# ROBUST MULTIVARIATE OUTLIER DETECTION USING MAHALANOBIS' DISTANCE AND MODIFIED STAHEL-DONOHO ESTIMATORS

**Sarah Franklin, Steve Thomas, Marie Brodeur Statistics Canada**
**Sarah Franklin, Statistics Canada, SSMD, R.H. Coats Bldg, 15th floor, Ottawa, Ontario, Canada, K1A 0T6**
**fransar@statcan.ca**

## ABSTRACT

This paper illustrates the practical application of a robust multivariate outlier detection method used to edit survey data. Outliers are identified by calculating Mahalanobis' distance, where the location vector and scatter matrix are robustly estimated using modified Stahel-Donoho estimators. This method of multivariate outlier detection has been successfully employed at Statistics Canada for several years by the Annual Wholesale and Retail Trade Survey and the Monthly Survey of Manufacturers. Currently, none of these surveys uses sampling weights during outlier detection. We propose a simple method for incorporating the weights. In order to compare outlier detection with and without the use of sampling weights, results are presented for a simulated contaminated bivariate population.

**Key Words: Covariance matrix; Mahalanobis distance; Projection pursuit; Editing; Robust estimators.**

## 1. INTRODUCTION

Barnett and Lewis (1995) define an outlier to be "an observation or subset of observations which appears to be inconsistent with the remainder" of the dataset. Outliers have different sources: they may be the result of an error, they may arise from the inherent variability of the dataset (i.e. extreme values from the tails of the distribution) or they may be generated from another model. Outliers are sometimes referred to as contaminants.

In sample surveys, since the sample is drawn from a finite population, outlier detection can be performed with or without the sampling weights. Thus, it is possible to make the distinction between extreme values versus influential observations. Influential observations are ones for which the combination of the reported value and the sampling weight have a large influence on the estimate. Extreme values may or may not be influential.

Outliers may be univariate or multivariate. Multivariate outliers are observations that are inconsistent with the correlational structure of the dataset. Thus, while univariate outlier detection is performed independently on each variable, multivariate methods investigate the relationship of several variables. A classical way of identifying multivariate outliers in a multivariate normal dataset is to calculate Mahalanobis' distance. This distance uses estimates of the location and scatter to identify values that are far away from the main cloud of data. Since outliers are present in the data, robust estimators of the multivariate location vector and scatter matrix should be used.

The first affine equivariant multivariate location estimator robust enough to tolerate up to 50% of outliers in the sample before it breaks down was independently discovered by Stahel (1981) and Donoho (1982). They proposed a weighted mean, where an observation's weight is calculated by looking at all possible univariate projections of the data. If there is a projection for which the observation is an outlier, then the weight is decreased. This method of robust estimation, which consists of analysing the structure of a multivariate dataset by looking at various projections, falls under projection pursuit techniques.

The Annual Wholesale and Retail Trade Survey (AWRTS) and the Monthly Survey of Manufacturers (MSM) use a robust multivariate outlier detection method. These two surveys perform outlier detection shortly after collection by calculating Mahalanobis' distance, where the mean vector and covariance matrix are robustly estimated using modified Stahel-Donoho estimators proposed by Patak (1990).

Until now, both surveys performed outlier detection without using sampling weights. Since influential values are of most interest to analysts, we have studied the impact of performing multivariate outlier detection with and without sampling weights. The results for MSM are presented in this paper.

The paper will begin with a general description of outlier detection, briefly describing some of the most popular univariate and multivariate methods used at Statistics Canada. The principle behind robust estimation and the modified Stahel-Donoho estimators used by AWRTS and MSM are described in sections 3 and 4. Section 5 presents results for simulated contaminated bivariate populations, comparing outlier detection with and without the use of sampling weights.

## 2. OUTLIER DETECTION

Outlier detection is usually part of the editing process of a sample survey. While most surveys collect multivariate data, univariate outlier detection methods are usually favoured for their simplicity. But these methods fail to detect observations that violate the correlational structure of the dataset. In surveys where the variables are highly correlated, this is a disadvantage. As Barnett and Lewis (1995) point out: 'A multivariate outlier need not be an extreme in any of its components. Someone who is short and fat need not be the shortest, or the fattest, person around. But that person can still be an outlier.'

Most outlier detection methods use some measure of distance to evaluate how far away an observation is from the centre of the data. To measure this distance, the sample mean and variance may be used but since they are not robust to outliers, they can mask the very observations we seek to detect. To avoid this masking effect, robust scale and location estimators, which are inherently resistant to outliers, may be used. It is for this reason that many outlier detection methods use order statistics, such as the median or quartile.

### 2.1. Univariate Outlier Detection Methods

Perhaps the most popular univariate outlier detection technique for survey data is the quartile method. This method creates an allowable range for the data using lower and upper quartiles: data falling outside of the range are outliers. The method is not only robust, but simple and non-parametric. Hidiroglou and Berthelot (1986) proposed an adaptation of the quartile method for trend data where the trends are first transformed to dampen a size masking effect.

These two quartile techniques are the ones most commonly used at Statistics Canada. Many surveys also use less formal ones, for example graphical methods. Lee, et al. (1992) provides a comprehensive review of outlier detection methods employed at Statistics Canada.

### 2.2. Some Multivariate Outlier Detection Methods

Let $\boldsymbol{X}=\{\boldsymbol{x}_1,..,\boldsymbol{x}_i,..,\boldsymbol{x}_n\}$ represent a set of $n$ data points in $R^p$, where the $i$th observation is $\boldsymbol{x}_i^T=(\boldsymbol{x}_{i1},...,\boldsymbol{x}_{ip})$. If $\boldsymbol{X}$ is a random sample from a multivariate normal distribution with mean vector $\boldsymbol{u}$ and covariance matrix $\boldsymbol{V}$ in $R^p$, a classical way of detecting outliers is to calculate Mahalanobis' distance for each observation using estimates of $\boldsymbol{u}$ and $\boldsymbol{V}$ as follows:

$$D_i=(\boldsymbol{x}_i-\hat{\boldsymbol{u}})^T\hat{\boldsymbol{V}}^{-1}(\boldsymbol{x}_i-\hat{\boldsymbol{u}})$$

Mahalanobis' distance identifies observations that lie far away from the centre of the data cloud, giving less weight to variables with large variances or to groups of highly correlated variables (Joliffe 1986). This distance is often preferred to the Euclidean distance which ignores the covariance structure and treats all variables equally.

A test statistic for $D_i$ can be created as follows

$$\frac{(n-p)n}{(n^2-1)p}D_i$$

which has an approximate $F$ distribution with $p$ and $n$-$p$ degrees of freedom (Afifi and Azen 1972).

Other currently popular multivariate outlier detection methods fall under projection pursuit techniques, originally proposed by Kruskal (1969)."The goal of projection pursuit procedures is to discover structure in a multivariate data set by projecting these data in a lower-dimensional space" (Rousseeuw and Leroy, 1987). Projection pursuit searches for interesting linear projections of multivariate data sets, where a projection is deemed interesting if it minimizes or maximizes a projection index, which is typically a variance.

For example, principal components may be considered to be interesting projections. Huber (1985) cites two main reasons why they are interesting: first, in the case of clustered data, the leading principal axes pick projections with good separations; secondly, the leading principal components collect the systematic structure of the data. Thus, the first principal component reflects the first major linear trend, the second principal component, the second major linear trend, etc. So, if an observation is located far away from any of the major linear trends it might be considered an outlier.

The multivariate outlier detection method presented in this paper uses Mahalanobis' distance to detect outliers and projection pursuit techniques to robustly estimate the covariance and mean matrix.

## 3. ROBUST ESTIMATION

This section discusses various popular methods of robust estimation.

### 3.1. M-Estimators

The purpose of robust estimation is to produce an efficient estimator in the presence of outliers, while minimizing bias. This is done by reducing the influence of the outliers on the estimator. To evaluate robust estimators, the usual properties such as bias and precision are of interest, as well as others that we will briefly define here. The breakdown point is the smallest fraction of contamination that can cause the bias of the estimator to become infinitely large, at which point the estimator is said to break down. Equivariance refers to whether an estimator is affected by location or scale transformations. If an estimator is unaffected by translations it is called translation or location equivariant. Affine equivariance means that a linear transformation of the explanatory variables transforms the estimator accordingly. An estimator that is scale and location equivariant for orthogonal transformations is called orthogonally equivariant. This property is useful when the only transformations performed are orthogonal, as we shall see in section 4.2. Rousseeuw and Leroy (1987), amongst others, provide formal definitions of all these concepts.

Some of the most popular robust estimators are M-estimators, first introduced by Huber (1964), where M stands for maximum likelihood. M-estimators are robustified maximum likelihood estimators that use a weight function to discount, or downweight, extreme values. In the univariate case, a robust M-estimator of location may be created as follows: for the observation, $x_i$, and location estimate, $T$, define the residual, $r_i = x_i - T$. Define a function, $\rho(x,T)=\rho(x_i-T)$ where $\rho(x,T)$ is symmetric (i.e. $\rho(x,T)=\rho(-x,T)$) with a unique minimum at zero. Typically, the role of this function is to decrease the influence of observations with large residuals, often resulting in a Winsorized estimator.

Once this function has been defined, minimization is performed. For example, for some function $\rho(x,T)$, a univariate M-estimate of location, $T$, could be obtained by solving the equation:

$$Minimize_T \sum_{i=1}^{n} \rho(x_i - T)$$

Differentiating this expression with respect to $T$ yields:

$$\sum_{i=1}^{n} \psi(x_i - T) = 0$$

Different $\rho(x,T)$, or its derivative $\psi(x,T)$, yield different M-estimators, including the usual maximum likelihood estimators. Huber (1964) proposed the following function:

$$\psi(r)=\begin{cases} -k, & r < -k \\ r, & -k \le r \le k \\ k, & r > k \end{cases}$$

where $r=x_i\text{-}T$ and $k$ is a constant called the tuning factor. This function determines the extent to which outliers are treated. Huber's M-estimator, above, results in a trimmed mean. In practice, M-estimators are often calculated using reweighted least squares formulas (Holland and Welsch, 1977).

### 3.2. Stahel-Donoho Estimators

M-estimators can be extended to multivariate data, however Huber (1977) has shown them to have a low breakdown point of at most $1/p$, where $p$ is the dimensionality of the data. Stahel (1981) and Donoho (1982) were the first to define a robust multivariate estimator with a high breakdown point of one-half for large data sets, regardless of the dimensions of the data. Their estimator is similar to an M-estimator. It is a weighted least squares estimator but is more robust because of the way in which the weights are calculated.

For each observation, a measure of outlyingness is calculated using projection pursuit methods. If there is any projection for which the observation is an outlier, its weight is reduced in the least squares formula. The method is best described by Maronna and Yohai (1995):" The outlyingness measure is based on the idea that if a point is a multivariate outlier, then there must be some one-dimensional projection of the data for which the point is a (univariate) outlier."

The Stahel-Donoho estimators of the location vector, $\boldsymbol{u}$, and the scale covariance matrix, $\boldsymbol{V}$, given robustness weight, $\delta_i$, can be expressed as follows:

$$\hat{\boldsymbol{u}}=\frac{\sum_i^n \delta_i \boldsymbol{x_i}}{\sum_i^n \delta_i} \qquad (1)$$

$$\hat{\boldsymbol{V}}=\frac{\sum_{i=1}^n \delta_i^2\,(\boldsymbol{x_i}-\hat{\boldsymbol{u}})(\boldsymbol{x_i}-\hat{\boldsymbol{u}})^T}{\sum_{i=1}^n \delta_i^2} \qquad (2)$$

Calculation of the robustness weights is described in the following section.

#### 3.2.1. Stahel-Donoho Robustness Weights

The Stahel-Donoho estimators look for a univariate projection that makes an observation an outlier. This is done by projecting the $p$-dimensional data set onto a randomly generated vector, $\boldsymbol{v}$ in $R^p$. The projected value, $\boldsymbol{v}^T\boldsymbol{x}_i$, of the ith observation, $\boldsymbol{x}_i$, is compared with the median value, *med*, and scaled by the median absolute deviation, *mad*. Thus, for each observation, $\boldsymbol{x}_i$, Stahel-Donoho estimators search for the largest one-dimensional residual, $r_i$, defined by:

$$r_i= \mathit{supremum}_{\boldsymbol{v}}\ \frac{|\boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{x_i}-med(\boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{X})|}{mad(\boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{X})} \qquad (3)$$

where the supremum is over all possible projections $\boldsymbol{v}$ in $R^p$, and $\mathrm{mad}(\boldsymbol{v}^T\boldsymbol{X})=\mathrm{med}|\boldsymbol{v}^T\boldsymbol{X}\text{-}\mathrm{med}(\boldsymbol{v}^T\boldsymbol{X})|$.

Next, a robustness weight function, $\delta(r)$, is defined where $\delta(r)$ is a strictly positive and decreasing function of $r\geq 0$, such that $r\,\delta(r)$ is bounded. These weights are used to calculate the weighted mean vector and covariance matrix in equations (1) and (2). Equation (3) is affine equivariant since $r_i$ does not change when the $\boldsymbol{x}_i$ are linearly transformed. Donoho (1982) showed that the estimate of the mean vector has a breakdown point of $[(n+1)/2\text{-}p]/n$, which tends to 1/2 for large *n*. The same can be shown for the estimate of the covariance matrix.

## 4. MODIFIED STAHEL-DONOHO ESTIMATOR

Patak (1990) proposed a modification of the Stahel-Donoho estimator that does not require looking at all possible projections $\boldsymbol{v}$ in $\mathrm{R}^p$. Instead, it looks at a subset of projections to get an initial covariance matrix. This covariance matrix is then used to perform principal component analysis. The final robustness weights in this modified Stahel-Donoho estimator downweight any observation that is an outlier when projected onto any of the $p$ principal components.

### 4.1. Initial Estimate of the Covariance Matrix

The initial estimate of the covariance matrix uses an iterative procedure. The method randomly generates $p$-dimensional orthogonal subspaces $\boldsymbol{V}$=($\boldsymbol{v}_1$,...,$\boldsymbol{v}_p$) in $\mathrm{R}^p$, where $\|\boldsymbol{v}_j\|$=1. For each subspace, the data are projected onto the $p$ basis vectors. For each observation, this method selects the basis that generates the smallest robustness weights.

The data are first robustly centered using an $\mathrm{L}_1$-estimator. Centering ensures that the final estimate of the covariance matrix is location invariant. The $\mathrm{L}_1$-estimator is a multivariate analogue of the median that is orthogonally equivariant. For a $p$-dimensional data set with observations $\boldsymbol{x}_i$ the $\mathrm{L}_1$ estimate of the location, $\boldsymbol{T}$, is defined as the solution to the minimization problem:

$$\min_T \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{T}\|$$

In the following formulas, $\boldsymbol{x}_i$ and $\boldsymbol{X}$ are centered. For each basis, define the one-dimensional residual, $r_{ij}$, for each observation, $\boldsymbol{x}_i$, as the projection onto the $j$th basis vector, $\boldsymbol{v}_j$, as follows:

$$r_{ij} = \frac{|\boldsymbol{v}_j^T \boldsymbol{x}_i - med(\boldsymbol{v}_j^T \boldsymbol{X})|}{mad(\boldsymbol{v}_j^T \boldsymbol{X})} \qquad (4)$$

and then trim the residuals as follows:

$$\tilde{r}_{ij} = \begin{cases} r_{ij}, & r_{ij} \le 2.5 \\ 2.5, & 2.5 < r_{ij} \le 4 \\ 0, & r_{ij} > 4 \end{cases} \qquad (5)$$

The cut-off points of 2.5 and 4 are based on the analysis of multivariate normal data sets and trim only extreme residuals.

For each basis, an observation's robustness weight, $\delta_i$, ranges from zero to one and combines the $p$ projections of $\boldsymbol{x}_i$ onto each basis vector as follows:

$$\delta_i = \Pi_{j=1}^{p} \frac{\tilde{r}_{ij}}{r_{ij}} \qquad (6)$$

Set the initial value of $\delta_i$ to 1, then for each basis calculate robustness weights, $\delta_i$, and select the smallest value. Apply these weights to formulas (1) and (2) to obtain an initial robust estimate of the covariance matrix.

To summarize, the sequence is as follows:

```
1. Center the data using an $L_1$ estimator of the location vector.
2. Set the initial weights to one: $\delta_i^0=1$, $i=1,...,n$.
3. For $k=1$ to $m$
        a. Randomly generate $\boldsymbol{v_1} \sim N(0,1)$ and normalize
        b. Calculate $\boldsymbol{v_2,...,v_p}$ in $R^p$ such that $\boldsymbol{v_i}^T\boldsymbol{v_j}=1$, $i=j$, $1 \le i < j \le p$.
        c. Calculate $\delta_i$ using formulas (4), (5) and (6).
        d. If $\delta_i^k<\delta_i$ then $\delta_i=\delta_i^k$.
4. Estimate the covariance matrix using $\delta_i$ and formulas (1) and (2).
```

Patak (1990) found that after ten iterations ($m=10$), the weights do not change significantly.

### 4.2. Final Estimate of the Mean and Covariance Matrix

The estimator of the final covariance matrix is very similar to the initial estimator, except that this time the weights are calculated by projecting the data onto the $p$ principal components. The purpose of this final step is to improve upon the initial estimate of the mean and covariance matrix which were generated by only looking at ten projections per dimension.

From the initial estimate of the robust covariance matrix, $p$ principal components are calculated. The principal components are used to calculate new weights using formulas (4), (5) and (6), where $\boldsymbol{v_j}$ now represents the $j$th principal component. These weights generate a new covariance matrix, using formulas (1) and (2), which in turn generates new principal components and new weights. The process is repeated until the weights do not change (in practice, one iteration has been found to be sufficient). Patak shows these modified Stahel-Donoho estimators to be orthogonally equivariant with a breakdown point of $(n/2\text{-}p)/n$. Orthogonal equivariance is sufficient since only orthogonal transformations are performed.

## 5. OUTLIER DETECTION FOR MSM

MSM collects shipments and inventories on Canadian manufacturers. The sample design is a stratified simple random sample of companies, stratified by industry, province and the size of the company. The size of the company is measured in revenue from shipments. MSM has four size strata: a take-all stratum comprised of the largest manufacturers and three take-some strata comprised of increasingly smaller sized manufacturers (i.e. large, medium and small).

Preliminary validity edits are performed during collection, but due to the large sample sizes, thorough micro editing of all questionnaires is not possible. Outliers are detected using Mahalanobis' distance, where the covariance and mean matrix are robustly estimated using the modified Stahel-Donoho estimators described in section 4. Performed directly after collection, outlier detection serves two purposes: to provide analysts with a list of outliers for editing and to prevent outliers from being used by imputation.

For a discussion of the results of outlier detection for AWRTS when sampling weights are not used see Franklin and Brodeur (1998).

### 5.1. Comparing Outlier Detection with and without the Sampling Weights

In production, sampling weights are not used during outlier detection. Instead, outlier detection is performed within domains that are similar to sampling strata. Since Mahalanobis' distance requires symmetrical data, a log transformation is applied to ensure symmetry.

There are two reasons why we are interested in using the sampling weights during outlier detection. The first is that analysts are most interested in influential values. One reason for this is that MSM is a continuous survey and consequently stratification deteriorates over time. Of particular interest are companies that were small at the time of stratification that grow over time. For these companies, the combination of the weight and reported value can be an outlier. So, we would like to perform outlier detection by collapsing across all size strata. The other reason why we want to use the sampling weights is to avoid performing outlier detection by sampling strata, which are often small.

Properly incorporating the complex sample design into the robust estimators of the covariance and mean matrix is complicated and would require burdensome reprogramming of the existing outlier program. We propose here a much simpler method: within each industry and geography domain, multiply each observation by its sampling weight. Apply a transformation, if necessary, to the weighted data in order to get a symmetrical distribution and perform outlier detection across all size strata within each industry and geography domain. We continue to perform outlier detection separately for each combination of industry and geography since these are estimation domains used to publish data.

### 5.1.1 Outlier Detection for MSM without the Sampling Weights (Method 1)

A log transformation is used to make the data symmetrical. The covariance and mean matrix are robustly estimated using the modified Stahel-Donoho estimators described in section 4, where robustness weights, $\delta_i$, are now calculated based on log($\boldsymbol{x}_i$) and where $\boldsymbol{x}_i$ in formulas (1) and (2) is now log($\boldsymbol{x}_i$). Mahalanobis' distance, $D_i$, is calculated for each domain defined by the observation's industry, geography and size:

$$D_i=(\log(\boldsymbol{x}_i)-\hat{\boldsymbol{u}})^T\hat{\boldsymbol{V}}^{-1}(\log(\boldsymbol{x}_i)-\hat{\boldsymbol{u}})$$

### 5.1.2. Outlier Detection for MSM with the Sampling Weights (Method 2)

For the weighted data, we are interested in the value $w_i\boldsymbol{x}_i$, where $w_i$ is an observation's sampling weight. A log transformation is applied to make the data symmetrical. The covariance and mean matrix are robustly estimated using the modified Stahel-Donoho estimators described in section 4, where robustness weights, $\delta_i$, are now calculated based on log($w_i\boldsymbol{x}_i$) and where $\boldsymbol{x}_i$ in formulas (1) and (2) is now log($w_i\boldsymbol{x}_i$). Mahalanobis' distance, $D_i$, is calculated for each domain defined by the observation's industry and geography:

$$D_i=(\log(w_i\boldsymbol{x}_i)-\hat{\boldsymbol{u}})^T\hat{\boldsymbol{V}}^{-1}(\log(w_i\boldsymbol{x}_i)-\hat{\boldsymbol{u}})$$

### 5.2. Creating a Contaminated Population

A simulation is used to compare the results of outlier detection for methods 1 and 2. To obtain the population, the MSM sample for shipments and raw materials for March 2000 for the province of Ontario and the industry NAICS = 332 (Fabricated Metal Products) was used to generate an uncontaminated bivariate lognormal population as follows:

1. A logarithmic transformation is applied to the two variables Goods of Own Manufacture, GOM (a shipments in revenue variable) and Raw Materials, RM (an inventory variable) in order to obtain a symmetrical distribution in each of the four size strata.
2. The population mean vector and covariance matrix are estimated for each stratum.
3. Lognormal bivariate populations for each size stratum are simulated using these estimates and the formulas to generate a multivariate normal distribution proposed by Rubinstein (1981).

Outliers are added to each size stratum by generating data from a bivariate lognormal distribution with the same mean vector but with nine times the variance. The idea for this contamination comes from Hulliger (1995). In each stratum, 5% of the population comes from this distribution. Thus, the population in stratum $h$ is log($\boldsymbol{x}_{i\,(h)}$)$\sim$0,95 N($\boldsymbol{\mu}_h$,$\boldsymbol{V}_h$)+0,05 N($\boldsymbol{\mu}_h$, $9\ \boldsymbol{V}_h$) where $\boldsymbol{x}_{i\,(h)}$ , $\boldsymbol{\mu}_h$ ,$\boldsymbol{V}_h$ are in R$^2$.

The simulated population has 1,240 units with 120 in the take-all stratum and 220, 360 and 540 units, respectively, in the large, medium and small take-some strata. From each size stratum, 1,000 simple random samples are drawn. The sampling fractions are approximately those of MSM: 100% for the take-all stratum, 20%, 8.3% and 5% for the large, medium and small take-some strata, respectively.

For each sample drawn from the simulated population, outlier detection with and without the sampling weights are compared as follows: without the sampling weights, outlier detection on the log($\boldsymbol{x}_{i\,(h)}$) is performed separately within each size stratum. With the sampling weights, we first apply an exponential transformation within each size stratum to obtain $\boldsymbol{x}_{i\,(h)}$, them multiply each observation by its sampling weight and combine all observations across size strata. A log transformation is applied to obtain a symmetrical distribution, then outlier detection is performed across all size strata.

### 5.3. Results

We compared the results of outlier detection with and without the sampling weights by looking at the following results: the number of outliers that are identified by both methods, the number of outliers identified by each method that come from the contaminated population and an explanation of those outliers that are unique to one method or the other. Observations identified as outliers are those for whom Mahalanobis' distance is greater than the 95th percentile of the appropriate F-distribution. Table 1 presents the results when we compared the two methods using real MSM data. Table 2 presents the results for the simulated MSM data.

**Table 1: Number of Outliers Detected by Methods 1 and 2 for Real MSM Data**

| Size Stratum | Sample Size, *n* | Sample Weight, $w_i$ | Outliers Detected by Method 1, without sampling weights | Outliers Detected by Method 2, with sampling weights | Outliers Common to Both Methods |
|---|---|---|---|---|---|
| Take-all | 108 | 1 | 10 | 10 | 9 |
| Large take-some | 37 | 5.85 | 3 | 4 | 3 |
| Medium take-some | 31 | 11.67 | 3 | 2 | 2 |
| Small take-some | 26 | 21.12 | 4 | 4 | 4 |

We can see that both methods detect the same number of outliers for observations belonging to the take-all stratum. Of the ten observations identified by both as outliers, nine are the same. The one that was identified by method 1 but not by method 2 is a small take-all (small RM value), which when compared with other weighted data, is no longer considered to be an outlier. The one that was identified by method 2 but not by method 1 is one which, when compared with the other weighted data, falls in the tail of the distribution and is considered to be an outlier.

For the large take-some stratum, Method 2 identifies one additional outlier. This is an observation whose combination of weight and reported value put it in the tail of the distribution. For the medium take-some, Method 1 identifies one additional outlier. This is a record with one of the smallest RM values which when multiplied by its weight no longer falls in the tail of the distribution. For the small take-some stratum, both methods identify the same four observations as outliers.

**Table 2: Average Number of Outliers Detected by Methods 1 and 2 from 1,000 Simulations**

| Size Stratum | Sample Size, *n* | Population Size, *N* | Number of Contaminants in the Population | Sample Weight, $w_i$ | Outliers Detected by Method 1, without sampling weights (Average) | Outliers Detected by Method 2, with sampling weights (Average) | Outliers Common to Both Methods (Average) |
|---|---|---|---|---|---|---|---|
| Take-all | 120 | 120 | 6 | 1 | 12 | 7.66 | 7.13 |
| Large take-some | 44 | 220 | 11 | 5 | 3.95 | 3.32 | 2.95 |
| Medium take-some | 30 | 360 | 18 | 12 | 2.12 | 3.09 | 1.79 |
| Small take-some | 27 | 540 | 27 | 20 | 2.02 | 4.16 | 1.69 |

For the simulated data, we find that Method 2 identifies fewer observations as outliers in the take-all stratum and more outliers in the take-some strata. Fewer are identified in the take-all stratum because the weighted method no longer identifies small take-alls as outliers, on the other hand it does identify as outliers those take-some companies for whom the combination of their GOM or RM and their weight makes them an outlier. From an analysts viewpoint, method 2 might be preferable, since small take-alls which are not influential when compared with the rest of the data are usually not of concern, but influential take-somes are of interest.

One thing we notice from these results is that we detect more outliers than expected: the simluated data were created so that 5% of the population was contaminated, but approximately 8% of the sample was identified as outliers. This is because since we are dealing with a contaminated distribution and robust estimators, we no longer know exactly the distribution of Mahalanobis' distance.

## 6. DISCUSSION

The majority of outliers detected with or without the sampling weights are similar, however some differences exist. The decision as to whether outlier detection should be performed with or without the sampling weights depends on the purpose of outlier detection. For example, if the analyst is interested in detecting small take-alls, then the sampling weights should not be used. However, if influential observations are of interest, then including the sampling weights, as illustrated is probably preferable.

In researching this outlier detection method, two main issues were identified: the treatment of missing values and the distribution of Mahalanobis' distance when using contaminated distributions and robust estimators.

In general, the MSM has excellent response rates for the main variables. However, the response rate for some inventory variables is low. We examined different options to deal with this problem. One is to impute the missing values before outlier detection is performed. The imputation method should be multivariate, since a univariate method can disrupt the multivariate structure of the data. Often, for MSM inventory data, the missing value is naturally zero. If we impute these observations with a zero, then we will create a spike in the distribution of the data. Also, these zero observations are different from units that reported for these characteristics, therefore outlier detection should probably be treated for them separately.

In order to avoid treating missing values, another option is to run the outlier routine several times. For instance, for AWRTS, the routine is first run on the observations that have responses for the three variables with the highest response rates. Then, of the three variables, we drop the one with the lowest response rate and run the routine with the observations having reported values for the remaining two variables. This way, the majority of observations are tested.

It is known that Mahalanobis' distance is F-distributed when the data are normally distributed. However, we are dealing with contaminated normal distributions, using robust estimators of the mean and covariance. Therefore, we are not certain what is the true distribution of Mahalanobis' distance. In the outlier detection routine, we continue to assume an F-distribution. Is this still valid?

## 7. ACKNOWLEDGEMENTS

The authors would like to thank Jean-François Beaumont and Charlie Arcaro for their very useful comments and insight.

## 8. REFERENCES

Afifi, A.A., and Azen, S.P. (1972), *Statistical analysis: a computer oriented approach*, New York: Academic Press, pp. 282.

Barnett, V and Lewis, T. (1995), *Outliers in Statistical Data*, Chichester: John Wiley and Sons. pp.7, 269.

Donoho, D. L. (1982), "Breakdown Properties of Multivariate Location Estimators," Ph.D. Qualifying Paper, Department of Statistics, Harvard University.

Hidiroglou, M. A., and Berthelot, J.-M. (1986), “Statistical Edit and Imputation for Periodic Surveys,” *Survey Methodology*, **12**, pp. 73-83.

Franklin, S. H., and Brodeur, M (1997), “A Practical Application of a Robust Multivariate Outlier Detection Method,”American Statistical Association, 1997 Proceedings of the Section on Survey Research Methods, pp. 186.

Holland, P.W., and Welsch, R.E. (1977), “Robust Regression Using Iteratively Reweighted Least Squares,” *Commun. Stat (Theory and Methods)*, **6**, pp.813-828.

Huber, P.J. (1964), “Robust Estimation of a Location Parameter,” *Annals of Mathematical Statistics*, **35**, pp. 73-101.

Huber, P.J. (1977), “Robust Covariances,” in Gupta, S.S. and Moore, D.S. (eds) *Statistical Decision Theory and Related Topics II*, New York: Academic Press, pp. 165-191.

Huber, P.J. (1985), “Projection Pursuit,” *The Annals of Statistics*, **13**(2), pp. 435-475.

Hulliger, Beat (1995), “Outlier Robust Horvitz-Thompson Estimators,” *Survey Methodology*, **21**(1), pp. 79-87.

Joliffe, I.T. (1986), *Principal Component Analysis*, New York: Springer-Verlag.

Kruskal, J.B. (1969), “Toward a Practical Method which Helps Uncover the Structure of a Set of Multivariate Observations by Finding the Linear Transformation which Optimizes a New ‘Index of Condensation’,” in R.C. Milton and J.A. Nelder (eds) *Statistical Computation*, New York: Academic Press.

Lee, H., Ghangurde, P.D., Mach, L. and Yung, W. (1992), “Outliers in Sample Surveys,” Working Paper No. BSMD-92-008E, Methodology Branch, Business Survey Methods Division, Ottawa, Canada: Statistics Canada.

Maronna, R.A., and Yohai, V.J. (1995), “The Behaviour of the Stahel-Donoho Robust Multivariate Estimator,” *Journal of the American Statistical Association*, **90** (429), pp. 330-341.

Patak, Z. (1990), Robust principal component analysis via projection pursuit, Master’s thesis, University of British Columbia, Canada.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, U.S.A: John Wiley and Sons.

Rubinstein, R. Y. (1981), *Simulation and the Monte Carlo Method*, U.S.A: John Wiley and Sons. pp. 65-68.

Stahel, W. A. (1981). Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen. Ph.D. Thesis, Swiss Federal Institute of Technology, Zurich, Switzerland.

# AN EVALUATION OF OUTLIER-RESISTANT PROCEDURES IN ESTABLISHMENT SURVEYS

Jean-Phillippe Gwet and Hyunshik Lee, Westat, Inc.
Jean-Phillippe Gwet, Westat, 1650 Reserch Blvd., Rockville, MD 20850
GwetJ1@westat.com

## ABSTRACT

Many outlier-resistant procedures have been proposed to handle the problem of outliers in establishment surveys. Only limited information is available on how they are compared in a wide range of situations that are expected to occur in establishment surveys. This paper reports some study results on comparative performance of various outlier procedures in terms of MSE-efficiency, bias, variance estimation, and coverage property of interval estimation. For interval estimation, not only the traditional method based on the normal theory but also methods based on the bootstrap technique was studied. It appears that these bootstrap methods can improve the coverage rate substantially. The features and limitations of the bootstrap approach will also be discussed.

**Key Words: Contaminated models, Skewed long tailed distributions, Winsorization, M-estimation, Adaptive methods.**

## 1. INTRODUCTION

The problem of outliers in establishment surveys has been given considerable attention lately. There have been some new developments since the International Conference on Establishment Surveys held in 1992 (see Lee, 1995 for a survey up to 1992). The problem has been around in sample surveys for a long time but systematic treatments of the problem started emerging in 1980's. Up until now, many outlier-resistant procedures have been proposed in the literature. However, a comprehensive study of various outlier-resistant alternatives has been lacking. It is our goal to fill some of this gap.

The outlier problem is more common in establishment surveys due to skewness of the target populations in establishment surveys. Therefore, ordinary statistics based on the least squares principle are not adequate. The main tool used in survey sampling in this situation is size-stratification or probability proportional to size sampling, which is fairly effective as long as the size measure is a good predictor of the actual survey values. However, the size measure becomes obsolete quickly over time and surprise values (i.e., outliers) occur almost always. Moreover, in a multi-purpose survey, the size measure good for some variables may not be good for others. When they occur, an ad-hoc treatment is often employed to treat the outliers. The nature of surprise makes survey practitioners somewhat reluctant to use a systematic treatment. Another obstacle of a routine use of outlier-resistant techniques is their complexities, lack of proper variance estimators, and their unknown inferential properties.

There have been some efforts made to address this kind of obstacles (Hulliger, 1999). Carrying the torch further, we hope in this paper to achieve the goal of providing practical suggestions and recommendations on how to implement currently available procedures rather than proposing new methods.

In Section 2, we discuss the outlier-resistant estimators to be studied in this paper. Section 3 provides the results of the study and in Section 4, some concluding remarks are presented.

## 2. OUTLIER-RESISTANT ESTIMATORS

We start with a simple situation, where the estimator is the sample mean and the sample design is simple random sampling, and discuss outlier-resistant alternatives applicable to this situation.

### 2.1 Estimation of the Population Mean

The familiar form of the sample mean for the estimation of the population mean ( $\overline{Y}$ ) is given by

$$\bar{y} = \frac{1}{n}\sum\nolimits_s y_i \tag{2.1}$$

where $n$ is the sample size, $y_i$ is $i$ -th observed sample value, $s$ is a simple random sample of size $n$ selected from the universe $U$ of size $N$. $\sum_s$ is shorthand for $\sum_{i\in s}$ or $\sum_{i=1}^{n}$ and this convention will be used throughout the paper unless specified otherwise. This estimator is known to be very vulnerable in the presence of outliers. Many alternatives have been proposed to cope with the problem especially under right skewed populations, which are often encountered in business surveys. A technique, simple, popular and well studied, is Winsorization. The simplest form is given as

$$\hat{\bar{Y}}_{\mathrm{W1}} = \frac{1}{n}\left(\sum_{i=1}^{n-k} y_{(i)} + k y_{(n-k)}\right), \tag{2.2}$$

where $y_{(i)}$ is the $i$-th order statistic and $1 \le k < n$. The *k*-times Winsorized mean is obtained by replacing the largest *k* values in the sample by the $(n-k)$-th order statistic, $y_{(n-k)}$. We used $k = 1$ in our simulation.

There are several variants of (2.2), one of which was studied by Rivest (1994) and is obtained by replacing the *k* largest observations in (2.2) by a linear combination of $y_{(n-k)}$ and $y_{(n-k+1)}$. The estimator is given by

$$\hat{\bar{Y}}_{\mathrm{W2}} = \frac{1}{n}\left(\sum_{i=1}^{n-k} y_{(i)} + kt\right), \quad t = \tau y_{(n-k)} + (1-\tau) y_{(n-k+1)}, \tag{2.2a}$$

where $0 \le \tau \le 1$. If $\tau = 1$, the estimator reduces to (2.2). He recommended using $\tau = 0.75$ with *k* = 1 for moderate skewness and this and $\tau = 1$ were used in our simulation.

As another variant, we also studied the following estimator:

$$\hat{\bar{Y}}_{\mathrm{W3}} = \frac{1}{n}\left(\sum_{i=1}^{n-k} y_{(i)} + \sum_{i=n-k+1}^{n} y^{*}_{(i)}\right), \tag{2.3}$$

where $y^{*}_{(i)} = f y_i + (1-f)t$ with the sampling fraction $f = n/N$ and *t* given in (2.2a). One nice feature of this estimator is that outlier treatment gets diminished as $f \to 1$.

Fuller (1991) studied the Winsorized mean under a super-population with a right long tailed distribution, where a finite population is treated as a random sample from the super-population. He found that the sample mean in (2.1) is more efficient in terms of the mean square error (MSE) when the right tail distribution is shorter than the exponential distribution but the Winsorized mean is more MSE-efficient if the right tail distribution is longer than the exponential. If the long tailedness is known a priori, the more efficient estimator can be chosen accordingly. For the case where the tail distribution is not known, Fuller proposed a strategy in which a test is first performed and then an estimator is selected according to the test result. The test statistic is given by

$$F_{Tj} = \frac{j^{-1}\sum_{i=n-j+1}^{n} Z_{yi}}{(T-j)^{-1}\sum_{i=n-T+1}^{n-j} Z_{yi}}, \tag{2.4}$$

where $Z_{yi} = (n-i+1)(y_{(i)} - y_{(i-1)})$, *j* is a number between 1 and *n* but it is usually 1 or 2, and *T,* the number of large observations used to construct the test, is chosen so that $j + 1 \le T \le n$. A class of estimators called the test-and-estimate procedure is then defined as

$$\hat{\mu}_{Tj} = \begin{cases} \bar{y} & \text{if } F_{Tj} < K_j \\ n^{-1}\left\{\sum_{i=1}^{n-j} y_{(i)} + j(y_{(n-j)} + K_j \bar{d}_{Tj})\right\}, & \text{otherwise} \end{cases} \tag{2.5}$$

where $K_j$ is the cut-off value that determines whether the sample mean or the other estimator is used and $d_{Tj}$ is a factor that ensures continuity. The procedure is a function of three numbers, *T*, *j*, and $K_j$, which have to be determined for each application. In our simulation, we used *j* = 1 and 2, and *T* = 5, 7, and 10 depending on the sample size, and $K_j$ = 5.8. For a finite population, Fuller proposed to estimate the population mean by

$$Y_F = f\bar{y} + (1-f)\hat{\mu}_{Tj}.$$

Huber (1964) proposed the M-estimator for long tailed distributions, which has been adopted in survey sampling. Huber's M-estimator is implicitly defined by the following estimating equation:

$$\sum_s \psi(y_i - \tilde{\mu}/\tilde{\sigma}) = 0, \tag{2.6}$$

where $\psi$ is the Huber's function, $\tilde{\mu}$ is the resulting M-estimator, and $\tilde{\sigma}$ is a resistant estimate of scale such as median absolute deviation (MAD). Huber's $\psi$-function is defined as $\psi(t) = \max(-c, \min(c,t))$, where $c$ is a (tuning) constant (we used $c = 3$ for our simulation). A closed form solution to equatioon (2.6) is not available and thus, an iterative procedure is used. An often used procedure is called the iterative reweighted least square (IRLS) procedure. If $r_i^{(k)} = (y_i - \tilde{\mu}^{(k)}) / \tilde{\sigma}^{(k)}$, $z_i^{(k)} = \psi(r_i^{(k)}) / r_i^{(k)}$, and $\bar{z}^{(k)}$ is the mean of the $z_i^{(k)}$, then the $k$-th IRLS estimator is given by $\tilde{\mu}^{(k)} = \sum_s \left( z_i^{(k-1)} \big/ (n\bar{z}^{(k-1)}) \right) y_i$. The first iteration requires initial estimates for $\tilde{\mu}$ and $\tilde{\sigma}$, which should be robust. We used the median and MAD in our simulation.

For symmetric long-tailed distributions, a resistant M-estimator can be very efficient without having a bias. However, for skewed distributions, the bias is often too large to accept and a bias correction is needed even though the estimator could still be more MSE-efficient due to a very small variance. Using the model-based approach, Chambers (1986) proposed such an outlier resistant estimator originally for the case where an auxiliary variable is available for estimation of the population mean. Without auxiliary variable, Chambers' estimator becomes

$$\hat{\bar{Y}}_{\mathrm{C}} = f\bar{y} + (1-f)(\tilde{u} + \tilde{r})$$

where $\tilde{r} = \tilde{\sigma}_r \sum_s \psi(r_i / \tilde{\sigma}_r) / n$, $r_i = y_i - \tilde{\mu}$ and $\tilde{\sigma}_r$ is a resistant scale estimate for $r_i$'s. The $\psi$-function is that of Huber with a large tuning factor (say, $c = 10$). Lee (1995) proposed a partial correction of the bias and it is given by

$$\hat{\bar{Y}}_{\mathrm{L}} = f\bar{y} + (1-f)(\tilde{\mu} + \theta\bar{r})$$

where $\bar{r} = \sum_s r_i / n$ and $0 \le \theta \le 1$. Note that if $\theta = 1$, then the estimator is fully bias-corrected but reduces to the nonresistant sample mean. The $\theta$ can be determined based on prior information or past data for a similar or the same population. With the consideration of consistency, $\theta$ is set to be $\theta = B^2 / (B^2 + V)$ where $B = -E(\bar{r})$ is the bias of $\tilde{\mu}$ and $V$ is the variance of $\bar{y}$. Usually, $\theta \to 1$ as $n \to \infty$ since $V \to 0$ so that the estimator is consistent. Of course, $B$ and $V$ are unknown in reality. However, if we use consistent estimates for $B$ and $V$, then the resulting Lee estimator is also consistent. Lee and Patak (1998) studied the use of an estimated $\theta$ to robustify the generalized regression estimator. In our simulation, we estimated the $B^2$ by $\max\left(0, \hat{B}^2 - v(\hat{B})\right)$ where $\hat{B} = \tilde{\mu} - \bar{y}$.

The tuning factor $c$ of the M-estimator defined in (2.6) determines the amount of outliers to be treated. It is difficult to determine the best $c$ that minimizes the MSE of the M-estimator without knowing the population distribution. Hulliger (1993) proposed a procedure in which one tries different $c$ values and pick a value among those tried that gives the minimum estimated MSE. This estimator is called the minimum estimated risk (MER) estimator. In the simulation, we used $c$ = 1, 2, 3, 4, 8, 12, 16, 20, and $\infty$ for the MER estimator ($c = \infty$ corresponds to the sample mean). To estimate the MSE for each $c$, we used the closed variance formula for the M-estimator studied by Gwet (1997) and the estimate of $B^2$ discussed above.

The iterative procedure to solve for the M-estimator given by (2.6) converges quickly with robust initial estimates for $\tilde{\mu}$ and $\tilde{\sigma}$ and the first iteration estimator is often sufficient. Moreover, the small difference between the fully-iterated and the first iteration does not really matter in sample surveys, anyway, as noticed by Lee (1991). Hulliger (1999) proposed to use the one-step or one-iteration M-estimator instead of fully-iterated M-estimator. We also studied the one-step M-estimator.

So far, we have considered estimators based on the value modification strategy (e.g., Winsorization), where the values of outliers are modified, and estimators based on the M-estimation technique. Another strategy frequently used and having an appealing feature from the design-based perspective is the weight modification strategy. The estimator has the following form:

$$\hat{\bar{Y}}_{\mathrm{HS}} = \frac{1}{N} \left\{ \sum_{i=1}^{n-j} \frac{N - \lambda n_2}{n_1} y_{(i)} + \sum_{i=n-j+1}^{n} \lambda y_{(i)} \right\}$$

Hidiroglou and Srinath (1981) provided the $\lambda$ that minimizes the MSE based on some population quantities, which are normally unknown. We used a weight reduction factor of 0.5 (i.e., $\lambda = 0.5\, N/n$ ) in our simulation. We detected outliers using the quartile method (see Lee, 1995). We tried in vain to use the optimal $\lambda$-formula with estimated parameters but often ran into a problem of having one outlier to estimate the variance of the outlier stratum.

### 2.2 Variance Estimation and Construction of Confidence Interval

Estimation of the variance of a survey estimator is a very important part of survey research. However, outlier resistant alternatives often do not have an appropriate variance estimator due to the complexity of the estimation procedure. Due to this complexity, the replication/resampling methods are easier to use and thus, we focus on those, particularly, the jackknife and the bootstrap. The BRR is awkward to apply for simple random samples and thus we did not include it in our study.

The jackknife variance estimator of $\hat{\theta}$ is given by:

$$v_{\mathrm{JK}}\left(\hat{\theta}\right) = (1-f)\frac{n-1}{n}\sum_{j=1}^{n}\left(\hat{\theta}^{(j)} - \hat{\theta}^{(\bullet)}\right)^2,$$

where, $\hat{\theta}^{(j)}$ is computed using the *j*-th jackknife replicate with *j*-th unit deleted and $\hat{\theta}^{(\bullet)}$ is the average of the $\hat{\theta}^{(j)}$'s. Gwet (1997) extensively studied the jackknife variance estimator of the M-estimator and he proved under some regularity conditions that this variance estimator is consistent for estimating the asymptotic variance of the M-estimator.

Using the jackknife variance estimate, one can construct $100(1-2\alpha)\%$ confidence interval with the standard normal reference points by $\left[\hat{\theta} + z^{(\alpha)}\sqrt{v_{\mathrm{JK}}(\hat{\theta})}, \hat{\theta} + z^{(1-\alpha)}\sqrt{v_{\mathrm{JK}}(\hat{\theta})}\right]$, where $z^{(\alpha)}$ and $z^{(1-\alpha)}$ are $100\alpha$ and $100(1-\alpha)$ percentile point of the standard normal distribution. Note that $-z^{(\alpha)} = z^{(1-\alpha)}$ and thus the interval is symmetric about the point estimator. When the distribution of the estimator is skewed as are the distributions of the estimators we are studying, the interval's rate of coverage of the true population parameter is far short of the nominal value especially for small or medium sample sizes. The bootstrap technique has an advantage for this situation. Moreover, the technique can better handle an estimator whose functional form is non-smooth as are some of the estimators discussed here. For an iid situation, a bootstrap sample of the same size as the original sample is selected with replacement. The same estimator ($\hat{\theta}$) is applied to the bootstrap sample to calculate the bootstrap estimate. Repeating this procedure a number of times (say, *L* times), we obtain *L* bootstrap estimates from which the Monte Carlo approximation of the bootstrap variance estimate can be computed as follows:

$$\frac{1}{L}\sum_{j=1}^{L}\left(\hat{\theta}_j^* - \bar{\hat{\theta}}^*\right)^2,$$

where $\hat{\theta}_j^*$ is the *j*-th bootstrap sample estimate and $\bar{\hat{\theta}}^*$ is the average of the *L* bootstrap estimates.

From the *L* bootstrap estimates, one also can estimate the distribution function of the estimator $\hat{\theta}$ and then construct a confidence interval from the estimated distribution function denoted by $\hat{G}$. Let the $100\alpha$ percentile point and the $100(1-\alpha)$ percentile point be denoted by $G^{-1}(\alpha)$ and $G^{-1}(1-\alpha)$. Then the interval $[G^{-1}(\alpha), G^{-1}(1-\alpha)]$ can be used as the $100(1-2\alpha)$ percent confidence interval for $\theta$ (Efron 1981, 1982). This interval is usually asymmetric and has a better coverage property than the symmetric interval. This method is called the percentile method. Chambers and Dorfman (1994) studied the percentile method and for the ratio and robust ratio estimators within the model-based framework. They concluded that the bootstrap percentile method does not improve the coverage.

However, when the skewness is severe, the percentile method needs a bias correction (Efron 1981, 1982). This Bias-Corrected (BC) interval can further be improved by using some "acceleration constant" *a*, where the resulting interval is called the BC$_a$ interval (Efron, 1987). The interval is given by

$$[\hat{G}^{-1}\{\Phi(z_\alpha)\}, \hat{G}^{-1}\{\Phi(z_{1-\alpha})\}],$$

where $z_\alpha = z_0 + \left(z_0 + z^{(\alpha)}\right)/\left\{1 - a\left(z_0 + z^{(\alpha)}\right)\right\}$ with $z_0 = \Phi^{-1}\left(\hat{G}(\hat{\theta})\right)$ for the standard normal cumulative distribution function $\Phi$, and $z_{1-\alpha}$ is similarly defined. If $a = 0$, then the interval reduces to the BC interval and if $z_0$ is also set to be zero, then the interval becomes the percentile method. To use the BC$_a$ method, we need the acceleration factor *a*, which can be computed using the nonparametric formula given in Efron (1988).

The bootstrap technique has been adapted for sample surveys by several authors (McCarthy and Snowden, 1985; Rao and Wu, 1988; Sitter, 1992a, 1992b). We implemented the modification proposed by McCarthy and Snowden (1985), which is called the with-replacement bootstrap (BWR). In this method each bootstrap sample is selected as a simple random sample with replacement of size *m* instead of *n* where $m=(n-1)/(1-f)$. If *m* is not an integer, the bootstrap sampling procedure is randomized so that the expected bootstrap sample size is *m*. The bootstrap samples are used to compute the bootstrap variance estimate and the three bootstrap confidence intervals (percentile, BC, and BC$_a$).

## 3. SIMULATION STUDY

The main objective of this simulation study is to compare the performance of the resistant estimators discussed earlier. To this end, we created eight artificial finite populations from four parametric distributions as described below. The four parametric distributions used to generate the study populations are the mixed normal, Pareto, lognormal, and Weibull distributions. Table 1 shows parameters of these distributions used to generate the study populations and their characteristics.

**Table 1. Parametric Distributions and Their Parameter Values Used to Generate the Study Populations and Characteristics of the Study Populations**

| Parametric Distribution | Distribution Function | Population Label | Parameter Values Used | Population Characteristics | | |
|---|---|---|---|---|---|---|
| | | | | Mean | Skewness | Kurtosis |
| Mixed Normal[1] | $\delta N(\mu_1,\sigma_1^2)+$<br>$(1-\delta)N(\mu_2,\sigma_2^2)$ | MIXN1 | $\mu_1=\mu_2=5$,<br>$\sigma_1=1,\ \sigma_2=3,$ | 4.98 | -0.47 | 0.69 |
| | | MIXN2 | $\mu_1=5$, $\mu_2=10$,<br>$\sigma_1=1,\ \sigma_2=3,$ | 5.24 | 1.87 | 6.65 |
| Log Normal | $F(x)=\Phi(\log x/v)$ | LogN1 | $v=1.27$ | 2.23 | 3.66 | 16.37 |
| | | LogN2 | $v=1.68$ | 2.82 | 4.20 | 20.98 |
| Pareto | $F(x)=1-(1+x)^{-\gamma}$,<br>$x>0$ | Pareto1 | $\gamma=2.67$ | 1.54 | 4.71 | 31.33 |
| | | Pareto2 | $\gamma=2.13$ | 1.78 | 5.76 | 44.93 |
| Weibull | $F(x)=1-\exp(-x^{1/\alpha})$,<br>$x>0$ | Weibull1 | $\alpha=1.84$ | 1.55 | 4.36 | 25.42 |
| | | Weibull2 | $\alpha=2.87$ | 4.24 | 7.34 | 65.59 |

[1] Note: $N(\mu_i,\sigma_i^2)$ is a normal distribution with mean $\mu_i$ and variance $\sigma_i^2$, $i=1,2$, $\delta$ is a random variable that takes a value of 0 or 1 with $P(\delta=1)=0.95$.

From each of even numbered study populations, we generated 2,000 samples with each of three sample sizes (10, 20, or 30), and evaluated each point estimator at that sample. We used the Monte Carlo approximations for the true statistics. This set of simulation is mainly used to see the effect of sample size on the performance of the estimators in terms of relative bias and relative efficiency with respect to the sample mean. Using populations MIXN1, LogN1, Pareto1 and Weibull1, we evaluated the estimators in terms of the relative bias and relative efficiency as well as variance estimation and interval estimation. Since this requires a lot more computing resources, we used only one sample size of 20. We computed all interval estimates discussed in Section 2 with $100(1-2\alpha)=95\%$ but we do not present the result for the BC interval to save the space. In the same vein, we also present only results for selective estimators, where only one Winsorized, one Fuller, and one M-estimator among their kind are kept in the presentation.

The performance of outlier-resistant estimators is compared with that of the sample mean. The relative efficiency measure of a given point estimator is defined as the ratio of the Monte Carlo variance of the sample mean to the Monte Carlo mean square error of the alternative estimator. Each point estimator is also evaluated with respect to its bias. To evaluate the variance estimators, we calculated the relative bias of a variance estimator as the relative difference of the Monte Carlo expectation of the variance estimator with respect to the Monte Carlo variance of the point estimator.

We studied closed variance formulae for the sample mean, the two M-estimators, and Chambers estimator. They worked reasonably well except the one for the one-step M-estimator. Since they are available only for these estimators (and more work is needed), we do not discuss their results here.

We calculated Fuller estimator with $j = 1$ and 2 but the results were virtually the same and thus, we present the results for $j = 1$ only.

When the confidence interval did not cover the true population mean, we checked whether the population mean fell on the right or left side of the interval. The percentage of times it fell on the left (right) side is called the left (right) side noncoverage rate. Since the normal theory confidence interval is constructed symmetrically, the left and right noncoverage rates are useful to check the asymmetry of the distribution of the point estimator in question. If the distribution is symmetric, the left and right noncoverage rates should be close each other.

The simulation results are presented in Tables 2 and 3. In the tables, the Winsorized mean given in (2.2a) with $\tau = 0.75$ and 1 are coded as Win0.75-T1 and Win1.00-T1, respectively. The Winsorized mean given in (2.3) with the same $\tau$ -values are coded similarly but suffixed by T2 as Win0.75-T2 and Win1.00-T2. The bootstrap variance estimator was not studied for the Hulliger estimator because of too excessive computing time it requires.

### 3.1 Performance (Relative Bias and Efficiency) of the Estimators

*Winsorized Estimators*: They perform generally well across the populations and different sample sizes. They work particularly well under the very skewed populations (Pareto and Weibull). Win0.75-T1 and Win0.75-T2 seems slightly better in general than the other Winsorized means. Since the Winsorized estimators are consistent, their relative bias decreases as $n$ increases. The simplicity of the estimators is another advantage.

*Fuller Estimator*: The test and estimate feature of this estimator makes it to be least biased among the outlier-resistant estimators but its MSE efficiency is mediocre in general except under the Weibull populations, where it performs quite well. It seems that the procedure is not sensitive to detect mild and moderate skewness and thus it tends to pick the sample mean more often than it should to have a good MSE-efficiency.

*M-estimators*: The fully-iterated M- and one-step M-estimators perform well under the mixed normal populations and when the sample size is small. However, when sample size is large and/or the distribution is very skewed, it is heavily biased and its MSE-efficiency can be much worse than the sample mean even though its variance is much smaller. This clearly demonstrates the necessity of bias correction under the heavily skewed populations. The two M-estimators are very similar in every aspect but the fully iterated one is slightly more efficient almost always.

*Chambers Estimator*: It performs generally well particularly under the lognormal populations. It controls the bias well for the sample sizes we studied. However, the relative bias stays at the same level even though the sample size increases and its inconsistency is its most serious disadvantage. It is possible to improve its MSE-efficiency by using different tuning factors for bias correction depending on the shape of the underlying distribution. However, we intended to do a blind-fold experiment assuming that nothing is known about the population.

*Lee estimator*: It performs remarkably well overall. It often has the best MSE-efficiency. Its relative bias is always decreases as the sample size increases demonstrating its consistency. Its adaptive nature is also appealing in application.

*Hulliger estimator*: Like the Fuller estimator, this is also less biased than other resistant estimators. Its performance is somewhat better than the Fuller but still mediocre compared to other resistant estimators. The computational burden to find the minimum estimated MSE is also its disadvantage.

*Hidiroglou-Srinath estimator*: The estimator performs quite well overall. The choice of $\lambda$ we used seems quite reasonable for those populations we studied. A serious drawback of the estimator is its bias, which does not decrease as the sample size increases.

### 3.2 Variance Estimation and Interval Estimation

The jackknife variance estimator works reasonably well overall. However, it tends to be more biased for the M-estimators, which is somewhat surprising since it works quite well for other more complicated estimators. The normal theory confidence interval based on the jackknife variance estimator has a good coverage only when the population is symmetric and slightly long-tailed. When the population is heavily skewed and long tailed, the coverage rate is usually far short of the nominal value due to skewness of the distribution of the point estimators we studied. In these cases, the right side noncoverage is large and the left side noncoverage is almost zero.

The bootstrap variance estimator works well except for the Winsorized estimator and the Lee estimator for which the bootstrap behaves somewhat erratically. We do not have a satisfactory explanation for this behavior since we expected that it should work similarly to or slightly better than the jackknife. It requires a further investigation.

The confidence intervals constructed using the bootstrap methods always improve the coverage rate and sometimes quite dramatically over the symmetric interval based on the jackknife variance estimator. The BC method studied but not shown here works better than the percentile method. The $BC_a$ method is in turn better than the BC method. The improved coverage, however, was achieved on the expense of a longer confidence interval. The improvement is also realized by a more even distribution of the right and left side noncoverage of the jackknife confidence interval. We can see clearly that bootstrap confidence intervals shift towards the right side and they cover the true mean much more often on the right side than the symmetric interval based on the jackknife variance estimator.

The coverage rate for the Lee estimator is generally lower than for other estimators, while the length of the confidence interval is the shortest. Whenever the point estimator has a non-negligible bias, the coverage rate suffers. Thus, the confidence interval with less biased estimators has a better coverage rate. It is noteworthy that the confidence interval based on the Chambers estimator has often the best coverage.

## 4. Summary and Concluding Remarks

We studied various outlier resistant estimators under simple random sampling from outlier prone right skewed populations. Although the sample design is simple, the results of the study are applicable to the stratified simple random sample design, which is often used for business establishment surveys, if the resistant estimators are applied stratum by stratum.

We focused on how differently the resistant estimators perform under the different types of skewed distributions and how their performance changes as the sample size changes. We also stressed the importance of variance estimation and the inferential statistic in the form of confidence interval.

One of the important characteristics of an estimator is whether the estimator is consistent or not because most outlier resistant estimators are biased and the bias becomes dominant for large sample sizes. The Winsorized, Fuller, Hulliger, Lee estimators are consistent. The latter three estimators are adaptive in the sense that the estimators have an automatic mechanism to adjust to the severity of outlier problem for a given sample so that the outliers present in the sample are more aggressively treated. The Chambers estimator works generally well but it is not consistent. The tuning factor for the bias correction in the Chambers estimator can be modified according to the underlying population if the shape of the population is known a priori but it is not automatic.

Robust estimation under skewed populations should be achieved by a bias-variance trade-off. The M-estimators which focus on the variance reduction results in a heavily biased and less MSE-efficient estimator when the underlying distribution is heavily skewed. Therefore, bias correction is needed to make an outlier-resistant estimator more MSE-efficient under various shapes of skewed populations. Three estimators (Chambers, Hulliger, and Lee) that are based on the M-estimation technique incorporate the bias-correction differently. It seems that the Chambers and Lee estimators correct the bias more effectively. The Fuller estimator has the bias-correction feature by using the sample mean if the test reveals less than severe skewness. However, it seems that the test is not sensitive to detect moderate skewness and corrects too much of the bias, resulting in a less MSE-efficient estimator. The simple and old Winsorization technique proves itself to be able to handle outliers under severely skewed populations. The Hidiroglou-Srinath estimator based on the weight reduction strategy works nicely over all with the simple choice of $\lambda$.

The jackknife variance estimator works reasonably well but the confidence interval constructed using the jackknife variance estimator has a poor coverage rate under skewed distributions. On the other hand, the bootstrap methods improve the coverage sometime dramatically but the length of the confidence interval is also increased. Even though the behavior of the bootstrap variance estimator for the Lee estimator and the Winsorized estimators is unexplicably erratic, the technique shows a great potential. The $BC_a$ confidence interval performs the best but its coverage rate is still far short of the nominal value. Therefore, a further improvement is needed. The method is second-order correct (i.e., the order of $\sqrt{n}$ ) when the point estimator is unbiased (see Efron, 1987), and thus, an improved method may need to be third-order correct (i.e., order of $n$).

## References

Chambers, R.L. (1986), "Outlier Resistant Finite Population Estimation," *Journal of American Statistical Association*, **81**, pp. 1163-1169.

Chambers, R.L., and Dorfman, A.H. (1994), "Robust Sample Survey Inference via Bootstrapping and Bias Correction: The Case of the Ratio Estimator", *Proc. of the Stat. Comp. Sect., Amer. Stat. Ass.*, pp.51-60.

Efron, B. (1981), "Nonparametric Standard Errors and Confidence Intervals" (with discussion), *Canadian Journal of Statistics*, **9**, pp. 139-172.

Efron, B. (1982), "The Jackknife, the Bootstrap, and Other Resampling Plans," CBMS **38**, SIAM-NSF.

Efron, B. (1987), "Better Bootstrap Confidence Intervals," *Jour. of Amer. Stat. Ass.*, **82**, pp. 171-185.

Fuller, W.A. (1991), "Simple Estimators of the Mean of Skewed Populations," *Statist. Sinica*, 1, 137-158.

Gross, W.F., G. Bade, J.M. Taylor, and C.W. Lloyd_Smith (1986), "Some Finite Population Estimators Which Reduce the Contribution of Outliers," *Proceedings of the Pacific Statistical Congress*, Auckland, New Zealand, pp. 386-390.

Gwet, J.-P. (1997), "Resistant Statistical Inference in Survey Sampling," Ph.D. Thesis, Carleton University.

Gwet, J.-P. and L.-P. Rivest (1992), "Outlier Resistant Alternative to the Ratio Estimator," *Journal of American Statistical Association*, **87**, pp. 1174-1182.

Hidiroglou, M.A., and K.P. Srinath (1981), "Some Estimation of Population Total from Simple Random Samples Containing Large Units," *Journal of American Statistical Association*, 76, pp. 690-695.

Huber, P.J. (1964), "Resistant Estimation of Location Parameter," *Ann. of Math. Statistics*, **35**, pp. 73-101.

Hulliger, B. (1995), "Resistantified Horvitz-Thompson Estimator," *Survey Methodology*, **21**, pp.79-87.

Hulliger, B. (1999), "Simple and Resistant Estimators for Sampling," *Proceedings of the Survey Research Methods Section of the Amrican Statistical Association* (to appear).

Lee, H. (1991), "Model-based Estimators That Are Resistant to Outliers," *Proceedings of the Annual Research Conference*, Washington, D.C.: Bureau of the Census.

Lee, H. (1995), "Chapter 26. Outliers in Business Surveys," *Business Survey Methods*, eds. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge, and P. Kott, New York: John Wiley.

Lee, H. and Z. Patak (1998), "Outlier Resistant Generalized Regression Estimator," *Proceedings of the Survey Methods Section*, *Statistical Society of Canada*, pp. 241-247.

McCarthy, P.J., and C.B. Snowden (1985), "The Bootstrap and Finite Population Sampling," in *Vital and Health Statistics*, **2-95**, Public Health Service Publication 85-1369, Washington, DC: U.S. Government Printing Office.

Rao, J.N.K., and C.F.J. Wu (1988), "Resampling Inference with Complex Survey Data," *Journal of American Statistical Association*, **83**, pp. 231-241.

Rivest, L.-P. (1994), "Statistical Properties of Winsorized Means for Skewed Distributions," *Biometrika*, **81**, pp. 373-384.

Searls, D.T. (1966), "The Estimator for a Population Mean Which Reduces the Effect of Large True Observations," *Journal of American Statistical Association*, **61**, pp. 1200-1205.

Sitter, R. (1992), "A Resampling Procedure for Complex Survey Data," *Journal of American Statistical Association*, **87**, pp. 755-765.

**Table 2. Simulation Results on the Performance of the Sample Mean and Resistant Alternative Estimators with Three Sample Sizes under Even Numbered Universes**

| Universe | Estimator | Relative Bias (Point) | | | Relative MSE-Efficiency | | |
|---|---|---|---|---|---|---|---|
| | | 10 | 20 | 30 | 10 | 20 | 30 |
| MIXN2 | Sample Mean | 0.00 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| | Win0.75-T1 | -0.02 | -0.02 | -0.01 | 1.19 | 1.02 | 0.98 |
| | Win1.00-T1 | -0.03 | -0.02 | -0.01 | 1.17 | 0.98 | 0.94 |
| | Win0.75-T2 | -0.02 | -0.01 | -0.01 | 1.18 | 1.02 | 0.99 |
| | Win1.00-T2 | -0.02 | -0.02 | -0.01 | 1.17 | 1.00 | 0.96 |
| | Fuller | 0.00 | 0.00 | 0.00 | 1.03 | 1.00 | 1.01 |
| | M-Estimator | -0.01 | -0.02 | -0.01 | 1.25 | 1.20 | 1.23 |
| | One-Step M. | -0.01 | -0.02 | -0.01 | 1.29 | 1.22 | 1.25 |
| | Chambers | 0.00 | 0.00 | 0.00 | 1.02 | 1.01 | 1.00 |
| | Lee | -0.01 | -0.02 | -0.01 | 1.23 | 1.18 | 1.20 |
| | Hulliger | -0.01 | -0.01 | 0.00 | 1.20 | 1.03 | 1.01 |
| | Hidiroglou-Srinath | -0.02 | -0.02 | -0.02 | 1.22 | 1.29 | 1.15 |
| LogN2 | Sample Mean | -0.02 | 0.00 | -0.01 | 1.00 | 1.00 | 1.00 |
| | Win0.75-T1 | -0.20 | -0.11 | -0.10 | 1.26 | 1.15 | 1.01 |
| | Win1.00-T1 | -0.26 | -0.15 | -0.13 | 1.19 | 1.10 | 0.94 |
| | Win0.75-T2 | -0.19 | -0.10 | -0.09 | 1.27 | 1.15 | 1.03 |
| | Win1.00-T2 | -0.25 | -0.14 | -0.11 | 1.21 | 1.12 | 0.98 |
| | Fuller | -0.07 | -0.02 | -0.02 | 1.06 | 1.00 | 0.99 |
| | M-Estimator | -0.32 | -0.33 | -0.36 | 1.35 | 0.92 | 0.56 |
| | One-Step M. | -0.34 | -0.35 | -0.37 | 1.33 | 0.88 | 0.53 |
| | Chambers | -0.14 | -0.12 | -0.13 | 1.35 | 1.39 | 1.20 |
| | Lee | -0.24 | -0.19 | -0.17 | 1.40 | 1.22 | 1.00 |
| | Hulliger | -0.11 | -0.04 | -0.06 | 1.07 | 1.09 | 1.06 |
| | Hidiroglou-Srinath | -0.32 | -0.32 | -0.32 | 1.76 | 1.62 | 1.27 |
| Pareto2 | Sample Mean | 0.00 | 0.02 | 0.01 | 1.00 | 1.00 | 1.00 |
| | Win0.75-T1 | -0.08 | -0.04 | -0.03 | 1.54 | 1.53 | 1.46 |
| | Win1.00-T1 | -0.11 | -0.06 | -0.05 | 1.49 | 1.54 | 1.45 |
| | Win0.75-T2 | -0.08 | -0.04 | -0.03 | 1.54 | 1.50 | 1.42 |
| | Win1.00-T2 | -0.10 | -0.05 | -0.04 | 1.51 | 1.55 | 1.47 |
| | Fuller | -0.02 | 0.00 | 0.00 | 1.16 | 1.09 | 1.08 |
| | M-Estimator | -0.11 | -0.10 | -0.11 | 1.52 | 1.47 | 1.08 |
| | One-Step M. | -0.12 | -0.11 | -0.11 | 1.51 | 1.41 | 1.00 |
| | Chambers | -0.04 | -0.03 | -0.03 | 1.44 | 1.63 | 1.66 |
| | Lee | -0.09 | -0.07 | -0.06 | 1.60 | 1.64 | 1.42 |
| | Hulliger | -0.05 | 0.00 | -0.01 | 1.39 | 1.23 | 1.28 |
| | Hidiroglou-Srinath | -0.16 | -0.16 | -0.16 | 1.51 | 1.54 | 1.32 |
| Weibull2 | Sample Mean | -0.02 | 0.06 | 0.04 | 1.00 | 1.00 | 1.00 |
| | Win0.75-T1 | -0.37 | -0.21 | -0.18 | 1.92 | 1.77 | 1.61 |
| | Win1.00-T1 | -0.49 | -0.31 | -0.25 | 1.84 | 1.76 | 1.57 |
| | Win0.75-T2 | -0.35 | -0.19 | -0.15 | 1.90 | 1.73 | 1.57 |
| | Win1.00-T2 | -0.47 | -0.27 | -0.21 | 1.88 | 1.79 | 1.61 |
| | Fuller | -0.28 | -0.09 | -0.07 | 1.52 | 1.30 | 1.23 |
| | M-Estimator | -0.72 | -0.76 | -0.78 | 1.52 | 0.99 | 0.55 |
| | One-Step M. | -0.73 | -0.77 | -0.80 | 1.50 | 0.97 | 0.54 |
| | Chambers | -0.54 | -0.53 | -0.53 | 1.75 | 1.59 | 1.05 |
| | Lee | -0.51 | -0.33 | -0.25 | 1.90 | 1.87 | 1.60 |
| | Hulliger | -0.46 | -0.44 | -0.47 | 1.69 | 1.74 | 1.19 |
| | Hidiroglou-Srinath | -0.33 | -0.34 | -0.34 | 1.86 | 1.91 | 1.51 |

**Table 3. Coverage and Length of the 95% Confidence Intervals Constructed by Various Methods with the Sample Mean and Resistant Estimators for Sample Size 20**

| Universe | Estimator | Relative Bias | Relative Efficiency | Rel Bias of Varince Estimator | | Coverage Rate (Jackknife) | | | Coverage Rate (Percentile) | | | Coverage Rate (BCA) | | | Length of Interval | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Jackknife | Bootstrap | Left | Middle | Right | Left | Middle | Right | Left | Middle | Right | Jackknife | Percentile | BCA |
| MIXN1 | Sample Mean | 0.00 | 1.00 | 0.03 | 0.02 | 0.023 | 0.938 | 0.039 | 0.024 | 0.938 | 0.039 | 0.028 | 0.937 | 0.036 | 0.89 | 0.88 | 0.89 |
| | Win1.00-T1 | -0.01 | 0.96 | 0.04 | 0.04 | 0.015 | 0.934 | 0.052 | 0.020 | 0.941 | 0.040 | 0.026 | 0.933 | 0.042 | 0.90 | 0.90 | 0.91 |
| | Fuller | 0.00 | 1.00 | 0.03 | 0.03 | 0.023 | 0.938 | 0.039 | 0.023 | 0.938 | 0.039 | 0.028 | 0.935 | 0.037 | 0.89 | 0.88 | 0.89 |
| | M-Estimator | 0.00 | 0.99 | 0.06 | 0.06 | 0.022 | 0.938 | 0.041 | 0.021 | 0.940 | 0.040 | 0.023 | 0.941 | 0.036 | 0.90 | 0.90 | 0.91 |
| | Chambers | 0.00 | 1.00 | 0.03 | 0.02 | 0.023 | 0.938 | 0.039 | 0.024 | 0.938 | 0.039 | 0.028 | 0.937 | 0.036 | 0.89 | 0.88 | 0.89 |
| | Lee | 0.00 | 0.99 | 0.06 | -0.49 | 0.022 | 0.939 | 0.040 | 0.046 | 0.881 | 0.073 | 0.063 | 0.877 | 0.061 | 0.90 | 0.72 | 0.73 |
| | Hulliger | 0.00 | 0.97 | 0.15 | N/A | 0.020 | 0.938 | 0.043 | N/A | N/A | N/A | N/A | N/A | N/A | 0.94 | N/A | N/A |
| | Hidroglou-Srinath | -0.01 | 0.89 | 0.11 | 0.04 | 0.02 | 0.92 | 0.06 | 0.01 | 0.94 | 0.05 | 0.02 | 0.94 | 0.04 | 0.94 | 0.93 | 0.94 |
| LogN1 | Sample Mean | 0.00 | 1.00 | 0.01 | 0.01 | 0.001 | 0.791 | 0.209 | 0.002 | 0.810 | 0.189 | 0.016 | 0.853 | 0.131 | 3.30 | 3.20 | 3.94 |
| | Win1.00-T1 | -0.17 | 1.13 | 0.13 | 0.64 | 0.002 | 0.668 | 0.331 | 0.001 | 0.809 | 0.190 | 0.002 | 0.804 | 0.195 | 2.74 | 3.27 | 3.17 |
| | Fuller | -0.02 | 1.01 | 0.12 | 0.07 | 0.001 | 0.786 | 0.214 | 0.001 | 0.810 | 0.189 | 0.014 | 0.859 | 0.128 | 3.36 | 3.24 | 3.96 |
| | M-Estimator | -0.37 | 0.89 | 0.25 | 0.47 | 0.000 | 0.463 | 0.537 | 0.000 | 0.614 | 0.386 | 0.001 | 0.751 | 0.249 | 2.00 | 2.17 | 2.60 |
| | Chambers | -0.16 | 1.42 | 0.14 | 0.19 | 0.001 | 0.746 | 0.254 | 0.001 | 0.791 | 0.209 | 0.008 | 0.860 | 0.133 | 2.67 | 2.70 | 3.28 |
| | Lee | -0.20 | 1.24 | 0.04 | 0.28 | 0.000 | 0.670 | 0.330 | 0.007 | 0.735 | 0.259 | 0.019 | 0.781 | 0.201 | 2.54 | 2.65 | 2.91 |
| | Hulliger | -0.07 | 1.19 | 0.08 | N/A | 0.001 | 0.787 | 0.213 | N/A | N/A | N/A | N/A | N/A | N/A | 3.09 | N/A | N/A |
| | Hidroglou-Srinath | -0.23 | 1.21 | 0.41 | 0.41 | 0.00 | 0.67 | 0.33 | 0.00 | 0.75 | 0.25 | 0.00 | 0.82 | 0.18 | 2.66 | 2.68 | 3.11 |
| Pareto1 | Sample Mean | 0.01 | 1.00 | -0.07 | -0.09 | 0.003 | 0.846 | 0.152 | 0.010 | 0.851 | 0.140 | 0.028 | 0.869 | 0.104 | 0.63 | 0.60 | 0.72 |
| | Win1.00-T1 | -0.12 | 1.29 | -0.01 | 0.54 | 0.002 | 0.734 | 0.265 | 0.007 | 0.853 | 0.141 | 0.006 | 0.860 | 0.135 | 0.53 | 0.62 | 0.60 |
| | Fuller | 0.00 | 1.03 | 0.00 | -0.04 | 0.003 | 0.846 | 0.152 | 0.009 | 0.851 | 0.141 | 0.033 | 0.865 | 0.103 | 0.63 | 0.61 | 0.73 |
| | M-Estimator | -0.19 | 1.27 | 0.01 | 0.08 | 0.003 | 0.713 | 0.285 | 0.001 | 0.793 | 0.206 | 0.006 | 0.868 | 0.126 | 0.48 | 0.50 | 0.58 |
| | Chambers | -0.04 | 1.29 | -0.03 | -0.03 | 0.003 | 0.843 | 0.155 | 0.008 | 0.852 | 0.141 | 0.021 | 0.877 | 0.102 | 0.58 | 0.57 | 0.67 |
| | Lee | -0.13 | 1.38 | -0.06 | -0.06 | 0.003 | 0.764 | 0.234 | 0.017 | 0.783 | 0.201 | 0.038 | 0.815 | 0.148 | 0.51 | 0.50 | 0.54 |
| | Hulliger | 0.00 | 1.05 | -0.03 | N/A | 0.003 | 0.848 | 0.149 | N/A | N/A | N/A | N/A | N/A | N/A | 0.62 | N/A | N/A |
| | Hidroglou-Srinath | -0.15 | 1.29 | 0.12 | 0.09 | 0.01 | 0.75 | 0.24 | 0.00 | 0.81 | 0.19 | 0.01 | 0.85 | 0.14 | 0.54 | 0.54 | 0.60 |
| Weibull1 | Sample Mean | 0.02 | 1.00 | -0.07 | -0.09 | 0.003 | 0.833 | 0.165 | 0.010 | 0.841 | 0.150 | 0.029 | 0.866 | 0.106 | 2.24 | 2.16 | 2.61 |
| | Win1.00-T1 | -0.15 | 1.18 | 0.00 | 0.39 | 0.001 | 0.709 | 0.291 | 0.005 | 0.846 | 0.150 | 0.005 | 0.849 | 0.147 | 1.94 | 2.22 | 2.19 |
| | Fuller | 0.01 | 1.01 | 0.00 | -0.05 | 0.003 | 0.833 | 0.165 | 0.009 | 0.842 | 0.150 | 0.027 | 0.869 | 0.105 | 2.28 | 2.19 | 2.65 |
| | M-Estimator | -0.36 | 0.86 | 0.10 | 0.14 | 0.001 | 0.563 | 0.436 | 0.001 | 0.689 | 0.311 | 0.001 | 0.806 | 0.193 | 1.65 | 1.75 | 2.05 |
| | Chambers | -0.12 | 1.22 | 0.02 | -0.05 | 0.003 | 0.785 | 0.212 | 0.007 | 0.827 | 0.167 | 0.017 | 0.877 | 0.107 | 2.04 | 1.96 | 2.37 |
| | Lee | -0.18 | 1.25 | -0.07 | 0.01 | 0.001 | 0.723 | 0.277 | 0.016 | 0.780 | 0.205 | 0.036 | 0.807 | 0.158 | 1.82 | 1.80 | 1.96 |
| | Hulliger | -0.03 | 1.10 | 0.02 | N/A | 0.003 | 0.824 | 0.173 | N/A | N/A | N/A | N/A | N/A | N/A | 2.21 | N/A | N/A |
| | Hidroglou-Srinath | -0.21 | 1.18 | 0.15 | 0.14 | 0.00 | 0.72 | 0.28 | 0.00 | 0.79 | 0.21 | 0.01 | 0.85 | 0.15 | 1.93 | 1.93 | 2.21 |

# WINSORIZATION FOR IDENTIFYING AND TREATING OUTLIERS IN BUSINESS SURVEYS

Ray Chambers, University of Southampton, Philip Kokic, Insiders GmbH and Paul Smith and Marie Cruddas,
Office for National Statistics

Ray Chambers, Dept of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK
rc@alcd.soton.ac.uk

## ABSTRACT

The issue of how to deal with representative outliers in finite populations is particularly important for business surveys where some sample elements can have values many times bigger than the mean. In particular it is useful to have a method which is automated, and which produces an estimator with demonstrable theoretical properties rather than relying on the judgement of a survey statistician. In the Office for National Statistics (ONS) in the UK the method which has been introduced recently is winsorization, which can be viewed equivalently as altering the value of an extreme observation or altering its weight so that it has less effect on the estimated total. The alteration is made in such a way as to give an estimate with good mean square error properties in repeating surveys.

The paper describes one particular approach to winsorization, called one-sided winsorization, that adjusts the value/weight of observations significantly larger than the fitted values implied by the estimation model. The definition of "significantly larger" relies on the calculation of a threshold parameter based on past survey information. It implicitly assumes that the auxiliary data used in fitting the model are correct and contain no outliers, which may be reasonable in a stratified business survey situation.

The paper also covers some of the practical issues that must be considered in using winsorization, and report some of the experiences obtained in using the method with ONS business surveys.

**Key Words: Minimum-mse estimator, robust methods**

## 1. INTRODUCTION

It is an unfortunate fact that the economic populations that are the targets for business surveys are typically skewed and prone to outliers. When outlying values are sampled, they can exercise an influence on the survey estimate that is quite out of proportion to their number in the sample. Chambers (1986) classified outliers into two groups – representative outliers, corresponding to data values that are "correct" in that they are valid observations from the target population, and non-representative outliers, corresponding to values that are actually errors introduced into the sample data at some stage prior to estimation. Strictly speaking, data values in the latter group are in the domain of editing systems rather than outlier robust estimation systems, but this distinction tends to be blurred in practice, with many outlier robust estimation systems used to identify and "correct" outlying values. In this paper we shall assume that any outliers in our sample data are representative.

Throughout this paper our focus will be on weighted linear estimation of finite population totals. That is, we consider estimation schemes where the estimate of the total T of a survey variable Y can be written in the form

$$\hat{T} = \sum_s w_i Y_i$$

Here $Y_i$ denotes the value of the survey variable Y for the ith population unit, and $w_i$ is the sample weight for this unit. The set s denotes the labels of those population units that are in the sample. Outliers can affect this type of estimate in two different ways:

(1) The outlier may be a very large Y-value, completely unlike any other sample Y-value, with a weight that implies that there is a substantial number of "similar" values in the population;

(2) The outlier Y-value may not be particularly large relative to other sample data values, but its associated weight may be very large relative to other sample weights.

The outcome for both of these situations is that the estimate $\hat{T}$ is dominated by the contribution $w_i Y_i$ from the outlying unit.

There are two conceptually quite distinct, but in practice often interchangeable, "solutions" to this problem. The first is to modify the weight associated with the outlying sample unit. The second is to modify the actual Y-value that has been recorded for the unit.

The first solution falls into the domain of "weight trimming" or weight constraining. There has been a substantial amount of work carried out over the last decade into optimal methods for weight modification. See Potter (1990) and Deville and Särndal (1992). We do not pursue this idea any further here, mainly because in the context of outlier adjustment this corresponds to using a different weight for the sample unit when estimating the population total for the variable containing the outlying value compared with the weight that would be used with the same unit when estimating population totals for variables that do not contain outliers. Issues of consistency and "transparency" of survey estimation usually dictate that a single weight be used with any sample unit (a so-called uniweight strategy) and consequently variable-specific weight modification as an outlier solution is, at least in theory, inappropriate.

The second solution, where the outlying Y-value for the unit is appropriately modified to lessen its impact on the estimate $\hat{T}$, represents the approach that we take in this paper. In the following section we consider a relatively simple and easily implemented version of this approach that is often referred to as "winsorization". In particular, we focus on one-sided winsorization, where a pre-defined rule is used to adjust unusually large values of Y downwards, leaving remaining values unchanged. One-sided winsorization makes sense where outliers always occur on one "side" of the data. In business survey applications where the survey variables are typically intrinsically positive and the extremes are the unusually large values, this type of outlier pattern is common. However there are situations where outliers can be on either positive or negative, and a two-sided winsorization approach may be more sensible. An experimental version of this idea is currently being investigated at ONS. However, the results of this investigation are still preliminary (Kokic and Smith, 1999).

A problem with the one-sided winsorization approach is that it will introduce a negative bias into estimation. The extent of this bias and its relationship to the variability of the weighted estimate based on winsorized data values is the focus of the discussion in the next section. Practical issues arising when a winsorization-based estimation strategy is adopted is the focus of section 3. Section 4 concludes the paper with a discussion of more general issues related to introduction of outlier robust estimation methods within the business survey operations of the UK Office for National Statistics (ONS).

## 2. ONE-SIDED WINSORIZATION

In survey estimation, one-sided winsorization is where a pre-defined rule is used to adjust an outlying (positive) value $Y_i$ of the survey variable Y downwards, leaving remaining values unchanged (Searls, 1966). For more complex estimation models an outlier is defined by its large (positive) residual when compared with its expected value under the estimation model in use. The value of the adjusted variable is denoted $Y_i^*$ and the corresponding winsorized estimator of the total is

$$\hat{T}_W = \sum_s w_i Y_i^* . \qquad (1)$$

There are two basic ways in which the value $Y_i^*$ can be defined. These correspond to different ways of thinking about the "winsorized" contribution of an outlying sample unit to the estimate of the population total. The first is where we see this contribution as "weighted up" in the same way (say by the inverse of the unit's sample inclusion probability) as we would "weight up" its value if it were not outlying. That is, we define $Y_i^*$ as

$$Y_i^* = \begin{cases} K & \text{if } Y_i > K \\ Y_i & \text{otherwise} \end{cases} \qquad (2)$$

where K > 0 is a pre-specified constant, the so-called "cut-off" value for the winsorization procedure.

The second way of defining the winsorized contribution of a particular unit to the estimate of the population total is where we think of this as being the sum of two parts – the first being the actual Y-value (unwinsorized) of the unit and the second being its (winsorized) contribution to the estimated sum of Y-values for the non-sampled population units. This second interpretation falls in naturally with the idea that all the outlying values in our sample data are "correct" and so outliers should be allowed to contribute their (unweighted) values to the estimate of the population total. Their contribution to the population total of the non-sampled units, however, cannot be left unchanged, and so their winsorized value is multiplied by their estimation weight minus one to define this contribution. That is, $Y_i^*$ is computed as

$$Y_i^* = \begin{cases} \frac{1}{w_i} Y_i + \left( \frac{w_i - 1}{w_i} \right) K & \text{if } Y_i > K \\ Y_i & \text{otherwise} \end{cases} \tag{3}$$

where again K > 0 is the cut-off value for the winsorization procedure. In the literature, (2) is sometimes referred to as Type 1 winsorization, while (3) is referred to as Type 2 winsorization. An obvious advantage of Type 2 winsorization is that as $w_i \to 1$, so an outlier basically only represents itself, the degree to which $Y_i$ is modified decreases substantially.

In one of the earliest applications of the winsorization idea to sample surveys, Ernst (1980) showed that for simple random sampling with replacement from a continuous distribution, there exists a cut-off K such that the Type 1 winsorized estimator has a smaller mean squared error than any other estimator that reduces sampled values. Gross (1985) extended this result to stratified simple random sampling without replacement, and Gross *et al.* (1986) introduced the concept of Type 2 winsorization and demonstrated very little difference in mean squared error between the Type 1 and 2 winsorized estimators. All of these results tackled the problem from a design-based viewpoint. Since Type 2 winsorization is the only approach that also makes sense from a prediction theory viewpoint, this is the method of winsorization that we consider in this paper.

## 2. 1. How to determine the cut-off value?

The key methodological issue for the application of winsorization is the choice of the cut-off value K. Tamby (1988) considered this, and suggested a method based on the interquartile distance of the sample Y-values. His research showed that with this choice the Type 2 winsorized estimator had smaller mean squared error than the expansion estimator under simple random sampling. Kokic and Bell (1994) tackled this problem within the context of stratified random sampling. They showed that the overall mean squared error of a stratified Type 2 winsorized estimator was minimised by choosing a value of K within stratum h of the form

$$K_h = (N_h/n_h - 1)^{-1} L + \bar{y}_h \tag{4}$$

where $\bar{y}_h$ is the sample mean of the Y-values within stratum h and L > 0 is a constant chosen so that the bias of the stratified winsorized estimator is –L. Their numerical results show that the stratified winsorized estimator using cut-offs defined by (4) nearly always has smaller mean squared error than the stratified expansion estimator.

Clarke (1995) obtained a model-based extension of this result. He considered the case where the population values underlying the sample data can be characterised in terms of the very general model where $Y_i; i = 1, \ldots, N$ are independent realisations of random variables satisfying:

$$\begin{aligned} E(Y_i) &= \mu_i \\ \operatorname{var}(Y_i) &= \sigma_i^2 \end{aligned} \tag{5}$$

and where the winsorized estimator of the population total of the Y-values is defined by

$$\hat{T}^* = \sum_s Y_i + \sum_s (w_i - 1) Y_i^* \tag{6}$$

Here $Y_i^* = \min(Y_i, K_i)$ and $K_i$ is the cut-off value for the ith population unit. Note that we can equivalently write

$$\hat{T}^* = \sum_s w_i Z_i$$

where $Z_i = \min\{Y_i, K_i + (Y_i - K_i)/w_i\}$. That is, the winsorized estimator (6) can be written as a weighted sum of the transformed variables $Z_i$. Clarke (1995) shows that the choice $K_i$ that minimises the mean squared error of (6) under the model (5) is then

$$K_i = \mu_i^* - B(w_i - 1)^{-1} \tag{7}$$

where $\mu_i^* = E(Y_i^*)$ and B is the bias of $\hat{T}^*$ under (5). That is

$$B = \sum_s (w_i - 1)(\mu_i^* - \mu_i). \tag{8}$$

In practice, the values $\mu_i^*$ will be difficult to estimate. Consequently, Clarke suggests use of the approximation

$$K_i \approx \mu_i + L(w_i - 1)^{-1} \tag{9}$$

where L > 0 is the negative of the bias B in (8) above.

Computation of the cut-off value (9) can be accomplished using the same approach as that outlined in Kokic and Bell (1994). To start, define the weighted residuals

$$D_i = (Y_i - \mu_i)(w_i - 1).$$

For a fixed value of the coefficient L, the bias (8) can then be written

$$\begin{aligned} B(L) &= \sum_{i \in s} (w_i - 1)(\mu_i^* - \mu_i) \\ &= \sum_{i \in s} (w_i - 1)\{E[\min(Y_i, K_i)] - \mu_i\} \\ &= \sum_{i \in s} E[\min\{(Y_i - \mu_i)(w_i - 1), (K_i - \mu_i)(w_i - 1)\}] \\ &= \sum_{i \in s} E[\min(D_i, L)] \\ &= \sum_{i \in s} E[\min(0, L - D_i) + D_i] \\ &= -E\left\{\sum_{i \in s} \max(D_i - L, 0)\right\}. \end{aligned}$$

Let $\hat{\mu}_i$ denote an estimate of $\mu_i$. An estimate of the above bias is then

$$\hat{B}(L) = -\sum_s \max(\hat{D}_i - L, 0)$$

where

$$\hat{D}_i = (Y_i - \hat{\mu}_i)(w_i - 1).$$

Since the optimal value of L is the value where $B(L)=-L$, we can find this value by solving the equation $\psi(L)=L+\hat{B}(L)=0$. This is straightforward to do once we recognise that $\psi(L)$ is piecewise linear and decreasing, with distinct values

$$\psi_k=\psi\left(\hat{D}_{(k)}\right)=\hat{D}_{(k)}-\sum_s \max\left(\hat{D}_{(i)}-\hat{D}_{(k)},0\right)=(k+1)\hat{D}_{(k)}-\sum_{j=1}^{k}\hat{D}_{(j)}\,.$$

Here $\hat{D}_{(1)}\geq\hat{D}_{(2)}\geq\ldots$ are the ordered values of $\hat{D}_i$. Hence all we need to do is to find the last value of k, $k^*$ say, for which $\psi_k$ is still positive, and then set $\hat{L}$ to

$$\hat{L}=(k^*+1)^{-1}\sum_{j=1}^{k^*}\hat{D}_{(j)}\,.$$

Note that $-\hat{L}$ is then an estimate of the winsorization bias under the model (5). Consequently an estimate of the mean squared error of (6) can be calculated by adding the square of this term to the estimated variance of (6).

The above procedure depends on having access to a "good" estimate $\hat{\mu}_i$ of the expected value $Y_i$ under (5) for each sampled unit $i\in s$. A little reflection shows that there is some circularity here, since one obvious unbiased estimator of $\mu_i$ is $Y_i$ itself, which leads to an infinite cut-off for unit i (i.e. no winsorization). Clarke (1995), following the lead provided by Kokic and Bell (1994), suggests that a more efficient estimate for $\mu_i$ can be obtained using data from several previous repeats of the survey. This requires that (i) such data exist and (ii) they can be used to estimate $\mu_i$. The second requirement is equivalent to requiring that a model can be specified for the distribution of Y-values across both the population units at the present time and the population units at the times of these previous survey repeats. A particular problem arises when $Y_i$ is an outlying value and either there are few (non-outlying) historical values of $Y_i$ or these values also contain outliers. In either case we would expect the value of $\mu_i$ to be estimated with a substantial positive bias, and hence $K_i$ set "too high".

In practice the above problems can be partly overcome by using an outlier robust method to estimate $\mu_i$. See Huber (1981) and Hampel *et al* (1986). These methods are now available as standard options in many statistical analysis packages. The problem of not having any "historical" data to estimate $\mu_i$ can also be resolved by using outlier robust estimation methods with the current survey data. For example, a common specification is $\mu_i=\beta X_i$, where $X_i$ is the value of a known auxiliary variable X, and $\beta$ is an unknown constant. In this context we can use outlier robust estimation methods to estimate $\beta$ from the sample data. Let b denote this robust estimate. The estimate of $\mu_i$ we then use to determine $K_i$ is $bX_i$.

A further issue that arises is whether the estimate $\hat{L}$ should be updated between repetitions of the survey. Our experience is that it is best not to change this value too often, and our recommendation is that $\hat{L}$ is held fixed between redesigns of the survey so that estimates of change are approximately unbiased and unaffected by the movement of cut-offs between surveys.

## 3. PRACTICAL ISSUES IN WINSORIZATION

### 3. 1. Level of Winsorization

Estimates derived from business survey data are typically released at various levels of aggregation. For example, estimates for quite "fine" industry strata (e.g. three digit level of the Standard Industrial Classification (SIC)) may be released at a national level, while estimates for more aggregated industry classes (e.g. two digit SIC) might be released at regional level. It is a standard requirement that these two sets of outputs be mutually consistent. A complex aggregation structure for outputs leads to problems when applying winsorization in practice, since one then has to decide the level at which winsorization is applied. That is, one has to decide which group of output strata form the "population" for calculation of the winsorization cut-off $K_i$ for a particular sample unit. One cannot just vary this cut-off depending on the aggregation level for the estimate, since then inconsistencies can arise when comparing optimally winsorized estimates at different aggregation levels. In particular, optimally winsorized estimates at lower aggregation levels (e.g. three digit SIC) will not aggregate to optimally winsorized estimates at higher levels of aggregation (e.g. two digit SIC) because the cut-off values that are optimal at the lower level will not be optimal at the higher level and vice versa. Typically we would expect the lower level cut-offs to be smaller than the higher level cut-offs.

To the best of our knowledge there is no way of resolving this problem while at the same time maintaining efficiency at both levels of aggregation. The approach adopted at the ONS is to decide which set of outputs is the most important and optimally winsorize at that level. Since greater importance is usually given to higher levels of aggregation rather than to lower levels in ONS business surveys, this compromise typically leads to larger cut-off values (and hence "less" winsorization) of the sample data than would be the case if the optimal cut-offs were calculated at lower levels. An empirical evaluation of this strategy is contained in Kokic and Smith (1998). An alternative approach is to introduce rescaling factors to lower level aggregates, so that they are consistent with higher levels, but remain in the proportions determined from winsorization with low-level cut-offs. Although superficially attractive, the system rapidly becomes extremely complex where a range of estimates is produced.

### 3. 2. Winsorization Of Derived Variables

Another important issue as far as using winsorization in practice is concerned is how one deals with so-called derived variables. These are variables that are defined in terms of the collected survey variables. In many cases such derived variables are in fact the only outputs from a survey.

To make things a little more specific, consider an arbitrary sample unit and suppose that the set of variables collected for it in the survey is denoted $Y_l, Y_2, \ldots, Y_p$, where now a subscript defines a different collected variable. Let $a_0, a_1, \ldots, a_p$ denote a set of known values and suppose a derived variable $X = a_0 + a_1 Y_1 + \ldots + a_p Y_p$ is calculated for this unit on the basis of its collected data. For example total sales and turnover are two derived variables that may be calculated by summing collected (component) variables while profit is typically calculated as the difference between receipts and costs, themselves typically obtained by summation of different sets of collected variables.

In theory, winsorization can be applied to each of $Y_l, Y_2, \ldots, Y_p$ as well as to X. However, it is easy to see that the linear relationship between X and $Y_l, Y_2, \ldots, Y_p$ will not hold in general for the resulting estimates.

An obvious alternative is to compute a winsorized version $Y_i^*$ of each component $Y_i$, then calculate the "winsorized" version of X as $X^* = a_0 + a_1 Y_1^* + \ldots + a_p Y_p^*$. This will not correspond to the "directly winsorized" version of X, and so will not lead to a mean squared error optimal estimator of the population total of this variable. In particular, the effectiveness of this approach will depend largely on the bias of this indirectly winsorized estimator.

Let $\hat{T}_i^*$ denote the winsorized estimate of the total of $Y_i$, with bias $B_i$, and define the estimate of the total of X obtained by summing these estimates (i.e. the indirectly winsorized estimator of the population total of X):

$$\hat{T}^{**} = a_0 + a_1 \hat{T}_1^* + \ldots + a_p \hat{T}_p^*$$

The mean squared error of this estimator is given by

$$\mathrm{MSE}(\hat{T}^{**}) = \sum_i a_i^2 \operatorname{var}\left(\hat{T}_i^*\right) + \sum_i \sum_j a_i a_j \operatorname{cov}\left(\hat{T}_i^*, \hat{T}_j^*\right) + \sum_i \sum_j a_i a_j B_i B_j$$

The first term in this mean squared error will be relatively small, since winsorization reduces variances. Furthermore winsorization typically has a relatively minor effect on the second covariance term. Consequently it is the third, bias dependent, term that can be important. In particular, if the $a_i$ are all positive, then since the bias $B_i$ is always non-positive, we see that the mean squared error of $\hat{T}^{**}$ is likely to be much larger than what would be achieved by winsorizing the variable X itself. In contrast, where some of the $a_i$ are negative, this third term can be negative and lead to a smaller mean squared error for $\hat{T}^{**}$ than would be obtained by directly winsorizing X.

Since the optimal winsorization procedure also produces an estimate of the bias of each $\hat{T}_i^*$, it is straightforward to compute an estimate of the third, bias determined, term in the mean squared error of $\hat{T}^{**}$. This suggests that we use $\hat{T}^{**}$ when this estimate is negative.

This still leaves the problem of what to do when the bias component of $\hat{T}^{**}$ is unacceptably large. Although there is no generally optimal solution to this problem, the following compromises seem reasonable. Both are based on winsorizing X first, and then combining this winsorized version of X with the relationship between X and $Y_1, Y_2, \ldots, Y_p$ to decide on winsorized versions of $Y_1, Y_2, \ldots, Y_p$ that "sum" to X. That is, the difference between X and its winsorized value $X^* = \min\{X, K + (X - K)/w\}$ is distributed among its components $Y_1, Y_2, \ldots, Y_p$ in such a way that the new components $Y_1^*, \ldots, Y_p^*$ satisfy $X^* = a_0 + a_1 Y_1^* + \ldots + a_p Y_p^*$.

(a) If X is an outlier and needs to be winsorized, this may have occurred due to inaccurate or out of date benchmark information. In this case it is quite likely that each of its components are also outliers and hence require roughly the same proportion of downward adjustment as X. Consequently we put $Y_i^* = \alpha Y_i$, where $\alpha = (X^* - a_0)/(X - a_0)$.

(b) It is possible that X is an outlier as a result of one or several, but not all, of its components being large. In this case it is more effective to distribute the difference between X and $X^*$ amongst the largest $Y_i$ values.

Option (b) can be implemented using the same approach as used in section 2.1 to determine the optimal value of L in (9). We set $Y_i^* = \min(Y_i, H)$ and then choose H so that $X^* = a_0 + a_1 Y_1^* + \ldots + a_p Y_p^*$. The value of H may be found by solving the equation

$$\psi(H) = X - X^* - \sum_{i=1}^{p} a_i Y_i I(Y_i \geq H) + H \sum_{i=1}^{p} a_i I(Y_i \geq H) = 0 .$$

The function $\psi(H)$ is piecewise linear and decreasing for H > 1. Let $Y_{(1)} \geq Y_{(2)} \geq \ldots \geq Y_{(p)}$ denote the sorted $Y_i$ values, with corresponding coefficients $a_{(1)}, a_{(2)}, \ldots, a_{(p)}$. Then for j = 2, 3, ... we successively compute

$$\psi_j = \psi(H = Y_{(j)}) = X - X^* - \sum_{i=1}^{j} a_{(i)} Y_{(i)} + Y_{(j)} \sum_{i=1}^{j} a_{(i)}$$

and let $j^*$ be the largest value of j for which $\psi_j \geq 0$. The solution for H is then

$$H = \left( \sum_{i=1}^{j^*} a_{(i)} Y_{(i)} + X^* - X \right) \left( \sum_{i=1}^{j^*} a_{(i)} \right)^{-1} .$$

Note that a new value of H must be computed for each sampled observation for which $X^* < X$. However, since there will usually only be a few such outliers, the computational burden should not be excessive.

Non-linear combinations of winsorized variables are not readily calculated from published estimates in general, so there is less of a challenge in ensuring consistency between the survey estimates and these derived variables. For operational reasons therefore these variables are derived after winsorization of their components.

### 3. 3. Non-representative outliers

The derivation of the optimal cut-offs is dependent on the assumption that all the values being used from previous datasets are valid observations from the target population. However, in practice the editing systems in use do not identify all the observations that are subject to measurement and other errors. Where these errors are not large, their effect on the winsorization procedure is small. However, larger errors have a masking effect in the same way that one outlier may mask another in classical outlier theory. This is because the bias introduced in one-sided winsorization quickly dominates the mean-squared error, so producing a minimum-mse solution requires a relatively small amount of bias to be introduced to the estimator in order to reduce the sampling variance. If this bias is "used up" correcting a Y value subject to measurement error, there is none left to treat real (representative) outliers.

A strict application of the minimum mean squared error approach to the Monthly Production Inquiry (a survey of production in the manufacturing sector in the UK) identified of the order of 2/10000 businesses per month on average as outliers using one-sided winsorization. This was quite different from previous experience with this survey. Some investigation demonstrated that businesses identified as outliers in this way often had data (measurement) problems, and it became clear that the previous outlier methods used by the ONS (e.g. poststratification) were adjusting for both representative and non-representative outliers in the data. To replicate this behaviour, the one-sided winsorization cut-offs were adjusted downwards, thus introducing more bias into the procedure, but now identifying a greater number of outliers.

The benefit of continuing to use winsorization in this situation is its objective "ordering" of the observations. By rewriting Z as defined by (7) above as

$$Z_i = \min\left\{1, \frac{1}{w_i}\left(1 + \frac{K}{Y_i}\left(w_i - 1\right)\right)\right\} Y_i = o_i Y_i \qquad (10)$$

we obtain values $0 < o_i < 1$ which indicate how much an observation has been winsorized. By sorting the $o_i$ for a particular value of K, we determine which observations should be made outliers first, so if we raise or lower the cut-off, K, we adjust respectively fewer or more sample observations.

## 4. DISCUSSION

A value modification strategy like winsorization overcomes many of the practical problems associated with outlier robust inference. It is fairly easily implemented in practice since it is straightforward to integrate pre-specified cut-offs into a processing and estimation system. A further advantage is that the adjustment of outliers is then separated from estimation, so that estimation can be performed using standard packages with the winsorization taking place prior to this phase. Note, however, that the weights to be used in estimation must be available before winsorized values can be obtained. A key point however is that variance estimates produced by standard methods or packages will not correspond to the actual mean squared error of the winsorized estimates, and so the coverage properties of the resulting confidence intervals will be seriously overstated. This is an area where further research is necessary.

An important practical issue for application of winsorization is the level at which this is carried out. The compromise chosen for ONS business surveys was initially to winsorize at the whole-survey level. This, however, provides little protection against outliers at detailed levels of disaggregation. The compromise solution in many instances has been to apply winsorization separately at both levels, and then to rescale disaggregate estimates to give consistency with higher levels of aggregation. This approach requires data to be divided into appropriate "independent" groups of variables that can be treated in this way, to avoid over-complicating the rescaling process.

The approach described in this paper relies on winsorizing each variable separately. Within a processing system this requires that different cut-offs and winsorized values are held for each variable (although with type 2 winsorization it is possible to deduce either the original or the winsorized value from the other if the cut-off and weights are known). A stricter adherence to the uniweight principle would be to look for a multiplier $o_i$ from (10) which could be used for all variables for a unit i. This would make processing and consistency issues easier to solve, but probably at the expense of accuracy for the key estimates of survey (rather than derived) variables.

An extension of the winsorization approach to allow cut-offs to be defined on both sides of a fitted estimation line is described in Kokic and Smith (1999). This relies on past data and some assumptions about distribution shape; it also produces an estimator where the bias from adjusting large observations downward is countered by adjusting small observations upward. Initial testing showed better bias properties than the one-sided estimator, but a similar mean squared error. Both winsorization methods have improved confidence interval coverage properties compared with the unadjusted ratio estimator, with two-sided winsorization being slightly better.

An important point to note about two-sided winsorization is that the number of smaller values which is adjusted upward is often large compared with the number of values adjusted downward, since the higher values are often further from the fitted estimation model. This means that the method is not useful for outlier detection – far too many observations are adjusted for all adjusted values to be outliers. Two-sided winsorization is perhaps best viewed as a robust method with a particular influence function.

## 5. REFERENCES

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association* **81**, pp. 1063-1069.

Clarke, R. G. (1995). Winsorization methods in sample surveys. Masters thesis, Department of Statistics, Australian National University.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, pp. 376-382.

Ernst, L.R. (1980). Comparison of estimators of the mean which adjust for large observations. *Sankhya* **42**, pp. 1-16.
Gross, W.F. (1985). An optimal estimator which reduces the contribution of outliers. *Unpublished manuscript*, Canberra, Australia: The Australian Bureau of Statistics.

Gross, W.F., Bode, G., Taylor, J.M. and Lloyd-Smith, C.W. (1986). Some finite population estimators which reduce the contribution of outliers. *Proceedings of the Pacific Statistical Congress*, Auckland, New Zealand, 20-24 May 1985.

Huber, P.J. (1981). Robust statistics. New York: John Wiley.
Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.E. (1986). Robust statistics: The approach based on influence functions. New York: Wiley.

Kokic, P. N. and Bell, P. A. (1994). Optimal Winsorising cut-offs for a stratified finite population estimator. *Journal of Official Statistics* **10**, pp. 419-435.

Kokic, P. N. and Smith, P. (1998). Winsorization of outliers in business surveys. *Submitted for publication*.
Kokic, P. N. and Smith, P. (1999). Outlier-robust estimation in sample surveys using two-sided winsorization. *Unpublished draft*.

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

Potter, F. J. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 235-230.

Tamby, J.L. (1988). An integrated approach for the treatment of outliers in sub-annual economic surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 229-234.

# ICES II, INVITED SESSION ON OUTLIERS: DISCUSSION

**Beat Hulliger, Swiss Federal Statistical Office**
**Swiss Federal Statistical Office, Espace de l'Europe 10, CH-2010 Neuchâtel, Switzerland**
Beat.Hulliger@bfs.admin.ch

## 1 INTRODUCTION

The session on outliers in establishment surveys is of particular importance since outliers are a continued concern in practice. Two of the three papers in this session on outliers are built on practical experience with robust methods in the national statistical institutes of Canada and UK. The third reports on a comparative simulation study. Simulations are the most realistic way to evaluate different estimators under various outlier situations. Thus the session covers the topic very well and my thanks go to the authors and the organizer for a very stimulating session. I will comment on the papers in turn, then try to synthesize aspects of them and draw some conclusions.

## 2 PAPER BY FRANKLIN, BRODEUR AND THOMAS

The paper "Robust Multivariate Outlier Detection Using Mahalanobis' Distance and Stahel-Donoho Estimators" shows, to my knowledge, the first application of multivariate robust estimation in a business survey. One of the problems of classical multivariate robust methods is that they are developed for i.i.d. data. Therefore classical multivariate robust methods have to be adapted to the production system of large business surveys and in particular they must take into account sample designs, which often use very different sampling rates in different strata, and other corrections for imperfect data. The paper investigates the important step from unweighted to weighted methods.

The authors propose to work with a Mahalanobis distance which is based on a robust estimate of the covariance matrix. This approach makes sense only if the bulk of the data is symmetrically distributed. Thus there is an important pre-processing step which is discussed only shortly in Section 5.2: Many variables in business statistics are heavily skewed and must be transformed to a more symmetric distribution. This transformation to symmetry may be crucial for the success of the outlier detection. Since outliers may have a large influence on estimates of asymmetry it would be interesting to know more about the symmetrisation step. Also the choice of the psi-function and its tuning constants would deserve a discussion.

The authors propose to use the Stahel-Donoho estimators for location and covariance. They are weighted versions of the usual estimators. The weights depend on a measure of outlyingness. And it is exactly this measure of outlyingness where the main effort concentrates. The original Stahel-Donoho weights need a search through "all" possible directions which is usually done by sampling directions at random. While Stahel-Donoho propose to use random subsets of the observations to find a most extreme weight of an observation, the authors use a random search through orthogonal basis' proposed by Patak. Whether and how this search is more advantageous is not clear from the article. Stahel-Donoho propose to look for the worst direction for a point and then they downweight according to this direction. Patak proposes to look for the basis which after downweighting gives least weight to an observation.

One of the crucial points in the article is how to combine sample weights $w_i$ with robustness weights $\delta_i$. The multivariate situation is more complex by the addition of a further variable, the survey weight. The approach of creating new variables which are the products of the survey weight times the original or log-transformed variables $y_i$ does not yield a consistent estimator of the covariance. The weighted covariance and location estimators

$$\hat{\mathrm{V}}' = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} w_i\, \delta_{\mathrm{i}}^2\, (\mathrm{y_i} - \hat{\mathrm{u}}')(\mathrm{y_i} - \hat{\mathrm{u}}')^{\mathrm{T}}}{\sum_{\mathrm{i=1}}^{\mathrm{n}} w_i\, \delta_{\mathrm{i}}^2}, \qquad \hat{\mathrm{u}}' = \frac{\sum_{\mathrm{i}}^{\mathrm{n}} w_i\, \delta_{\mathrm{i}}\, \mathrm{y_i}}{\sum_{\mathrm{i}}^{\mathrm{n}} w_i\, \delta_{\mathrm{i}}}$$

are consistent alternatives to the ones proposed by the authors. They give less weight to the units with small inclusion probabilities if the weight is derived from the inverse of inclusion probabilities.

## 3 PAPER BY GWET AND LEE

The paper "An Evaluation of Outlier-Resistant Procedures in Establishment Surveys" sets out with explaining a range of possible estimators for the univariate case. It pays special attention to the fundamental problem of bias and how the different estimators tackle it. For example many estimators reduce the effect of the robustification when the sampling fraction becomes large. Lees estimators directly uses the relative weight of the bias in a MSE-estimate as a parameter. Hulligers Minimum Estimated Risk estimator estimates the mean squared error of a set of M-estimators and chooses the one with minimum MSE. Chambers estimator corrects an initial robust estimator by adding a less robust mean of residuals. The Hidiroglou-Srinath estimator minimizes a theoretical MSE-function. However, the minimum depends on robust estimates of certain population parameters.

The authors check various proposals for variance estimation. Closed form variance estimators rely on approximations which seem rather bad in some cases like for the one-step formula. It would have been interesting to know more about the reasons and conditions for this failure. Possible reasons might be that the usual variance estimator for M-estimators does not take into account the combined effect of asymmetry and preliminary scale estimation, which results in a bias (Carroll 1979). The problem becomes explicit with the complete influence function of the M-estimator $\theta$ with preliminary scale estimator $\delta$ under asymmetry (Huber 1981): $IF(x;\theta) = (\psi(r)\delta - IF(x;\delta)B)/A$, where $r$ is the standardised residual and $A$ and $B$ depend on the psi-function and the distribution. The corresponding asymptotic variance is

$$V(\theta) = E(IF(x;\theta)^2) = \frac{1}{A^2}\left(\delta^2 \int\psi(r)^2 df + B^2V(\delta) + \delta B\int\psi(r)IF(x;\delta)df\right).$$

Under assumption of symmetry only the first summand is non-zero. Approximate calculations show that for the Weibull distribution with parameter 2.87 and a tuning constant *c=3* the first summand contributes only 23% to the true variance. Thus the corresponding variance estimator largely underestimates the true variance. Furthermore the optimal tuning constant is about *c=100*. This is well beyond the region considered for tuning constants up to now. A further reason for the underestimation of the variance of the one-step estimator is that the variance of the initial estimate, the median, is not included in the usual formula either.

For the simulations I concentrate my comments on the case *n=30* because I think that robust estimators need samples of size 30 or maybe 20. Also the Jackknife variance estimator behaves much better with *n=30*.

The One-step M-estimator and the fully iterated M-estimator behave very similar. The one-step estimator is always a little bit less efficient. This might be due to the high variability of the median, which serves as the starting value. Both estimators use a tuning constant of *c=3* for all distributions. This tuning constant is too small for the lognormal distributions (rel. bias -0.26 and -0.36) and for the Weibull distributions (rel. bias -0.36 and -0.78).

The bias of the Hulliger estimator is usually small but the efficiency gain is only modest. A possible reason might be the underestimation of the variance due to a simultaneous scale estimation combined with a fixed choice of trial constants. Also, since the minimum of the MSE usually is attained nearby the mean, the set of tuning constants of the authors with largest *c=20* may not allow a good approximation to the minimum. Nevertheless the minimum search may be too conservative and choosing the minimal constant which yields smaller estimated MSE than the mean would be a better choice.

The Chambers estimator performs remarkably well. What is the reason? For a negligible sampling rate the estimator is the sum of a robust M-estimator *$\mu$(c)* with tuning constant *c=3* plus the average of the scores for an M-estimator with a tuning constant *c'=10*, the bias correction. In a symmetric situation the average score at *$\mu$(c)* is approximately *0* irrespective the tuning constant. In an asymmetric situation the bias correction is bounded by *$\mu$(c')-$\mu$(c)*. Therefore Chambers estimator is, roughly speaking, an M-estimator whose tuning constant is *c=3* if the underlying distribution is symmetrical and tends to *c'=10* if the underlying distribution is asymmetric. The clever trick is to use the average score for the correction. This avoids estimation of the variance.

Like the Hulliger estimator the Lee estimator seems to suffer from an underestimate of the variance which tends to inflate $\theta$ and correspondingly the bias correction. At *n=30* the Lee estimator looses the advantage it had at *n=10* and *20* over its close relative, the Chambers estimator (except for Weibull 2). It seems that the bias then becomes too large.

## 4 PAPER BY CHAMBERS, KOKIC, CRUDDAS AND SMITH

In the paper "Winsorization" for Identifying and Treating Outliers in Business Surveys" the authors mention influence of outlying weights $w_i$ and outlying variables $y_i$ on weighted sums. They opt for a value modification strategy, i.e. an outlying value $y_i$ should be replaced by $y_i$*, because outliers are tied to variables. Their Winsorization procedures do not consider outlying weights. One-sided Winsorization seems to works well if only one-sided outliers have to be considered. Also two-sided Winsorization introduces bias if the underlying distribution is skew. The advantage of type 2 Winsorization, i.e. the property that no Winsorization occurs if an observation has a weight of *1*, may be a disadvantage if also non-representative outliers have to be treated, as is the case in practice.

The choice of the cut-off value *K* is based on an estimate of the bias, which itself is based on an estimate of the "uncontaminated" mean of the observations. If this estimate is in order, then there is no need for a further step. Therefore, as the authors note, there is some circularity in the choice of *K*. Of course, if we can use more auxiliary information, e.g. from the past, we may enhance our estimation. But this is true for many other procedures, too. In practice the procedure comes down to a sort of one-step procedure: Take a robust estimator $\mu$. Then use the weighted residuals to decide on a good choice of a cut-off value *K*. Then use a winsorized mean for prediction.

The paper assumes first that any outliers in the sample data are representative. The distinction of representative and non-representative outliers, introduced in Chambers (1986), is useful at a theoretical level but should be made continuous in practice. The problem is that in the face of a possible outlier one often cannot tell whether it is representative or not. In the same way as plain rejection of outliers is usually not the most efficient robust procedure plain acceptance of representativity is usually too optimistic. The discussion on non-representative outliers in Section 3.3 of the paper goes in this direction. In fact it argues that not all observations may be considered as representative. The authors propose to write the Winsorized total as $\Sigma\ w_i\ o_i(K)\ y_i$ and to choose the cut-off value *K* such that a certain number of outliers are downweighted. This is very much in line with Hulliger (1999a), where in addition to the sample weights $w_i$ robustness weights $u_i(c)$, corresponding to $o_i(K)$, are introduced and where the amount of downweighting can be controlled by looking at the average robustness weight.

The considerations on the level of Winsorization are very important in practice. The authors propose to make the choice in the light of the importance of a partition for the analysis of the data. This is certainly sensible. An additional criterium should be that the population is sufficiently homogeneous to allow to speak of something like "the bulk" of the data and outliers.

The problem of Winsorization of derived variables is interesting since many business surveys use derived variables for analysis. It is similar to a problem of error localisation in editing. In practice one usually does not like to touch on benchmark information. But it may not be reasonable to leave it untouched if by changing a few variables (the benchmark) we can make many variables non-outlying. The authors propose some practical rules on how to proceed. However, the problem is properly multivariate and general solutions will need a proper multivariate treatment.

## 5 CONCLUSION

All the estimators considered in the papers, except the Hulliger estimator, depend on the choice of one or more tuning constants. Winsorizing a fixed number of observations works remarkably well. The reason might be that these estimators implicitely adapt to the asymmetry of the distribution and are consistent, i.e. with growing sample size the winsorisation is negligible. However, their breakdown point is low and tending to 0 necessarily, like for the explicitely adaptive estimators by Lee and Hulliger. Chambers estimator is a good way to choose between two M-estimators with too different tuning constants. However, if the optimal tuning constant is beyond the range considered it cannot cope with the asymmetry either.

The papers show a move to consideration of practical problems which I warmly welcome. More work on the practical problems of application of robust estimators must be done. On the other hand I think that there is still work open in understanding the (theoretical) behaviour of these estimators. Theoretical comparisons of these estimators are difficult. Sensitivity curves, even numerically calculated, of them could clarify their behaviour below breakdown. E.g. the sensitivity curve of the Hulliger estimator is linear but with a smaller slope than the mean.

The problem of outlying weights was only treated in the first paper. Chambers (1997) and Duchesne (1999) propose robustifications based on GREG estimators which modify outlying weights. Adhoc solutions in practice are Winsorization of weights and shrinking of the Lorenz-curve of weights towards identity (Hulliger 1999b).

Multivariate robust methods are just emerging in business surveys and need more attention in the future. However, the simplicity of estimators is a serious concern because too complicated procedures will not be applied in practice.

## 6 REFERENCES

Carroll, R.J. (1979), "On Estimating Variances of Robust Estimators When the Errors Are Asymmetric," Journal of the American Statistical Association, **74**, pp. 674-679.

Chambers, R.L. (1986), "Outlier Robust Finite Population Estimation," *Journal of American Statistical Association*, **81**, pp. 1163-1169.

Chambers, R.L. (1997), "Weighting and Calibration in Sample Survey Estimation," Conference on Statistical Science Honouring the Bicentennial of Stefano Franscini's Birth, Birkhäuser Verlag.

Duchesne, P. (1999), "Robust Calibration Estimators," Survey Methodology, **25**, pp. 43-56.

Huber, P.J. (1981), "Robust Statistics", Wiley.

Hulliger, B. (1999a), "Simple and Robust Estimators for Sampling," *Proceedings of the Survey Research Methods Section of the American Statistical Association* (to appear).

Hulliger, B. (1999b), "A proposal for treatment of extreme weights," Technical Note, SFSO.