ON THE INCLUSION OF LARGE UNITS IN SIMPLE RANDOM SAMPLING

M.A. Hidiroglou, Statistics Canada

1. Introduction

The stratification of a population is a technique often used in survey sampling. This technique may produce gains in precision in estimating characteristics of the population. The problem considered in this paper is the stratification of a population into two strata, the take-all and take-some strata. The take-all stratum contains some of the largest units in the population while the take-some stratum contains the remaining units. The take-all stratum units are surveyed entirely while a sample random sample is drawn from the take-some stratum. This type of stratification is particularly useful for populations whose distribution exhibits a marked positive skewness, with a few large units and many small units. Failure to recognize that such highly skewed populations should be stratified in the above manner may result in over-estimation of the population characteristics. This last point has been studied by Hidiroglou and Srinath (1977).

Approximate cutoff rules for stratifying a population into take-all and take-some universes have been given by Dalenius (1950) and Glasser (1962). Glasser (1962) expressed the cutoff value (that value which delineates the boundary of the take-all and take-some subuniverses) as a function of the mean, the sampling weight and the population variance. Their cutoff values were derived on the assumption that a prespecified sample size n was to be drawn without replacement from a population of size N. In the present context, exact and approximate cutoff rules have been worked out for a similar situation. Rather than providing the sample size, the desired level of precision c (coefficient of variation) of the estimates is given. Note that in many sampling situations, the sampler is given a set of objectives in terms of reliability of the estimates.

2. The Sampling Procedure and Method of Estimation

Consider a finite population ϕ_N consisting of N units labelled $y_1, y_2, \ldots, y_N$. Define ordered statistics $y_{(N)}, y_{(N-1)}, \ldots, y_{(1)}$ where $y_{(N)} \leq y_{(N-1)} \leq \ldots \leq y_{(1)}$.

Let a simple random sample of size $n(\ell)$ be selected. Note that $n(\ell)$ is no longer a fixed size. Rather, it is a variable which depends on the number of take-all units ℓ to be included in the sample. Assume that the desired level of precision for the estimated total is given as c. The total Y may be written as:

$$Y = \sum_{i=1}^{\ell} y_{(i)} + \sum_{i=\ell+1}^{N} y_{(i)} \quad . \tag{2.1}$$

Given that ℓ units are take-all and $n(\ell)-\ell$ units are take-some, an estimator of the total Y would be:

$$\hat{Y} = \sum_{i=1}^{\ell} y_{(i)} + \frac{N-\ell}{n(\ell)-\ell} \sum_{i=1}^{n(\ell)-\ell} z_i \tag{2.2}$$

where $y_{(N)} \leq z_i \leq y_{(\ell+1)}$ for $i=1,2,\ldots,n(\ell)-\ell$.

The variance of $\hat{Y}$ is:

$$V(\hat{Y}) = \frac{(N-\ell)\{N-n(\ell)\}}{n(\ell)-\ell} S^2_{N-\ell} \tag{2.3}$$

where

$$S^2_{N-\ell} = \frac{1}{N-\ell-1} \sum_{i=\ell+1}^{N} (y_{(i)} - \mu_{N-\ell})^2$$

$$\mu_{N-\ell} = \frac{1}{N-\ell} \sum_{i=\ell+1}^{N} y_{(i)} \quad .$$

In terms of reliability c, $V(\hat{Y})$ may be re-expressed as $V(\hat{Y}) = c^2\hat{Y}^2$. Substituting $V(\hat{Y}) = c^2Y^2$ into (2.3) and solving for $n(\ell)$:

$$n(\ell) = \ell + \frac{(N-\ell)^2 S^2_{N-\ell}}{c^2Y^2 + (N-\ell)S^2_{N-\ell}} \quad . \tag{2.4}$$

3. The Optimum Point

The objective is to find the optimum value of y which minimizes the sample size $n(\ell)$ for the given level of precision c. A necessary condition for the optimum point is that (2.4) with $\ell=m$ shall not exceed (2.4) with $\ell=m-1$ or $\ell=m+1$. This means that the optimum value of $y(y^*)$ is found whenever $n(m-1) \geq n(m)$ and $n(m) \leq n(m+1)$. This condition can be made more flexible if we introduce a real number b into the inequalities, that is,

$$n(m-1) \geq n(m) + b - 1 \tag{3.1}$$

and

$$n(m) \leq n(m+1) + b - 1,$$

where b can be used to control the number of units to include in the take-all stratum.

Stopping rule (3.1) is the exact one for finding the optimal cutoff for a given b. To express (3.1) in terms of the optimal cutoff neighbouring values $y_{(m)}$ and $y_{(m+1)}$, we need the following two relations.

$$(N-m)S^2_{\nu+1} = (N-m-1)S^2_{\nu} + \frac{N-m}{N-m+1}(y_{(m)} - \mu_{\nu})^2 \tag{3.2}$$

and

$$(N-m-2)S^2_{\nu-1} = (N-m-1)S^2_{\nu} - \frac{N-m}{N-m-1}(y_{(m+1)} - \mu_{\nu})^2,$$

where

$$S^2_{\nu+k} = \frac{1}{N-m+k-1} \sum_{i=m-k+1}^{N} (y_{(i)} - \mu_{\nu+k})^2$$

and

$$\mu_{\nu+k} = \frac{1}{N-m+k} \sum_{i=m-k+1}^{N} y_{(i)}$$

for $k = -1,0,1$ and $\nu = N-m$.

Substituting (2.4), (3.2) into (3.1) one can show that:

$$(y_{(m)}-\mu_\nu)^2 \geq \{\frac{[bN-n_m-(b-1)m](N-m)}{(n_m-m)(N-n_m-b+1)} + \frac{1}{N-m}\}S_\nu^2$$

and (3.3)

$$(y_{(m+1)}-\mu_\nu)^2 \leq \{\frac{[bN-n_m-(b-1)m](N-m-2)}{(n_m-m)(N-n_m+b-1)} + \frac{1}{N-m}\}S_\nu^2 .$$

The compromise for (3.3) if m is the optimum number of units to include with certainty is

$$(y^*-\mu_\nu)^2 = \{\frac{b(N-m-1)}{n_m-m} + \frac{1}{2}[\frac{(b-1)(N-m)}{N-n_m-b+1} + \frac{(b-1)(N-m-2)}{N-n_m+b-1}]$$

$$+ \frac{1}{2}\frac{b(b-1)}{(n_m-m)}[\frac{N-m}{N-n_m-b+1} - \frac{N-m-2}{N-n_m+b-1}]\}. \quad (3.4)$$

Note that if m is the optimum number of units to be included in the sample with certainty, then $y_{(m+1)} < y^* \leq y_m$. Also, equation (3.4) is one solution of the system in inequalities given by (3.3). While (3.4) is a necessary condition for an optimum, it is not necessarily sufficient. More than one solution may exist, in which case the one that minimizes $n(\ell)$ for given b would be chosen. As Glasser (1962) points out, while it may not pay to include with certainty a given unit by itself, it may pay to include it with several other units.

Noting that

$$n_m = m + \frac{(N-m)^2 S_\nu^2}{c^2Y^2+(N-m)S_\nu^2} \quad (3.5)$$

and

$$N-n_m = \frac{(N-m)c^2Y^2}{c^2Y^2+(N-m)S_\nu^2} \quad (3.6)$$

substitution of (3.5) and (3.6) into (3.4) yields

$$(y^*-\mu_\nu)^2 \doteq \frac{bc^2Y^2}{N-m} + (2b-1)S_\nu^2 + \frac{(N-m)(b-1)S_\nu^4}{c^2Y^2}, \quad (3.7)$$

provided that b-1 is very much smaller in magnitude than $NcY^2/(cY^2+NS_\nu^2)$.

An upper limit for y^* can be obtained in terms of the population variance S_N^2, population size N and mean μ_N by using the following inequalities:

$$N(y^*-\mu_N) \leq (N-m)(y^*-\mu_\nu), \quad (3.8)$$

$$(N-m-1)S_\nu^2 \leq (N-1)S_N^2 - \frac{mN(y^*-\mu_N)^2}{N-m} \quad (3.9)$$

where it is true that for $m>0$, $y^*<\mu_m$, μ_m being the mean of the m largest units in the population. Substituting inequalities (3.8) and (3.9) into (3.7), we obtain after some simplification the approximate cutoff rule

$$y^* < \mu_N + [\frac{bc^2Y^2}{N} + S_N^2\{(2b-1)+\frac{N(b-1)S_N^2}{c^2Y^2}\}]^{1/2}. \quad (3.10)$$

This inequality depends only on the population size, the coefficient of variation c,b,μ_N and S_N. This approximation will be good only when m is relatively small compared to N. The more extreme and the more variable the large units, the less well the limit approximates the exact solution. Although the computer programming and time involved in obtaining the exact cutoff point is quite minimal, it is nevertheless instructive to characterize the bound in terms of known population values.

Approximation (3.10) reveals one point about b's effect on the boundary point. If $b_2 > b_1$, then the boundary point associated with b_2 will be higher than the one associated with b_1. Note that the converse also follows. The choice of b is user dependent. Under various situations, the number of units in the take-all stratum may be varied. For instance, in business surveys, a possible determining factor affecting the cutoff rule could be the portion of the population that the take-all units represent in terms of the study variable. In this case, the user would probably take $b \leq 1$. Another factor could be response burden. The user would most likely introduce a rotation scheme which would permit some of the large units to rotate in and out of the sample. For this case, fewer units would be included in the take-all by choosing $b \geq 2$.

4. Some Practical Illustrations

The use of the procedure given in section 2 presumes that the population from which the sample is to be drawn, is to be a good proxy for the target population. An example where such a procedure may be used is the following. All the values associated with the units of a business universe are known at time t_1. A sample is drawn from this universe at time t_1+k_1, $k_1 \geq 0$, and to be used as a basis for inference to the universe characteristics from t_1+k_1 to t_1+k_2 where $k_2>k_1$. In this instance, the universe at time t_1, $\phi(t_1)$, may be different from the universe at time t_2, $\phi(t_2)$, $t_2>t_1$. However, if it can be assumed that the cutoff value computed at time t_1 is not too different from the one that would be computed at time t_2, then partitioning of the population $\phi(t_1)$ will still yield gains.

The data used to illustrate the results given in section 3 is from the 1976 Food and Beverage Annual Survey. This survey is essentially a census of all eating and drinking establishments covered by the Merchandising and Services Division of Statistics Canada. Establishments covered in this survey includes all known businesses with establishments classified to the

Standard Industrial Classification code 886(1970). The Standard Industrial Classification code is broken down further into seven kinds of businesses that range from licenced restaurants to beverage rooms, bars and night clubs. Data for this survey is presently being published at a subprovincial by kind of business cross-classification. The example takes a situation where the business universe is known at time t_1 (the 1976 Food and Beverage Restaurant Survey) and a sample is to be drawn at time t_1+k_1 (the projected Monthly, Tavern, Caterers and Restaurant Survey).

The cutoff rule that is illustrated is the one given by (3.1) with b chosen equal to 1 and 2 respectively. Four subprovincial by kind of business strata have been chosen to provide the examples. They are respectively: Beverage Rooms, Bars and Night Clubs in Newfoundland (stratum 1), Beverage Rooms, Bars and Night Clubs in the non-metropolitan areas of New Brunswick (stratum 2), Licenced Restaurants in Halifax-Dartmouth (stratum 3), and Beverage Rooms, Bars, Night Clubs in the non-metropolitan areas of Quebec (stratum 4). Some of the statistical characteristics for those strata are given in Table 5.1. These are the minimum, maximum and mean sales for each of the strata. The standard deviation, S_N, is also provided with the associated population size. Note that these statistics imply that the associated frequency distributions are positively skewed.

For each of the strata in question and given the coefficient of variation desired, we provide the number of units to be included in the take-all substratum, the exact and approximate cutoff and the sample that would have been selected had no take-all substratum been formed. This information is displayed in Table 5.2. Note that the approximate cutoff point is given by inequality (3.10) and the exact cutoff point by equation (3.3) with b=1.

In Table 5.2, m is equal to the number of units to be included in the sample with certainty and n(m) is the corresponding overall sample size required to achieve the desired reliability. Note that $n(m) < n(0)$ for all strata considered, where n(0) is the sample size with no take-all units. Hence, if take-all units are to be found, the overall sample size will be smaller than that of the sample with no take-all units. Note that the approximate cutoff given by (3.10) is quite close to the exact cutoff given by (3.3). Results for b=2 provided in Table 5.3 highlight the effect of b on the boundary points.

Note that p stands for the number of units in the take-all stratum and n(p) is the correspond ing overall sample size. Again, the approximate cutoff given by (3.10) is quite close to the exact cutoff given by (3.3) with b=2. The exact bound with b=2 tends to yield fewer take-all units than the exact bound with b=1. The same conclusion is reached if the approximate bound is used.

Table 5.1: Statistical Characteristics for the Strata of interest

Stratum	Minimum	Maximum	Mean	Standard Dev.	N
1	3,000	476,141	139,380	67,800	170
2	4,000	463,000	181,930	90,160	61
3	15,045	1223,360	350,250	263,830	63
4	3,345	885,333	132,770	72,520	632

Table 5.2: Information Concerning the Take-all Procedure Given by Inequalities (3.1) for b=1

Stratum	c	Exact Cutoff	Approximate Cutoff	m	n(m)	n(0)
1	0.1116	353,351	353,230	3	15	17
2	0.1063	339,071	357,840	4	13	16
3	0.1359	806,999	811,060	6	10	21
4	0.1209	598,192	542,450	3	12	20

Table 5.3: Information Concerning the Take-all Procedure Given by Inequalities (3.1) for b=2

Stratum	c	Exact Cutoff	Approximate Cutoff	p	n(p)	n(0)
1	0.1116	476,141	450,133	1	16	17
2	0.1063	462,303	451,950	1	15	16
3	0.1359	1,223,360	1,077,048	1	19	21
4	0.1209	832,991	717,265	2	15	20

5. Conclusion

It is desirable to stratify highly skewed populations on the basis of the size of the units. The approach suggested in the present paper is to put a certain number of large units into a take-all stratum and sample those with certainty. The remaining units, those attached to the take-some stratum, are sampled at an appropriate rate. The number of units to include with certainty depends on the desired level of precision c, the scalar b, the population mean μ_N, the population variance S_N^2 and the number of units N when criteria (3.1) is used. Note that the sampler may vary the number of units in the take-all stratum by varying b. The approximate stopping rule (3.10) may be used as an initial estimate for the corresponding exact cutoff given by (3.1) provided that the necessary information on the population of interest is available.

There are several advantages in stratifying a highly skewed population for the given method. For a fixed level of reliability, the overall sample size associated with this procedure will invariably be lower than the sample size associated with no stratification. Cochran (1963, p. 38-39) points out that for frequency distributions that are not reasonably close to normality, it is risky to use the normal approximation as a basis for constructing confidence intervals. By separating some of the largest observations from highly skewed distributions, confidence intervals are essentially based on populations which are less skewed. This last point should encourage the sampler in having more confidence in using the normal approximation. Finally, this type of stratification guards against overestimation of population characteristics when highly skewed distributions are sampled.

Glasser (1962) has pointed out that the definition one should assign to large units depends on the method of sampling the remainder of the population and the method of estimation.

References

[1] Cochran, W.G. (1963), Sampling Techniques, Wiley, 2nd ed. New York.

[2] Dalenius, T. (1950), "The Problem of Optimum Stratification in a Special Type of Design", Skandinavisk Aktuariediskrift, p. 61-70.

[3] Glasser, G.J. (1962), "On the Complete Coverage of Large Units in a Statistical Study", Review of the International Statistical Institute, Vol. 30, p. 28-32

[4] Hidiroglou, M.A. and Srinath, K.P. (1977), "Some Estimators of Population Totals from Simple Random Samples Containing Large Units". Proceedings of the American Statistical Society, Chicago, Social Statistics Section, p. 903-908.