Journal of Official Statistics
Vol. 10. No. 4, 1994, pp. 419–435

Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator

P.N. Kokic[1] *and P.A. Bell*[2]

Abstract: Within stratum expansion estimation is a popular method of estimating totals in a stratified finite population. However, if by chance several unusually large observations should fall in the sample, then the expansion estimator may grossly overestimate population totals. One technique to deal with this problem is to reduce sampled observations greater than a cutoff to a value closer to that cutoff, and then estimate the total using the new adjusted values. The resulting estimator is called the Winsorized estimator of a total. Although the Winsorized estimator is biased, it may have considerably smaller mean squared error than the expansion estimator. Necessarily, different Winsorizing cutoffs should be used for different strata. In this paper we examine the problem of estimating the optimum set of Winsorizing cutoffs for repeated surveys, where the cutoffs will be used for Winsorizing samples in future repeats of the survey. It is shown that an approximation to the optimum set of Winsorizing cutoffs may be obtained by expressing the cutoffs for all strata in terms of a single parameter L, and then searching for that value of L where a particular function equals zero. The function is a weighted sum of the tail probabilities and tail means above the Winsorizing cutoffs. It is found that by using simple estimates of these quantities, a good estimator of the optimum L value and hence of the optimum set of Winsorizing cutoffs may be constructed. In a computer simulation study the estimator is found to have considerably smaller mean squared error than the expansion estimator.

Key words: Finite population; expansion estimator; Winsorized estimator; simple random sampling.

1. Introduction

Consider a stratified finite population such that in stratum h there are N_h values of a variable $X_{hi} \in R$ taking the values X_{h1}, $X_{h2}, \ldots, X_h N_h$, where $h = 1, \ldots, H$. We consider only non-negative variables. To estimate the population total

$$T = \sum_{h=1}^{H} \sum_{i=1}^{N_h} X_{hi}$$

we select without replacement from each stratum h, a simple random sample, s_h, of size n_h, from the set $\{1, 2, \ldots, N_h\}$ and construct the Winsorized estimator of the

[1] Australian Bureau of Agricultural and Resource Economics, G.P.O. Box 1563, Canberra, A.C.T., 2601, Australia.
[2] Australian Bureau of Statistics, P.O. Box 10, Belconnen, A.C.T., 2616, Australia.
Acknowledgements: The authors wish to thank Edward Szoldra and Lyndall Adam for their computer programming assistance. Views expressed in this paper are those of the authors and do not necessarily reflect those of the Australian Bureau of Statistics or the Australian Bureau of Agricultural and Resource Economics. Where quoted or used they should be attributed to the authors.

total

$$\hat{T} = \sum_{h=1}^{H} (N_h/n_h) \sum_{i \in s_h} X_{hi}(K_h)$$

where K_h are non-negative numbers,

$$X_{hi}(K_h) = X_{hi}, \quad \text{if } X_{hi} < K_h,$$
$$= f_h X_{hi} + (1 - f_h)K_h, \quad \text{otherwise,}$$

and $0 \leqslant f_h \leqslant 1$. The values K_h are called the Winsorizing cutoffs. Following Gross, Bode, Taylor, and Lloyd-Smith (1986), when $f_h = 0$, $\hat{T}$ is called the Winsorized Type I estimator, and when $f_h = n_h/N_h$ it is called the Winsorized Type II estimator.

If for all h, $K_h = \infty$, then $\hat{T}$ reduces to the simple expansion estimator in a stratified population

$$\sum_{h=1}^{H} (N_h/n_h) \sum_{i \in s_h} X_{hi}.$$

This estimator is unbiased under the simple random sampling scheme mentioned above. In a finite population containing a few unusually large values of X_{hi} which we shall call outliers, there is a chance that the sample contains some of these outliers, in which case the expansion estimator may grossly overestimate the true value T.

The Winsorized estimator, $\hat{T}$, reduces sampled units greater than the cutoffs to a value between the cutoff and their original value so that their effect is not so great. We know that $\hat{T}$ is a biased estimator of the total, but at the same time it has significantly less variance than the expansion estimator. The aim in the present paper is to show how to estimate the cutoffs $\{K_h\}$ reliably and simply from survey data so as to minimise mean squared error of $\hat{T}$.

The Winsorized estimator is robust in the sense that outlying observations do not inflate its mean squared error as much as that of the expansion estimator. Winsorizing has the advantage over other robust techniques in that it is simple to perform in practice and straightforward to explain to users of statistical results, both important considerations when running large scale surveys. We examine the problem of estimating the optimum set of cutoffs for repeated surveys, where the cutoffs will be used for Winsorizing samples in future repeats of the survey.

Ernst (1980) showed that for simple random sampling with replacement from a continuous distribution, a Winsorizing cutoff may be chosen so that the Winsorized Type I estimator has less mean squared error than any other estimator that reduces sampled values. However, the Winsorized Type II estimator has several attractive practical properties. In particular, as n_h approaches N_h the contribution of an outlier to the Type II Winsorized estimator tends to the value of the outlier, whereas for Type I it tends to the cutoff K_h. Furthermore, Gross et al. (1986) has demonstrated that there is little difference in practice between the mean squared errors of both estimators. Hence, only Winsorized Type II estimators will be considered in this paper, and so from now on we set $f_h = n_h/N_h$.

Dalén (1987) and Tambay (1988) have also investigated properties of the Winsorized Type II estimator. Tambay set the cutoffs of the Winsorized estimator using a quartile distance method and found that in a number of practical situations it sometimes, although not always, had smaller mean squared error than the simple expansion estimator. He chose these cutoff values on the basis that they were robust against large outliers in the sample. We argue that the tail of the distribution must, to some degree, be taken into account when choosing the cutoffs.

It is possible to show that the Winsorized Type II estimator is a special case of

Chambers (1986) robust finite population estimator. Chambers, however, concentrates on the issue of how to achieve robustness in a finite population setting when auxiliary information is taken into account. The relationship between Chamber's estimator and ours provides a method of generalising the Winsorized estimator to the case when auxiliary size information is available for all units in the finite population. This generalisation will not be examined in this paper.

Other related works that have considered Winsorizing in the context of finite population sampling include Searls (1966), and Hidiroglou and Srinath (1981). The approach in these papers was to fix the Winsorizing cutoff prior to sampling. In repeated surveys this approach is attractive, as previous samples may be used to determine the cutoffs for Winsorizing in future repeats of the survey, but the problem still remains of how best to estimate the cutoffs. In the present article we address this problem for a stratified random sample in which different cutoffs are used for different strata.

Exact optimal Winsorizing cutoffs can be chosen if the underlying distribution for each stratum is known. The algorithm to find the exact cutoffs requires the evaluation of the tail probability and tail mean above a given value for each stratum. In the case where there is only one stratum and a large sample, a distribution can be accurately fitted and hence the optimum cutoff can be accurately estimated.

However, in most surveys there are many strata, a large number often with sparse sample. We are usually trying to select a large number of cutoffs simultaneously while controlling bias at a broad stratum group level. Consequently, we cannot be confident about the distribution we fit to each stratum, and hence we lack confidence about our estimate of the optimum set of cutoffs.

The new approach developed in Section 2 is to represent the H cutoffs $\{K_h\}$ by a single parameter L, thus reducing the dimension of the problem. A simple method of estimating L from a stratified random sample is developed. In Section 3 we investigate, in a computer simulation study, how well our estimator performs.

2. Theoretical Developments

2.1. *Model and notation*

Assume that for each h, $\{X_{hi}, i = 1, \ldots, N_h\}$ is a sequence of uncorrelated and identically distributed random variables with

$$E_m(X_{hi}) = \mu_h \tag{1}$$

and

$$Var_m(X_{hi}) = \sigma_h^2 < \infty$$

and assume that X_{hi} possesses a density $g_h(x) > 0$ for all $x > 0$. Furthermore, assume that $\{X_{hi}; i = 1, \ldots, N_h, h = 1, \ldots, H\}$ is statistically independent of the sample selection process. Here and throughout E_m denotes expectation over the model at (1) and E_d denotes the design expectation, that is, over the sample selection process. If no subscript is used then the expectation is firstly over the design and then the model.

Our aim is to choose cutoffs $\{K_h\}$ so that the Winsorized estimator $\hat{T}$ has minimum total mean squared error, $E\{(\hat{T} - T)^2\}$. The total mean squared error not only captures the variability of $\hat{T}$ due to the modelling assumptions given above, but also the variability resulting from the stratified simple random sampling scheme. That is, $E = E_m E_d$.

Let

$$J_{hi} = 1, \text{ if } X_{hi} \geqslant K_h$$
$$= 0, \quad \text{otherwise.}$$

From now on, unless otherwise indicated,

let $\sum_h$ denote the sum over $h = 1, \ldots, H$, $\sum_i$ the sum over $i = 1, \ldots, N_h$, and to simplify notation the second subscript, i, has been deleted from the variables X and J, except where it is needed for reasons of clarity. Also let $n = \sum_h n_h$ be the total sample size.

2.2. *Main results*

Under the assumptions made in the previous section we have the following result.

Theorem 1. *The minimum of the total mean squared error* $E\{(\hat{T} - T)^2\}$ *occurs for those values of* K_h *satisfying*

$$(N_h/n_h - 1)(K_h - \mu_h) + B = (N_h/n_h - 1)\{K_h E_m(J_h) - E_m(X_h J_h)\} \quad (2)$$

where

$$B = \sum_h N_h(1 - f_h) \times \{K_h E_m(J_h) - E_m(X_h J_h)\} \quad (3)$$

is the bias. Furthermore, at the minimum

$$B = -(n-1)^{-1} \sum_h N_h(1 - f_h)(K_h - \mu_h).$$

The proof of this theorem is given in the Appendix.

Notice that equation (2) depends on the tail probabilities and tail means in each stratum. Any algorithm to find the optimum set of cutoffs $\{K_h\}$ would therefore involve calculating, or estimating, these quantities. The theorem below leads us to a simple way of estimating $\{K_h\}$.

Theorem 2. *Suppose that for all* h, $n_h > 1$ *and* $n_h/N_h < 1$. *Then*

$$0 \leqslant 1 + (N_h/n_h - 1)(K_h - \mu_h)/B \leqslant n_h^{-1}. \quad (4)$$

Furthermore, suppose that $N, n \to \infty$ *in such a way that for all* h, $n_h > 1$ *and* $\varepsilon < n_h/N_h < 1 - \varepsilon$, *where* $0 < \varepsilon < 1/2$. *Then as* $N, n \to \infty$

$$(N_h/n_h - 1)(K_h - \mu_h) \sim -B \to \infty. \quad (5)$$

It should be noted that in the statement of this theorem the total number of strata is assumed to remain fixed as $N, n \to \infty$, although it may be large as is often the case in practice. The proof of this theorem is also given in the Appendix.

According to expressions (4) and (5), for the optimum set of cutoffs $\{K_h\}$, $(N_h/n_h - 1)(K_h - \mu_h)$ will be approximately constant for large n. Note that the distributions of data in individual strata are important in determining the Winsorizing cutoffs. According to (5) if there are large outliers in a particular stratum, then that stratum will affect the values of the optimum cutoffs through the bias term B.

Let $L = (N_h/n_h - 1)(K_h - \mu_h)$. Since equation (5) holds at the optimum set of cutoffs $\{K_h\}$, an approximate solution to (2) would be obtained by setting $L = L_0 = -B$ and

$$K_h = (N_h/n_h - 1)^{-1} L + \mu_h. \quad (6)$$

From (3) it follows that

$$L = \sum_h N_h(1 - f_h) \times \{E_m(X_h J_h) - K_h E_m(J_h)\}. \quad (7)$$

In Section 3 below we show in a particular situation that this approximation provides an answer almost as good as the optimum.

The approach outlined at (6) and (7) is attractive from several points of view. Firstly, we have reduced the problem of finding the optimum set of cutoffs $\{K_h\}$ from a search in H-dimensional space to a 1-dimensional search. Furthermore, we do not need accurate estimates of the tail means and probabilities in each stratum, but rather of a linear combination of them

across all strata as described by equation (7). We expect that this quantity can be more accurately estimated than the individual tail means and probabilities. In the following subsection we introduce a simple method of estimating the quantity L_0 from sample data.

2.3. *Estimation*

Let $X^*_{hi} = (X_{hi} - \mu_h)(N_h/n_h - 1)$, and $J^*_{hi} = 1$ if $X^*_{hi} \geqslant L$ and $J^*_{hi} = 0$ otherwise. Notice from (6) that $J^*_{hi} = J_{hi}$. From equations (6) and (7), to estimate L_0 we must find the value of L such that $F(L) = 0$, where

$$\begin{aligned} F(L) &= L - \sum_h N_h(1 - f_h) \\ &\quad \times \{(N_h/n_h - 1)^{-1} E_m(X^*_h J^*_h) \\ &\quad - (K_h - \mu_h) E_m(J^*_h)\} \\ &= L\left\{1 + \sum_h n_h E_m(J^*_h)\right\} \\ &\quad - \sum_h n_h E_m(X^*_h J^*_h). \end{aligned} \tag{8}$$

Notice that in expression (8), $E_m(J^*_h)$ and $E_m(X^*_h J^*_h)$ are, respectively, the tail probabilities and tail means in stratum h, and μ_h in the formula for $\{X^*_{hi}\}$ is the stratum mean. It would be natural to estimate these unknown quantities by their sample counterparts. Note that we focus on the problem of estimating the Winsorizing cutoffs for use in future repeats of the same survey. Therefore, the random variables $\{X_{hi}\}$ in the above formulation refer to the future time point when the sample is to be drawn and the cutoffs defined by (8) are approximately optimum for that sample. Let $\{y_{hi} : i = 1, \ldots, m_h, h = 1, \ldots, H\}$ be a sample taken from a previous run of the survey. Assume that these values were generated by the same random process that will determine the future sample values as described in Section 2.1. Then we obtain the following estimator of the function $F(L)$ based on information collected in a previous run of the survey

$$\begin{aligned} f(L) = L\Bigg\{1 + \sum_h (n_h/m_h) \\ \times \sum_i{}^* I(y^*_{hi} \geqslant L)\Bigg\} - \sum_h (n_h/m_h) \\ \times \sum_i{}^* y^*_{hi} I(y^*_{hi} \geqslant L) \end{aligned}$$

where $\sum^*$ is the sum over $i = 1, \ldots, m_h$, I is the indicator function

$$y^*_{hi} = (N_h/n_h - 1)(y_{hi} - \bar{y}_h) \tag{9}$$

and $\bar{y}_h = m_h^{-1} \sum_{i \in s_h} y_{hi}$. Our proposed technique for estimating L_0 is to search for the value of L such that $f(L) = 0$.

Let $y^*_{(1)} \geqslant y^*_{(2)} \geqslant \cdots \geqslant y^*_{(n)}$ be the values of $\{y^*_{hi}\}$ sorted in descending order. Notice that $f(y^*_{(1)}) = y^*_{(1)}$, and $f(y^*_{(n)})$ is negative. Furthermore, the function f is piecewise linear, continuous and decreasing. Therefore, one way to estimate L_0 is to sequentially determine the values $f(y^*_{(j)})$, $j = 1, 2, \ldots$ and stop the first time $f(y^*_{(j)})$ is negative. The estimated value of L_0 is then obtained by linear interpolation between $y^*_{(j-1)}$ and $y^*_{(j)}$. That is, estimate L_0 by evaluating

$$\begin{aligned} \hat{L} = \{y^*_{(j)} f(y^*_{(j-1)}) - y^*_{(j-1)} f(y^*_{(j)})\}/ \\ \{f(y^*_{(j-1)}) - f(y^*_{(j)})\} \end{aligned}$$

and set $\hat{K}_h = (N_h/n_h - 1)^{-1}\hat{L} + \bar{y}_h$.

As can be seen from the process above, estimating L_0 is particularly simple. The values in each stratum are relocated by subtracting the stratum mean and then scaled by multiplying by the stratum weight minus one. We then search for the point where $f(L) = 0$. This would typically involve only the few largest y^*_{hi} values.

In the following section, in a computer simulation study, we evaluate how well performs our method of estimating the optimum set of cutoffs.

3. Evaluation

Our primary interest in the estimation scheme introduced in Section 2 lies in how it may be applied to repeated surveys. In particular, data from one survey are used to estimate the optimum Winsorizing cutoffs that are to be used in subsequent collections of the same survey. In our analysis it is assumed that superpopulation models governing the distribution of the population units in a future repeat of the survey and for the sample used to estimate the cutoffs are identical, and statistically independent. Typically both these assumptions are made in practice when using data collected in a previous repeat of the survey to design future collections of the same survey.

To evaluate the approach of Section 2 we simulate data using a known model of stratum distributions based on a real survey. For repeated samples from the modelled population, the estimates $\hat{K}_h$ are obtained and the root mean squared error (RMSE) and standard error (SE) of the corresponding Winsorized estimator based on an independent sample in a future repeat of the survey is calculated. All RMSEs in this analysis were calculated from formulas and hence are exact. The formulas for the RMSE were obtained from expressions (A.1), (A.2) and those just below (A.2) in the Appendix. The exact RMSEs were evaluated by substituting $\hat{K}_h$ into these expressions. The performance of the estimator based on $\{K_h\}$ is gauged by comparison of these RMSEs with the standard error of the simple expansion estimator.

Data from the quarterly survey of stocks run by the Australian Bureau of Statistics (ABS) were used. This survey aims to measure the value of all stocks of materials, work in progress and finished goods owned by private sector businesses in Australia, see ABS (1990).

The population consists of about 463,000 businesses obtained from a list maintained by ABS. The survey is stratified by industry and each industry is stratified by up to five employment size strata, the top size strata being completely enumerated. The sample size is 8,500. Since there are no Winsorizing cutoffs in completely enumerated strata, these strata have been excluded from any further analysis.

The industries were grouped together into four broad industry groups: Mining, Manufacturing, Construction and Other. We selected optimum cutoffs to minimise mean squared error of the four broad industry level estimates and another set of cutoffs for the all industries level estimate. The population was simulated from the following model using a program written in SAS.

For simplicity, we shall assume that $n_h = m_h$ for all h and also that the sizes of the corresponding populations are the same in all strata. Let $\{Y_{hi}\}$ denote the population values from which the sample used to estimate the cutoffs was drawn. As noted before, assume that $\{Y_{hi}\}$ and $\{X_{hi}\}$ are independent and identically distributed. Also assume that $\{X_{hi}\}$ are independent and

$$X_{hi} = I_{hi} G_{hi} V_{hi} + (1 - G_{hi}) Z_{hi} \qquad (10)$$

where

$$P(I_{hi} = 1) = p_h, \quad P(I_{hi} = 0) = 1 - p_h,$$
$$0 \leqslant p_h \leqslant 1$$
$$P(G_{hi} = 1) = 1 - \varepsilon, \quad P(G_{hi} = 0) = \varepsilon,$$
$$0 \leqslant \varepsilon \leqslant 1$$

$\log(V_{hi})$ is distributed as a normal random

variable with mean ν_h and standard error τ_h and $\log(Z_{hi})$ is a normal random variable independent of V_{hi}. That is, the stratum distributions g_h are mixtures of two lognormal distributions with a point mass at zero. The variable X_{hi} represents the level of stocks held at a business. Frequently, businesses do not hold any stocks at all, and hence the value of p_h in some strata was quite small. The stratum level distributions of the positive stock values were found to be well approximated by a lognormal distribution. Under these circumstances the mixture model at (10) fitted the data quite well. The parameters p_h, ν_h and τ_h were estimated for each stratum using 11 quarters of survey data. Their estimates are given in Table 1. The variable Z_{hi} in the model is a gross error term introduced to analyse the performance of the Winsorized estimator in the presence of unusually large observations. For this study, the mean of $\log(Z_{hi})$ was set to 3 units larger than the mean of $\log(V_{hi})$, ν_h. This implies that the gross error term for stratum h has approximately 20 times the mean of the variable V_{hi}. For the simulation study the values $\varepsilon = 0$, 0.01 were chosen to cover most situations of practical interest. When $\varepsilon = 0.01$ very large observations will occur infrequently in the population. In practice it is very difficult to accurately estimate totals from samples drawn from this type of population. Notice that even when $\varepsilon = 0$, since the lognormal distribution can be highly skewed, there was still a high probability of moderately large observations occurring in the sample of some strata.

To evaluate the performance of the estimator of the cutoffs $\{K_h\}$ a computer simulation study was performed. To produce a single value of the set of cutoffs $\{\hat{K}_h\}$, a complete sample of data was generated using the model at (10). Using the simulated data, the $\hat{L}$ values and hence the cutoffs $\{\hat{K}_h\}$ were then calculated using the algorithm in Section 2.3.

For each of the industry level estimates, 1,000 independent replicates were produced. Simulated values of $\{\hat{K}_h\}$ corresponding to the 5th, 25th, 50th, 75th and 95th percentiles of the $\hat{L}$ values were obtained. For each of these sets of $\hat{K}_h$ values, the total mean squared errors of the corresponding Winsorized estimator based on the independent sample $\{X_{hi}, i \in s_h, h = 1, \ldots, H\}$ were calculated using an exact formula derived for the model (10). Tables 2 and 3 list the percentiles of the simulated values of $\hat{L}$, and the exact RMSEs, SEs and biases under model (10) of the corresponding Winsorized estimators, relative to its expected value under the model.

From the results in Tables 2 and 3 we see that in most cases the Winsorized estimator has considerably less RMSE than the expansion estimator. In nearly all cases the optimum value, L_0, was close to the 75th percentile of the simulated values of $\hat{L}$ and at the optimum the mean squared error of the Winsorized estimator was always considerably less than the variance of the expansion estimator.

When there was no gross error term in the model, that is when $\varepsilon = 0$, the cutoffs were estimated with a reasonable degree of accuracy, see Table 2. In fact, between the 25th and 95th percentiles of $\hat{L}$, the resulting estimators were only slightly worse than the optimum. In the case of Mining this range of $\hat{L}$ values gave a Winsorized estimator with RMSE within 14% of the optimum, for Manufacturing this range gave an estimator with RMSE within 3% of the optimum and for Other within 2% of the optimum. The effects of Winsorizing in the Mining sector seem to have been most dramatic probably due to the very long tails of the stratum distributions. In this

Table 1. Values of N_h, n_h, p_h, ν_h and τ_h by stratum. The first two digits of h represent industry and the final digit indicates the size of the stratum within industry

h	N_h	n_h	p_h	ν_h	τ_h	h	N_h	n_h	p_h	ν_h	τ_h
Mining											
011	360	6	0.17	7.8	0.34	022	175	6	0.47	12.0	0.32
012	29	6	0.47	12.0	1.00	023	71	18	0.64	13.0	0.86
013	20	8	0.87	14.2	0.96	024	28	6	0.55	13.8	0.36
021	2123	55	0.26	10.3	1.44						
Manufacturing											
031	2505	6	0.66	7.5	1.05	142	189	6	0.94	11.4	0.96
032	410	6	0.75	9.7	0.97	143	55	6	0.65	13.4	0.79
033	175	31	0.81	12.6	1.03	144	18	6	0.98	14.1	0.64
034	62	29	0.75	13.2	0.88	151	369	6	0.41	9.6	0.87
041	111	6	0.75	9.7	1.07	152	54	6	0.89	11.6	0.99
042	28	6	1.00	13.0	1.02	153	23	7	0.58	14.4	0.32
043	16	10	0.75	13.8	1.12	154	9	6	0.83	13.3	0.80
051	596	6	0.67	9.2	0.76	161	195	6	0.63	9.9	0.54
052	96	6	0.98	11.8	0.54	162	42	6	0.98	11.0	0.92
053	47	19	0.92	13.3	0.65	163	11	6	0.65	13.8	0.32
054	20	13	0.79	14.5	0.59	164	8	6	0.78	13.3	1.22
061	346	6	0.67	9.7	0.75	171	6815	428	0.68	9.4	1.11
062	56	9	0.85	12.5	0.72	172	899	132	0.83	11.4	0.95
063	21	19	0.77	14.9	0.72	173	308	99	0.81	12.8	0.87
064	11	10	0.94	14.5	0.79	174	59	20	0.72	14.1	0.57
071	815	6	0.65	9.8	0.81	181	1179	6	0.59	9.4	0.92
072	137	10	0.85	12.3	0.81	182	178	7	0.74	11.4	0.67
073	56	6	0.82	13.0	0.63	183	69	18	0.93	13.3	1.01
074	26	9	0.99	13.8	0.82	184	21	14	0.74	15.1	0.40
081	2954	8	0.56	8.6	0.78	191	1055	47	0.51	10.2	1.45
082	569	61	0.78	12.0	1.06	192	66	6	0.83	11.5	0.60
083	235	58	0.82	12.9	0.87	193	32	6	0.83	13.3	0.67
084	76	14	0.78	13.5	0.84	194	10	7	0.86	13.5	1.05
091	8192	250	0.71	9.5	1.08	201	824	6	0.70	8.1	0.78
092	868	105	0.83	11.6	0.77	202	76	6	1.00	12.1	1.28
093	288	45	0.93	12.8	1.06	203	24	6	0.72	11.5	1.04
094	53	12	0.86	14.5	0.70	211	1512	31	0.71	9.9	0.96
101	154	6	0.65	11.0	1.27	212	244	15	0.79	12.0	0.77
102	47	6	0.82	12.3	0.89	213	120	28	0.88	13.2	0.83
103	20	6	0.95	13.9	0.65	214	37	16	0.90	14.5	0.64
111	4218	110	0.68	8.4	0.93	221	2862	66	0.56	10.2	1.07
112	581	39	0.84	10.4	0.87	222	504	15	0.86	11.0	1.08
113	239	67	0.69	11.5	0.78	223	182	110	0.90	13.2	1.09
121	755	11	0.65	10.7	0.74	224	34	19	0.86	14.6	0.67
122	129	11	1.00	12.4	0.81	231	4528	186	0.71	9.1	1.54
123	75	18	0.91	13.9	0.60	232	457	66	0.88	12.0	0.91
131	52	6	0.51	10.0	1.07	233	213	95	0.81	13.0	1.18
141	1603	8	0.71	9.2	0.82	234	45	22	0.94	13.6	0.89
Construction											
241	20106	682	0.31	10.5	1.33	262	242	24	0.65	12.0	1.34
242	199	34	0.62	12.1	1.15	263	94	28	0.72	13.7	1.17
243	42	6	0.51	13.4	1.01	264	23	6	0.80	13.5	0.97

Table 1. Continued

h	N_h	n_h	p_h	ν_h	τ_h	h	N_h	n_h	p_h	ν_h	τ_h
244	9	8	0.59	15.9	0.62	271	4459	35	0.26	9.3	1.22
251	1509	58	0.23	11.4	0.95	272	182	6	0.46	10.3	0.89
252	41	9	0.73	12.9	1.29	273	84	26	0.65	12.7	1.26
253	8	6	0.44	11.5	1.10	274	15	6	0.54	13.2	0.78
261	2420	103	0.34	10.6	1.25						
Other											
281	19	6	0.63	9.5	1.15	372	1331	12	0.80	10.1	1.31
282	7	6	0.75	13.0	0.66	373	296	9	0.83	11.2	0.92
291	1577	29	0.54	10.3	1.12	374	52	6	1.00	12.1	1.11
292	93	6	0.78	12.1	0.83	381	16915	430	0.66	10.4	1.22
293	18	6	0.42	12.4	0.90	382	1049	194	0.84	12.7	0.93
301	1478	97	0.61	10.7	1.31	383	395	162	0.88	13.8	0.83
302	116	29	0.82	13.3	0.97	384	72	53	0.92	14.8	0.63
303	28	10	0.75	14.1	0.98	391	5854	172	0.69	10.9	1.09
311	3500	122	0.54	10.6	1.13	392	232	31	0.86	12.2	0.92
321	8326	647	0.46	10.7	1.50	393	42	28	0.86	13.4	1.32
322	660	179	0.72	12.6	1.16	401	9229	78	0.75	10.0	0.85
323	176	134	0.71	13.7	1.21	402	626	6	0.86	9.3	0.88
341	7281	235	0.57	10.5	1.24	403	41	6	0.77	10.0	1.56
342	481	109	0.81	12.9	1.04	411	7518	51	0.34	8.2	0.71
343	118	49	0.80	14.1	0.83	412	345	8	0.17	8.4	1.40
351	21725	1992	0.61	10.9	1.37	421	46688	1271	0.69	10.3	1.07
352	1792	476	0.79	12.5	1.31	422	1295	80	0.73	11.2	1.03
353	516	345	0.82	13.9	1.09	423	155	58	0.78	12.9	1.30
361	22057	398	0.68	10.6	0.84	424	20	13	0.74	14.2	0.79
362	639	46	0.87	12.4	0.72	431	27550	54	0.67	8.3	0.91
363	157	25	0.76	13.2	0.79	432	3630	72	0.78	10.2	0.80
364	33	12	0.80	14.6	0.42	433	1061	15	0.78	10.4	0.91
371	53720	291	0.64	8.8	1.19	434	158	6	0.88	11.4	0.59

case the possible gains from Winsorizing were much greater. The amount of bias induced by Winsorizing is for most cases virtually insignificant except in the case of Mining for the reasons described above. There was only a small chance of significantly underestimating the cutoffs and actually obtaining a Winsorized estimator with RMSE larger than the expansion estimator.

When ε was set to 0.01, see Table 3, the performance of the Winsorized estimator was considerably better than the simple expansion estimator. For most values of the estimated cutoffs, the Winsorized estimator significantly outperformed the expansion estimator. In fact, between the 5th and 95th percentiles of $\hat{L}$ the Winsorized estimator nearly always had considerably smaller RMSE than the expansion estimator. Furthermore, the ratio of the relative RMSEs of the Winsorized estimator at the 50th percentile of $\hat{L}$ to the RMSE of the expansion estimator was considerably less when $\varepsilon = 0.01$ than when $\varepsilon = 0$. In other words, when the population is contaminated with a small proportion of very large outlying observations, there is a very high chance that the estimated cutoffs will result in a Winsorized estimator with considerably smaller RMSE than the expansion estimator. As demonstrated in Table 2, even when there are not many out-

Table 2. Exact RMSEs, SEs and biases relative to expected total of the Winsorized estimators corresponding to percentiles of the simulated $\hat{L}$ values, $\varepsilon = 0$

Percentile	$\hat{L}$ ($\div 10^6$)	Relative RMSE %	Relative SE %	Relative bias %
Mining				
5	3.3	25.4	12.0	–22.4
25	5.2	19.9	13.4	–14.7
50	7.6	18.3	14.5	–11.2
75	12.2	17.8	16.4	–6.9
95	29.0	19.5	19.3	–2.9
Expansion estimator	∞	26.0	26.0	0.0
Manufacturing				
5	12.6	3.3	2.7	–1.9
25	16.0	3.0	2.7	–1.3
50	19.2	3.0	2.8	–1.0
75	24.4	2.9	2.8	–0.6
95	39.8	3.0	3.0	–0.2
Expansion estimator	∞	3.2	3.2	0.0
Construction				
5	15.8	10.9	7.3	–8.1
25	20.6	9.6	7.6	–5.8
50	27.1	9.1	8.2	–4.0
75	36.5	9.1	8.7	–2.5
95	65.7	9.5	9.4	–1.0
Expansion estimator	∞	10.5	10.5	0.0
Other				
5	36.1	2.2	2.0	–0.9
25	40.9	2.1	2.0	–0.7
50	48.4	2.1	2.0	–0.5
75	65.0	2.1	2.1	–0.3
95	122.0	2.2	2.2	–0.1
Expansion estimator	∞	2.2	2.2	0.0

lying observations in the population there is still a high probability that the estimated cutoffs will be reasonably good.

Notice also that the percentile values of $\hat{L}$ and hence of the cutoffs are greater when $\varepsilon = 0.01$. This illustrates the fact that the optimum Winsorizing cutoffs depend to some extent on the size of the outlying observations.

There is a very wide range of $\hat{L}$ values which produce a Winsorized estimator with smaller mean squared error than the expansion estimator. Thus, even though $\hat{L}$ depends on the largest values from each stratum, and so is estimated with a high degree of variability, the flatness of the optimum ensures that with high probability the resulting Winsorized estimator has smaller mean squared error than the expansion estimator.

The RMSE of the Winsorized estimator was also minimised over all possible values

Table 3. Exact RMSEs, SEs and biases relative to expected total of the Winsorized estimators corresponding to percentiles of the simulated $\hat{L}$ values, $\varepsilon = 0.01$

Percentile	$\hat{L}$ ($\div 10^6$)	Relative RMSE %	Relative SE %	Relative bias %
Mining				
5	4.1	42.4	19.6	–37.6
25	7.3	37.5	20.2	–31.6
50	14.7	33.3	21.8	–25.2
75	33.3	30.9	25.1	–18.0
95	113.0	36.1	35.3	–7.5
Expansion estimator	∞	70.0	70.0	0.0
Manufacturing				
5	33.7	7.6	4.5	–6.1
25	50.6	6.5	4.8	–4.4
50	70.0	6.1	5.2	–3.2
75	98.2	5.9	5.5	–2.2
95	190.9	6.3	6.2	–0.9
Expansion estimator	∞	7.4	7.4	0.0
Construction				
5	33.2	22.5	9.5	–20.4
25	57.1	18.5	11.0	–14.9
50	89.5	16.6	12.5	–10.9
75	133.3	16.0	13.9	–7.9
95	280.9	17.5	17.1	–3.6
Expansion estimator	∞	24.6	24.6	0.0
Other				
5	103.7	5.1	3.3	–3.9
25	139.6	4.9	3.4	–3.5
50	183.3	4.3	3.8	–2.0
75	254.5	4.2	4.0	–1.3
95	472.6	4.4	4.4	–0.5
Expansion estimator	∞	5.1	5.1	0.0

of K_h by solving (2) exactly. It was found that the RMSE of this overall "best" possible Winsorized estimator was only slightly smaller than the minimum RMSEs achieved by the Winsorized estimators in Tables 2 and 3. This result indicates that the degree of error associated with using the approximate formula for the optimum K_h is relatively minor and can usually be ignored.

It is clear from the analysis above that the Winsorized estimator nearly always outperforms the expansion estimator in terms of mean squared error. It is also useful to examine the risk, or frequency, with which unusually large errors are made by each estimator. To perform this analysis, for each of the industries 1000 independent replicates of the absolute relative error, $|\hat{T}/T - 1|$, of both the Winsorized and expansion estimators were produced. To be consistent with the previous analysis, the sample used

Table 4. Estimated percentiles of the absolute relative errors of the Winsorized and expansion estimators. Based on 1000 independent simulations

Percentile	$\varepsilon = 0$		$\varepsilon = 0.01$	
	Expansion estimator	Winsorized estimator	Expansion estimator	Winsorized estimator
Mining				
10	0.028	0.029	0.056	0.067
25	0.072	0.067	0.129	0.150
50	0.139	0.139	0.259	0.266
75	0.242	0.225	0.400	0.374
90	0.342	0.313	0.597	0.473
Manufacturing				
10	0.004	0.003	0.010	0.011
25	0.010	0.009	0.023	0.023
50	0.021	0.020	0.047	0.047
75	0.036	0.035	0.079	0.076
90	0.053	0.051	0.113	0.104
Construction				
10	0.012	0.012	0.021	0.033
25	0.032	0.032	0.063	0.078
50	0.066	0.066	0.134	0.139
75	0.117	0.114	0.218	0.213
90	0.164	0.154	0.305	0.279
Other				
10	0.002	0.003	0.006	0.007
25	0.007	0.007	0.015	0.017
50	0.015	0.015	0.031	0.031
75	0.025	0.024	0.055	0.052
90	0.035	0.035	0.079	0.076

to calculate the estimates of the total was drawn independently of the sample used to calculate the Winsorizing cutoffs. Various quantiles of these absolute relative errors, as given in Table 4, provide a useful comparison of the risk associated with the estimators.

In general the Winsorized estimator nearly always outperforms the expansion estimator above the 50th percentile of absolute relative error. These benefits are more clearly illustrated when comparing the results for the population with more outliers to the one with fewer extreme observations. At the 75th and 90th percentiles of absolute relative error, when $\varepsilon = 0.01$, the relative improvement obtained from Winsorizing is quite dramatic and much greater than when $\varepsilon = 0$. However, at the same time the Winsorized estimator performs slightly worse at the lower percentiles of absolute relative error when $\varepsilon = 0.01$. These facts illustrate the trade-off between bias and variance of Winsorizing. The protection that is obtained against extremely large errors from Winsorizing is usually at the price of introducing a small amount of bias in estimation.

4. Conclusions

Winsorizing in sample surveys is a practical

and effective tool for improving the efficiency of estimation. As large outlying observations are a common and serious problem in many sample surveys, their treatment is usually necessary. However, there is a hidden danger in Winsorizing: that is, if the cutoffs are set to the wrong value there is a significant chance that the resulting estimator will be worse in terms of its mean squared error than the original estimator. The technique presented in Section 2 of this paper provides a reliable method of estimating the optimum Winsorizing cutoffs for the expansion estimator in repeated surveys. If the cutoffs are estimated from a previous independent run of the survey, there is a high probability that the Winsorized estimator will have considerably smaller mean squared error than the simple expansion estimator.

The results described in the evaluation section of this paper indicate that the technique works effectively in practice. As expected, when outliers are not present in the sample data the advantages of Winsorizing are only minor. In this situation there is also a small chance that the resulting estimator has slightly larger mean squared error than the simple expansion estimator. However, when a small proportion of extreme outliers is present in the sample, Winsorizing the expansion estimator according to the technique described in this paper leads to a considerable improvement in the general performance of the estimator.

The Winsorizing technique described in this paper is also applicable to one-time surveys. Computer simulation results (not presented here) indicate that if the cutoffs are estimated from the current sample data, then the Winsorized estimate of a total based on the same sample has considerably smaller mean squared error than the corresponding expansion estimator. For model (10), the improvements are of a similar magnitude to those for the Winsorized estimator given in Tables 2 and 3.

There are a number of practical issues that arise when Winsorizing survey data, which have not been dealt with in this paper, but that may be worthy of future investigations. Firstly, the cutoffs are optimum only at the level which estimates are being formed. If estimates are also required at a finer level of categorisation, then either the cutoffs will have to be recalculated or the original cutoffs used. In the former case more efficient estimates will result. However, the estimates of totals at the finer level of categorisation will not sum to the estimate of the total at the higher level, which often is of serious concern to users of survey results. In the latter case this is not a problem. However, estimates of the total in the subcategory may be less efficient than those obtained using the optimum Winsorizing cutoffs. This may not be a major concern as according to our analysis there appears to be a wide range of cutoffs that result in reasonably efficient estimators.

Secondly, the Winsorizing cutoffs in this paper have been chosen to produce optimum estimates of level. In repeated surveys, estimates of change in level or movement are also important. It is not clear how the individual movements can be Winsorized in a natural way. The most sensible course of action seems to be to Winsorize each survey's data as described in this paper, but to choose cutoffs that minimise the mean squared error of movement. Such cutoffs could be expected to be lower than those for minimising the mean squared error of level since the biases created by Winsorizing the level estimates would partly cancel each other in the Winsorized estimate of movement.

Finally, it is possible to generalise

Winsorizing when auxiliary information is available for all units in the population. As mentioned in the Introduction, the estimator, $\hat{T}$, is in fact a special case of the robust finite population estimator examined in Chambers (1986). This relationship suggests a fairly straightforward extension of Winsorizing to the case where the survey variable is related to the auxiliary variables according to a linear regression model. The cutoff will then be a function of the auxiliary variables. The theory developed in this paper may be appropriate for finding the optimum cutoff function in this case.

Appendix

Proof of Theorem 1

To find the minimum we differentiate $E\{(\hat{T}-T)^2\}$ with respect to K_h. From Cochran (1977, p. 23), the design mean squared error is

$$E_d\{(\hat{T}-T)^2\} = \sum_h N_h(N_h/n_h - 1)S_h^2(K_h) + \left[\sum_h N_h\{\bar{X}_h(K_h) - \bar{X}_h\}\right]^2$$

where

$$\bar{X}_h = N_h^{-1}\sum_i X_{hi}, \quad \bar{X}_h(K_h) = N_h^{-1}\sum_i X_{hi}(K_h)$$

and

$$S_h^2 = (N_h-1)^{-1}\sum_i\{X_{hi}(K_h) - \bar{X}_h(K_h)\}^2.$$

Hence the total mean squared error is

$$E\{(\hat{T}-T)^2\} = \sum_h N_h(N_h/n_h - 1)\mathrm{Var}_m\{X_h(K_h)\} + \sum_h N_h\mathrm{Var}_m\{X_h(K_h) - X_h\} + \left[\sum_h N_h E_m\{X_h(K_h) - X_h\}\right]^2. \tag{A.1}$$

The final term inside the square brackets above is the total bias B. We may calculate an expression for each term in (A.1) under our modelling assumptions at (1) as follows. Note that

$$X_h(K_h) = X_h + (1-f_h)(K_h - X_h)J_h.$$

Hence the bias

$$B = \sum_h N_h E_m\{X_h(K_h) - X_h\}$$

$$= \sum_h N_h(1-f_h)\{K_h E_m(J_h) - E_m(X_h J_h)\}. \tag{A.2}$$

Also, since

$$E_m\{X_h^2(K_h)\} = \sigma_h^2 + \mu_h^2 + (1-f_h)^2\{K_h^2 E_m(J_h) - 2K_h E_m(X_h J_h) + E_m(X_h^2 J_h)\} + 2(1-f_h)\{K_h E_m(X_h J_h) - E_m(X_h^2 J_h)\}$$

$$\begin{aligned}\mathrm{Var}_m\{X_h(K_h)\} = \sigma_h^2 + (1-f_h)^2[&K_h^2 E_m(J_h) - 2K_h E_m(X_h J_h) + E_m(X_h^2 J_h) \\ &- \{K_h E_m(J_h) - E_m(X_h J_h)\}^2] + 2(1-f_h)\{K_h E_m(X_h J_h) \\ &- K_h \mu_h E_m(J_h) - E_m(X_h^2 J_h) + \mu_h E_m(X_h J_h)\}.\end{aligned}$$

Likewise,

$$\begin{aligned}\mathrm{Var}_m\{X_h(K_h) - X_h\} = (1-f_h)^2[&K_h^2 E_m(J_h) - 2K_h E_m(X_h J_h) + E_m(X_h^2 J_h) \\ &- \{K_h E_m(J_h) - E_m(X_h J_h)\}^2].\end{aligned}$$

To find the derivative of (A.1) note that for $p > 0$

$$E_m(X_h^p J_h) = \int_{K_h}^{\infty} x^p g_h(x)\, dx$$

and so

$$\frac{\partial}{\partial K_h} E_m(X_h^p J_h) = -K_h^p g_h(K_h).$$

Thus, after some simple algebra, we may show that

$$\frac{\partial B}{\partial K_h} = N_h(1-f_h)E_m(J_h)$$

$$\begin{aligned}\frac{\partial}{\partial K_h}\mathrm{Var}_m\{X_h(K_h)\} &= 2(1-f_h)^2\{K_h E_m(J_h) - E_m(X_h J_h)\}\{1 - E_m(J_h)\} \\ &\quad + 2(1-f_h)\{E_m(X_h J_h) - \mu_h E_m(J_h)\}\end{aligned}$$

and

$$\frac{\partial}{\partial K_h}\mathrm{Var}_m\{X_h(K_h) - X_h\} = 2(1-f_h)^2\{K_h E_m(J_h) - E_m(X_h J_h)\}\{1 - E_m(J_h)\}.$$

Consequently,

$$\begin{aligned}\frac{\partial}{\partial K_h} E\{(\hat{T} - T)^2\} &= 2N_h(N_h/n_h - 1)(1-f_h)^2\{K_h E_m(J_h) - E_m(X_h J_h)\}\{1 - E_m(J_h)\} \\ &\quad + 2(N_h^2/n_h)(1-f_h)^2\{E_m(X_h J_h) - \mu_h E_m(J_h)\} \\ &\quad + N_h(1-f_h)^2\{K_h E_m(J_h) - E_m(X_h J_h)\}\{1 - E_m(J_h)\} \\ &\quad + 2BN_h(1-f_h)E_m(J_h) \\ &= 2(N_h^2/n_h)(1-f_h)^2 E_m(J_h)\{K_h - \mu_h - K_h E_m(J_h) + E_m(X_h J_h)\} \\ &\quad + 2BN_h(1-f_h)E_m(J_h). \qquad \text{(A.3)}\end{aligned}$$

As $E_m(J_h) > 0$, on setting the RHS of (A.3) to zero we obtain the relationship at (2). We now show that the set of K_h values satisfying (2) minimise $E\{(\hat{T} - T)^2\}$.

Note that by (A.2), $B \geqslant -\sum_h N_h(1-f_h)\mu_h$, and so

$$\frac{\partial}{\partial K_h} E\{(\hat{T} - T)^2\} > 0$$

for all $K_h > \mu_h + (N_h/n_h - 1)^{-1}\sum_h N_h(1-f_h)\mu_h$. Also as $K_h \to 0$

$$\frac{\partial}{\partial K_h} E\{(\hat{T} - T)^2\} \to 2N_h(1 - f_h)B < 0.$$

Thus, since the derivative is continuous, the minimum of $E\{(\hat{T} - T)^2\}$ must occur at a local minimum. The only candidate for a local minimum is the set $\{K_h\}$ satisfying (2). Let us evaluate the second partial derivatives at this point. Now

$$\frac{\partial^2}{\partial K_h^2} E\{(\hat{T} - T)^2\} = 2(N_h^2/n_h)(1 - f_h)^2 E_m(J_h)\{1 - E_m(J_h)\} + 2\{N_h(1 - f_h)E_m(J_h)\}^2$$

at the solution of (2), and for $h \neq k$

$$\frac{\partial^2}{\partial K_h \partial K_k} E\{(\hat{T} - T)^2\} = 2N_h N_k(1 - f_h)(1 - f_k)E_m(J_h)E_m(J_k).$$

If $\mathbf{A}$ is a matrix with the kth element in the hth row equal to $(\partial^2/\partial K_h \partial K_k)E\{(\hat{T} - T)^2\}$ at the solution of (2) and $\mathbf{x}^T = (x_1, \ldots, x_H)$, then for all $\mathbf{x}$,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \left\{\sum_h x_h N_h(1 - f_h)E_m(J_h)\right\}^2 + \sum_h 2x_h^2(N_h/n_h)(1 - f_h)^2 E_m(J_h)\{1 - E_m(J_h)\} > 0.$$

That is, $\mathbf{A}$ is positive definite, so the solution of (2) is a local, and hence a global minimum of $E\{(\hat{T} - T)^2\}$.

Finally, the expression immediately after (3) follows by multiplying both sides of (2) by n_h, summing over the subscript h and applying expression (A.2). This completes the proof of Theorem 1.

Proof of Theorem 2

From (A.2) it follows that $B \leqslant N_h(1 - f_h)\{K_h E_m(J_h) - E_m(X_h J_h)\}$. By this relationship, $n_h/N_h < 1$, and (2) it can be seen that

$$K_h - \mu_h \geqslant (n_h - 1)\{E_m(X_h J_h) - K_h E_m(J_h)\} \geqslant 0. \tag{A.4}$$

By (A.4), since X_h has a non-degenerate distribution and $n_h > 1$, $K_h - \mu_h > 0$. Thus from (A.4) and (2)

$$\begin{aligned} 0 &\geqslant 1 + B/\{(N_h/n_h - 1)(K_h - \mu_h)\} \\ &= \{K_h E_m(J_h) - E_m(X_h J_h)\}/(K_h - \mu_h) \geqslant -(n_h - 1)^{-1}. \end{aligned} \tag{A.5}$$

Since for any $0 < \delta < 1$, if $0 \geqslant 1 + y \geqslant -\delta$ then $0 \leqslant 1 + y^{-1} \leqslant 1 - (1 + \delta)^{-1}$, we see that (A.5) implies (4).

To prove the second part of the theorem we begin by showing that

$$\max_h (K_h - \mu_h) \to \infty \quad \text{as} \quad N, n \to \infty. \tag{A.6}$$

We shall prove this by contradiction. Assume that for all h, $K_h - \mu_h < M$. Since

$$\frac{\partial}{\partial K_h}\{E_m(X_h J_h) - K_h E_m(J_h)\} = -E_m(J_h) < 0$$

$$E_m(X_h J_h) - K_h E_m(J_h) \geqslant E_m[\{X_h - (M + \mu_h)\}I(X_h > M + \mu_h)] > 0.$$

Therefore there exists a $\delta > 0$ such that for all h

$$E_m(X_h J_h) - K_h E_m(J_h) > \delta$$

and hence it follows from (A.4) that for all sufficiently large n,

$$\max_h(K_h - \mu_h) \geqslant \delta \max_h(n_h - 1) \geqslant \delta(n/H - 1) > M$$

where H is the number of strata. This contradicts our original assumption and so the result at (A.6) must hold. It is now a simple matter to establish (5). Since $B < 0$ and $n_h/N_h < 1 - \varepsilon$, it follows from (4) that for all h,

$$\begin{aligned} -B &\geqslant (N_h/n_h - 1)(K_h - \mu_h) \\ &> \{(1-\varepsilon)^{-1} - 1\}(K_h - \mu_h). \end{aligned}$$

Therefore by (A.6), $B \to -\infty$ as $N, n \to \infty$. Also since $n_h/N_h > \varepsilon$ and $n_h > 1$

$$\begin{aligned} (\varepsilon^{-1} - 1)(K_h - \mu_h) &> (N_h/n_h - 1)(K_h - \mu_h) \\ &\geqslant B(n_h^{-1} - 1) \\ &\geqslant -B/2. \end{aligned}$$

Therefore, for each h, $K_h - \mu_h \to \infty$ as $N, n \to \infty$. This result, the relationship at (A.5) and since $E(X_h) < \infty$ implies that

$$1 + B/\{(N_h/n_h - 1)(K_h - \mu_h)\} \to 0$$

as $N, n \to \infty$. Result (5) of the theorem follows from this expression.

5. References

Australian Bureau of Statistics (ABS) (1990). Stocks, Manufacturers' Sales December Quarter 1989 and Expected Sales to December 1990. Australia, cat. no. 5629.0, Canberra.

Chambers, R.L. (1986). Outlier Robust Finite Population Estimation. Journal of the American Statistical Association, 81, 1063–69.

Cochran, W.G. (1977). Sampling Techniques (3rd ed.). New York: John Wiley.

Dalén, J. (1987). Practical Estimators of a Population Total which Reduce the Impact of Large Observations. Research and Development Reports, Stockholm: Statistics Sweden.

Ernst, L.R. (1980). Comparison of Estimators of the Mean which Adjust for Large Observations. Sankhya, Ser. C, 42, 1–16.

Gross, W.F., Bode, G., Taylor, J.M., and Lloyd-Smith, C.W. (1986). Some Finite Population Estimators which Reduce the Contribution of Outliers. Pacific Statistical Conference: Proceedings of the Congress, Auckland, New Zealand, 20–24 May, 1985, 386–390.

Hidiroglou, M.A. and Srinath, K.P. (1981). Some Estimators of a Population Total from Simple Random Samples Containing Large Units. Journal of the American Statistical Association, 76, 690–695.

Searls, D.T. (1966). An Estimator for a Population Mean which Reduces the Effect of Large True Observations. Journal of the American Statistical Association, 61, 1200–1204.

Tambay, J.L. (1988). An Integrated Approach for the Treatment of Outliers in Sub-Annual Economic Surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 229–234.

Received November 1993
Revised July 1994