Survey Methodology, December 1995
Vol. 21, No. 2, pp. 107–116
Statistics Canada

On Searls' Winsorized Mean for Skewed Populations

LOUIS-PAUL RIVEST and DANIEL HURTUBISE[1]

ABSTRACT

This paper considers the winsorized mean as an estimator of the mean of a positive skewed population. A winsorized mean is obtained by replacing all the observations larger than some cut-off value R by R before averaging. The optimal cut-off value, as defined by Searls (1966), minimizes the mean square error of the winsorized estimator. Techniques are proposed for the evaluation of this optimal cut-off in several sampling designs including simple random sampling, stratified sampling and sampling with probability proportional to size. For most skewed distributions, the optimal winsorization strategy is shown, on average, to modify the value of about one data point in the sample. Closed form approximations to the efficiency of Searls' winsorized mean are derived using the theory of extreme order statistics. Various estimators reducing the impact of large data values are compared in a Monte Carlo experiment.

KEY WORDS: Outliers; Max domain of attraction; Mean square error; Simple random sampling; Stratified sampling.

1. INTRODUCTION

Samples drawn from positively skewed populations often contain outliers with values that are much larger than most sampled values. One usually tries to accomodate these large values when designing the survey (Glasser 1962; Hidiroglou 1987). However, given the multipurpose nature of most surveys, statisticians are often faced with outliers at the estimation stage. These data points make classical survey estimators, such as the sample mean, unstable. It is therefore of interest to study alternative estimators that lower the impact of large data values. Winsorization (Searls 1966) consists in replacing the data values larger than a cut-off value R by R before averaging. Searls suggested to select the value of R which minimizes the mean square error of the winsorized mean. One can also take R equal to the second largest data value in the sample (Rivest 1994). Searls' estimator was best among all the methods to adjust large data values studied by Ernst (1980). Hicks and Fetter (1993) implement Searls' winsorization strategy in an agriculture survey. Other strategies have been proposed for dealing with large observations in survey sampling. Chambers and Kokic (1993) review estimators derived from the theory of "Robust Statistics" (Huber 1981). Fuller (1991, 1993) proposes a preliminary test to detect the presence of extreme values in the sample; the impact of these values is lowered only in samples for which this test is significant. Lee (1994) provides a good review of this expanding literature.

The key to the implementation of Searls' winsorization method is the selection of the cut-off R. A simple algorithm for calculating the optimal cut-off for a known population in simple random sampling and in pps sample is proposed in Section 2. Repeated calculations of the optimal cut-off for several populations and several sample sizes reveal that, in most cases, the optimal scheme winsorizes one data point on average, regardless of the sample size. Section 3 extends the result of Section 2 to stratified sampling. A simple algorithm for the calculation of cut-off values in each stratum is proposed. The rule of winsorizing an average of one data point per sample regardless of sample size is shown to hold also in stratified samples. The efficiencies, with respect to the sample mean, of various winsorized estimators are calculated in Sections 4 and 5. Section 4 derives analytic large sample approximations to the efficiency of Searls' estimator using the theory of extreme order statistics while Section 5 compares, in a Monte Carlo study, estimators for reducing the impact of large data values.

2. SAMPLING PROPERTIES OF THE WINSORIZED MEAN

This section studies winsorized means for data drawn from either a continuous or a discrete distribution. Several families of continuous distributions are available to model positive skewed data. One has the Weibull family, $F_\alpha(x) = 1 - \exp(-(x/\beta)^{1/\alpha})$ for $x > 0$, the log-normal family, $F_\nu(x) = \Phi(\log(x/\beta)/\nu)$ for $x > 0$, and the Pareto family, $F_\gamma(x) = 1 - (1 + x/\beta)^{-\gamma}$ for $x > 0$, where β is a positive scale parameter and α, ν, and γ are positive shape parameters. Discrete skewed distributions arise in survey samping. Let $\{y_1, \ldots, y_N\}$ represent the values of the

[1] Louis-Paul Rivest and Daniel Hurtubise, Département de mathématiques et de statistique, Université Laval, Cité Universitaire, Québec, Canada, G1K 7P4.

variable of interest for the N units of a population to be sampled. If a simple random sample with replacement is drawn, then one can take $F(x) = \sum I(y_i \leq x)/N$ as the underlying distribution where $I(\cdot)$ represents the indicator function. In pps sampling, *i.e.*, sampling with replacement and with probabilities given by $\{p_i, i = 1, \ldots, N\}$, one would take $F(x) = \sum p_i I(y_i/(Np_i) \leq x)$. The standard estimator of $\bar{y}$ under pps sampling,

$$\bar{y}_s = \frac{1}{n} \sum_s \frac{y_i}{Np_i}$$

can then be regarded as the mean of a random sample of size n drawn from distribution F. Fuller (1991) provides examples of survey data having skewed distributions.

Let $X_1, X_2, \ldots, X_n$ denote a sample drawn from $F(x)$. In pps sampling, one would have $X_i = y_i/(Np_i)$ where p_i and y_i are the selection probability and the value of the y-variable for the i-th unit selected in the sample. The population mean μ is to be estimated by a winsorized mean,

$$\bar{X}_R = \frac{1}{n} \sum_1^n \min(X_i,R) = \bar{X} - \frac{1}{n} \sum_{i=1}^n \max(X_i - R,0), \tag{2.1}$$

where $\bar{X}$ is the mean of the X_i's. The expectation of $\bar{X}_R$ is equal to

$$E(\bar{X}_R) = \mu - \int_R^\infty (x - R)dF(x) = \mu - \int_R^\infty \int_R^X dydF(x).$$

Changing the order of integration in the above integral proves that $E(\bar{X}_R) = \mu + B(\bar{X}_R)$ where

$$B(\bar{X}_R) = - \int_R^\infty [1 - F(x)]dx \tag{2.2}$$

is the bias of the winsorized mean.

By (2.1), an expression for the variance of $\bar{X}_R$ is

$$n\mathrm{Var}(\bar{X}_R) = \sigma^2 - 2\mathrm{cov}[X_1,\max(X_1 - R,0)] + \mathrm{Var}[\max(X_1 - R,0)]$$

where X_1 is the first random variable in the sample and σ^2 is the variance of $F(x)$. Manipulations similar to those yielding (2.2) show that

$$E[\max(X_1 - R,0)^2] = 2\int_R^\infty (x - R)[1 - F(x)]dx,$$

and

$$E[\max(X_1 - R,0)X_1] = 2\int_R^\infty (x - R)[1 - F(x)]dx - RB(\bar{X}_R).$$

Thus

$$\mathrm{Var}(\bar{X}_R) = \frac{1}{n}\left\{\sigma^2 - 2\int_R^\infty (x - \mu)[1 - F(x)]dx - B^2(\bar{X}_R)\right\},$$

and

$$\mathrm{MSE}(\bar{X}_R) = \frac{\sigma^2}{n} - \frac{2}{n}\int_R^\infty (x - \mu)[1 - F(x)]dx + \frac{n-1}{n} B^2(\bar{X}_R). \tag{2.3}$$

Searls (1966) showed that the mean square error of $\bar{X}_R$ has a unique minimum which can be obtained by equating the derivative, with respect to R, of $\mathrm{MSE}(\bar{X}_R)$ to 0. This yields the following equation for the optimal winsorization constant $R(F,n)$,

$$\frac{R - \mu}{n - 1} - \int_R^\infty [1 - F(x)]dx = 0. \tag{2.4}$$

This is equivalent to equation (14) in Searls (1966). In the remainder of this work, $\bar{X}_R$ denotes the optimal winsorized mean obtained with the winsorization constant $R(F,n)$ which solves (2.4). Observe that the optimal cut-off point $R(F,n)$ is location and scale equivariant, *i.e.*, if $G(x) = F[(x - b)/a]$, then $R(G,n) = aR(F,n) + b$.

A general algorithm for solving (2.4) is easily constructed. First observe that as a function of R, the left hand side of equation (2.4) is increasing and concave in R since its derivative, $1/(n - 1) + 1 - F(R)$, is positive and decreasing. Therefore, the Newton-Raphson algorithm (Thisted 1988, 164-167) given by

$$R_{j+1} = R_j - \frac{(R_j - \mu) - (n - 1)\int_{Rj}^\infty [1 - F(x)]dx}{1 + (n - 1)[1 - F(R_j)]}, \tag{2.5}$$

with $R_0 = 2\mu$ as starting value converges smoothly to the solution of (2.4). For discrete distributions the computations are easily implemented by noting that

$$\int_R^\infty [1 - F(x)]dx = E[\max(X - R,0)].$$

Exact calculations of the optimal cut-off points $R(F,n)$ were carried out for the Weibull, the log-normal, and the Pareto families for samples of size s ranging between 5 and 200. Three distributions, corresponding to coefficients of variation (CV) of 1, 2, and 4, were considered in each family except for the Pareto family where only coefficients of variation of 2 and 4 were considered. The CV measures the skewness of a distribution, with large CVs corresponding to heavy skewness. The corresponding parameter values are given in Table 1.

Table 1

Parameter values of the distributions for which optimal cut-off values $R(F,n)$ were evaluated

CV	Weibull(α)	Log-normal(ν)	Pareto(γ)
1	1	0.83	-
2	1.84	1.27	2.67
4	2.87	1.68	2.13

For each distribution and each sample size, the optimal cut-off point was calculated using algorithm (2.5). Figure 1 presents the expected number of winsorized observations, $m(F,n) = n\{1 - F[R(F,n)]\}$ as a function of n while the corresponding efficiencies are reported in Figure 2. The efficiency of $\bar{X}_R$ is defined as $\mathrm{Var}(\bar{X})/\mathrm{MSE}(\bar{X}_R)$.

In Figure 1 the expected number of winsorized data values under the optimal scheme is, for most skewed distributions, close to 1 even for large sample sizes. Approximating this number by a Poisson distribution with parameter $m(F,n)$ shows there is a non-negligible probability that, under the optimal winsorization scheme, none of the data points is winsorized. This probability increases with the skewness of the distribution since $m(F,n)$ decreases with the CV. Thus, in samples from a highly skewed distribution, it is not always appropriate to winsorize the largest values. Such values should be winsorized only when they are large. As expected, in Figure 2, the largest gains in efficiency are obtained when the skewness is heavy. Therefore monitoring the two or three largest data values in a sample and curtailing their impact when these values are large is the key to a good winsorization strategy.

Figure 1 shows that the expected number of winsorized data values $m(F,n)$ decreases with the skewness of the distribution. This observation can be turned into a rigorous mathematical result. To this end, random variable Y is said to be more skewed than random variable X if Y has the same distribution as $\psi(X)$ where $\psi(x)$ is a convex function

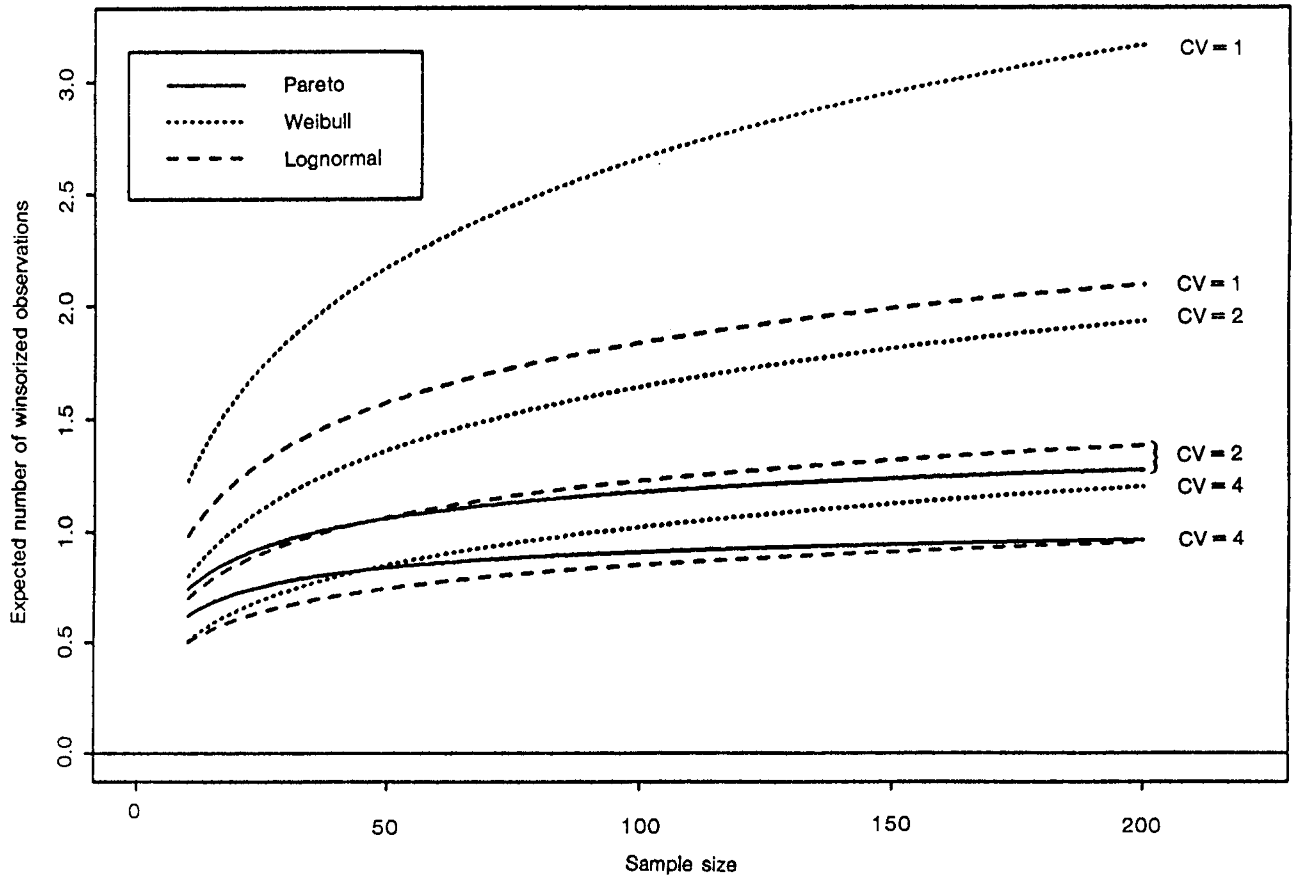

Figure 1. Expected number of winsorized observations for simple and stratified random sampling.

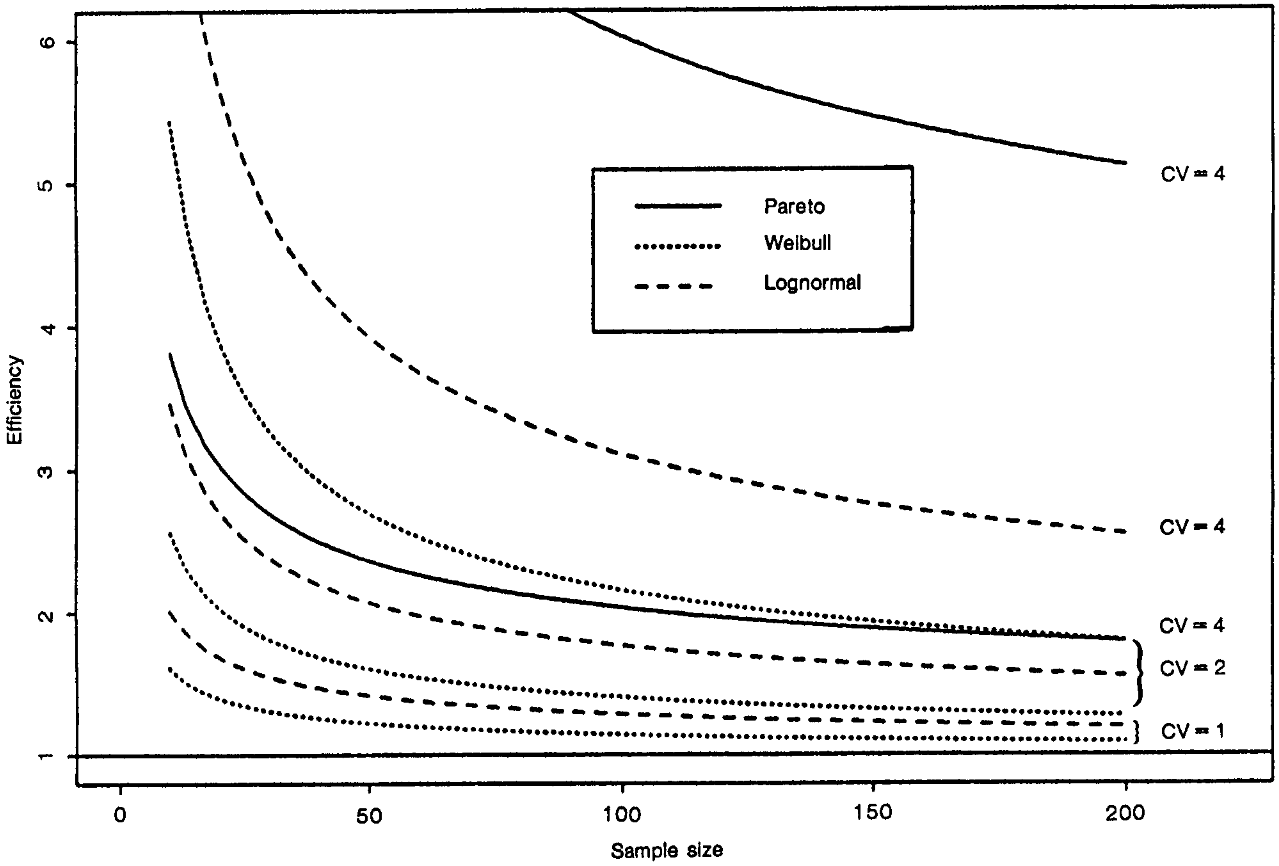

Figure 2. Efficiency of Searls winsorized mean.

of x. Under this definition X^2 is, as should be expected, more skewed than X. This notion of skewness corresponds to the convex partial ordering of van Zwet (Barlow and Proschan 1981). With this definition of skewness, one has the following proposition which is proved in the Appendix together with Propositions 2 and 3.

Proposition 1 If Y is more skewed than X then $m(F_X,n) > m(F_Y,n)$ where F_X and F_Y are the distributions of X and Y respectively.

The results of this section also apply to simple random sampling without replacement. For this design the mean square error of $\bar{X}_R$ is given by formula (2.3) with n replaced by $n/(1-f)$ where f is the sampling fraction. Algorithm (2.5), with n divided by $(1-f)$, can be used for calculating optimal cut-off values for without replacement simple random sampling.

3. WINSORIZATION IN STRATIFIED SAMPLING

There are many ways to generalize Searls' winsorization strategy to stratified sampling. In this section each stratum has its own cut-off value. Let R_h be the cut-off value in stratum h. The optimal values of $R_1, R_2, \ldots, R_L$, where L is the number of strata, are the ones that minimize the mean square error of $\bar{X}_R = \sum W_h \bar{X}_{Rh}$, where $\bar{X}_{Rh} = \sum \min(X_{hi}, R_h)/n_h$, $W_h = N_h/N$ and N_h is the size of stratum h and $N = \sum N_h$. An algorithm for determining these optimal cut-off values is proposed in this section.

Let $F_h(x)$, for $h = 1, \ldots, L$ be the distribution of X in stratum h, and μ_h and σ_h^2 be the mean and the variance of F_h. The derivation of the mean square error of $\bar{X}_R$, under with replacement stratified random sampling, follows that presented in Section 2, it gives

$$\mathrm{MSE}(\bar{X}_R) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left(\sigma_h^2 - 2\int_{R_h}^{\infty} (x-\mu_h)[1-F_h(x)]\,dx - B^2(\bar{X}_{Rh})\right) + \left(\sum_{h=1}^{L} W_h B(\bar{X}_{Rh})\right)^2 \qquad (3.1)$$

where $B(\bar{X}_{Rh})$ is the bias of $\bar{X}_{Rh}$ as an estimator of μ_h

$$B(\bar{X}_{Rh}) = -\int_{R_h}^{\infty} [1 - F_h(x)]dx.$$

Taking the partial derivatives with respect to R_h, $h = 1, \ldots, L$ yields the following equations for the optimal values:

$$\frac{W_h}{n_h}[R_h - \mu_h - B(\bar{X}_{Rh})] = -\sum_{h=1}^{L} W_h B(\bar{X}_{Rh}), \quad (3.2)$$

for $h = 1, \ldots, L$.

There is no simple way to solve (3.2). An approximate solution can be obtained by noting that $B(\bar{X}_{Rh})/n_h$ is, for all values of h, usually small as compared to the other terms. Dropping these terms leads to

$$\frac{W_h}{n_h}(R_h - \mu_h) = -\sum_{h=1}^{L} W_h B(\bar{X}_{Rh}), \quad (3.3)$$

for $h = 1, \ldots, L$. The solutions to (3.3) overestimate slightly the optimal values satisfying (3.2) since at these solutions the partial derivatives of (3.1) are all positive and since these partial derivatives are increasing functions of R_h, for $h = 1, \ldots, L$. Thus by solving (3.3) to estimate the cut-off values one does not run the risk of winsorizing too many data values. Equations (3.3) imply that $R_h = \mu_h + n_h R/(nW_h)$ where R is some positive constant. A simple equation for R is obtained by changing variable $y = nW_h(x - \mu_h)/n_h$ in the integrals for $B(\bar{X}_{Rh})$, $h = 1, \ldots, L$ where $n = \sum n_h$. This gives

$$-\sum_{h=1}^{L} W_h B(\bar{X}_{Rh}) = \frac{R}{n} = \int_{R}^{\infty}[1 - F(y)]dy = -B(\bar{X}_R), \quad (3.4)$$

where $F(y) = \sum n_h F_h[\mu_h + n_h y/(nW_h)]/n$. Equation (3.4) is easily solved using algorithm (2.5) proposed in Section 2 for the single sample case. Therefore simple approximations for Searls' optimal cut-off values in stratified sampling are easily calculated.

Since the distribution F defined above has a zero expectation, the mean square error of the stratified winsorized mean obtained by solving (3.3) is equal to:

$$\mathrm{MSE}(\bar{X}_R) = \frac{1}{n}\left(\sigma_F^2 - 2\int_R^{\infty} y[1 - F(y)]dy - B(\bar{X}_R)^2\right) + B(\bar{X}_R)^2 + \left(\frac{1}{n}B(\bar{X}_R)^2 - \sum_{h=1}^{L}\frac{W_h^2 B^2(\bar{X}_{Rh})}{n_h}\right) \quad (3.5)$$

where σ_F^2 is the variance of F. The last term of (3.5) is easily shown to be negative or null; it is null when $B(\bar{X}_R) = nW_h B(\bar{X}_{Rh})/n_h$ for $h = 1, \ldots, L$. The variance of the stratified mean, $\bar{X} = \sum W_h \bar{X}_h$, is equal to σ_F^2/n. Thus a conservative approximation to the efficiency of $\bar{X}_R$ with respect to $\bar{X}$ in stratified sampling is equal to the corresponding efficiency for a random sample of size n drawn from F. Note also that $n[1 - F(R)]$ represents the expectation of the total number of winsorized data points in the L strata.

The optimal winsorization scheme obtained by solving (3.3) has a simple form for many allocation rules. Under proportional allocation, *i.e.*, $n_h = nW_h$ for $h = 1, \ldots, L$, one gets $R_h = \mu_h + R$. Under Neyman optimal allocation, with $n_h = nW_h\sigma_h/(\sum W_h\sigma_h)$ where σ_h is stratum h's standard deviation, one gets $R_h = \mu_h + \sigma_h R/(\sum W_h\sigma_h)$. If in addition, the distributions of X within the strata are equal up to a change in location and scale, *i.e.*, $F_h = F_0[(x - \mu_h)/\sigma_h]$ for some distribution F_0, then $F(x) = F_0[x/(\sum W_h\sigma_h)]$. In this case the characteristics of optimal winsorized means in stratified sampling and in simple random sampling are the same. Thus Figure 1 presents the expected total number of winsorized data points in the L strata as a function of the total sample size n, under Neyman allocation, when F_0 is one of the distributions of Table 1. Figure 2 gives the corresponding efficiencies.

The results of this section are easily generalized to without replacement stratified sampling by replacing n_h by $n_h/(1 - f_h)$ throughout the calculations. The derivation of optimal cut-off values for stratified pps sampling is easily carried out by taking $F_h(x) = \sum p_{hi} I(y_{hi}/(N_h p_{hi}) \leq x)$ where p_{hi} denotes the selection probability for unit the i-th unit of stratum h.

4. LARGE SAMPLE APPROXIMATIONS TO THE EFFICIENCY OF THE WINSORIZED MEAN

For most distributions, equation (2.3) defining the optimal cut-off does not have an explicit solution. This section derives closed form approximations to this solution using the theory of extreme order statistics. This will permit the derivation of explicit approximations to the efficiency of the optimal winsorized mean. Searls' optimal winsorization strategy will then be compared to a simple non parametric winsorization scheme where the largest order statistic is replaced by the second largest (Rivest 1994).

The form of the approximation to $R(F,n)$ depends on the limiting distribution, as the sample size n goes to infinity, of the largest order statistic suitably normalized. For distributions whose support is the positive axis, there are only two possible limiting distributions which are given by Galambos (1987, p. 53-54)

$$H_{1,\alpha}(x) = \exp(-x^{-\alpha}) \quad \text{for} \quad x > 0 \quad \text{and} \quad \alpha > 0$$

and

$$H_{3,0}(x) = \exp[-\exp(-x)] \quad \text{for} \quad x \quad \text{in} \quad R.$$

For many distributions used for the statistical analysis of positive random variables, for example the Weibull and the log-normal families, the sample maximum suitably normalized converges to $H_{3,0}(x)$. Distributions whose sample maxima converge to $H_{1,\alpha}(x)$ for some $\alpha > 0$ have heavy tails. For such distributions $1 - F(x)$ goes to 0 at a rate of $O(x^{-\alpha})$. The Pareto and the F distributions are in this class.

Distributions whose sample maxima converge to $H_{3,0}(x)$ are considered first.The following characterization is due to von Mises (1964): the sample maximum of a twice differentiable distribution $F(x)$ converges to $H_{3,0}(x)$ if, as x goes to ∞,

$$\lim_x \frac{g'(x)}{g^2(x)} = 0 \tag{4.1}$$

where $f(x)$ is the density of F, $g(x) = f(x)/[1 - F(x)]$ is the failure rate of F, and g' is the derivative of g. An approximation to winsorization constant $R(F,n)$ for this class of distributions is presented next.

Proposition 2 If $F(x)$ is such that (4.1) holds and if, for large values of x, it satisfies:

i) $xg(x)$ increases;
ii) $xg'(x)/g(x)$ is less than some positive constant c;

then the optimal winsorization constant $R(F,n)$ satisfies

$$R(F,n) =$$

$$F^{-1}\left(1 - \frac{g[F^{-1}(1-1/n)]F^{-1}(1-1/n)[1+o(1)]}{n}\right);$$

and $m(F,n) = g(F^{-1}(1 - 1/n))F^{-1}(1 - 1/n)$ $[1 + o(1)]$. Furthermore, the mean squared error of Searls' winsorized mean is approximately equal to

$$\text{MSE}(\bar{X}_R) \approx \frac{\sigma^2}{n} - \frac{R(F,n)^2}{n^2}.$$

In the Weibull family, $F_\alpha^{-1}(1 - t) = [-\log(t)]^\alpha$, $g(x) = x^{1/\alpha - 1}/\alpha$. The hypotheses of Proposition 2 are met and $m(F_\alpha, n)$, the expected number of winsorized observations in a large Weibull sample, is $\log(n)[1 + o(1)]/\alpha$ which goes to ∞ as n increases. Figure 1 suggests that the convergence is very slow, especially for large coefficients of variation.

Now consider distributions whose sample maxima converge to $H_{1,\alpha}(x)$. This class of distributions has been characterized by Gnedenko (1962): the sample maximum of F converges to $H_{1,\alpha}(x)$ if one can write

$$1 - F(x) = L(x)/x^\alpha \tag{4.2}$$

where as x goes to ∞, $L(x)/L(kx)$ converges to 1, for any constant k. Note that for F to have a finite second moment, one needs $\alpha > 2$ in (4.2). The Pareto distribution satisfies (4.2) with $\alpha = \gamma$.

Proposition 3 If F satisfies (4.2) with parameter α where $\alpha > 2$, then as n goes to infinity, $R(F,n) = F^{-1}[1 - (\alpha - 1)/n][1 + o(1)]$, *i.e.*, $m(F,n) \approx \alpha - 1$. Furthermore,

$$\text{MSE}(\bar{X}_R) \approx \frac{\sigma^2}{n} - \frac{\alpha R(F,n)^2}{n^2(\alpha - 2)}.$$

For distributions satisfying (4.2) a finite number of data points are on average winsorized as the sample size goes to ∞. To some extent, this can be seen in Figure 1 where the curves of $m(F_\gamma, n)$ for the Pareto distribution have $m(F_{2.33}, n) = 1.33$, and $m(F_{2.67}, n) = 1.67$ as asymptotes.

Propositions 2 and 3 shed some light on the estimation of the optimal cut-off value. When F is unknown, a possible estimator for $R(F,n)$ is the value that minimizes an estimator of the mean square error of $\bar{X}_R$. This leads to

$$\frac{R - \bar{X}}{n - 1} = \frac{1}{n}\sum_{i=1}^{n} \max(X_i - R, 0) \tag{4.3}$$

as an estimating function for R. This procedure is questionable when the underlying distribution is highly skewed, *i.e.*, when F satisfies the assumption of Proposition 3. On average, there will only be $\alpha - 1$ non-null terms in the right hand side of equation (3). Thus $\hat{R}$ will, on average be determined by the $\alpha - 1$ largest data values and the sample maximum will have the largest influence on $\hat{R}$. This will make $\hat{R}$ highly unstable and, considering the findings of Figure 1, the second largest sample order statistic should be a better estimator of $R(F,n)$ than the solution of (3.3). This is exemplified in the Monte Carlo simulations of Section 5.

Table 2 compares approximations to the bias and to the mean square error of Searls' winsorized mean $\bar{X}_R$ to those of the once winsorized mean $\bar{X}_1$ obtained by taking the cut-off value R equal to the second largest observation. Rivest (1994) shows this choice of cut-off value yields the optimal non-parametric winsorized mean. He also derives the large sample approximations for the bias and the mean square error of $\bar{X}_1$ appearing in Table 2. The corresponding expressions for $\bar{X}_R$ are taken in Propositions 2 and 3.

Table 2

Approximations to the bias and to the mean square error of the once winsorized mean $\bar{X}_1$ and of Searls' optimal winsorized mean, $\bar{X}_R$, for the Weibull and for the Pareto distribution ($\Gamma(\cdot)$ stands for the gamma function)

		WEIBULL	PARETO
$\bar{X}_R$	MSE	$\frac{\sigma^2}{n} - \frac{(\log n)^{2\alpha}}{n^2}$	$\frac{\sigma^2}{n} - \frac{\gamma}{(\gamma - 2)(\gamma - 1)^{2/\gamma} n^{2-2/\gamma}}$
	bias	$-\frac{(\log n)^{\alpha}}{n}$	$-\frac{1}{(\gamma - 1)^{1/\gamma} n^{1-1/\gamma}}$
$\bar{X}_1$	MSE	$\frac{\sigma^2}{n} - \frac{2\alpha(\alpha - 1)(\log n)^{2\alpha-2}}{n^2}$	$\frac{\sigma^2}{n} - \frac{2\Gamma(1 - 2/\gamma)}{\gamma(\gamma - 1)n^{2-2/\gamma}}$
	bias	$-\frac{\alpha(\log n)^{\alpha-1}}{n}$	$-\frac{\Gamma(1 - 1/\gamma)}{\gamma n^{1-1/\gamma}}$

In Table 2 the mean square error of $\bar{X}_R$ is much smaller than that of $\bar{X}_1$. Indeed, for the Weibull distribution the large sample efficiency of $\bar{X}_R$ with respect to $\bar{X}_1$ is equal to that of $\bar{X}_R$ with respect to $\bar{X}$. Thus non-parametric winsorization reduces the mean square error of estimators of the mean of a skewed population however further reductions in mean square error can be obtained if information concerning the underlying distribution is available. This is illustrated in the Monte Carlo comparisons presented in the next section.

The results of this section apply to stratified sampling. For this design, the large sample solution to equation (3.4) is determined by the stratum with the most skewed distribution. If F_1 is the most skewed distribution then $nW_1R(F_1,n_1)/n_1$ is an approximate solution to (3.4) where an approximation to $R(F_1,n_1)$ is found in Proposition 2 or in Proposition 3 depending on the tail of F_1. In this case only data points in stratum 1 are winsorized in large stratified samples. Searls' winsorized mean is then equal to W_1 times the optimal winsorized mean for stratum one plus a weighted sum of the sample means in the other strata.

5. MONTE CARLO COMPARISONS OF ESTIMATORS OF THE MEAN OF A SKEWED DISTRIBUTION

This section presents Monte Carlo comparisons of the mean square error and of the biases of five estimators of the mean of population CHICKEN of Fuller (1991). This population has 2000 units; its coefficient of variation is 4.46. Further numerical comparisons of the five estimators considered in this section for other distributions, either finite or infinite, are presented in Rivest (1993a and b).

The five estimators under consideration are:

- Searls' winsorized estimator, $\bar{X}_R$, calculated as if the the underlying distribution was known;
- A winsorized estimator where the cut-off value is set equal to the second largest data value of an auxiliary sample of size $2n$; this is an instance where limited auxiliary information concerning the underlying distribution F is available (in the Monte Carlo simulations each simulated sample had its own auxiliary sample);
- The once winsorized mean, $\bar{X}_1$, introduced in Section 4;
- A winsorized estimator where R is estimated from the sample by solving equation (4.3);
- Fuller preliminary test estimator with $j = 3$ (*i.e.*, the numerator of the preliminary test involves the three largest observations), T (the total number of data points involved in the preliminary test) equal to $[4n^{1/2} - 10]$ and K_3, the cut-off value equal to 3.5. A detailed description of this estimator appears in Fuller (1991) and in Rivest (1993a and b). This estimator curtails the largest data values only when a test statistic for detecting extreme data values is significant.

The biases and the efficiencies of $\bar{X}_R$ were calculated exactly. For the other estimators, the biases and the efficiencies presented in Figures 1 and 2 were obtained in Monte Carlo simulations based on 100,000 repetitions.

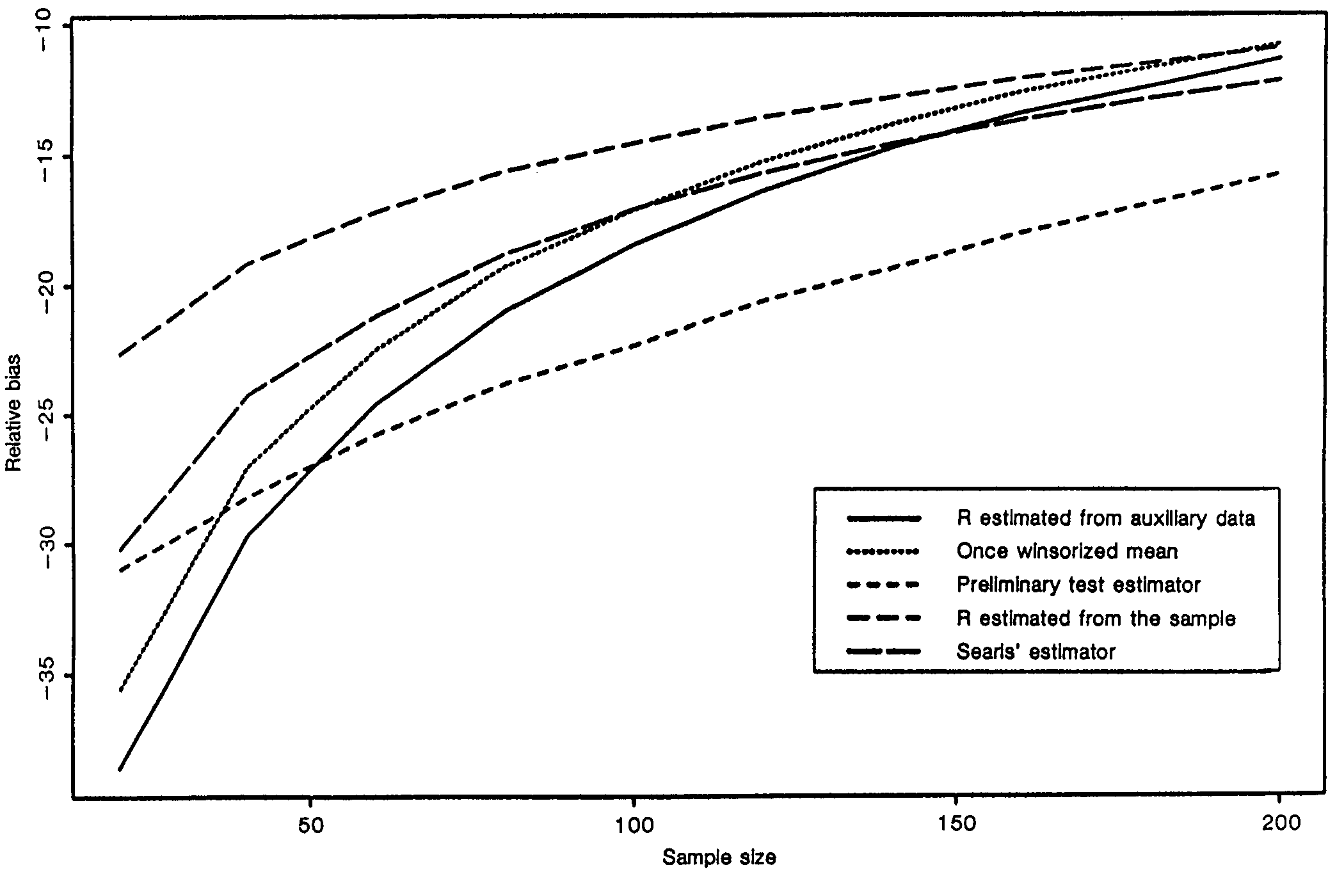

Figure 3. Relative bias of five estimators for the mean of CHICKEN.

Figure 3 indicates that the biases of winsorized estimators are important, even in large samples. Several interesting conclusions can be drawn from Figure 4. First, as expected from Table 2 Searls' estimator is much more efficient than the once winsorized mean. Estimating the optimal cut-off value using limited auxiliary information is highly efficient. This holds true as long as the study variable can be modeled by a superpopulation distribution having a finite variance, see Rivest (1993a) for further discussions. In a sampling context, the auxiliary samples could be data from previous surveys standardized to account for possible changes over time in the distribution of the variable under study.

Among the three estimators of Figure 4 that do not rely on auxiliary information, Fuller estimator is the best. This is in agreement with the simulation results of Fuller (1991). Estimating the cut-off value by minimizing an estimate of the mean square error does poorly especially in small samples. Thus, as shown in Section 4, the resulting estimator is highly sensitive to the wild data values that sometimes appear in small samples. This estimator is not recommended.

6. CONCLUSIONS

Many strategies can be used to accomodate the large values that sometimes arise in surveys. If auxiliary information, such as census data, is available then one can use Searls' estimator in either simple random sampling, stratified sampling, or pps sampling. Since the cut-off values are fixed constant mean square error estimators can be derived from formulae (2.3) and (3.1).

When extra information is not available, the once winsorized mean and Fuller preliminary test estimator can be used. Research is now under way to generalize these estimators to stratified designs. An estimator for the mean square error of the once winsorized mean is proposed in Rivest (1994),

$$v(\bar{X}_1) = \frac{1}{n} S^2 - \frac{1}{n^2} (X_n + X_{n-1} - 2\bar{X}_1) \\ (X_n - 3X_{n-1} + 2X_{n-2})$$

where S^2 denotes the variance of the X-sample and $X_n > X_{n-1} > X_{n-2}$ denote the three largest data values in

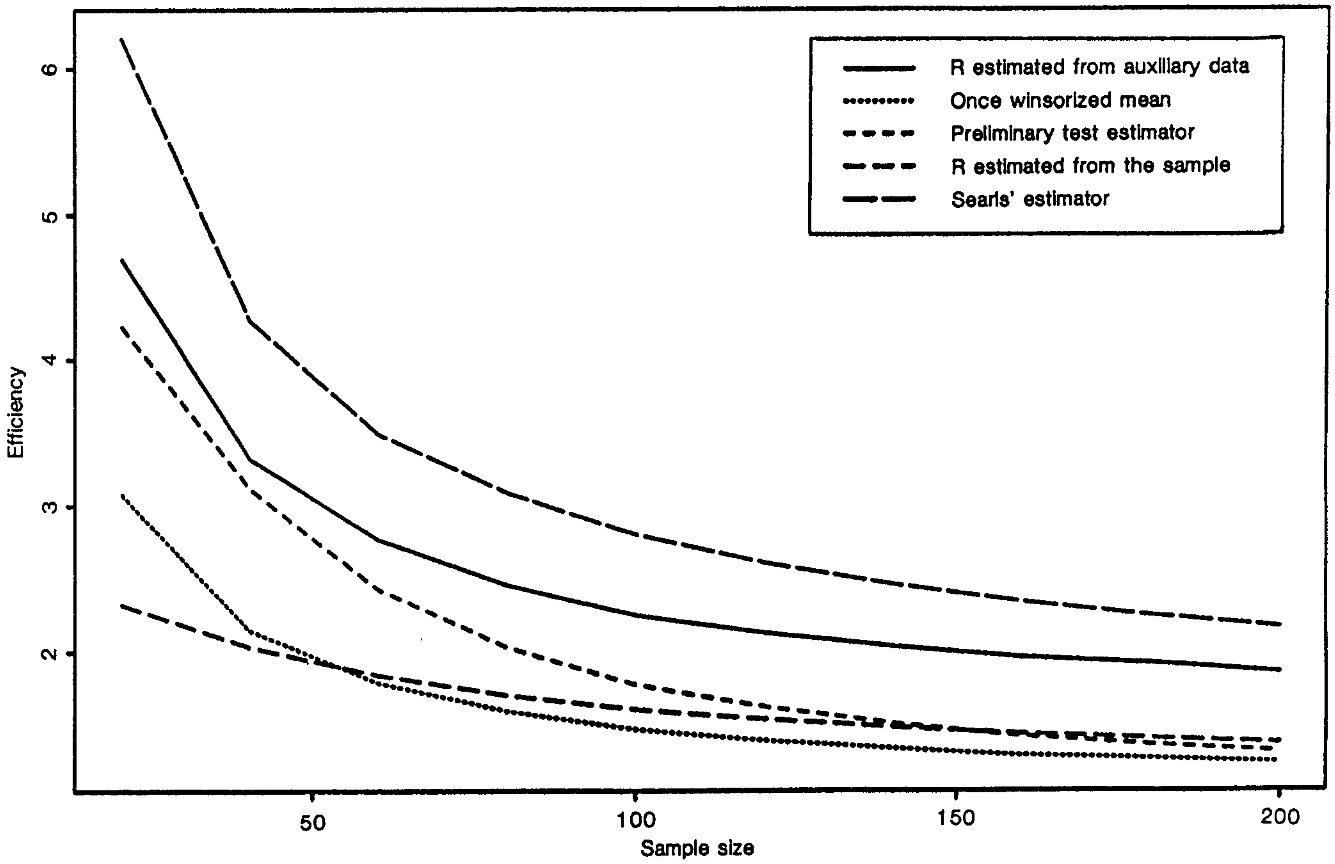

Figure 4. Efficiency of five estimators for the mean of CHICKEN.

that sample. This estimator has a small bias in infinite populations. However the coverage of the standard confidence interval $\bar{X}_1 \pm z_{1-\alpha/2}\sqrt{v(\bar{X}_1)}$ is often well below the nominal $100(1 - \alpha)\%$ level especially when the underlying distribution is skewed. Further research is needed to obtain reliable confidence intervals for estimators of the mean of skewed populations.

ACKNOWLEDGEMENTS

This research was supported by the Natural Science and Engineering Research Council and by the Fond pour la formation des chercheurs et l'aide à la recherche of Québec.

APPENDIX 1

Proof of Proposition 1 The assumption that Y is more skewed than X implies that there exists a convex function ψ such that $\psi(X)$ and Y have the same distribution. Let R denote $R(F_X,n)$. To prove the result, it suffices to show that $\psi(R) < R(F_Y,n)$. This is equivalent to

$$\frac{\psi(R) - E(Y)}{n - 1} < \int_{\psi(R)}^{\infty} [1 - F_Y(x)]dx. \qquad \text{(A.1)}$$

By Jensen's inequality, $E(Y) = E[\psi(X)] > \psi[E(X)]$. Thus using (2.3), the left hand side of (A.1) is less than or equal to

$$\frac{R - E(X)}{n - 1}\,\frac{\psi(R) - \psi[E(X)]}{R - E(X)} < \frac{R - E(X)}{n - 1}\,\psi'(R) =$$

$$\int_{R}^{\infty} [1 - F_X(y)]dy \cdot \psi'(R)$$

where ψ' is the derivative of ψ. Since ψ' is increasing, the left hand side of the above inequality is less than or equal to:

$$\int_{R}^{\infty} \psi'(y)[1 - F_X(y)]dy = \int_{\psi(R)}^{\infty} [1 - F_Y(x)]dx.$$

This shows that (A.1) holds.

Proof of Proposition 2 The following result obtained by applying Theorems 2.7.5 and 2.7.11 of Galambos (1987) to the distribution $F(z^{p+1})$ is used extensively. If the sample maxima of distribution $F(x)$ converges to $H_{3,0}(x)$, then all the moments of F exist and

$$\int_x^\infty y^p[1 - F(y)]dy \sim \frac{[1 - F(x)]\,x^p}{g(x)} \tag{A.2}$$

where $g(x) \sim h(x)$ means that $g(x)/h(x)$ converges to 1 as x goes to infinity. Using (A.2), $R(F,n)$ is obtained by solving

$$\frac{R - \mu}{n - 1} = \frac{1 - F(R)}{g(R)}\,(1 + o(1)).$$

Let $R = F^{-1}(1 - a/n)$, then, up to $(1 + o(1))$, the above equation becomes

$$a = g\left[F^{-1}\left(1 - \frac{a}{n}\right)\right]F^{-1}\left(1 - \frac{a}{n}\right). \tag{A.3}$$

Let $a_0 = g[F^{-1}(1 - 1/n)]F^{-1}(1 - 1/n)$ and $a_1 = g$ $[F^{-1}(1 - a_0/n)]F^{-1}(1 - a_0/n)$. Since for large values of $x,xg(x)$ is increasing, $a_0 > a_1$ and the solution to (A.3) belongs to the interval (a_1,a_0). In order to prove the result, one has to show that a_1/a_0 converges to 1 as n goes to ∞.

Since $g(x) = f(x)/[1 - F(x)]$, one can write

$$a_0 = \exp\left[\int_{F^{-1}(1-a_0/n)}^{F^{-1}(1-1/n)} g(t)dt\right] =$$

$$\exp\left[a_0 - a_1 - \int_{F^{-1}(1-a_0/n)}^{F^{-1}(1-1/n)} tg'(t)dt\right],$$

where the second expression is obtained by integrating by parts. Since $tg'(t)/g(t)$ is less then c, one has $a_0 >$ $\exp(a_0 - a_1)a_0^{-c}$. If a_1/a_0 does not converge to 1, say $a_1/a_0 < 1 - \epsilon < 1$ for an infinite sequence of sample sizes, the previous inequality implies that $a_0^{1+c} > \exp(a_0\epsilon)$. This is a contradiction since a_0 tends to ∞ as n becomes large. The approximation for MSE$(\bar{X}_R)$ is obtained by using (A.2) with $p = 2$.

Proof of Proposition 3 If the sample maxima of distribution $F(x)$ converges to $H_{1,\alpha}(x)$ then F satisfies the following properties (Feller 1971, p. 281):

$$\int_x^\infty y^p[1 - F(y)]dy \sim \frac{[1 - F(x)]x^{p+1}}{\alpha - p - 1} \tag{A.4}$$

for any p such that $\alpha - p - 1 \geq 0$. By (A.4), $R(F,n)$ is obtained by solving $F(R) = 1 - [\alpha - 1 + o(1)]/n$. This leads to the approximation for $R(F,n)$. To derive the approximation for MSE$(\bar{X}_R)$, one applies (A.4) with $p = 1$.

REFERENCES

BARLOW, R.E., and PROSCHAN, F. (1981). *Statistical Theory of Reliability and Life Testing*. Silver Spring MD: To Begin With.

CHAMBERS, R.L., and KOKIC, P.N. (1993). Outlier robust sample survey inference. *Bulletin of the International Statistical Institute. Proceedings of the 49th session*, book 2, 54-72.

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

FELLER, W. (1971). *An Introduction to Probability Theory and its Applications*. Volume II. Second Edition. New York: Wiley.

FULLER, W.A. (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica*, 1, 137-158.

FULLER, W.A. (1993). Estimators for long-tailed distributions. *Bulletin of the International Statistical Institute. Proceedings of the 49th session*, book 2, 39-54.

GALAMBOS, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*. Second edition. Malabar FL: Krieger.

GNEDENKO, B.V. (1962). *The Theory of Probability*. New York: Chelsea.

GLASSER, G.J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, 30, 28-32.

HICKS, S., and FETTER, M. (1993). An evaluation of robust estimation techniques for improving estimates of total hogs. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 385-389.

HIDIROGLOU, M.A. (1987). The construction of a self-representing stratum of large units in survey design. *The American Statistician*. 40, 27-31.

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

LEE, H. (1994). Outliers in Survey Data. Statistics Canada paper.

RIVEST, L.-P. (1994). Some sampling properties of winsorized means for skewed distributions. *Biometrika*, 81, 373-384.

RIVEST, L.-P. (1993a). Winsorization of survey data. *Bulletin of the International Statistical Institute. Proceedings of the 49th session*, book 2, 73-89.

RIVEST, L.-P. (1993b). Winsorization of survey data. *Proceedings of the International Conference on Establishment Surveys*. Alexandria, Virginia: American Statistical Association, 396-401.

SEARLS, D.T. (1966). An estimator which reduces large true observations. *Journal of the American Statistical Association*, 61, 1200-1204.

THISTED, R.A. (1988). *Elements of Statistical Computing*. New York: Chapman and Hall.

VON MISES, R. (1964). *Selected Papers of Richard von Mises*. Volume II. Providence: American Mathematical Society.