VARIANCES FOR A ROTATING SAMPLE FROM A CHANGING POPULATION

Normand Laniel, Statistics Canada

R.H. Coats Building, 11th Floor, Tunney's Pasture, Ottawa, K1A 0T6, CANADA

1. INTRODUCTION

In repeated surveys, samples are periodically drawn from a changing population to provide estimates of the level of a characteristic of interest and its change in level between two successive occasions. This change in level is, among other things, due to three causes. These are: the deletion of units (deaths) from the population, the addition of new units (births) to the population and the evolution of the characteristic of interest itself for the units common to both periods.

The above situation occurs in many repeated surveys conducted by Statistical Agencies, for example, at Statistics Canada the redesigned Monthly Wholesale Trade and Monthly Retail Trade Surveys. These two surveys will have their sampling frames updated at regular time intervals with independent sources (other surveys and some administrative files) and, each month, some businesses will rotate out of the sample while others will rotate in the sample.

Kish (1965, pp 457-458) gave an expression for the sampling variance of a change in sample mean based on overlapping samples. He assumed that the population was the same over time (e.g. not affected by births nor deaths) and that its size was sufficiently large for the finite population corrections to be ignored. Tam (1984) removed the assumption of a large population and obtained an exact expression for the sampling variance. This paper further removes the assumption of the population being the same over time and deals with the problem of calculating the sampling variance of an estimate of change based on samples affected by rotation, deaths and births and drawn on the first and second occasion of a repeated survey.

After giving some definitions, two sampling plans for the rotation of the sample and their corresponding estimators for the change in level are presented. Then, the sampling variances of the estimated change are given. This involves the derivation of the sampling covariance between the estimated levels of both occasions. Finally, the results and their extension to more than two occasions are discussed.

2. DEFINITIONS

Let denote as P_x the N_x units of a finite population from which a sample is drawn on the first occasion and as P_y the N_y units of the changed population from which a sample is drawn on the second occasion. Associated with unit i, $i \varepsilon P_x$, is the value x_i which denotes the observation on this unit on the first occasion, and with unit j, $j \varepsilon P_y$, the value y_j which denotes the observation on this unit on the second occasion. Also, let $P_c = P_x \cap P_y$ be the N_c units common to both populations, $P_d = P_x - P_c$ the N_d units deathed between the two occasions and $P_b = P_y - P_c$ the N_b units birthed between the two occasions.

Now, the population variances and covariance can be denoted as

$$S_x^2 = \frac{1}{N_x - 1} \sum_{i \varepsilon P_x} (x_i - \bar{X})^2, \tag{1}$$

$$S_{xy} = \frac{1}{N_c - 1} \sum_{i \varepsilon P_c} (x_i - \bar{X}_c)\ (y_i - \bar{Y}_c), \tag{2}$$

$$S_{y,c}^2 = \frac{1}{N_c - 1} \sum_{j \varepsilon P_c} (y_j - \bar{Y}_c)^2, \tag{3}$$

$$S_{y,b}^2 = \frac{1}{N_b - 1} \sum_{j \varepsilon P_b} (y_j - \bar{Y}_b)^2, \tag{4}$$

where $\bar{X} = \sum_{i \varepsilon P_x} x_i / N_x$, $\bar{X}_c = \sum_{i \varepsilon P_c} x_i / N_c$,

$\bar{Y}_c = \sum_{j \varepsilon P_c} y_j / N_c$ and $\bar{Y}_b = \sum_{j \varepsilon P_b} y_j / N_b$

The estimate of the change in level considered here is defined as follows. Denote as $\hat{X}$ the estimate of the population total obtained from the sample drawn on the first occasion and as $\hat{Y}$ the estimate of the population total obtained from the sample drawn on the second occasion. Then, the estimate of change is the difference between the estimates of the population totals on the two occasions, which is denoted by $\hat{Y} - \hat{X}$. Note that once the sampling variance formula of this estimate of change is obtained, it is straightforward to get the same expression for a change in sample mean.

3. SAMPLING PLANS AND ESTIMATORS

Assume that a simple random sample of size n_x has been selected without replacement from P_x on the first occasion. Then, between the first and second occasions, all the dead units within P_x are identified by a source independent to the sample and are deleted from the surveyed population. Thus, a random number of units in the first sample, denoted by n_{xc}, are kept in the remaining population, P_c. Note that if the source of deaths was the sample, the results of Tam would be used with the approach of Cochran (1977, p. 35-37) to estimate the required total over the subpopulation of common units on the second occasion. Also, assume that N_b births are added to the reduced population.

On the second occasion, the two subpopulations P_c and P_b of population P_y are sampled independently. A simple random sample of size n_b is selected without replacement from the subpopulation of births, P_b. For the selection of the sample from P_c two sampling plans with sample rotation are considered. They are modified versions of Tam's Sampling Plans A and B. The modification is to retain a fixed proportion, denoted by r, of the first sample units which fell in P_c instead of a fixed sample size. The reason is that n_{xc} is a random variable with the consequence that not enough units in the sample may be left in the population after the deletion of deaths to satisfy a fixed sample size. The first sampling plan is as follows.

Sampling Plan A. A random subsample of size equal to a predetermined proportion r times n_{xc} of the first sample units which fell in P_c is retained as part of the second sample. For the remaining part of the second sample, a simple random sample of size $(1-r)$ n_{xc} is selected without replacement from P_c excluding the first sample units which fell in it, that is, from a population comprising $N_c - n_{xc}$ units.

The above sampling plan will work satisfactorily if two conditions are satisfied. These are:

1. the sample size n_x should be large enough such that the probability that $n_{xc}=0$ is very small, and
2. n_x should be small enough such that the probability that $n_{xc} > \frac{N_c}{(2-r)}$ is very small.

The first condition is necessary to ensure with high probability that some of the first sample units fell in P_c and that an estimate can be produced for this sub-population. The second condition gives a high probability that N_c-n_{xc} will be large enough so that $(1-r)$ n_{xc} units can be selected for the remaining part of the second sample from P_c excluding the first sample units.

If n_x cannot be chosen small enough to satisfy condition 2 above, then the following sampling plan would be used instead.

Sampling Plan B. A random subsample of size equal to a predetermined proportion r times n_{xc} of the first sample units which fell in P_c is retained as part of the second sample. For the remaining part of the second sample, a simple random sample of size $(1-r)$ n_{xc} is selected without replacement from P_c excluding only the rn_{xc} retained units of the first sample, that is, from a population comprising $N_c - rn_{xc}$ units.

This sampling plan also requires condition 1, provided above, in order to work satisfactorily. Condition 2 is unnecessary since Sampling Plan B ensures with probability 1 that $(1-r)$ n_{xc} units can be selected for the remaining part of the second sample. However, the inconvenience with Sampling Plan B is that some of the $(1-r)$ n_{xc} first sample units initially left aside from the second sample may be selected for the remaining part of the second sample, thus, inflating the effective number of first sample units retained.

On the first occasion, the usual expansion estimator is used to estimate the population total X, with $\frac{N_x}{n_x}$ as the expansion factor multiplying the sample total.

On the second occasion the two components of the population total Y are estimated separately. These two components are Y_c, the subpopulation total for the units common to both occasions, and Y_b, the subpopulation total for the births. The estimate of Y_b is given by $\frac{N_b}{n_b}$ times the total for the births in the sample. For both Sampling Plans A and B, the estimate of Y_c is also the expansion estimator with $\frac{N_c}{n_{xc}}$ as the expansion factor, which is a random variable. This last estimate is unbiased, as long as n_{xc} is positive.

In the next section the variances of the above estimates of total and the covariance between $\hat{X}$ and $\hat{Y}$ are derived for the two sampling plans. The expression for the sampling variance of $\hat{Y} - \hat{X}$ is then obtained.

4. RESULTS

Since n_{xc} is a random variable, only approximate variance formulas can be derived for the estimate of change $\hat{Y} - \hat{X}$. The derivation below is done by first calculating the expectations conditional on n_{xc} and then by averaging on all possible values of n_{xc}.

The sampling variance of $\hat{Y} - \hat{X}$ can be written as

$$\mathrm{Var}\ (\hat{Y}-\hat{X}) = \mathrm{Var}\ (\hat{Y}_b) + \mathrm{Var}\ (\hat{Y}_c) + \mathrm{Var}\ (\hat{X}) - 2\ \mathrm{Cov}\ (\hat{Y}_c, \hat{X}) \tag{5}$$

since the births are sampled independently of the other units. The formulas for the first and third right hand terms are known to be

$$\mathrm{Var}(\hat{Y}_b) = N_b^2 \left(\frac{1}{n_b} - \frac{1}{N_b}\right) S_{y,b}^2 \tag{6}$$

$$\text{and } \mathrm{Var}\ (\hat{X}) = N_x^2 \left(\frac{1}{n_x} - \frac{1}{N_x}\right) S_x^2. \tag{7}$$

To derive formulas for the second and fourth terms of equation (5), we rewrite them as

$$\mathrm{Var}\ (\hat{Y}_c) = E\ (\mathrm{Var}(\hat{Y}_c \mid n_{xc})) + \mathrm{Var}\ (E(\hat{Y}_c \mid n_{xc})) \tag{8}$$

$$\text{and } \mathrm{Cov}\ (\hat{Y}_c, \hat{X}) = E\ (\mathrm{Cov}(\hat{Y}_c, \hat{X} \mid n_{xc})) + \mathrm{Cov}\ (E(\hat{Y}_c \mid n_{xc}), E(\hat{X} \mid n_{xc})). \tag{9}$$

Under condition 1 given in section 3, we have that $E\ (\hat{Y}_c \mid n_{xc})$ can be assumed equal to Y_c. Thus, the second right hand terms of both equations (8) and (9) are negligeable. Hence, only the first right hand terms then need to be evaluated. Their evaluation is next presented.

Let I_i and I_j' denote, respectively, the inclusion indicators of unit i, $i \varepsilon P_x$, for the first sample and of unit j, $j \varepsilon P_y$, for the second sample; and let $E(\cdot \mid n_{xc})$ denote the conditional expectation on n_{xc} over repeated selections of the first and second samples. This will allow to obtain the inclusion conditional probabilities with Tam's results. These conditional probabilities are as follows.

Under Sampling Plan A or B and from Lemma 1 of Tam, the conditional probability of including unit j, $j \varepsilon P_c$, in the second sample is

$$E(I_i' \mid n_{xc}) = \frac{n_{xc}}{N_c} \tag{10}$$

and, the conditional probability of including units j, $j \varepsilon P_c$, and j', $j' \neq j$ and $j' \varepsilon P_c$, jointly in the second sample is

$$E(I_j' I_{j'}' \mid n_{xc}) = \frac{n_{xc}\ (n_{xc} - 1)}{N_c\ (N_c - 1)}. \tag{11}$$

Using equations (8), (10) and (11), it is easily shown that the unconditional variance of $\hat{Y}_c$ is

$$\mathrm{Var}(\hat{Y}_c) = N_c^2 \left(E\left[\frac{1}{n_{xc}}\right] - \frac{1}{N_c}\right) S_{y,c}^2. \tag{12}$$

The parameters of the probability distribution of $\frac{1}{n_{xc}}$ are not known exactly. However, following Sukhatme and Sukhatme (1970), it is possible to derive an approximate expression for its mean. Under the assumption that n_x is large enough such that n_{xc} does not deviate much from $\frac{n_x N_c}{N_x}$, its expectation, they showed that, up to terms of order 2,

$$E\left[\frac{1}{n_{xc}}\right] = \frac{N_x}{n_x N_c}\left[1 + \frac{(N_x - n_x)\ (N_x - N_c)}{(N_x - 1)\ n_x N_c}\right]. \tag{13}$$

In substituting the above expression into equation (12), an approximate formula for the variance of $\hat{Y}_c$ is obtained.

To derive a formula for the covariance term, additional conditional inclusion probabilities need to be provided. These are obtained as follows.

From Lemma 2 of Tam, the joint conditional probability of including unit i, $i \varepsilon P_c$, in both the first and second samples is

$$E(I_i I_i' \mid n_{xc}) = \frac{r n_{xc}}{N_c} \text{ for Sampling Plan A,}$$

$$= \frac{r n_{xc}}{N_c} + \frac{(1 - r)^2 n_{xc}^2}{N_c (N_c - r\ n_{xc})} \tag{14}$$

for Sampling Plan B.

Also from Lemma 2 of Tam, the joint conditional probability of including unit i, $i \varepsilon P_c$, in the first sample and unit j, $j \varepsilon P_c$ and $j \neq i$, in the second sample is

$$E(I_i I_j' \mid n_{xc}) = \frac{n_{xc}^2 - r n_{xc}}{N_c (N_c - 1)} \tag{15}$$

for Sampling Plan A,

$$= \frac{n_{xc}^2 - rn_{xc}}{N_c(N_c-1)} - \frac{(1-r)^2\, n_{xc}^2}{N_c(N_c-1)\,(N_c-rn_{xc})}$$

for Sampling Plan B.

Finally, it is straightforward to show that the joint conditional probability of including unit i, $i\varepsilon P_d$, in the first sample and unit j, $j\varepsilon P_c$, in the second sample is

$$E(I_i I_j' | n_{xc}) = \frac{(n_x - n_{xc})\, n_{xc}}{N_d\, N_c} \tag{16}$$

for Sampling Plan A or B.

Using the above results, the sampling covariance can be obtained. The conditional covariance is first expressed in terms of the inclusion indicators. That is

$$\begin{aligned} \mathrm{Cov}(\hat{Y}_c, \hat{X} | n_{xc}) &= E\left(\left[\frac{N_x}{n_x} \sum_{i\varepsilon P_x} I_i x_i - \sum_{i\varepsilon P_x} x_i\right]\right. \\ &\qquad \left.\left[\frac{N_c}{n_{xc}} \sum_{j\varepsilon P_c} I_j' y_j - \sum_{j\varepsilon P_c} y_j\right] \middle| n_{xc}\right) \\ &= \frac{N_c}{n_x} \frac{N_c}{n_{xc}} \sum_{i\varepsilon P_x} \sum_{j\varepsilon P_c} \Big[E\,(I_i I_j' | n_{xc}) \\ &\qquad - E(I_i | n_{xc})\, E(I_j' | n_{xc}) \Big] x_i y_j . \end{aligned} \tag{17}$$

Using the relationships

$$\sum_{i\varepsilon P_x} \sum_{j\varepsilon P_c} = \sum_{i\varepsilon P_d} \sum_{j\varepsilon P_c} + \sum_{i\varepsilon P_c} \sum_{\substack{j\varepsilon P_c \\ j=i}} + \sum_{i\varepsilon P_c} \sum_{\substack{j\varepsilon P_c \\ j\neq i}} ,$$

$$\sum_{i\varepsilon P_c} \sum_{\substack{j\varepsilon P_c \\ j\neq i}} x_i y_j = \sum_{i\varepsilon P_c} \sum_{j\varepsilon P_c} x_i y_j - \sum_{i\varepsilon P_c} x_i y_i ,$$

Tam's Lemma 3 and equations (14), (15) and (16) with (17) to get the unconditional covariance, we obtain that

$$\mathrm{Cov}(\hat{Y}_c, \hat{X}) = N_x N_c \left(\frac{r}{n_x} - \frac{E\,[n_{xc}]}{n_x\, N_c} \right) S_{xy}$$

for Sampling Plan A.

$$= N_x N_c \left(\frac{r}{n_x} - \frac{E\,[n_{xc}]}{n_x\, N_c} + \frac{(1-r)}{n_x}\, E \left[\frac{n_{xc}}{N_c - rn_{xc}} \right] \right) S_{xy} \tag{18}$$

for Sampling Plan B.

The probability distribution of n_{xc} is hypergeometric and its mean is $\frac{n_x N_c}{N_x}$. The probability distribution of $\frac{n_{xc}}{N_c - rn_{xc}}$ is unknown but an approximation for its mean can be derived using the same approach as for $\frac{1}{n_{xc}}$. Recalling the assumption, which was made earlier, that n_x is large enough so that n_{xc} does not deviate much from $\frac{n_x N_c}{N_x}$, then using the order 2 Taylor's formula one gets

$$E \left[\frac{n_{xc}}{N_c - rn_{xc}} \right] = \frac{n_x}{(N_x - rn_x)} \left[1 + \frac{(N_x - n_x)\,(N_x - N_c)\, N_x rn_x}{(N_x - 1)\,(N_x - rn_x)^2\, N_c} \right] \tag{19}$$

Now in substituting (19) and the expectation of n_{xc} in (18), the unconditional covariance is obtained.

Finally, the variance of $\hat{Y} - \hat{X}$ is given by the substitution of equations (6), (7), (12) and (18) into (5).

5. CONCLUSION AND DISCUSSION

This paper has given approximate sampling variance formulas under two sampling plans for an estimate of change in population totals between the first two occasions of a repeated survey. It was assumed that the change in population total was due to births and deaths in the population and the evolution of the characteristic of interest itself. The estimate was based on simple random samples without replacement affected by rotation.

Using a conditional approach and the sampling plans given in this paper it is possible to derive the sampling variance formula for an estimate of change between any two successive occasions, say h and h+1. Assuming that births and deaths occur between any two occasions, the variance formula would then comprise 2h+1 variance terms and h covariance terms. This, since there would be 2h+1 subpopulations to consider. A practician will not use such an approach because too many terms are involved in the computation. He will prefer to make further approximations.

An approximation that can be made is to assume that the situation of the first two occasions always occur. This means that the formulas derived in this paper are used for any two successive occasions. However, the consequences of this assumption need some investigation.

ACKNOWLEDGEMENTS

The author would like to thank Benoit Quenneville and Michael Hidiroglou, both from Statistics Canada, for their helpful comments and suggestions.

REFERENCES

COCHRAN, W.G. (1977), "Sampling Techniques", New York: John Wiley, 3rd edition.

KISH, L. (1965), "Survey Sampling", New York: John Wiley.

SUKHATME, P.V. and SUKHATME, B.V. (1970), "Sampling Theory of Surveys with Applications", p 27-29, Iowa State University Press, Ames, Iowa, USA.

TAM, S.M. (1984), "On Covariances From Overlapping Samples", The American Statistician, 38, 288-289.