SOME ESTIMATORS OF POPULATION TOTALS FROM SIMPLE RANDOM SAMPLES CONTAINING LARGE UNITS

M.A. Hidiroglou and K.P. Srinath, Statistics Canada

1. INTRODUCTION

The problem considered in this paper is the estimation of the population total of some characteristic from a simple random sample containing a few large or extreme observations. These observations are true observations belonging to the population that is being sampled. The presence of these observations in the sample will tend to make the usual estimate of the population total $\hat{Y}_0 = N\bar{y}$ (where $\bar{y}$ is the sample mean and N the population size) exceed the population total Y by a considerable amount though the estimation procedure itself is unbiased. It is therefore important to deflate the weights for such units at the estimation stage once they have been sampled and identified.

Several techniques have been proposed to handle unusually large values. Tukey and McLaughlin (1963) considered trimmed and Winsorized sample means from symmetric distributions. Crow (1964) has studied weighting procedures for observations. Fuller (1960) studied one-sided Winsorized means, Winsorization being applied to the largest observations only, assuming that the right tail of the distribution is well approximated by the tail of a Weibull distribution. Censored sample procedures have been considered by numerous authors (see for example Dixon (1960)). Searls (1966) proposed an estimator that used information external to the sample to predetermine a point, T, which separates "large" sample observations from the rest.

Recently, in studying estimators for skewed populations, Jenkins, Ringer and Hartley (1973) have adopted biased estimators which were preferable to $N\bar{y}$. Their quadratic loss function incorporated both the squared bias and the variance of the estimators, i.e., the mean square error (MSE).

We confine our attention to estimators which involve only a change of the usual weights as this seems a realistic and practical approach in sample surveys. No knowledge of the number of large units (outliners) in the population is assumed. We propose three estimators which are designed to reduce the effect of these large observations. The efficiencies of these estimators are empirically investigated along with the efficiency of the post-stratified estimator which involves a knowledge of the number of outliers in the population. The criterion for comparison of the proposed estimators with the usual estimator $N\bar{y}$ is the ratio of the variance of the unbiased estimator to the mean square error of these estimators. It is shown that, in certain situations, these estimators will have a smaller mean square error than the usual estimator $N\bar{y}$.

2. THE ESTIMATORS

We assume that a population $\{Y_1, Y_2, \ldots, Y_N\}$ of size N contains T large units. It is assumed that T is unknown. These outliers are are elements of the population whose Y-value exceeds a prespecified value T. A simple random sample of size n is drawn without replacement from the population and t outliers are identified. The estimators which we consider are:

$$\hat{Y}_1 = \sum_{i=1}^{t} y_i + \frac{N-t}{n-t} \sum_{i=t+1}^{n} y_i, \qquad (2.1)$$

$$\hat{Y}_2 = \frac{N}{n} \sum_{i=1}^{n} y_i - \frac{Nt}{2n} \left(\frac{n-t}{n}\right) \left(\sum_{i=1}^{t} \frac{y_i}{t} - \sum_{i=t+1}^{n} \frac{y_i}{n-t} \right) \qquad (2.2)$$

and

$$\hat{Y}_3 = r \sum_{i=1}^{t} y_i + \frac{N-rt}{n-t} \sum_{i=t+1}^{n} y_i . \qquad (2.3)$$

Estimator (2.1) assigns weight one to the outlier units and adjusts the weights of the non-outliers so that the sum of the sample weights adds up to N. Estimator (2.2) assigns a weight to the outlier units which is dependent upon the number of outliers in the sample. Finally, estimator (2.3) generalizes estimator (2.1) in that it assigns an optimal weight r to the outlier and non-outlier units.

If T is known a priori, the post-stratified estimator is:

$$\hat{Y}_4 = \frac{T}{t} \sum_{i=1}^{t} y_i + \frac{N-T}{n-t} \sum_{i=t+1}^{n} y_i . \qquad (2.4)$$

The bias and the mean square error (MSE) of these estimators are given in the following section.

3. THE MSE OF THE ESTIMATORS

We shall first consider the usual estimator of the population total $\hat{Y}_0$. $\hat{Y}_0$ may be expressed as the sum of outlier units and non-outlier units as:

$$\hat{Y}_0 = \frac{N}{n} \left\{ \sum_{i=1}^{t} y_i + \sum_{i=t+1}^{n} y_i \right\} . \qquad (3.1)$$

The variance of Y_o in the form given in (3.1) is

$$V(\hat{Y}_0) = \{ f^{-1} T\left(\frac{N-T}{N-1}\right) (1-\delta)^2 + N(f^{-1} - 1) \frac{T-1}{N-1} C_2^2 \delta^2 + N(f^{-1} - 1) \frac{N-T-1}{N-1} C_1^2 \delta^2 \} \bar{Y}_\nu^2 \qquad (3.2)$$

where f is the sampling fraction, δ is the ratio of the mean of the outlier units $\bar{Y}_\mu$ in the population to the mean of the non-outlier $\bar{Y}_\nu$ units in the population, C_1 and C_2 are the coefficients

of variation for the non-oulier and outlier units in the population respectively.

It can easily be shown that the biases of $\hat{Y}_1$, $\hat{Y}_2$ and $\hat{Y}_3$, for $T \geq 1$ are

$$B(\hat{Y}_1) = -T(1-f)(\delta-1)\,\bar{Y}_\nu \,, \qquad (3.3)$$

$$B(\hat{Y}_2) = \frac{-T(\delta-1)(N-T)\,\bar{Y}_\nu}{2N} \,, \qquad (3.4)$$

$$B(\hat{Y}_3) = -T(1-rf)(\delta-1)\,\bar{Y}_\nu \,. \qquad (3.5)$$

Note that estimators (2.1) and (2.3) are consistent whereas estimator (2.2) is not. The mean square error (MSE) of these estimators can be presented in two ways, depending on T. For T=1, the mean square error can be derived exactly. For $T > 1$, the approximate MSE for $\hat{Y}_1$ and $\hat{Y}_3$ is obtained using $E(t) \doteq 1/E(t)$. For $T > 1$, the exact MSE for $\hat{Y}_2$ has been derived.

We first present the exact mean square errors associated with T = 1. Details of the derivations are not given here.

$$MSE(\hat{Y}_1) = \{(1-f)(1-\delta)^2 + [\frac{f(N-1)}{n-1}(N-n) + N(1-f)(f^{-1} - \frac{N}{N-1})]\, C_1^2\, \bar{Y}_\nu^2\} \qquad (3.6)$$

$$MSE(\hat{Y}_2) = \{\frac{N^2 n(1-f)}{f(N-1)(n-1)}[(1-f)(1 - \frac{n+1}{2n})^2 + 1]\, C_1^2 + (1-\delta)^2\,[(1-f)[1 - \frac{f^{-1}}{2}(1 + \frac{1}{n})]^2 + f]\}\, \bar{Y}_\nu^2 \qquad (3.7)$$

and

$$MSE(\hat{Y}_3) = \{[(1-f) + f(1-r)^2](1-\delta)^2 + [\frac{f(N-r)^2}{n-1}\,\frac{N-n}{N-1} + N(1-f)(f^{-1} - \frac{N}{N-1})]\, C_1^2\}\, \bar{Y}_\nu^2 \,. \qquad (3.8)$$

The optimal value of r for (3.8) is given as

$$r_o = \frac{(1-f)\,C_1^2 + f(1-\delta)^2}{\frac{(1-f)}{N}\,C_1^2 + f(1-\delta)^2} \,.$$

Next, we provide expressions for MSE for $T > 1$.

$$MSE(\hat{Y}_1) \doteq (1-\delta)^2 f(1-f)\,T\,(1 - \frac{T}{N}) + (T-1)\, f(1-f)\, C_2^2\, \delta^2 + \frac{(1-f)}{f(N-T)}[(N-fT)^2 - f^2T]\, C_1^2 + T^2(1-f)^2(1-\delta)^2\}\, \bar{Y}_\nu^2 \,. \qquad (3.9)$$

$$MSE(\hat{Y}_2) = [\frac{T(1-\delta)(N-T)(n-1)}{2n(N-1)}]^2 + \left(\frac{C_2\delta}{2f}\right)^2 [Et + \frac{2Et^2}{n} + \frac{Et^3}{n^2} - \frac{1}{T}(Et^2 + \frac{2Et^3}{n} + \frac{Et^4}{n^2})] + \left(\frac{C_1}{2f}\right)^2 [(4n - \frac{3\,Et^2}{n} - \frac{Et^3}{n^2} - \frac{1}{N-T}(4n^2 - 4nEt - 3Et^2 + \frac{2Et^3}{n} + \frac{Et^4}{n^2})] + (\frac{1}{2nf})^2 (1-\delta)^2\, V(nt + t^2)\}\, \bar{Y}_\nu^2, \qquad (3.10)$$

where $V(nt + t^2) = n^2\, V(t) + 2n\, Cov\,(t,t^2) + V(t^2)$ and Et^k, k=1, 2,3,4, are moments obtained from the hypergeometric distribution given by

$$H(t|N,n,T) = \frac{\binom{N-T}{n-t}\binom{T}{t}}{\binom{N}{n}} \,, \quad 0 \leq t \leq T,\ N-T > n.$$

The mean square error of $\hat{Y}_3$ for $T > 1$ is

$$MSE(\hat{Y}_3) \doteq \{r^2(1-\delta)^2\, f(1-f)\, T(1 - \frac{T}{N}) + r^2(T-1)\, f(1-f)\, C_2^2\, \delta^2 + \frac{(1-f)}{f(N-T)}[(N-rfT)^2 - r^2f^2T]\, C_1^2 + T^2(1-rf)^2\,(1-\delta)^2\}\, \bar{Y}_\nu^2 \,. \qquad (3.11)$$

The optimal value of r for $T > 1$ is obtained by minimizing (3.11). Differentiating (3.11) with respect to r and solving for r, we obtain

$$r_o = \frac{g_1(N,f,T,\delta,C_1)}{g_2(N,f,T,\delta,C_1,C_2)} \qquad (3.12)$$

where

$$g_1(N,f,T,\delta,C_1) = (1-\delta)^2\, fT^2 + \frac{(1-f)TN}{N-T}\, C_1^2$$

and

$$g_2(N,f,T,\delta,C_1,C_2) = (1-\delta)^2\, fT\,[(1-f)(1 - \frac{T}{N}) + fT]$$

$$+ f(1-f)(T-1)[C_2^2 \delta^2 + \frac{T}{N-T} C_1^2].$$

The variance of the post-stratified estimator $\hat{Y}_4$ for T > 1 is given by

$$V(\hat{Y}_4) \doteq \{C_1^2 [f^{-1}(1-f)(N-T) + \frac{T}{nf}] + C_2^2 \delta^2 [f^{-1}(1-f)T + \frac{N-T}{nf}]\} \bar{Y}_\nu^2 . \quad (3.13)$$

4. AN EMPRICAL INVESTIGATION OF THE ESTIMATORS

To investigate the efficiency and utility of the proposed estimators, we have used a variety of artificial populations. We have studied the relative efficiency of these estimators for various values of C_1, C_2, δ, f, N and T. The relative efficiency is defined as the ratio of the variance of the usual estimator of the total $\hat{Y}_0$ to the mean square error of $\hat{Y}_i$, i=1, 2,3,4.

The empirical investigation has been extensive and in view of the difficulty of presenting a great number of tables, only six tables are presented. Tables 1 through 5 are constructed to reveal a difference in the behaviour of the estimators $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$ for various values of C_1, C_2, δ, f and T for a given value of N. Within each of these tables C_2 and T vary while, δ, f and C_1 are fixed. The tables differ from each other by having one of the variables δ, f or C_1 vary while the other two variables are fixed. Table 6 differs from the others in that a large value of N and a small sampling fraction f have been used. The conclusions drawn from these tables, in general, should apply to other populations.

Tables of Relative Efficiencies

Estimators	$\hat{Y}_1$		$\hat{Y}_2$		$\hat{Y}_3$		$\hat{Y}_4$	
1.			δ=5	f=0.3		C_1= 0.5	N=500	
T \ C_2	1.0	2.0	1.0	2.0	1.0	2.0	1.0	2.0
2	1.26	-	1.16	-	1.26 (1.10)	-	0.62	-
4	1.37	2.32	1.41	2.00	1.41 (1.49)	2.32 (0.78)	0.77	0.49
10	1.02	2.13	1.30	2.08	1.37 (2.15)	2.17 (1.32)	1.03	0.74
15	0.75	1.69	1.10	1.89	1.30 (2.43)	1.92 (1.62)	1.15	0.84
25	0.48	1.18	0.81	1.56	1.20 (2.72)	1.61 (2.02)	1.28	0.94
80	0.14	0.40	0.35	0.83	1.06 (3.12)	1.20 (2.77)	1.43	1.07
2.	δ=5	f=0.1		C_1=0.5			N=500	
2	1.37	-	1.28	-	1.37 (1.19)	-	0.40	-
4	1.75	3.22	1.56	2.13	1.75 (1.76)	3.22 (0.75)	0.53	0.30
10	1.85	4.17	1.85	2.63	2.04 (3.22)	4.17 (1.46)	0.78	0.53
15	1.53	3.70	1.85	2.63	1.96 (4.11)	3.84 (2.00)	0.92	0.65
25	1.06	2.78	1.67	2.00	1.69 (5.34)	3.12 (2.87)	1.09	0.79
3.	δ=10	f=0.3		C_1=0.5			N=500	
2	1.78	-	1.72	-	1.78 (1.16)	-	0.48	-
4	1.64	3.12	1.78	2.56	1.78 (1.58)	3.12 (0.89)	0.75	0.46
10	0.92	2.04	1.30	2.13	1.45 (2.24)	2.17 (1.47)	1.15	0.76
15	0.64	1.51	1.02	1.82	1.31 (2.51)	1.85 (1.79)	1.30	0.87
25	0.40	0.99	0.71	1.41	1.19 (2.78)	1.51 (2.18)	1.45	0.98
80	0.12	0.33	0.31	0.71	1.06 (3.14)	1.16 (2.86)	1.59	1.10

4.	δ=10		f=0.1		C_1=0.5			N=500	
2	2.43	-	1.92	-	2.43 (1.23)	-		0.27	-
4	3.03	6.25	2.32	2.94	3.12 (1.89)	6.25	(0.87)	0.46	0.27
10	2.08	5.00	2.27	2.94	2.56 (3.48)	5.26	(1.70)	0.82	0.53
15	1.51	3.70	2.08	2.78	2.13 (4.41)	4.17	(2.30)	1.00	0.66
25	0.94	2.44	1.69	2.50	1.72 (5.66)	3.03	(3.28)	1.20	0.81
5.	δ=5		f=0.3		C_1=1.0			N=500	
2	1.06	-	1.06	-	1.06 (1.16)	-		0.83	-
4	1.12	1.47	1.12	1.39	1.14 (1.53)	1.47	(0.80)	0.88	0.63
10	1.00	1.67	1.14	1.64	1.18 (2.17)	1.69	(1.33)	1.02	0.79
15	0.81	1.52	1.05	1.64	1.16 (2.44)	1.64	(1.64)	1.10	0.87
25	0.54	1.14	0.84	1.45	1.14 (2.73)	1.49	(2.03)	1.20	0.95
80	0.16	0.41	0.37	0.83	1.06 (3.12)	1.19	(2.77)	1.37	1.06
6.	δ=5		f=.01		C_1=0.5			N=10,000	
5	1.06	1.19	1.05	1.14	1.07 (2.88)	1.19	(1.09)	0.52	0.23
15	1.22	1.64	1.16	1.41	1.22 (6.41)	1.64	(2.39)	0.57	0.29
25	1.33	2.04	1.25	1.64	1.35 (9.75)	2.04	(3.70)	0.62	0.35
25	1.51	2.70	1.41	1.96	1.54(15.79)	2.70	(6.22)	0.70	0.45
65	1.61	3.12	1.54	2.17	1.67(21.08)	3.12	(8.61)	0.77	0.52
85	1.61	3.33	1.61	2.32	1.75(25.76)	3.45	(10.89)	0.83	0.58

Note: Dashes indicate that C_2 is non-existent for these cases. The numbers in brackets are the optimal r_o values given by (3.12).

It is seen from the above tables that, for fixed δ, f, C_1, C_2, and N, the efficiencies of of the estimators decrease after an initial improvement as T increases. The efficiency gain in using these estimators increases as the coefficient of variation C_2 of the outlier units increases. Comparing the values in Table 1 with those in Table 5, we see that as C_1 increases, the efficiencies of the estimators decrease for small values of T and increase after a certain number of outliers has been reached. Comparing values in Tables 1 and 3, we see that as δ increases from 5 to 10, gains in efficiency are not uniform. In fact, for large T, there is a greater loss in efficiency. This is due to the fact that the bias term of the estimators dominates the mean square error as δ increases. Referring to Tables 1 and 2, 3 and 4, it is seen that as f decreases, gains in efficiencies of the estimators increase.

To stress the effectiveness of these estimators, a fairly large population of N=10,000 and a small sampling fraction of f=0.01 have been used. The results are given in Table 6. Note that for a few number of outliers in the population, the gain in using these estimators is not very considerable. However, as the number of outliers in the population increases, the effectiveness of these estimators improves quite significantly.

It is possible to make the following general observations. The best estimator to use with respect to efficiency is $\hat{Y}_3$. $\hat{Y}_2$ has lower efficiency than $\hat{Y}_1$ for a small number of outliers, however, after a certain number of outliers has been reached, $\hat{Y}_2$ is superior to $\hat{Y}_1$. Hence, $\hat{Y}_2$ is to be preferred to $\hat{Y}_1$ in the presence of a moderate number of outliers. For a small number of outliers, the post-stratified estimator $\hat{Y}_4$ is not as good as the other estimators because the allocation between the post-strata is likely to be poor, being very different from the optimum allocation in such cases. But, as expected, once a certain number of outliers is reached, it is superior to all estimators including $\hat{Y}_0$.

$\hat{Y}_3$, the optimal estimator, requires a knowledge of T, C_1, C_2 and δ from the sample. We use these in the expression (3.12). Estimating r_o using sample values could imply a departure from optimal efficiency of $\hat{Y}_4$. To study this possible departure, the efficiency of $\hat{Y}_3$ has been investigated for different values of $r_o(1+\Delta)$, where $0.0 \leq \Delta < 1.0$. Two situations have been invest-

TABLES OF RELATIVE EFFICIENCIES OF $\hat{Y}_3$ FOR $r_o(1+\Delta)$

7.	δ=5	f=0.3	C_1=0.5		C_2=1.0	N=500
T	2	4	10	15	25	80
$\lvert\Delta\rvert$ \ r_o	1.10	1.49	2.5	2.43	2.72	3.12
0.0	1.26	1.41	1.37	1.30	1.20	1.06
0.1	1.26	1.41	1.35	1.26	1.15	0.94
0.2	1.26	1.40	1.31	1.19	1.04	0.69
0.3	1.26	1.38	1.23	1.09	0.89	0.47
0.4	1.25	1.35	1.15	0.97	0.74	0.33
0.5	1.25	1.32	1.05	0.85	0.61	0.24
0.6	1.23	1.29	0.95	0.74	0.51	0.18
0.7	1.23	1.25	0.86	0.64	0.42	0.14
0.8	1.21	1.21	0.77	0.56	0.35	0.11
0.9	1.20	1.16	0.69	0.48	0.29	0.09

8.	δ=5	f=0.01	C_1=0.5		C_2=1.0	N=10,000
T	5	15	25	45	65	85
$\lvert\Delta\rvert$ \ r_o	2.88	6.41	9.75	15.79	21.08	25.76
0.0	1.069	1.216	1.346	1.546	1.678	1.755
0.2	1.069	1.216	1.345	1.545	1.673	1.746
0.4	1.069	1.216	1.344	1.541	1.541	1.664
0.6	1.069	1.216	1.343	1.535	1.648	1.697
0.8	1.069	1.215	1.341	1.527	1.626	1.656

igated. The first one being a large population of size 10,000 with an associated small sampling fraction of 0.01 and the second being a small population size of 500 with a fairly high sampling fraction of 0.3. The results are given in Tables 7 and 8. From the preceding tables, it is seen that when there is a low number of outliers, the efficiency of $\hat{Y}_3$ is not significantly affected by departures from optimal r_o. As the number of outliers increases in the first (N=500) population, even small departures from optimal r_o result in low efficiency. Note that in the case of the second population (N=10,000), departures from optimal r_o are not significant even for large number of outliers in the population.

5. CONCLUSIONS

When the sampling fraction f and the number of outliers T are small, use of the estimator $\hat{Y}_1$ would result in substantial gains in efficiency. If f and T are moderately large, use of $\hat{Y}_2$ is recommended. $\hat{Y}_3$ can be used to advantage if values of C_1, C_2, δ and T are approximately known from previous surveys. Deviations from the optimal r_o associated with $\hat{Y}_3$ will not affect the efficiency if T is small. If T is large and known, it is obvious that the post-stratified estimator $\hat{Y}_4$ should be used.

REFERENCES

Bershad, M., "Some Observation On Outliers", Unpublished dittoed memorandum, 1960, Statistical Research Division, U.S. Bureau of Census.

Chinnappa, N., "A Preliminary Note On Methods of Dealing With Unusually Large Units In Sampling from Skew Populations", Unpublished, IASMD technical memorandum, February 1976.

Crow, Edwin L., "The Statistical Construction of a Single Standard from Several Available Standards", IEEEE Transactions On Instrumentation and Measurement, 13 (1964), 180-5.

Dixon, W.J., "Simplified Estimation From Censored Normal Samples", Annals of Mathematical

Statistics, 31 (1960), 385-91.

Fuller, W.A., "Simple Estimation for the Mean of Skewed Populations" 1960, U.S. Bureau of Census.

Hartley, Herman O., and Rao, J.N.K., "A New Estimation Theory for Sample Surveys", Biometrika, 55 (March 1968), 547-57.

Jenkins, O.C., Ringer, L.G., and Hartley, H.O., "Root Estimators", Journal of the American Statistical Association, 68 (1973) 414-19.

Rao, C.R., "Some Aspects of Statistical Inference in Problems of Sampling from Finite Populations", in Foundation of Statistical Inference. Holt, Rinehart and Winston of Canada Ltd., (1971) 177-202.

Searls, D.T., "An Estimator which Reduces Large True Observations", Journal of the American Statistical Association, (1966), 1200-4.

Tukey, J.W., and McLaughlin, D.H., "Less Vulnerable Confidence and Significance Procedures for Location Based On A Single Sample: trimming/ Winsorization 1", Sankhya, Series A, 25 (1963), 331-52.