimator superior to the mean for the Weibull distribution with parameter $\gamma > 1$. In Section 4 we

described a simple test of the hypothesis that $\gamma = 1$ against the alternative that $\gamma > 1$. In practice it is the tail of the distribution that produces the skewness in the distribution of the sample mean. Also we have found it difficult to model the entire distribution of variables encountered in survey sampling practice. These results suggest the following estimation procedure.
i) Test the hypothesis that the order statistics were selected from an exponential ($\gamma = 1$) versus the alternative $\gamma > 1$ using the $T$ largest order statistics.
ii) If the exponential hypothesis is accepted, use the sample mean as the estimator of the population mean. If the exponential hypothesis is rejected, estimate the largest order statistic(s) by a procedure appropriate for the Weibull with $\gamma > 1$.

We shall study the class of estimators defined by

$$\begin{aligned}\hat{\mu}_{Tj} &= \bar{y} && \text{if } F_{Tj} < K_j \\ &= n^{-1}\left\{\sum_{i=1}^{n-j} Y_{(i)} + j\left[Y_{(n-j)} + K_j\bar{d}_{Tj}\right]\right\} && \text{otherwise,}\end{aligned} \tag{5.1}$$

where $F_{Tj}$ is defined in (4.1), $K_j$ is the cut-off value that determines when the alternative estimator of the large values is used and

$$\bar{d}_{Tj} = (T-j)^{-1}\left\{\sum_{i=j}^{T-1}\left[Y_{(n-i)} - Y_{(n-T)}\right] + j\left[Y_{(n-j)} - Y_{(n-T)}\right]\right\}.$$

The estimator of (5.1) is a test-and-estimate procedure in which the estimator is a continuous function of the sums formed from different sets of order statistics. The sample mean and the Winsorized mean are special cases of estimator (5.1) obtained by setting $K_j$ equal to infinity and zero respectively.

It is difficult to specify the number of tail observations, $T$, the number of large order statistics, $j$, and the cut-off values, $K_j$, to use in constructing the estimator for the tail portion. It would seem that $T$ approximately equal to one fifth to one third of the observations is reasonable for many populations and sample sizes. It also seems that one can reduce this fraction in large ($n > 200$) samples. When the sample is large, setting $T \doteq 30$ seems to perform well. Table 1 suggests that the optimum $j$ depends on the sample size, with larger $j$ possible for larger samples. We consider $j = 1, 2, 3$ in the Monte Carlo section. Larger $K_j$ will give higher efficiency for populations in which the mean is optimum. We now investigate the effect of varying $K_j$ for the exponential distribution.

For the exponential distribution with $T = n$, $n > 1$, and $K_1 = (n-1)h$,

$$\begin{aligned}E\left(\hat{\mu}_{n1}\right) &= \lambda\left[1 - n^{-1}(1+h)^{1-n}\right], \\ \operatorname{Var}\left(\hat{\mu}_{n1}\right) &= \lambda^2\left[n^{-1} - n^{-2}(1+h)^{-2(n-1)}\right].\end{aligned}$$

The bias is negative for all $h < \infty$ and the absolute bias is a decreasing function of $h$ reaching a maximum for $h = 0$. The variance is decreased by the square of the bias and hence the mean square error of the estimator is a constant function of $h$, being always equal to the variance of the mean.

Huang (1970) has investigated the relative efficiency of this and similar estimators under the exponential model. Huang derived the multiple, $a_m$, of $\hat{\mu}_{n1}$ which minimizes the mean square error of the estimator and the multiple, $a_u$, required to produce an unbiased estimator. He showed that

$$\left[n(1+h)^{n-1} - 1\right] \left[(n+1)(1+h)^{n-1} - 2\right]^{-1} \hat{\mu}_{n1} = a_m \hat{\mu}_{n1}$$

minimizes the mean square error for a fixed $h$ and that

$$n(1+h)^{n-1} \left[n(1+h)^{n-1} - 1\right] \hat{\mu}_{n1} = a_u \hat{\mu}_{n1}$$

is unbiased for $\lambda$. Huang studied the ratio of the mean square error of $n(n+1)^{-1}\overline{y}$ to the mean square error of $a_m\hat{\mu}_{n1}$ and the ratio of the variance of $\overline{y}$ to the variance of $a_u\hat{\mu}_{n1}$. Huang demonstrated that these ratios are monotone increasing in $h$ for fixed $n > 1$ and monotone increasing in $n$ for fixed $h$, with one as the limiting value in both cases. The ratios for $h = 0$ are

$$\frac{\text{MSE}\left(n(n+1)^{-1}\overline{y}\right)}{\text{MSE}(a_m\hat{\mu}_{n1})} = \frac{n}{n+1}$$

and

$$\frac{\text{Var}(\overline{y})}{\text{Var}(a_u\hat{\mu}_{n1})} = \frac{n-1}{n}.$$

A few ratios for $n = 10$ and $n = 21$ for the exponential distribution are given in Table 2. The $\delta$ of Table 2 is the percentage level for the $F$-test of (5.1). The loss in efficiency associated with $h = 0$ relative to $h = 1$ is about $n^{-1}$. The loss is less than $0.5n^{-1}$ for the $h$-value associated with the 25 percent level of $F$. The loss in efficiency is less than 0.1 percent if the preliminary test is performed at the 0.5 percent level.

## 6. A Monte Carlo Study

The procedures discussed in the preceeding sections were applied to samples selected from two real populations and from a Weibull distribution with $\gamma = 3$. The first population is the chickens per segment (in tens less ten) for segments with chickens observed in the United States Department of Agriculture area survey of the Southeastern states 1959-60-61. The segment is the primary sampling

Table 2. Relative efficiency of preliminary test estimators for the exponential

| Sample size | $\delta$ | $h$ | $\frac{\text{MSE}\,(n(n+1)^{-1}\bar{y})}{\text{MSE}\,(a_m\hat{\mu}_{n1})}$ | $\frac{\text{Var}(\bar{y})}{\text{Var}(a_u\hat{\mu}_{n1})}$ |
|---|---|---|---|---|
| 10 | 100 | 0.0 | 0.9091 | 0.9000 |
| | 25 | 0.1667 | 0.9610 | 0.9571 |
| | 10 | 0.2911 | 0.9827 | 0.9810 |
| | 2.5 | 0.5067 | 0.9954 | 0.9950 |
| | 0.5 | 0.8011 | 0.9991 | 0.9990 |
| 21 | 100 | 0.0 | 0.9545 | 0.9524 |
| | 25 | 0.0720 | 0.9805 | 0.9793 |
| | 10 | 0.1220 | 0.9914 | 0.9910 |
| | 2.5 | 0.2025 | 0.9977 | 0.9977 |
| | 0.5 | 0.3035 | 0.9995 | 0.9995 |

unit of the area frame. The second population is the size of farm, "Acres in Place," (in tens of acres) published in the *Agricultural Census* (1959). In both cases the populations were simplified so that the smallest relative frequency for any particular $y$-value is 1/2000. The populations are given in the Appendix. The chicken population has a mean of 14.69 and a variance of 4293, the acre population has a mean of 301 and a variance of 2,338,204, and the Weibull distribution with $\gamma = 3$ has a mean of 6 and a variance of 684. The coefficients of variation are 4.46, 5.08, and 4.36 for chickens, acres, and Weibull with $\gamma = 3$, respectively. To facilitate comparisons all populations have been coded by dividing the observations by the population mean. Thus, all three populations used in the simulations have a mean of one and a variance that is equal to the squared coefficient of variation.

The distribution functions for chickens and acres deviate markedly from the Weibull model in the left part of the distribution. On the other hand, the upper portion of the distribution, say the largest ten percent, is well approximated by the tail of a Weibull distribution with a slope corresponding to a $\gamma$ of about 3. Also the coefficients of variation for chickens and acres approximate that of a Weibull distribution with shape parameter, $\gamma = 3$.

Six estimators of the population mean were compared in the Monte Carlo study. In addition to the simple mean, the once-Winsorized mean and the twice-Winsorized mean, three estimators of the test-and-estimate type were constructed. The three test-and-estimate procedures are the estimators defined in (5.1) for $j = 1$, 2, and 3. The cut-off values are $K_1 = 5.8$, $K_2 = 4.2$ and $K_3 = 3.5$. The cut-off levels for the $F$-tests were set so that less than one half of one percent of the samples with $T = 30$ from an exponential would be modified.

Simple random replacement samples of size 25, 100, and 200 were selected from three populations. Two thousand samples were used for $n = 25$ and one thousand samples were used for $n = 100$ and $n = 200$. The samples were restricted so that they contained, approximately, the correct fraction of the 1/1000 largest observations in the population. A $T$ of 10, 30, and 30 was used to construct the test-and-estimate procedures for $n = 25$, 100, and 200, respectively. The Monte Carlo results for samples of size 25 are given in Table 3.

Table 3. Monte Carlo comparison of alternative estimators for the mean (2000 samples of size 25, $T = 10$)

| Estimator | Mean | $nV\{\hat{\mu}\}$ | $n\text{MSE}\{\hat{\mu}\}$ | $V\{\bar{y}\}/\text{MSE}\{\hat{\mu}\}$ |
|---|---|---|---|---|
| | | **Chickens** | | |
| Mean | 1.008 | 20.03 | 20.03 | 1.00 |
| Once-Winsorized | 0.674 | 4.64 | 7.30 | 2.74 |
| Twice-Winsorized | 0.549 | 1.29 | 6.38 | 3.14 |
| Prelim. ($F = 5.8$) | 0.803 | 8.10 | 9.07 | 2.21 |
| 2-Prelim. ($F = 4.2$) | 0.729 | 3.67 | 5.51 | 3.64 |
| 3-Prelim. ($F = 3.5$) | 0.694 | 2.38 | 4.72 | 4.24 |
| | | **Acres** | | |
| Mean | 1.016 | 27.14 | 27.14 | 1.00 |
| Once-Winsorized | 0.682 | 4.06 | 6.59 | 4.12 |
| Twice-Winsorized | 0.582 | 0.99 | 5.36 | 5.06 |
| Prelim. ($F = 5.8$) | 0.794 | 6.94 | 8.00 | 3.39 |
| 2-Prelim. ($F = 4.2$) | 0.743 | 2.59 | 4.24 | 6.40 |
| 3-Prelim. ($F = 3.5$) | 0.721 | 1.93 | 3.88 | 6.99 |
| | | **Weibull ($\gamma = 3$)** | | |
| Mean | 1.009 | 18.08 | 18.08 | 1.00 |
| Once-Winsorized | 0.678 | 6.77 | 9.36 | 1.93 |
| Twice-Winsorized | 0.503 | 3.26 | 9.44 | 1.92 |
| Prelim. ($F = 5.8$) | 0.860 | 10.98 | 11.47 | 1.58 |
| 2-Prelim. ($F = 4.2$) | 0.748 | 7.95 | 9.54 | 1.90 |
| 3-Prelim. ($F = 3.5$) | 0.715 | 6.25 | 8.28 | 2.18 |

The contribution of the large observations to the variance of the sample mean is demonstrated by the small variance of the Winsorized means relative to the sample mean. For the Weibull distribution, the mean square error of the once-Winsorized mean and of the twice-Winsorized mean are about one-half of the variance of the sample mean. This is consistent with the Weibull theory used to construct Table 1. The efficiency of the mean relative to Winsorization is even smaller for the chickens and acres populations than for the Weibull.

While once-Winsorized means have much smaller mean square errors than the simple mean, they have a bias that is roughly one-third of the population mean. The test-and-estimate procedures defined by (5.1) have a smaller bias and a larger variance than the Winsorized means. The large value for the test ($F =$

5.8) used with the estimator that modifies only the largest observation resulted in a relatively small bias but a larger mean square error than the corresponding once-Winsorized mean. The procedure based on the comparison of the largest two observations to the next eight largest had a smaller bias and a smaller mean square error than the twice-Winsorized mean for chickens and acres. The bias of the test-and-estimate procedure based on the largest two observations was smaller than the bias of the twice-Winsorized mean but the mean square errors were comparable.

The best procedure with respect to mean square error was the procedure that modified the three largest observations. This was slightly surprising, given the small sample size.

Our discussion assumes that we are dealing with a single sample. If samples are to be summed, this must be recognized in evaluating the estimators. For example, the mean square error (= variance) of the sum of six independent simple means from samples of 25 for the chicken population is 4.78 while the mean square error of the sum of six independent once-Winsorized means is estimated to be

$$[6(0.326)]^2 + 6(0.204) = 5.05$$

and the estimated mean square error of the sum of six test-and-estimate procedures modifying only the largest observation, $\hat{\mu}_{10,1}$, is

$$[6(0.190)]^2 + 6(0.361) = 3.46.$$

In this example the variance gains of the biased estimators is such that they retain an advantage for the sum of several estimators. In our case the biased estimators are always biased towards zero. Therefore, there is always a number $L$ such that the sum of $L$ unbiased estimators is superior to the sum of $L$ biased estimators. The test-and-estimate procedures are less biased than the Winsorized estimators. Hence, a sum of test-and-estimate procedures will retain an advantage over a sum of simple means for a larger $L$, than will a sum of Winsorized means.

The Monte Carlo properties of the estimators for samples of size 100 and 200 are given in Table 4 and Table 5, respectively. The basic ordering of the mean square errors of the estimators is the same for the three sample sizes. The efficiency of the five estimators that modify the largest observations relative to the simple mean declines as the sample size increases. This is to be expected because the relative importance of the largest observations declines as the sample size increases. Associated with this phenomenon is the fact that the variances of all of the modification procedures decline at a rate smaller than $n^{-1}$. Thus, the entries in the table, $n\text{Var}(\hat{\mu})$, are uniformly larger for $n = 200$ than for $n = 100$. On the other hand, the bias shows a marked reduction from $n = 100$ to $n = 200$.

Table 4. Monte Carlo comparison of alternative estimators for the mean (1000 samples of size 100, $T = 30$)

| Estimator | Mean | $nV\{\hat{\mu}\}$ | $n\text{MSE}\{\hat{\mu}\}$ | $V\{\bar{y}\}/\text{MSE}\{\hat{\mu}\}$ |
|---|---|---|---|---|
| | | **Chickens** | | |
| Mean | 1.005 | 20.16 | 20.16 | 1.00 |
| Once-Winsorized | 0.835 | 10.17 | 12.91 | 1.56 |
| Twice-Winsorized | 0.734 | 5.89 | 12.95 | 1.56 |
| Prelim. ($F = 5.8$) | 0.900 | 12.76 | 13.75 | 1.47 |
| 2-Prelim. ($F = 4.2$) | 0.832 | 9.12 | 11.94 | 1.69 |
| 3-Prelim. ($F = 3.5$) | 0.786 | 7.09 | 11.68 | 1.73 |
| | | **Acres** | | |
| Mean | 1.000 | 25.56 | 25.56 | 1.00 |
| Once-Winsorized | 0.803 | 8.73 | 12.63 | 2.02 |
| Twice-Winsorized | 0.723 | 3.17 | 10.86 | 2.35 |
| Prelim. ($F = 5.8$) | 0.857 | 11.11 | 13.15 | 1.94 |
| 2-Prelim. ($F = 4.2$) | 0.803 | 5.03 | 8.90 | 2.87 |
| 3-Prelim. ($F = 3.5$) | 0.778 | 4.01 | 8.93 | 2.86 |
| | | **Weibull ($\gamma = 3$)** | | |
| Mean | 1.007 | 19.33 | 19.33 | 1.00 |
| Once-Winsorized | 0.862 | 10.83 | 12.73 | 1.52 |
| Twice-Winsorized | 0.772 | 8.16 | 13.37 | 1.45 |
| Prelim. ($F = 5.8$) | 0.943 | 13.62 | 13.95 | 1.39 |
| 2-Prelim. ($F = 4.2$) | 0.906 | 11.79 | 12.67 | 1.53 |
| 3-Prelim. ($F = 3.5$) | 0.876 | 10.80 | 12.34 | 1.57 |

There is still a considerable gain in efficiency from using the modifications at samples of size 200 for all three populations. There are gains of about 20% for the Weibull and chicken distributions, while larger gains are attained for the acres population.

One form of a reweighted estimator was introduced in (1.1). The mean square error of that reweighted estimator is

$$\sum_{r=0}^{n} P_r \left\{a^2(r)\sigma_1^2 + [1 - a(r)]^2\sigma^2 + [a(r) - p_1]^2(\mu_1 - \mu_2)^2\right\}$$

where

$$P_r = \binom{n}{r} p_1^{n-r}(1 - p_1)^r.$$

Using this formula, the mean square error of the estimator based on the optimum values of $a(r)$ defined in (1.2) was computed for $n = 100$ and $p_1 = 0.97$. The mean square error relative to that of the sample mean is given in line three of Table 6.

Table 5. Monte Carlo comparison of alternative estimators for the mean (1000 samples of size 200, $T = 30$)

| Estimator | Mean | $nV\{\hat{\mu}\}$ | $n\text{MSE}\{\hat{\mu}\}$ | $V\{\bar{y}\}/\text{MSE}\{\hat{\mu}\}$ |
|---|---|---|---|---|
| | | **Chickens** | | |
| Mean | 0.995 | 19.77 | 19.77 | 1.00 |
| Once-Winsorized | 0.888 | 13.89 | 16.42 | 1.20 |
| Twice-Winsorized | 0.809 | 9.37 | 16.65 | 1.19 |
| Prelim. ($F = 5.8$) | 0.938 | 16.05 | 16.81 | 1.18 |
| 2-Prelim. ($F = 4.2$) | 0.895 | 13.13 | 15.34 | 1.29 |
| 3-Prelim. ($F = 3.5$) | 0.859 | 11.49 | 15.32 | 1.29 |
| | | **Acres** | | |
| Mean | 0.996 | 25.42 | 25.42 | 1.00 |
| Once-Winsorized | 0.853 | 14.79 | 19.13 | 1.33 |
| Twice-Winsorized | 0.776 | 6.19 | 16.26 | 1.56 |
| Prelim. ($F = 5.8$) | 0.898 | 17.65 | 19.73 | 1.29 |
| 2-Prelim. ($F = 4.2$) | 0.843 | 9.14 | 14.09 | 1.80 |
| 3-Prelim. ($F = 3.5$) | 0.816 | 5.92 | 12.68 | 2.00 |
| | | **Weibull** ($\gamma = 3$) | | |
| Mean | 0.995 | 18.93 | 18.93 | 1.00 |
| Once-Winsorized | 0.906 | 13.73 | 13.73 | 1.22 |
| Twice-Winsorized | 0.842 | 10.64 | 10.64 | 1.21 |
| Prelim. ($F = 5.8$) | 0.962 | 15.82 | 15.82 | 1.18 |
| 2-Prelim. ($F = 4.2$) | 0.943 | 14.28 | 14.94 | 1.27 |
| 3-Prelim. ($F = 3.5$) | 0.929 | 13.47 | 13.47 | 1.31 |

Table 6. Relative mean square errors of alternative estimators of the mean ($n = 100$ and $p_1 = 0.97$)

| Estimator | Population | | |
|---|---|---|---|
| | Chickens | Acres | Weibull($\gamma = 3$) |
| Mean | 1.00 | 1.00 | 1.00 |
| 3-Prelim. | 0.58 | 0.35 | 0.64 |
| Reweighted, $a(r)$ | 0.49 | 0.42 | 0.51 |
| Reweighted, $a(3)$ | 0.53 | 0.46 | 0.55 |

A simpler form of the reweighted estimator uses a value of $a$ that does not depend on $r$. An estimator of this form is obtained for our example by fixing $a$ at $a(3)$. The values of $a(3)$ are 0.9805, 0.9850, and 0.9775 for chickens, acres, and Weibull, respectively. Thus the estimator weights the large observations by about one third to two thirds of the post stratified weight. Table 6 contains the relative mean square errors of the simple mean, the two reweighted estimators, and the test-and-estimate procedure that modifies the three largest observations. The relative mean square error of the test-and-estimate procedure is taken from

Table 4. The comparison is for a sample size of 100 so that the expected fraction greater than $A$ for the reweighted estimator is the same as the fraction of observations modified by the test-and-estimate procedure.

The test-and-estimate procedure is superior for the acres population. The reweighted estimators are superior for the Weibull and for chickens. These results are very supportive of the simple test-and-estimate procedure because the test-and-estimate procedure uses no information outside of the sample. The first reweighted estimator is based upon the optimum value of $a(r)$ which used knowledge of the means and variances of the two parts of the population and of the fractions falling in the two parts. The second reweighted estimator requires almost as much knowledge of the population.

The efficiency of the preliminary test procedures is greater than 99% for the exponential distribution, no outside information is required for these estimators, and large gains in efficiency are possible for skewed distributions. Therefore these procedures are highly recommended for samples selected from skewed populations.

## 7. Estimators for Finite Populations

In the preceeding sections we have discussed estimation for an infinite population mean. To extend the results to a finite population mean, we assume the finite universe, $U_N$, to be a random sample of size $N$ selected from the distribution, $F_X(x)$. Assume that a simple random nonreplacement sample of size $n$ is selected from $U_N$, and that an estimator of the mean of the finite population is desired.

We shall use $x_i$ to denote an element of the sample of $n$ or of the population of $N$ and adopt the convention that the indexing is such that the sample is composed of elements indexed by $i = 1, 2, \ldots, n$. Let $\mu$ and $\sigma^2$ be the superpopulation mean and variance. Let $\bar{x}_n$ be the sample mean and let $\bar{X}_N$ be the finite population mean.

The following lemma enables us to choose an estimator for the finite population mean given estimators for $\mu$. The origin of the estimator of the finite population mean that is a linear combination of the sample mean and an estimator of the superpopulation mean is uncertain. Brewer and Ferrier (1966) cite a book review by Fisher (1942) in which Fisher states that if $\hat{\mu}_m$ is the maximum likelihood estimator of $\mu$ then

$$\tilde{\mu}_{mN} = N^{-1} n \bar{x}_n + N^{-1}(N - n)\hat{\mu}_m \tag{7.1}$$

is the maximum likelihood estimator of $\bar{X}_N$.

**Lemma 5.** *Let the finite population, $U_N$, be a random sample from $F_X(x)$. Let $\widehat{\mu}_1$ and $\widehat{\mu}_2$ be two estimators of the superpopulation mean $\mu$ defined for samples of size $n$. If $E\{(\widehat{\mu}_1 - \mu)^2\} \le E\{(\widehat{\mu}_2 - \mu)^2\}$ then*

$$E\{N^{-2}\left[n\overline{x}_n + (N-n)\widehat{\mu}_1 - \overline{X}_N\right]^2\} \le E\{N^{-2}\left[n\overline{x}_n + (N-n)\widehat{\mu}_2 - \overline{X}_N\right]^2\}.$$

**Proof.** We have

$$\begin{aligned} E(\{N^{-1}[n\overline{x}_n + (N-n)\widehat{\mu}_1] - \overline{X}_N\}^2) &= E([N^{-1}(N-n)]^2(\widehat{\mu}_1 - \overline{X}_{N-n})^2) \\ &= N^{-2}(N-n)^2\left(E\{(\widehat{\mu}_1-\mu)^2\} + E\{(\widehat{\mu}_1-\mu)(X_{N-n}-\mu)\} + E\{(X_{N-n}-\mu)^2\}\right), \end{aligned}$$

where $\overline{X}_{N-n} = (N-n)^{-1}\sum_{i=n+1}^{N} x_i$. Since $\widehat{\mu}_1$ and $\overline{X}_{N-n}$ are unconditionally independent and $E\{\overline{X}_{N-n}\} = \mu$, we have

$$\begin{aligned} \text{MSE}(\widetilde{\mu}_{iN}) &= E(\{N^{-1}[n\overline{x}_n + (N-n)\widehat{\mu}_i] - \overline{X}_N\}^2) \\ &= (N^{-1}(N-n))^2\left(E\{(\widehat{\mu}_i - \mu)^2 + (N-n)\sigma^2\}\right) \end{aligned} \tag{7.2}$$

and the result follows.

**Corollary.** *If $\widehat{\mu}$ is the minimum mean square error estimator for $\mu$, then*

$$\widetilde{\mu}_N = N^{-1}[n\overline{x}_n + (N-n)\widehat{\mu}] \tag{7.3}$$

*is the minimum mean square error estimator for $\overline{X}_N$.*

The unconditional mean square error of $\widetilde{\mu}_N$, defined in (7.3) as an estimator $\overline{X}_N$, is given in (7.2). Let $n\widehat{V}(\widehat{\mu})$ be an unbiased (or consistent) estimator for $nE\{(\hat{\mu} - \mu)^2\}$ and let $s_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$. Then

$$n[N^{-1}(N-n)]^2\left\{\hat{V}(\hat{\mu}) + (N-n)^{-1}s_n^2\right\} \tag{7.4}$$

is an unbiased (or consistent) estimator of $nE\{(\widetilde{\mu}_N - \overline{X}_N)^2\}$. See Bellhouse (1987) for extensions of some of these results.

**Acknowledgements**

The majority of this work was done in the late 1960's and was partly supported by joint statistical agreements with the U.S. Bureau of the Census. See Fuller (1970). Students at Iowa State University at that time who conducted research on the topic include Martin Rosenzweig, D. H. McElhone and H. T. Huang. William Hawkes introduced procedures based upon this research at A. C. Nielsen in 1978. Recent research was partly supported by J.S.A. 89-2 with the

U.S. Bureau of the Census. I thank Joseph Croos for the recent Monte Carlo work. I also thank the referees and editors for useful comments.

## Appendix: Experimental Populations

| | Probability of $Y$-value | Chickens | Acres | | Probability of $Y$-value | Chickens | Acres |
|---|---|---|---|---|---|---|---|
| 1 | .0665 | 0 | 5 | 47 | .0005 | 88 | 1520 |
| 2 | .2195 | 1 | 25 | 48 | .0005 | 90 | 1540 |
| 3 | .0700 | 3 | 60 | 49 | .0005 | 92 | 1560 |
| 4 | .1080 | 4 | 90 | 50 | .0005 | 94 | 1580 |
| 5 | .1060 | 5 | 120 | 51 | .0005 | 96 | 1600 |
| 6 | .1020 | 7 | 160 | 52 | .0005 | 98 | 1640 |
| 7 | .0610 | 9 | 200 | 53 | .0005 | 100 | 1670 |
| 8 | .0510 | 11 | 240 | 54 | .0005 | 102 | 1700 |
| 9 | .0400 | 13 | 280 | 55 | .0005 | 105 | 1740 |
| 10 | .0365 | 15 | 350 | 56 | .0005 | 110 | 1770 |
| 11 | .0300 | 18 | 400 | 57 | .0005 | 115 | 1800 |
| 12 | .0180 | 22 | 500 | 58 | .0005 | 120 | 1850 |
| 13 | .0170 | 27 | 550 | 59 | .0005 | 125 | 1900 |
| 14 | .0135 | 32 | 650 | 60 | .0005 | 130 | 1950 |
| 15 | .0110 | 37 | 750 | 61 | .0005 | 135 | 2000 |
| 16 | .0075 | 42 | 850 | 62 | .0005 | 140 | 2100 |
| 17 | .0055 | 47 | 950 | 63 | .0005 | 145 | 2150 |
| 18 | .0005 | 51 | 1000 | 64 | .0005 | 150 | 2200 |
| 19 | .0005 | 52 | 1010 | 65 | .0005 | 155 | 2300 |
| 20 | .0005 | 53 | 1020 | 66 | .0005 | 160 | 2400 |
| 21 | .0005 | 54 | 1030 | 67 | .0005 | 165 | 2500 |
| 22 | .0005 | 55 | 1040 | 68 | .0005 | 170 | 2600 |
| 23 | .0005 | 56 | 1050 | 69 | .0005 | 175 | 2700 |
| 24 | .0005 | 57 | 1070 | 70 | .0005 | 180 | 2800 |
| 25 | .0005 | 58 | 1090 | 71 | .0005 | 185 | 2900 |
| 26 | .0005 | 59 | 1100 | 72 | .0005 | 190 | 3000 |
| 27 | .0005 | 60 | 1120 | 73 | .0005 | 200 | 3100 |
| 28 | .0005 | 61 | 1140 | 74 | .0005 | 220 | 3300 |
| 29 | .0005 | 62 | 1160 | 75 | .0005 | 240 | 3500 |
| 30 | .0005 | 63 | 1180 | 76 | .0005 | 260 | 3700 |
| 31 | .0005 | 64 | 1200 | 77 | .0005 | 280 | 3900 |
| 32 | .0005 | 65 | 1220 | 78 | .0005 | 300 | 4100 |
| 33 | .0005 | 66 | 1240 | 79 | .0005 | 340 | 4500 |
| 34 | .0005 | 67 | 1260 | 80 | .0005 | 360 | 4800 |
| 35 | .0005 | 68 | 1280 | 81 | .0005 | 380 | 5200 |
| 36 | .0005 | 69 | 1300 | 82 | .0005 | 400 | 5800 |
| 37 | .0005 | 70 | 1320 | 83 | .0005 | 440 | 6200 |
| 38 | .0005 | 71 | 1340 | 84 | .0005 | 470 | 6800 |
| 39 | .0005 | 72 | 1360 | 85 | .0005 | 500 | 7500 |
| 40 | .0005 | 73 | 1380 | 86 | .0005 | 550 | 8500 |
| 41 | .0005 | 74 | 1400 | 87 | .0005 | 600 | 9600 |
| 42 | .0005 | 76 | 1420 | 88 | .0005 | 650 | 15000 |
| 43 | .0005 | 78 | 1440 | 89 | .0005 | 700 | 25000 |
| 44 | .0005 | 80 | 1460 | 90 | .0005 | 1000 | 35000 |
| 45 | .0005 | 84 | 1480 | 91 | .0005 | 2000 | 45000 |
| 46 | .0005 | 86 | 1500 | | | | |

## References


Anscombe, F. J. (1960). Rejection of outliers. *Technometrics* **2**, 123-147.

Anscombe, F. J. (1961). Examination of residuals. *Proc. 4th Berkeley Symp. Math. Statist. Probab.* **1**, 1-36.

Barnett, V. and Lewis, T. (1984). *Outliers in Statistical Data*, 2nd edition. John Wiley, New York.

Bellhouse, D. R. (1987). Model-based estimation in finite population sampling. *Amer. Statist.* **41**, 260-262.

Bershad, M. (1961). Some observations on outliers. Unpublished memorandum, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

Bickel, P. J. (1967). Some contributions to the theory of order statistics. *Proc. 5th Berkeley Symp. Math. Statist. Probab.* **1** (Edited by LeCam, L. and Neyman, J.), 575-591. University of California Press, Berkeley.

Brewer, K. R. W. and Ferrier, A. E. (1966). Optimum weighting of outlying observations when sampling from economic populations. Unpublished manuscript, Commonwealth Bureau of Census and Statistics, Canberra, Australia.

Chambers, R. L. (1986). Outlier robust finite population estimation. *J. Amer. Statist. Assoc.* **81**, 1063-1069.

David, H. A. (1981). *Order Statistics*, 2nd edition. John Wiley, New York.

Dixon, W. J. (1950). Analysis of extreme values. *Ann. Math. Statist.* **21**, 488–506.

Dixon, W. J. and Tukey, J. W. (1968). Approximate behavior of the distribution of Winsorized *t*. *Technometrics* **10**, 83–98.

Ernst, L. R. (1980). Comparison of estimators of the mean which adjust for large observations. *Sankhyā Ser.C* **42**, 1–16.

Fisher, R. A. (1942). Book review. *Nature* **150**, 196.

Fuller, W. A. (1970). Simple estimators for the mean of skewed populations. Report to the U.S. Bureau of the Census, Iowa State University, Ames, Iowa.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* John Wiley, New York.

Hawkins, D. M. (1980). *Identification of Outliers.* Chapman and Hall, London.

Hidiroglou, M. A. and Srinath, K. P. (1981). Some estimators of a population total from simple random samples containing large units. *J. Amer. Statist. Assoc.* **76**, 690–695.

Huang, H. T. (1970). Relative efficiency of some estimators for the exponential employing preliminary tests. Unpublished report, Department of Statistics, Iowa State University, Ames, Iowa.

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

Jackson, O. A. Y. (1967). An analysis of departures from the exponential distribution. *J. Roy. Statist. Soc. Ser.B* **29**, 540–549.

Kale, B. K. and Sinha, S. K. (1971). Estimation of expected life in the presence of an outlier observation. *Technometrics* **13**, 755–759.

Lin, G. D. (1989). The product moments of order statistics with applications to characterizations of distributions. *J. Statist. Plann. Inference* **21**, 395-406.

Mann, N. R., Schafer, R. E. and Singpurwalla, N. D. (1974). *Methods for Statistical Analysis of Reliability and Life Data.* John Wiley, New York.

McElhone, D. H. (1970). Estimation of the mean of skewed distributions using systematic statistics. Unpublished Ph.D. thesis, Ames, Iowa, Library, Iowa State Uiversity.
Tukey, J. W. (1962). The future of data analysis. *Ann. Math. Statist.* **33**, 1–67.
Tukey, J. W. and McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample. *Sankhyā Ser.A* **25**, 331–352.
Veale, J. R. and Huntsberger, D. V. (1969). Estimation of a mean when one observation may be spurious. *Technometrics* **11**, 331-339.

Department of Statistics, Iowa State University, IA 50011, U.S.A.

(Received August 1989; accepted June 1990)