J. King Saud Univ., Vol. 4, *Admin. Sci.* (1), pp. 65-81 (A.H. 1412/1992)

A Composite Estimator for the Total in a Finite Population

Z. A. El-Beshir
Associate Professor, Department of Quantitative Methods, College of Administrative Sciences, King Saud University, Riyadh, Saudi Arabia

(Received 4/3/1411; Accepted for Publication 24/1/1412)

Abstract. We consider the situation in which the sampled population is a subset of a larger finite population that is "exposed to sampling" according to a certain probability distribution. The exposure probabilities are determined independently of the sampling method used. A composite estimator is proposed for the population total together with approximate formulae for its expectation and variance. An unbiased estimator for the apporximate variance is also suggested.

1. Introduction

It sometimes happens that the actually sampled population is itself a subset of a finite taraget population that gets 'exposed' to sampling according to some probability distribution. The probability distribution is independent of the sampling method used, and the sampler has no means for controlling it. For example, in sampling the population of a town, we may decide, for convenience, to take the sample from among those individuals who attended an outpatient clinic such as E.N.T. or Eye clinic. The group of individuals attending the clinic at a specified period of time may be looked at as a subgroup of the inhabitants of the town that is exposed to the clinic in accordance with a chance mechanism. And although it is possible, through a properly chosen sampling procedure, to control the selection probabilities, it is not possible to do the same for exposure probabilities of units in the target population. As a result, if a target population parameter is to be estimated, we can not rely soley on a design-based estimator since a model specifying the manner in which units in the actual population are generated from the traget population has to be set. A sort of model-design-based estimator must therefore be resorted to.

The main problem with modelling in the analysis of surveys stems from the fact that it is usually difficult, in practice, to estimate bias in the estimator or detect departures from the model [1]. Attempts for combining model-based and design-based estimators into a composite estimator were made in the field of small area estimation, by Schaible [2] and Fay and Herriot [3]. In this paper an estimator, based partly on a model and partly on design, is suggested for the population total. We shall call it composite because of its mixed nature although this term is usually used for an estimator that represents a combination of two separate estimators.

2. The Problem

We have a finite target population U of N units. A sample of size n is to be taken from U by some method. The sampler, however, cannot sample U itself but can only sample a subset of it, S_k say, of size k, that happened to be exposed to sampling. The value of k is fixed in advance and satisfies $n < k < N$. The exposed subset S_k is a random realization following the conditional (on k) distribution $P(S_k)$.

Our problem is to estimate the total T of a characteristic y in the population U. Of course if the sample size is not fixed, the exposed subset which we may then denote by S will follow the unconditional distribution P(S), and the size of S, K say, will be a random variable satisfying $0 \leqslant K \leqslant N$. In the example of the outpatients clinic, individuals in the town form U, those in the clinic S_k, while fixing K at k will mean sampling only subsets of U of size k.

3. Exposure Probability and Inclusion Probability

Let I_i be an exposure indicator taking the value 1 if unit i is exposed but 0 otherwise. The exposure probability of unit i, p_i say, is

$$p_i = \text{prob}\,(I_i = 1) = E(I_i) \qquad i = 1,\ldots, N$$

while the conditional exposure probability for unit i, given that the number exposed is fixed at k, is

$$p_i' = \text{prob}\,(I_i = 1/K = k) = E(I_i = 1/K = k) \qquad i = 1,\ldots,N$$

Similarly, the conditional exposure probability for units i and j is

$$p_{ij}' = \text{prob}\,(I_i\, I_j = 1/K = k) \qquad i = 1,\ldots,N$$

The p_i' and p_{ij}' can be expressed in terms of the p_i. For if we let

$$\sum_i^N p_i = N^*,\ d = \sum_i^N p_i$$

(1 - p$_i$), then it can be shown (Hājek [4] pp. 83 and 142) that

$$p'_i = p_i \left[1 + \frac{(k - N^*)(1 - p_i)}{d} + \frac{(k - N^*)^2 - d}{d^2} x \, (1 - p_i)\left(\bar{\bar{p}} - p_i\right) + o\left(d^{-1}\right) \right]$$

$$\ldots\ 1 \leq i \leq N \qquad\qquad \ldots (3.\ 1)$$

Where $\bar{\bar{p}} = \sum_i^N p_i^2 (1 - p_i) / d$, and where $o(d^{-1})d \to 0$ as $d \to \infty$ provided that $| k - N^* | d^{-\frac{1}{2}}$ remains bounded by a constant independent of d. Also,

$$p'_{ij} = p_i \ p_j \left[1 + \frac{2 - p_i - p_j}{d} (k - N^*) + \frac{(k - N^*)^2 - d}{d^2} \right.$$

$$\left. \left[(1 - p_i)\left(\bar{\bar{p}} - p_i\right) + (1 - p_j)\left(\bar{\bar{p}} - p_j\right) + (1 - p_i)(1 - p_j) \right] + o\left(d^{-1}\right) \right]$$

$$\ldots\ 1 \leq i \neq j \leq N \qquad\qquad \ldots (3.\ 2)$$

where d o(d^{-1}) → 0 in probability.

Now if we assumed that:

$$p_i = p \qquad \text{all i} \cdot\cdot \qquad \ldots (3.3)$$

then since $\sum_i^N p_i = N^*$, it follows that $p = N^* / N$ so that on using (3.1) and (3.2) we have the asymptoticaly valid results:

$$p_i = p_1' \sim \frac{k}{N} \qquad \text{all i} \qquad \ldots (3.4)$$

$$p'_{ij} = p'_2 \sim \frac{k^2 - N^* + N^{*2} / N}{N^2} \qquad \text{for all } i \ \& \ j \quad (i \neq j) \qquad (3.\ 5)$$

Suppose a sample of n units is to be drawn from S$_k$. Denote by π_i (k) the overall inclusion (in the sample) probability for unit i given that k units are exposed, and by π_{ij}(k) the overall inclusion probability, given k, for units i and j (i ≠ j). Define the inclusion indicator t_i that takes the value 1 or 0 depending on whether or not unit i(i = 1,…,N) is included in the sample. It follows that:

$$\pi_i(k) = \text{Prob. } (t_i = l/K = k)$$

$$= \text{Prob. (i exposed and selected } /K = k)$$
$$= \text{Prob. (i selected/i exposed and } K = k) \text{ Prob. (i exposed/}K = k)$$

Assuming (3.3) and using (3.4), we have

$$\pi_i(k) = m_i \, p_1' \qquad \dots (3.6)$$

where m_i = Prob (i selected /i exposed and K = k). By a similar argument:

$$\pi_{ij}(k) = m_{ij} \, p_2' \qquad \dots (3.7)$$

where m_{ij} is the probability that both i and j (i≠j) are selected given that they are both exposed and K = k. The inclusion probabilities (3.6) and (3.7), each consists of two factors: One factor (p_1' and p_2') is determined independently of the sampling method used and can not be controlled. The other factor (m_i or m_{ij}) is dependent on the sampling method, and can thus be controlled by the sampler. For example the units can be selected so that the m_i are proportional to a measure of size.

Finally to obtain the explicit forms of P(S) and P(S_k) we note that if $p_1, p_2, \dots p_N$ is any sequence of numbers, here exposure probabilities, satisfying $0 \leqslant p_i \leqslant 1$ for all i and if S is an arbitrary subset of U and r its complement in U, then assuming the I_i independent

$$P_{(s)} = \prod_{i \in s} p_i \prod_{i \in r} (i - p_i) \qquad \dots(3.8)$$

If the number exposed is fixed at k, we have

$$P_{(s_k)} = C^* \prod_{i \in s} p_i \prod_{i \in r} (1 - p_i) \qquad if \;\; K = k \qquad \dots(3.9)$$
$$= 0 \qquad if \;\; K \neq k$$

with $C^* = C^* (p_1, \dots, p_N)$, where C^* is such that $\sum^{C} P(S_k) = 1$ and the summation is over all $C = \binom{N}{k}$ distinct subsets of size k. In the terminology of Hājek [4] (3.8) and (3.9) define Possion sampling and rejective sampling respectively.

As a result if we assume that the number exposed is fixed at k and that $p_i = p$ for all i, the non zero value in (3.9) becomes

$$C^* \, p^k \, (1 - p)^{N-k}$$

Summing this over all subsets of size k and equating to 1 we have

$$C^* = \frac{1}{C\ p^k (1-p)^{N-k}}$$

so that

$$p(s_k) = \frac{1}{C} \qquad \text{... (3. 10)}$$

if K = k but equals 0 otherwise.

4. Estimation of the Population Total

In this section we estimate the total T of characteristic y in U on the basis of a sample of size n from the exposed subset S_k. We assume that exposures are independent and that the total number of units exposed is fixed at k. Furthermore, we make the simplifying assumption that the conditional exposure probabilities p_1' and p_{1j}' of a unit and of two different units respectively, are as given by (3.4) and (3.5), and that the conditional distribution of S_k is given by (3.10).

Define the subclass of linear estimators

$$\hat{T}_1^* = \sum_i^n \beta_i\ y_i \qquad \text{.... (4. 1)}$$

where the β_i(i = 1, ..., N) are constants. For $\hat{T}^*_1$ to be unbiased given K = k, we must have

$$E\left(\hat{T}_1^* / k\right) = \sum_i^N \pi_i(k)\ \beta_i y_i = \sum_i^N y_i = T \qquad \text{... (4. 2)}$$

or, in view of (3.6),

$$\sum_i^N m_i\ p_1'\ \beta_i\ y_i = T$$

This requries that

$$m_i\ p_1'\ \beta_i = 1 \qquad i = 1, ..., N$$

so that

$$\hat{T}^{*} = \sum_{i}^{n} \frac{y_i}{p_1' \, m_i} = \sum_{i}^{n} \frac{y_i}{\pi_i(k)} \qquad \text{... (4. 3)}$$

is an unbiased estimator of T. It is the only unbiased estimator in the subclass (4.1) and is thus the 'best' estimator in that subclass. Note however that the expectation in (4.2) is conditional since averaging is made over all subsets of size k. The estimator $\hat{T}^*$ will have zero variance if

$$\pi_i \ (k) = \frac{ny_i}{T}$$

i.e if the $\pi_i(k)$ are proportional to the y_i. Let x be a characteristic the value of which is positive and known for all units in the population. If x_i, the value of x for unit i, is approximately proportional to the y_i, we may expect considerable reduction in the variance of $\hat{T}^*$ if the $\pi_i(k)$ are made proportional to the x_i. If for example we managed to make

$$\pi_i \ (k) = \frac{nx_i}{\sum_{i}^{N} x_i}$$

and if x and y are strictly proportional, (*e.g.* Cochron [5], p. 259) $x_i / \sum_{i}^{N} x_i$ will equal y_i / T and $\hat{T}^* = T$ which has zero variance.

The characteristic x can be any variable positively correlated with y provided that its values are positive and known for all units in U. It may, for example, be a measure of size, an eye estimate of y, or the value of y at a previous occasion. It may also be a characteristic known to be positively correlated with the phenomenon under study such as age.

Now, if k = N, then $S_k = U$ and $p_1' = 1$. The estimator $\hat{T}^*$ is then the usual Horvitz - Thompson estimator $\hat{T}_{HT} = \sum_{i}^{n} y_i / m_i$. For this estimator, several methods are available for satisfying the condition $m_i = nx_i / \sum_{i}^{N} x_i$ when sampling without replacement. Among these are the methods of Durbin [6], Sampford [7], Samiuddin and Asad [8], Chao [9], Chakrabarti [10], Gupta *et al.* [11], Lahiri [12], and Midzuno [13]. When, however, k < N as is here assumed, $\pi_i(k)$ contains the factor p_1' which is uncontrollable. On the other hand the sampler is sampling S_k rather than U itself. By a suitable method of selection we may achieve

$$m_i = \frac{nx_i}{\sum_i^k x_i} \qquad \text{... (4. 4)}$$

but then $\hat{T}^*$ will not equal T even under strict proportionality between x and y, and hence no reduction in variance may be gained. Let us set

$$X = \sum_{i=1}^{N} x_i, \quad X_k = \sum_{i=1}^{N} x_i$$

and put

$$X_k = \alpha X \qquad \text{... (4. 5)}$$

Using (3.6), (4.4) and (4.5) we have

$$\pi_i(k) = \frac{nx_i\, p_1'}{\alpha X} \qquad \text{... (4. 6)}$$

so that, in view of (4.6) and (4.3) and assuming strict proportionality between x and y we get:

$$\hat{T}^* = \frac{\alpha}{p_1'} T$$

Hence, if we require an estimator with zero variance under strict proportionality, we may take

$$\hat{T} = \frac{p_1'}{\alpha} \sum_i^n \frac{y_i}{m_i\, p_1'} = \frac{1}{\alpha} \sum_i^n \frac{y_i}{m_i} \qquad \text{... (4. 7)}$$

If x and y are strictly proportional, and the sample is selected such that (4.4) is satisfied, $\hat{T}$ is a constant and thus has zero variance.

As a result, if x is approximately proportional to y we may expect considerable precision in $\hat{T}$ assuming our assumption about the p_1' is valid.

Note, however, that $\hat{T}$ is not in general, unbiased. It is unbiased if $\alpha = 1$. In the special case k = N, $\alpha = 1$ and $\hat{T}$ is again the Horvitz - Thompson estimator, which can be shown to be unbiased [14].

Although $\hat{T}$ is not a combination of two separate estimators: one design-based and the other model-based, we shall still describe it as 'composite' being partly a design-based and partly a model-based (through the assumption concerning exposure probabilities).

5. Approximate Mean and Variance of $\hat{T}$

Since X_k is a random variable, so is α. Hence the estimator $\hat{T}$ is a ratio of two random variables. As a result, to obtain the expectation and variance of $\hat{T}$ we resort to approximation. Put

$$\hat{T} = X \; \frac{y^*}{x_k} = X \quad z\left(Y^*, X_k\right)$$

where $Y^* = \sum_i^n y_i / m_i$. Using the inclusion indicator t_i, it can be readily established that:

$$E\left(\sum_i^n y_i / m_i\right) = \sum_i^N p_1' \, y_i \sim \frac{k}{N} T \qquad \text{... (5. 1)}$$

and that

$$V\left(t_i\right) = \pi_i(k)\left(1 - \pi_i(k)\right)$$

$$\underset{i \neq j}{cov}\left(t_i, t_j\right) = \pi_{ij}(k) - \pi_i(k)\,\pi_j(k)$$

so that

$$V\left(\sum_i^n y_i / m_i\right) = V\left(\sum_i^N t_i y_i / m_i\right) = \sum_i^n \pi_i(k)\left(1 - \pi_i(k)\right) \frac{y_i^2}{m_i^2}$$

$$+ \sum_{i \neq j}^N \frac{y_i}{m_i} \frac{y_i}{m_j}\left(\pi_{ij}(k) - \pi_i(k)\,\pi_j(k)\right) \qquad \text{... (5. 2)}$$

Furthermorte, denote by S_{ik} the i th ($i = 1, \ldots, c$) group of size k and by X_{ik} its total. Also, define the indicator a_i such that it takes the value 1 if S_{ik} is realized but 0 otherwise. Then from (3.10), and noting that one S_{ik} can occur at a time, we have for $i = 1, \ldots, c$ and $j \neq i$:

$$E\left(a_i\right)=\frac{1}{C} \quad ,V\left(a_i\right)=\frac{1}{C}-\frac{1}{C^2}\ ,\ cov\left(a_i,\ a_j\right)=-\frac{1}{C^2}$$

It follows that

$$E\left(X_k\right)=E\left(\sum_i^C a_i\ X_{ik}\right)=\frac{1}{C}\sum_i^C X_{ik}$$

And because each x_i appears $\binom{N-1}{k-1}$ times in the total $\sum_i^C X_{ik}$ we have

$$E\left(X_k\right)=\frac{k}{N}\ X \qquad \text{... (5. 3)}$$

It can also be verified that:

$$V\left(X_k\right)=\frac{1}{C}\left[\sum_i^C X_{ik}^2-\frac{1}{C}\left(\sum_i^C X_{ik}\right)^2\right]=S^2 \qquad \text{... (5. 4)}$$

and

$$Cov\left(\sum_i^n y_i\ /\ m_i,\ X_k\right)\sim\frac{1}{C}\sum_j^N y_j\ \tilde{X}_k^j-\left(\frac{k}{N}\right)^2\ XT \qquad \text{... (5. 5)}$$

where $\bar{X}_k^j$ is the sum of the sums of all X_{ik} for all groups S_{ik} in which unit j occurred.

Approximate Mean

It is well known (*e.g.* Lindley [15] p. 135) that if z = (Y*, X_k) is a function of Y* and X_k, and if E(Y*) and E(X_k) are the means of Y* and X_k respectively, and V(Y*) and V(X_k) their respective variances, then for sufficiently ,small $\sqrt{V(Y^*)}$ and $\sqrt{V(X_k)}$ and well-behaved z:

$$E\left(z\right)=z\left(E\left(Y^*\right),\ E\left(X^*\right)\right)+\frac{1}{2}\ V\left(Y^*\right)\frac{\partial^2 z}{\partial\ Y^{*2}}+$$

$$Cov\left(Y^*,\ X_k\right)\frac{\partial^2 z}{\partial\ Y^*\partial X_k}+\frac{1}{2}V\left(X_k\right)\frac{\partial^2 z}{\partial X_k^2}\quad\ldots \qquad \text{... (5.6)}$$

$$V\left(z\right)=V\left(Y^*\right)\left(\frac{\partial z}{\partial Y^*}\right)^2+2\ Cov\left(Y^*,\ X_k\right)\left(\frac{\partial z}{\partial X_k}\ \frac{\partial z}{\partial Y^{*2}}\right)+V\left(X_k\right)\left(\frac{\partial z}{\partial X_k}\right)^2\ldots \text{ (5. 7)}$$

where all the partial differentials are evaluated at $X_k = E(X_k)$ and $Y^* = E(Y^*)$
But:

$$z\left(E\left(Y^*\right), E\left(X_k\right)\right) \sim \frac{T}{X}, \quad \frac{\partial^2 z}{\partial Y^{*2}} = 0, \quad \frac{1}{2} V\left(X_k\right) \frac{\partial^2 z}{\partial X_k^2} \sim \frac{N^2 S^2 T}{k^2 X^3}$$

and

$$Cov\left(Y^*, X_k\right) \frac{\partial^2 z}{\partial Y^* \partial X_k} \sim - \frac{N^2}{k^2 X^2 C} \sum_j^N y_j \tilde{X}_k^j + \frac{T}{X}$$

and from (5.6)

$$E(z) = \frac{T}{X} - \left(\frac{N}{k}\right)^2 \frac{1}{X^2 C} \sum_j^N y_j \tilde{X}_k^j + \frac{T}{X} + \left(\frac{N}{k}\right)^2 \frac{S^2 T}{X^3}$$

hence,

$$E(\hat{T}) = X \;\; E(z) = T + \left(\frac{N^2}{k^2 X^2} \left[S^2 T - \frac{X}{C} \sum_j^N y_i \tilde{X}_k^j \right] + T \right) \quad \dots (5.8)$$

Since $\tilde{X}_k^j$ is the sum of the sums of $\binom{N-1}{k-1}$ groups each of size k, it follows that

$$\frac{N^2}{k^2 XC} \sum_j^N y_i \tilde{X}_k^j = \frac{N}{X} \sum_j^N y_i \tilde{\tilde{X}}_k^j$$

where $\tilde{\tilde{X}}_k^j$ is the mean per unit for the units contained in all subsets of size k containing j. The mean $\tilde{\tilde{X}}_k^j$ should not differ much from the overall mean per unit $\bar{X}$ and we can thus put

$$\frac{N^2}{k^2 XC} \sum_j^N y_i \tilde{X}_k^j \simeq T$$

so that (5.8) simplifies to

$$E(\hat{T}) \simeq T \left(1 + \frac{S^2}{k^2 X^2} \right)$$

This result shows that, subject to the validity of the exposure model, and the order of the approximation used, the bias in $\hat{T}$ approaches zero as $S^2 \to 0$. Of course, this bias is zero if k = N for then $S^2 = 0$.

Approximate Variance

By an argument similar to that used in arriving at (5.8) and with reference to (5.7) it can be shown that

$$V(\hat{T}) \simeq \frac{N^2}{k^2}\left[\sum_i^N \pi_i(k)(1-\pi_i(k))\frac{y_i^2}{m_i^2} + \sum_i^N \sum_{\neq j}^N \frac{y_i}{m_i}\frac{y_j}{m_j}(\pi_{ij}(k) - \pi_i(k)\pi_j(k))\right]$$

$$+\frac{N^2T^2S^2}{k^2X^2} - \frac{2N^2T}{k^2X}\left\{\frac{1}{C}\sum_j^N y_j\,\tilde{X}^{j}_{k} - \frac{k^2}{N^2}XT\right\} \qquad \text{... (5. 9)}$$

the last term in (5.9) can be written as

$$-2\frac{N^2T}{k^2X}\left\{\frac{1}{C}\sum_j^N y_j\,\tilde{X}^{j}_{k} - \frac{k^2}{N^2}XT\right\} = \frac{NT}{X}\sum_j^N y_j\,\bar{\bar{X}}^{j}_{k} - T^2$$

Again if $\bar{\bar{X}}^{j}_{k} = \bar{X}$ this term vanishes, and since $\bar{\bar{X}}^{j}_{k}$ is expected to be close to $\bar{X}$ we may eliminate this term from (5.9) which then takes the relatively simpler form

$$V(\hat{T}) \simeq \frac{N^2}{k^2}\left[\sum_i^N \pi_i(k)(1-\pi_i(k))\frac{y_i^2}{m_i^2}\right.$$

$$\left.+\sum_i^N \sum_{\neq j}^N \frac{y_i}{m_i}\frac{y_j}{m_j}(\pi_{ij}(k) - \pi_i(k)\pi_j(k))\right]$$

$$+\frac{N^2T^2S^2}{k^2X^2} = V^*(\hat{T}) \quad say.$$

Note that if k = N, then $\pi_i(k) = m_i$ and $S^2 = 0$ so that $V^*(\hat{T})$ becomes the variance of the Horvitz - Thompson estimator $\hat{T}_{HT}$.

We may think of the quantity within the square brackets as giving the variance of $\hat{T}$ when the sampled population is the target population. When the actual population is a subset of the target population, a component representing the between actual populations variability must be included in the expression of variance. This is supplied by the last term in the above expression.

6. Estimation of V* ($\hat{T}$)

To estimate the approximate variance V* ($\hat{T}$) we first estimate S^2. Suppose C' ($C' < C$) subsets of U, of size k each are exposed independently. For ease of reference, let these be $S_{1k}, S_{2k}, \ldots, S'_{ck}$. Then using the indicator variable of section (5) we get

$$E\left(\sum_{i}^{c'} X_{ik}^{2}\right)=\sum_{i}^{c} X_{ik}^{2}\, E\,(a_i)=\frac{1}{C}\sum_{i}^{c} X_{ik}^{2}$$

So an unbiased estimator of $\sum_{i}^{C'} X_{ik}^2$ is $C\sum_{i}^{C'} X_{ik}$.

On the other hand, since each unit appears $\underset{C}{\text{in}}$ $({}^{N-1}_{k-1})$ of the X_{ik}'s each unit contributes its value that number of times in the sum $\sum_i X_{ik}$. Therefore

$$\sum_{i}^{c} X_{ik}=\binom{N-1}{k-1} X$$

and S^2 can be written as

$$S^{2}=\frac{1}{C}\left\{\sum_{i}^{c} X_{ik}^{2}-\frac{1}{C}\binom{N-1}{k-1}^{2} X^{2}\right\}$$

Because X is assumed known, an unbiased estimator of S^2 is

$$\hat{S}^{2}=\frac{1}{C}\ \{\, C\sum_{i}^{c'} X_{ik}^{2}-\frac{1}{C}\binom{N-1}{k-1}^{2} X^{2}\}$$

$$=\sum_{i}^{c'} X_{ik}^{2}-\left(\frac{k}{N}\, X\right)^{2}$$

Similarly, using the indicator t_i we note that

$$E\left(\sum_{i}^{n} y_{i}\right)^{2}=\sum_{i}^{N} y_{i}^{2}\pi_{i}\,(k)+\sum_{i}^{N}\sum_{\neq j}^{N} y_{i} y_{j}\,\pi_{ij}\,(k)$$

Hence

$$\hat{y} = \sum_{i}^{n} \frac{y_i^2}{\pi_i(k)} + \sum_{i}^{n} \sum_{\neq j}^{n} \frac{y_i y_j}{\pi_{ij}(k)}$$

is an unbiased estimator of T^2. Assuming $\hat{y}$ and $\hat{S}^2$ independent, and employing t_i we finally arrive at

$$v^*(\hat{T}) = \frac{N^2}{k^2}\left[\sum_{i}^{n}(1-\pi_i(k))\frac{y_i^2}{m_i^2} + \sum_{i}^{n}\sum_{\neq j}^{n}\frac{y_i}{m_j}(\pi_{ij}(k) - \pi_i(k)\pi_j(k))\right] + \frac{N^2 y \hat{S}^2}{k^2 X^2}$$

which is an unbiased estimator of $V^*(\hat{T})$.

7. Other Versions of $\hat{T}$

Other estimators for T that are versions of $\hat{T}$ may also be suggested. The first version is obtained by substituting $\frac{nx_i}{X_k}$ for m_i in (4.7) to get the estimator.

$$\hat{T}' = \frac{X}{n}\sum_{i=1}^{n}\frac{y_i}{X_i} \qquad \text{... (7. 1)}$$

The advantage of this estimator is that it enables us to make use of standard results in the theory of ratio estimators. Thus the expectation of $\hat{T}'$ (within and over the S_k's) will take the form:

$$E(\hat{T}') = X\hat{R}$$

where $\hat{R}$ is the expected value of a ratio in simple random sample of size k from the target population (U). The variance of $\hat{T}'$ may be obtained by using

$$V_k(\hat{T}') = E_k \frac{X^2}{X_k^2} V_k\left(\hat{T}'_{(k)}\right) + V_k \frac{X}{X_k} T_k \qquad (7. 2)$$

where E_k & V_k means taking expectation and variance over all groups of size k respectively, and

$$\hat{T}'_{(k)} = \frac{X_k}{n} \sum_{i=1}^{n} \frac{y_i}{X_i}$$

$$T_k = \sum_{i=1}^{k} y_i$$

A second estimator, also a version of $\hat{T}$, is obtained by assuming Lahiri-Midzuna [12,13] scheme in which the inclusion probabilities are proportional to $\sum_{i}^{n} x_i$. If in (4.7) we substitute $\frac{\sum_{i}^{n} X_i}{X_k}$ for m_i we get the estimator.

$$\hat{T}'' = \frac{\sum_{i}^{n} y_i}{\sum_{i}^{n} X_i} X \qquad (7.3)$$

The expected value of $\hat{T}''$ is approximately $X\hat{R}$.

A third version of $\hat{T}$ is obtained by simply substituting k/N for α and nx_i/X_k for m_i in (4.7) to get

$$\hat{T}''' = \frac{N}{k} \sum_{i}^{n} y_i / m_i$$

which, though slightly inferior, yet unbiased.

The main problem with the estimators $\hat{T}'$, $\hat{T}''$ and $\hat{T}'''$ is that the explicit forms of their variance have not yet been derived. It is therefore difficult to assess their relative precision. The three estimators are suggested to me by a referee to whom I am grateful.

8. Concluding Remarks

In the preceding discussion, an attempt is made to generalize the Horvitz - Thompson estimator, $\hat{T}_{HT}$, so as to incorporate the situation in which units are exposed to sampling with preassigned probabilities. The proposed estimator $\hat{T}$, seems to provide small bias and high precision, assuming x and y proportional, if S^2 is small *i.e* if there is little variability between the exposed (actual) populations S_{ik}. It is however important to remember that, apart from the approximation used, the extend to which we can rely on results about bias and precision of $\hat{T}$ depends on the validity of the assumed model.

It is informative to view $\hat{T}$ from a different angle. The estimator $\hat{T}_{HT} = \sum_{i}^{n} y_i / m_i$, assuming (4.4), estimates the total of S_k and not U, and hence it underestimates T. By dividing it by α, the ratio of the total of x in S_k to that in U, underestimation is reduced. The factor $\frac{1}{\alpha}$ in $\hat{T}$ thus, in a sense, works as an inflator that helps compensate for the underestimation incurred when using $\hat{T}_{HT}$ alone.

A practical implication of the preceding discussion is that if the actual population differs from the target population, and if exposure probabilities are equal, then $\hat{T}_{HT}$ will no longer be unbiased with small variance. This is so even if (4.4) is satisfied and x and y proportional. The extend of the bias and precision in $\hat{T}$ depends on the relative sizes of the actual and target populations as well as on inter-actual population variability.

References

[1] Kalton, G. "Models in the practice of survey sampling" *Int. Statist. Rev.* 51 (1983), 175-188.

[2] Schible, W.L. "A composite estimator for small area statistics". In: *Synthetic Estimates for Small Areas,* Ed.J. Steomberh, pp. 36-53. National Institute on Drug Abuse Research monograph 24, Washington D.C.: U.S. Government Printing Office, 1979.

[3] Fay, R.E. and Herriot, R.A. "Estimates of income for small place: an application of James-Stein procedures to Census data". *J. Am. Statist. Assoc.* 74 (1979), 269-277.

[4] Hajek, J. *Sampling from a Finite Population.* New York: Marcel Dekker, 1981.

[5] Cochran W.G. *Sampling Techniques.* John Wiley & Sons Inc. 1977.

[6] Durbin. J. "Design of multistage surveys for the estimation of sampling errors" *Appl. Statist.,* 16 (1967), 152-64.

[7] Sampford, M.R. "On sampling without replacement with unequal probabilities of selection" *Biometrika,* 54 (1967) pp. 119-27.

[8] Samiuddin, M. and Asad, H. "A simple procedure of unequal probability sampling" *Biometrika*, 68, 3 (1981), 728-31.

[9] Chao, M.T. "A general purpose unequal probability sampling plan" *Biometrika*, 69, 3, (1982), 653-6.

[10] Chakrabarti, M.C. "On the use of incidence matrices of designs in sampling from finite populations". *J. Ind. Statist. Assoc.* 1 (1963), 78-85.

[11] Gupta, V.K., Nigam, A.K., Kumar, P. "On a family of sampling schemes with inclusion probability proportional to size" *Biometrika* 69 (1982) , 191-196.

[12] Lahiri, D.B. "A method for sample selection providing unbiased ratio estimates." *Bull. Int. Stat. Inst.* 33, 2, (1951) 133-140.

[13] Midzuno, H. "On the sampling system with probability proportionate to sum of sizes" *Ann. Inst. Stat. Math.*, 2 (1951), 99-108.

[14] Horvitz, D.G. and Thompson, D.J. "A generalization of sampling without replacement from a finite universe". *J. Am. Statist. Assoc.* 47 (1952) 663-85.

[15] Lindley, D.V. *Introduction to Probability and Statistics from Bayesion Viewpoint. Part I. Probability.* Cambridge University Press (1964).

مقدار مركب للمجموع في مجتمع محدود

زين العابدين عبدالرحيم البشير

أستاذ مشارك، قسم الأساليب الكمية، جامعة الملك سعود، الرياض، المملكة العربية السعودية

(قُدم للنشر في ١٤١١/٣/٤هـ وقُبل للنشر في ١٤١٢/١/٢٤هـ)

ملخص البحث. نتناول الحالة التي يكون فيها المجتمع الذي تجرى معاينته جزء من مجتمع محدود أكبر تعرض لعملية المعاينة وفق توزيع احتمالي معين. احتمالات التعرض للمعاينة تتحدد بشكل مستقل عن طريقة المعاينة المستخدمة. تم اقتراح مقدار مركب لمجموع المجتمع مع صيغة تقريبية لتوقعه وتباينه. كذلك تم اقتراح مقدر غير متحير للتباين.