Journal of Official Statistics, Vol. 24, No. 1, 2008, pp. 53–78

On the Covariance Between Related Horvitz-Thompson Estimators

John Wood[1]

National Statistical Institutes often use repeating surveys to estimate measures of change in the levels of the statistics estimated by the surveys. In order to assess the significance of any measured change it is necessary to have an estimate of the variance of this change. An important component of this variance is the covariance between successive estimates of level. This covariance depends very much on the rotation scheme used to determine the overlap between successive samples. This article presents a general solution for this covariance in situations where measures of change are based on Horvitz-Thompson estimators. This solution applies exactly or approximately for a very large class of rotation schemes, including the most commonly applied ones. The article also presents an unbiased estimator for the covariance and compares the general solution with specific solutions previously published.

Key words: Measures of change; repeating survey; rotating survey; rotation scheme; variance.

1. Introduction

Many National Statistical Institutes produce estimates of economic or other statistics on a regular basis, monthly, quarterly or annually. These series of estimates are usually based on rotating surveys, for both practical and methodological reasons. Very often, the focus of interest in these series is the change in the estimate from one period to the next (the change in the volume of retail sales, for example).

Given this interest in the change between two estimates, the literature on estimating the variances of such changes is surprisingly sparse. Economic and social decisions are often made on the basis of estimates of change, and lack of knowledge of the accuracy of these estimates could well lead to erroneous decisions being made, if they are based on measures which are insufficiently accurate.

There is some literature on estimating the variance of changes in index numbers, for example Valliant (1991). However, index numbers are a special case because they are usually based on a fixed panel of units observed over an extended period of time and do not suffer from the complications of changing but overlapping samples within a changing population. A similar restriction applies in area probability surveys, where a fixed panel of primary selection units allows covariances over time to be approximated from the ultimate cluster estimates (see for example, Hansen et al. 1953).

[1] Office for National Statistics, Cardiff Road, Newport NP10 8XG, UK. Email: john.wood@ons.gsi.gov.uk

Acknowledgments: I am grateful to Mike Hidiroglou of Statistics Canada for useful comments that helped to clarify the exposition in this article, to anonymous referees, whose comments engendered considerable improvements to the content and presentation, and to Dr Markus Šova for assistance in deriving the results of Appendix D.

This article addresses the issue of covariances over time in the context of rotating samples but confines its attention to Horvitz-Thompson estimators (Horvitz and Thompson 1952). These estimators are not the most efficient estimators of either levels or changes in levels. Model-based estimators, for example, are usually much more efficient, although they may be subject to model bias and estimation of their variances and covariances depends on the model specified. However, a large and frequently used class of efficient estimators such as ratio estimators and generalised regression estimators is based on the application of functions of Horvitz-Thompson estimators (see for example, Särndal et al. 1992).

As Nordberg (2000) describes, the variances of these more efficient estimators and the variances of changes in these estimators may be approximately expressed as linear combinations of the variances and covariances of the component Horvitz-Thompson estimators. For these, formulae for variances and for covariances between different response variables from the same population and time period are well known (Särndal et al. 1992) but there are no general formulae for the covariances between Horvitz-Thompson estimators for different periods. This is an essential and still missing component for estimating the variances of changes in the general class of efficient estimators based on Horvitz-Thompson estimators. This article concentrates on this specific aspect.

There is some published literature with regard to special cases. Tam (1984) presented covariance formulae for two rotation schemes ("Sampling Plans" in his terminology) in the context of simple random sampling from a fixed population. Laniel (1987) extended these to allow for a changing population, although some approximations were applied. Nordberg (2000) used a different technique, applying a computer program to generate estimates of covariance by averaging over a set of conditional covariances. Berger (2004a, 2004b) based his estimates of the variance of change also on the aggregation of conditional covariances, using a Poisson sampling approximation. However, his method involved a variety of matrix operations and no explicit covariance formulae were presented.

This article presents general formulae for the covariance and an unbiased estimator of the covariance between two related Horvitz-Thompson estimators. These formulae apply exactly for a special class of rotation schemes and approximately for a much larger class that includes the majority of commonly used rotation schemes. Section 2 describes the problem and derives the required formulae. Section 3 analyses the effect of deviations from the conditions under which the covariance formulae are exact and demonstrates the approximate validity of the formulae for a much larger class of rotation schemes. Section 4 compares the general formulae in Section 2 with the more specific formulae produced by Tam (1984), Laniel (1987) and Nordberg (2000) and examines the effect on these special cases of deviations from the conditions required for exact application of the formulae. Section 5 presents a brief discussion and conclusion.

2. Derivation of Covariance Formulae

We consider the situation where a National Statistical Institute, or other organisation, conducts repeating surveys of a finite population. Repetition of the survey may be monthly, quarterly, annually or with some other periodicity. During the course of these surveys, the study population itself may change, through the departure of units that cease to exist in the population (deaths) and the appearance of new units in the population (births). The survey samples will be affected by these changes in the population and by actions of the surveying organisation to replenish the sample to compensate for the losses due to deaths, to ensure that a representative proportion of births is included in the sample and to apply controlled rotation of units out of and into the sample, in order to spread the respondent burden across the whole population.

Often, users wish to compare estimates of population characteristics for different periods. Usually, this comparison will be for adjacent periods but comparisons may also be applied between nonadjacent periods. The development in this article is sufficiently general that it applies to either case (that is, for comparisons between adjacent or nonadjacent periods).

Clearly, these comparisons will be affected by changes in the population and sample, as discussed above, as well as by changes in the observed characteristics themselves. Consider, then, two different periods, which may not be adjacent. We use the subscripts 1 and 2 to distinguish these two periods. Denote the (different) study populations for these two periods as U_1 and U_2. Denote the common population, the set of units that exist in both periods, as $U_c = U_1 \cap U_2$. Denote the corresponding samples as s_1, s_2 and $s_c = s_1 \cap s_2$.

I assume complete knowledge at the time Sample s_1 is drawn of which units exist in Population U_1 and their corresponding inclusion probabilities. I also assume complete knowledge at the time Sample s_2 is drawn of which units exist in Population U_2 and their corresponding inclusion probabilities, both marginal and conditioned on the responses for Sample s_1. I make no other assumptions. In general, I allow the inclusion probabilities for Period 2 to depend on the observed responses for Sample s_1 but the results of this article are exactly valid only when the conditional inclusion probabilities conform to the constraints defined by Conditions (3) and (4) below. Section 3 discusses more fully the effects of varying the extent to which the Period 2 inclusion probabilities depend on the observed responses for Sample s_1.

Consider now a set of survey estimates based on Horvitz-Thompson estimators, as discussed above. For some measurable characteristic, we have, from Samples s_1 and s_2, Estimators $\hat{T}_1$ and $\hat{T}_2$ of population totals T_1 and T_2. We have

$$T_1 = \sum_{k \in U_1} y_{1k}$$

where y_{1k} is the value of the response variable in Period 1 for unit k in Population U_1 and

$$\hat{T}_1 = \sum_{k \in s_1} \frac{y_{1k}}{\pi_{1k}} = \sum_{k \in U_1} \frac{y_{1k}}{\pi_{1k}} I_{1k}$$

where π_{1k} is the probability of including unit k in Sample s_1 and I_{1k} is a random indicator variable, taking the value 1 when unit k is in Sample s_1 and 0 otherwise.

Similar expressions apply to the population total T_2 and its Estimator $\hat{T}_2$.

The variance of the difference between Estimators $\hat{T}_1$ and $\hat{T}_2$ is

$$\operatorname{Var}(\hat{T}_2 - \hat{T}_1) = \operatorname{Var}(\hat{T}_1) + \operatorname{Var}(\hat{T}_2) - 2\operatorname{Cov}(\hat{T}_1, \hat{T}_2) \tag{1}$$

As discussed above, formulae for $\operatorname{Var}(\hat{T}_1)$ and $\operatorname{Var}(\hat{T}_2)$ and their estimators are well known and we concentrate on the covariance term and its estimator.

We have

$$\mathrm{Cov}(\hat{T}_1, \hat{T}_2) = \mathrm{Cov}\left(\sum_{k \in U_1} \frac{y_{1k}}{\pi_{1k}} I_{1k}, \sum_{l \in U_2} \frac{y_{2l}}{\pi_{2l}} I_{2l}\right) = \sum_{k \in U_1} \sum_{l \in U_2} \frac{y_{1k} y_{2l}}{\pi_{1k} \pi_{2l}} \mathrm{Cov}(I_{1k}, I_{2l})$$

$$= \sum_{k \in U_1} \sum_{l \in U_2} \frac{y_{1k} y_{2l}}{\pi_{1k} \pi_{2l}} (\pi_{1k2l} - \pi_{1k} \pi_{2l}) \qquad (2)$$

where we use the term π_{1k2l} to denote the joint inclusion probability that unit k from Population U_1 is in Sample s_1 and unit l from Population U_2 is in Sample s_2.

Determination of π_{1k2l} depends on the rotation scheme used to generate the overlap between Sample s_1 and Sample s_2. In this article, I base the analysis on a central class of rotation schemes which have the following two properties:

$$\Pr(l \in s_2 | k \in s_1 \ \& \ l \in s_1) = \Pr(l \in s_2 | k \notin s_1 \ \& \ l \in s_1) = \Pr(l \in s_2 | l \in s_1) \qquad (3)$$

and

$$\Pr(l \in s_2 | k \in s_1 \ \& \ l \notin s_1) = \Pr(l \in s_2 | k \notin s_1 \ \& \ l \notin s_1) = \Pr(l \in s_2 | l \notin s_1) \qquad (4)$$

for $k \neq l$.

For this class of rotation schemes, the conditional probability that any unit is in Sample s_2, conditioned on Sample s_1, depends only on the presence or absence of that same unit in Sample s_1, not on the presence or absence of any other unit in Sample s_1. Conditions (3) and (4) apply, exactly or approximately, to most rotation schemes in common use by National Statistical Institutes. Section 3 contains a discussion on the general significance of Conditions (3) and (4). Section 4 discusses in more detail the significance of these conditions with regard to the rotation schemes presented in Tam (1984), Laniel (1987) and Nordberg (2000).

For the class of rotation schemes defined by Conditions (3) and (4), we have the result:

Result 2.1

$$\pi_{1k2l} = \begin{cases} \pi_{1k}\pi_{2l} & : l \notin U_1 \ \text{or} \ \pi_{1l} = 1 \\ \pi_{1kl}\pi_{2|1,l} + \dfrac{(\pi_{1k} - \pi_{1kl})(\pi_{2l} - \pi_{1l}\pi_{2|1,l})}{(1 - \pi_{1l})} & : l \in U_1 \ \& \ \pi_{1l} < 1 \end{cases} \qquad (5)$$

where π_{1kl} is the joint inclusion probability that both units k and l are in Sample s_1 and $\pi_{2|1,l} = \Pr(l \in s_2 | l \in s_1)$ is the conditional inclusion probability that unit l is in Sample s_2 given that it was previously selected for Sample s_1.

To demonstrate this, we first express the joint inclusion probability π_{1k2l} as:

$$\begin{aligned} \pi_{1k2l} = & \ \Pr(k \in s_1 \ \& \ l \in s_2) \\ = & \ \Pr(k \in s_1 \ \& \ l \in s_1 \ \& \ l \in s_2) \\ & + \Pr(k \in s_1 \ \& \ l \notin s_1 \ \& \ l \in s_2) \\ = & \Pr(l \in s_2 | k \in s_1 \ \& \ l \in s_1) \qquad (6) \\ & \Pr(k \in s_1 \ \& \ l \in s_1) + \Pr(l \in s_2 | k \in s_1 \ \& \ l \notin s_1) \Pr(k \in s_1 \ \& \ l \notin s_1) \end{aligned}$$

So, if Conditions (3) and (4) apply

$$\pi_{1k2l} = \Pr(l \in s_2 | l \in s_1)\Pr(k \in s_1 \ \& \ l \in s_1) + \Pr(l \in s_2 | l \notin s_1)\Pr(k \in s_1 \ \& \ l \notin s_1) \quad (7)$$

Note that Equation (7) also applies when $k = l$ because $\Pr(l \in s_1 \ \& \ l \in s_1) = \Pr(l \in s_1)$ and $\Pr(l \in s_1 \ \& \ l \notin s_1) = 0$. Moreover, Equation (7) applies for $k = l$ under all circumstances, regardless of whether Conditions (3) and (4) are met for $k \neq l$, because it follows directly from Equation (6).

We may therefore write

$$\pi_{1k2l} = \Pr(l \in s_2 | l \in s_1)\Pr(k \in s_1 \ \& \ l \in s_1) + \Pr(l \in s_2 | l \notin s_1)\Pr(k \in s_1 \ \& \ l \notin s_1) + \delta_{kl} \quad (8)$$

where

$$\begin{aligned} \delta_{kl} = & \{\Pr(l \in s_2 | k \in s_1 \ \& \ l \in s_1) - \Pr(l \in s_2 | l \in s_1)\}\Pr(k \in s_1 \ \& \ l \in s_1) \\ & + \{\Pr(l \in s_2 | k \in s_1 \ \& \ l \notin s_1) - \Pr(l \in s_2 | l \notin s_1)\}\Pr(k \in s_1 \ \& \ l \notin s_1) \\ = & \pi_{1k2l} - \Pr(l \in s_2 | l \in s_1)\Pr(k \in s_1 \ \& \ l \in s_1) - \Pr(l \in s_2 | l \notin s_1)\Pr(k \in s_1 \ \& \ l \notin s_1) \end{aligned}$$

is the difference between the joint inclusion probability π_{1k2l} and the expression in Equation (7). Clearly, $\delta_{kl} = 0 (k \neq l)$ only when Conditions (3) and (4) apply but note that $\delta_{kk} = 0$ is always true, even if Conditions (3) and (4) do not apply, as is obvious from Equation (6).

Under the assumption that Conditions (3) and (4) apply and Equation (7) holds, consider first the special cases $l \notin U_1$ and $\pi_{1l} = 1$. If $l \notin U_1$ then $\Pr(l \in s_1) = 0$ and, from Equation (7), $\pi_{1k2l} = 0 + \pi_{2l}\pi_{1k} = \pi_{1k}\pi_{2l}$. If unit l exists in Population U_1 and $\pi_{1l} = 1$ then $\pi_{1k2l} = \pi_{2l}\pi_{1k} + 0 = \pi_{1k}\pi_{2l}$. This proves the first component of Result 2.1.

For each of these special cases, the associated term in the double summation of Equation (2) contributes nothing to the covariance. Nonzero contributions to the covariance therefore arise only from those units l for which $l \in U_1$ and $\pi_{1l} < 1$. Since, by definition, $l \in U_2$ we also have $l \in U_c$. For these units:

$$\pi_{2l} = \pi_{2|1,l}\pi_{1l} + \Pr(l \in s_2 | l \notin s_1)(1 - \pi_{1l})$$

Hence, since $\pi_{1l} < 1$

$$\Pr(l \in s_2 | l \notin s_1) = \frac{\pi_{2l} - \pi_{2|1,l}\pi_{1l}}{(1 - \pi_{1l})} \quad (9)$$

Also

$$\Pr(k \in s_1 \ \& \ l \notin s_1) = \Pr(k \in s_1) - \Pr(k \in s_1 \ \& \ l \in s_1) = \pi_{1k} - \pi_{1kl} \quad (10)$$

Substituting Expressions (9) and (10) and the definition $\pi_{2|1,l} = \Pr(l \in s_2 | l \in s_1)$ into Equation (7) gives:

$$\pi_{1k2l} = \pi_{2|1,l}\pi_{1kl} + \frac{(\pi_{2l} - \pi_{2|1,l}\pi_{1l})}{(1 - \pi_{1l})}(\pi_{1k} - \pi_{1kl}) \quad (11)$$

With some minor rearrangement, this proves the second component of Result 2.1.

Result 2.1 demonstrates that, for the class of rotation schemes which satisfy Conditions (3) and (4), the joint inclusion probability across the two samples may be expressed in terms of single and joint inclusion probabilities within each sample and the conditional inclusion probabilities $\pi_{2|1,l}$. For this class of rotation schemes, $\pi_{2|1,l}$ does not depend on which units, other than unit l, are included in Sample s_1. The conditional inclusion probabilities $\pi_{2|1,l}$ should therefore be easily determined from the specification of the rotation scheme. There may be some complications in calculating $\pi_{2|1,l}$ if Periods 1 and 2 are very far apart, if Conditions (3) and (4) are not met or if it is desired to allow for unit nonresponse or other uncontrolled effects. In general, this article does not consider such complications and treats the conditional inclusion probabilities $\pi_{2|1,l}$ as known but Section 4 contains examples of these complications relating to the rotation schemes examined in Tam (1984) and Laniel (1987). See also the discussion after Result 2.5 below.

It is now straightforward to apply Result 2.1 in Equation (2) and so obtain the required expression for the population covariance:

Result 2.2

$$\begin{aligned}\mathrm{Cov}(\hat{T}_1, \hat{T}_2) &= \sum_{k \in U_1} \sum_{\substack{l \in U_c \\ (\pi_{1l} \neq 1)}} \frac{y_{1k} y_{2l}}{\pi_{1k} \pi_{2l}} \frac{(\pi_{1kl} - \pi_{1k}\pi_{1l})(\pi_{2|1,l} - \pi_{2l})}{1 - \pi_{1l}} + \sum_{k \in U_1} \sum_{\substack{l \in U_c \\ (l \neq k)}} \frac{y_{1k} y_{2l}}{\pi_{1k} \pi_{2l}} \delta_{kl} \\ &\approx \sum_{k \in U_1} \sum_{\substack{l \in U_c \\ (\pi_{1l} \neq 1)}} \frac{y_{1k} y_{2l}}{\pi_{1k} \pi_{2l}} \frac{(\pi_{1kl} - \pi_{1k}\pi_{1l})(\pi_{2|1,l} - \pi_{2l})}{1 - \pi_{1l}} \qquad (12)\end{aligned}$$

with equality if Conditions (3) and (4) apply.

The first line of this result is exact but it contains the residual covariance term $\sum_{k \in U_1} \sum_{\substack{l \in U_c \\ (l \neq k)}} \frac{y_{1k} y_{2l}}{\pi_{1k} \pi_{2l}} \delta_{kl}$, which is zero only when Conditions (3) and (4) apply and is difficult to estimate when these conditions do not apply. The next section demonstrates that the designation "residual covariance" is appropriate because, under normal circumstances, this term is negligible and the second line of Result 2.2 may be used. Appendix A presents the detailed derivation of the second line and also the derivation of the following unbiased estimator.

Result 2.3

$$\hat{\mathrm{Cov}}(\hat{T}_1, \hat{T}_2) = \sum_{k \in s_1} \sum_{\substack{l \in s_{2(1)} \\ (\pi_{1l} \neq 1)}} \frac{y_{1k} y_{2l} (\pi_{1kl} - \pi_{1k}\pi_{1l})(\pi_{2|1,l} - \pi_{2l})}{\pi_{1k} \pi_{2l} \left[\pi_{1kl}(\pi_{2|1,l} - \pi_{2l}) + \pi_{1k}(\pi_{2l} - \pi_{1l}\pi_{2|1,l}) \right]} \qquad (13)$$

where $s_{2(1)}$ contains those units in Sample s_2 which are in Population U_1 (that is, $s_{2(1)} = s_2 \cap U_1$).

The following alternative version of Equation (13) highlights more clearly the relationship between the estimator and the population covariance of Equation (12):

$$\hat{\mathrm{Cov}}(\hat{T}_1, \hat{T}_2) = \sum_{k \in s_1} \sum_{\substack{l \in s_{2(1)} \\ (\pi_{1l} \neq 1)}} \frac{y_{1k} y_{2l} (\pi_{1kl} - \pi_{1k}\pi_{1l})(\pi_{2|1,l} - \pi_{2l})}{\pi_{1k} \pi_{2l} \left[\pi_{1k}\pi_{2l}(1 - \pi_{1l}) + (\pi_{1kl} - \pi_{1k}\pi_{1l})(\pi_{2|1,l} - \pi_{2l}) \right]}$$

Under simple random sampling without replacement (SRSWOR), we may drop the subscripts k and l because $\pi_{1k} = \pi_{1l} = \pi_1$, $\pi_{2l} = \pi_2$ and $\pi_{2|1,l} = \pi_{2|1}$. We then obtain the following simplified expression for the population covariance, ignoring the residual covariance

Result 2.4

$$\mathrm{Cov}_{\mathrm{SRS}}(\hat{T}_1, \hat{T}_2) = N_1\left(\frac{\pi_{2|1}}{\pi_2} - 1\right)\frac{\sum_{k \in U_c}(y_{1k} - \bar{y}_1)y_{2k}}{N_1 - 1} \tag{14}$$

where N_1 is the number of units in Population U_1 and $\bar{y}_1$ is the mean response in Period 1 for Population U_1.

The corresponding unbiased estimator is

Result 2.5

$$\widehat{\mathrm{Cov}}_{\mathrm{SRS}}(\hat{T}_1, \hat{T}_2) = \frac{n_c\left(1 - \dfrac{\pi_2}{\pi_{2|1}}\right)}{\pi_1\pi_2\left(1 - \dfrac{\pi_1\pi_{2|1}}{\pi_2 n_1}\right)}\left\{\left(\overline{y_1y_2}\right)_{s_c} - \frac{\pi_1\pi_{2|1}n_{2(1)}}{\pi_2 n_c}\bar{y}_{s_1}\bar{y}_{s_{2(1)}}\right\} \tag{15}$$

where n_1 is the number of units in Sample s_1, n_c is the number of units in the common Sample $s_c = s_1 \cap s_2$, $n_{2(1)}$ is the number of units in Sample $s_{2(1)}$, $\bar{y}_{s_1}$ is the mean response for Sample s_1, $\bar{y}_{s_{2(1)}}$ is the mean response for Sample $s_{2(1)}$ and $(\overline{y_1y_2})_{s_c}$ is the mean cross-product of the response variables over the common Sample $s_c = s_1 \cap s_2$.

Appendix B presents the detailed derivation of Equations (14) and (15).

It is worth noting that it is not possible, for the special case of SRSWOR, to express the ratios $\pi_2/\pi_{2|1}$ and $\pi_{2|1}/\pi_2$ as ratios of known sample and population sizes because $\pi_{2|1} = \mathrm{E}[n_c]/\pi_1 N_c = N_1\mathrm{E}[n_c]/n_1 N_c$. Hidiroglou et al. (1995) present an estimator for $\pi_2/\pi_{2|1}$ in this form: $n_1 n_2 N_c/N_1 N_2 n_c$. This or other estimators may be useful in situations where the conditional inclusion probabilities $\pi_{2|1,l}$ are difficult to calculate exactly. Investigation of such estimators is worth further study.

Note that, in Equation (15), the finite population correction factor is $(1 - (\pi_2/\pi_{2|1}))$ and the bias correction factor is $1/(1 - (\pi_1\pi_{2|1}/\pi_2 n_1))$. Hence the implicit number of degrees of freedom for the estimator is $(n_c - (\pi_1\pi_{2|1}n_c/\pi_2 n_1))$, rather than the more familiar $n_c - 1$. This arises because the mean responses in the second term of the covariance expression are based not on the common Sample s_c but on the larger Samples s_1 and $s_{2(1)}$. Note that $\pi_1\pi_{2|1}n_c/\pi_2 n_1 \leq 1$ because $n_c \leq n_1$ and $\pi_2 \geq \pi_1\pi_{2|1}$. Also, the expected value of the premultiplying factor of the product of sample means, $\mathrm{E}[\pi_1\pi_{2|1}n_{2(1)}/\pi_2 n_c]$, is approximately equal to one because $\mathrm{E}[n_{2(1)}] = \pi_2 N_c$ and $\mathrm{E}[n_c] = \pi_1\pi_{2|1}N_c$.

3. Conditions for Valid Application of the Covariance Formulae

The results of Section 2 are strictly valid only when Conditions (3) and (4) apply. We now consider the significance of these conditions and under what circumstances the formulae may be used as approximations when the conditions are not met.

First, note that Condition (3) clearly applies when the conditional inclusion probability $\Pr(l \in s_2|s_1)$ is a constant for all Samples s_1 which contain unit l. Similarly, Condition (4) applies when the conditional inclusion probability $\Pr(l \in s_2|s_1)$ is a constant for all Samples s_1 which do not contain unit l.

It is not necessary that $\Pr(l \in s_2|s_1)$ be a constant over the relevant set of Samples s_1 for Conditions (3) and (4) to apply. For example, if Sample s_1 has a random sample size, it is possible for $\Pr(l \in s_2|s_1)$ to depend on the observed Sample size n_1, provided that the expected value of $\Pr(l \in s_2|s_1)$ over the subset of samples which contain unit k $(k \neq l)$ is the same for all units $k \in U_1 (k \neq l)$. This is most easily achieved by applying simple random sampling with replacement. Devising an unequal probability sampling scheme with this property would be extremely difficult. Note that dependence on the observed sample size n_1 is not equivalent to dependence on the retained sample size $n_{1(2)}$. Dependence on $n_{1(2)}$ does not allow conformity with Conditions (3) and (4). Examples of this are examined in the discussion in Section 4 on the rotation schemes in Laniel (1987) and Nordberg (2000).

It is also possible for Condition (3) to apply when the conditional inclusion probabilities depend on the observed responses for Sample s_1. For example, we may define a fixed, cut-off limit L for which $\pi_{2|1,l}$ depends on whether $y_{1l} < L$ or $y_{1l} \geq L$. Since the resultant, conditioned probabilities do not depend on whether any other unit is in the sample, Condition (3) applies. However, in this situation, the marginal probabilities $\{\pi_{2l}\}$ are not known for those units not included in Sample s_1 because the conditional probabilities $\{\pi_{2|1,l}\}$ are not known. Since such a situation is not acceptable in the context of Horvitz-Thompson estimation, it follows that the imposition of Conditions (3) and (4), and so the exact validity of the results of Section 2, requires that the conditional probabilities $\Pr(l \in s_2|l \in s_1)$ and $\Pr(l \in s_2|l \notin s_1)$ do not depend on the observed responses for Sample s_1.

Unfortunately, it is common for the conditional inclusion probabilities for the selection of Sample s_2 to depend on the outcome of Sample s_1. This may arise as a matter of policy, when the size of Sample s_2 or the marginal inclusion probabilities $\{\pi_{2l}\}$, possibly determined by population auxiliary variables, are dependent on the observed results for Sample s_1 – that is, if Sample s_2 is treated as the second stage of a two-stage sampling scheme. More subtly, the conditional inclusion probabilities may be affected by the random inclusion of deaths in Sample s_1 or births in Sample s_2. The examination of the rotation schemes from Laniel (1987) and Nordberg (2000) in Section 4 identifies examples of these effects.

It is therefore important to assess the effect of deviations from Conditions (3) and (4). This effect is expressed in the residual covariance $\sum_{k\in U_1}\sum_{l\in U_2}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\delta_{kl}$.

Note that, because $\delta_{ll} = 0$, as discussed in the comment after Equation (8), the residual covariance may also be expressed as:

$$\sum_{k\in U_1}\sum_{\substack{l\in U_2\\ l\neq k}}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\delta_{kl} = \sum_{k\in U_1}\sum_{l\in U_2}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\delta_{kl} \tag{16}$$

To assess the magnitude of the residual covariance, note that:

$$\sum_{k\in U_1}\sum_{\substack{l\in U_2\\ l\neq k}}\delta_{kl} = \sum_{k\in U_1}\sum_{l\in U_2}\delta_{kl} = \mathrm{Cov}\left[n_1, \left\{n_2 - \sum_{l\in s_{1(2)}}\frac{\pi_{2|1,l} - \pi_{2l}}{1 - \pi_{1l}}\right\}\right] \tag{17}$$

as demonstrated in Appendix C.

If Sample s_1 has a fixed size, Expression (17) is zero. If n_1 is random, Expression (17) may not be zero because the second covariate in the expression is certainly random.

However, Expression (17) is likely to be very small. First, the dominant covariance terms $\mathrm{Cov}[I_{1l}, I_{2l}]$ are missing from Expression (17) – that is the relevance of the adjustment term $\{-\sum_{l \in s_{1(2)}} (\pi_{2|1,l} - \pi_{2l})/(1 - \pi_{1l})\}$ in the second covariate. Also, each term $(\pi_{2|1,l} - \pi_{2l})/(1 - \pi_{1l})$ may be expressed as

$$\frac{\pi_{1l}(\pi_{2|1,l} - \pi_{2l})}{\pi_{1l}(1 - \pi_{1l})} = \frac{\mathrm{Cov}[I_{1l}, I_{2l}]}{\mathrm{Var}[I_{1l}]}$$

In this light, Expression (17) is likely to be small in most circumstances. If the inclusion indicators I_{1l} and I_{2l} are highly correlated, the second covariate in Expression (17) is close to $n_2 - n_{1(2)}$, which is the additional number of units required to make up the second sample and is unlikely to be correlated with the total number of units in the first sample (unless, for example, n_2 is fixed, in which case there would be a negative correlation between n_1 and $n_2 - n_{1(2)}$, but this would be extremely unusual in practice). If the inclusion indicators I_{1l} and I_{2l} are not highly correlated, the selection of Sample s_2 is almost unrelated to the selection of Sample s_1 and Expression (17) is also small.

Obviously, the actual value of Expression (17) depends on the sample designs and rotation scheme used and there may be occasions when this value is not negligible, but there is clearly a wide range of circumstances in which the value of Expression (17) is exactly or approximately zero. This provides an aid to assessing the magnitude of the residual covariance in Expression (16). If Expression (17) is zero or negligible and there is negligible correlation between δ_{kl} and $y_{1k}y_{2l}/\pi_{1k}\pi_{2l}$, the residual covariance is also negligible. In these circumstances, the results of Section 2 may be applied as approximations.

Further empirical research will be necessary to identify the circumstances under which the results of Section 2 are not reliable but such circumstances are likely to be rare. If the marginal inclusion probabilities are approximately proportional to size, the terms $\{y_{1k}y_{2l}/\pi_{1k}\pi_{2l}\}$ are approximately constant, the correlation between δ_{kl} and $y_{1k}y_{2l}/\pi_{1k}\pi_{2l}$ is negligible and the residual covariance should also be negligible. If the marginal inclusion probabilities are constant, as under SRSWOR, the correlation between δ_{kl} and $y_{1k}y_{2l}/\pi_{1k}\pi_{2l}$ should also be negligible, again leading to a negligible residual covariance. The examples in the next section contain explicit formulae for the residual covariance in the context of SRSWOR sampling. These formulae demonstrate that, for these cases at least, the residual covariances are zero or negligible.

4. Comparison of Covariance Formulae

4.1. Comparison with Tam (1984)

Tam (1984) considered the special case of SRSWOR for a constant population (that is, $U_1 = U_2 = U$ and $N_1 = N_2 = N$) under three different rotation schemes (or *Sampling Plans*, in his terminology). These are described briefly below, to highlight the relevant attributes for this comparison. See Tam (1984) for the full descriptions.

For each sampling plan, a fixed n_1 units in Sample s_1 are first selected from Population U by SRSWOR. A fixed n_2 units for Sample s_2 are then selected according to the sampling plan, as follows.

Sampling Plan A: a fixed n_c units are selected by SRSWOR from Sample s_1 and a further $n_2 - n_c$ units are selected by SRSWOR from the population excluding those units in Sample s_1.

Sampling Plan B: a fixed n_c units are selected by SRSWOR from Sample s_1 and a further $n_2 - n_c$ units are selected by SRSWOR from the population excluding only the previously selected n_c units. That is, the $n_1 - n_c$ units rejected from Sample s_1 are eligible for reselection in Sample s_2, which is not the case in *Sampling Plan A*.

Sampling Plan C: the fixed n_2 units are selected by SRSWOR from Population U, independently of Sample s_1.

We compare the general result given by Equation (14), under SRSWOR, with the results from Tam (1984), under the assumption that Equation (14) is valid. The validity of Equation (14) is considered at the end of this subsection.

Assuming that $N_1 = N_2 = N$, Equation (14) becomes:

$$\mathrm{Cov}_{\mathrm{SRS}}(\hat{T}_1, \hat{T}_2) = N\left(\frac{\pi_{2|1}}{\pi_2} - 1\right)\frac{\sum\limits_{k \in U}(y_{1k} - \bar{y}_1)(y_{2k} - \bar{y}_2)}{(N-1)} \tag{18}$$

Modifying Tam's notation to conform to the notation used above and multiplying by N^2 to allow for the fact that Tam's formulae are expressed in terms of means, Tam (1984) presents the covariance term in the general form

$$(1 - f)\frac{N^2 n_c}{n_1 n_2} S_{xy} \tag{19}$$

where $S_{xy} = \sum\limits_{k \in U}(y_{1k} - \bar{y}_1)(y_{2k} - \bar{y}_2)/(N-1)$ and f depends on the sampling plan.

Clearly, the term S_{xy} is identical to the rightmost component of Equation (18) and the premultiplying factor in Equation (19) may be written as $N((n_c/n_1\pi_2) - (fNn_c/n_1 n_2))$, since $\pi_2 = n_2/N$.

Under *Sampling Plan A*, $f = n_1 n_2/Nn_c$ and the premultiplying factor is $N((n_c/n_1\pi_2) - 1)$. Clearly, in this plan $\pi_{2|1} = n_c/n_1$ and Tam's result accords with Equation (18). Note also that $\Pr(l \in s_2 | l \notin s_1) = (n_2 - n_c)/(N - n_1)$.

Under *Sampling Plan B*, $f = (n_1 n_2/Nn_c) - ((n_1 - n_c)(n_2 - n_c)/(N - n_c)n_c)$ and the premultiplying factor is

$$N\left\{\frac{n_c}{n_1\pi_2} - 1 + \frac{N(n_1 - n_c)(n_2 - n_c)}{(N - n_c)n_1 n_2}\right\} = N\left\{\frac{n_c}{n_1\pi_2} + \frac{\left(1 - \frac{n_c}{n_1}\right)(n_2 - n_c)}{(N - n_c)\pi_2} - 1\right\}$$

In this plan, $\pi_{2|1} = (n_c/n_1) + (1 - n_c/n_1)((n_2 - n_c)/(N - n_c))$ and, again, Tam's result accords with Equation (18). Note also that $\Pr(l \in s_2 | l \notin s_1) = (n_2 - n_c)/(N - n_c)$.

Under *Sampling Plan C*, $f = 1$ and the covariance is zero. In this case, $\pi_{2|1} = \pi_2$ and the result is also in accord with Equation (18).

It is clear that the comparisons above are exactly valid because the relevant conditional inclusion probabilities are constant, as shown above, Conditions (3) and (4) are met and Equation (14) applies exactly.

4.2. Comparison with Laniel (1987)

Laniel (1987) considered the special case of SRSWOR for a changing population using two different rotation schemes, or *sampling plans*. These are described briefly below, using the notation of this article, to highlight the relevant attributes for this comparison. See Laniel (1987) for the full descriptions.

For each sampling plan, a fixed n_1 units in Sample s_1 are first selected from Population U_1 by SRSWOR. For Sample s_2, a fixed n_b units are selected from the births by SRSWOR and $n_{1(2)}$ units are selected from the common Population U_c according to the sampling plan, as follows.

> *Sampling Plan A*: a predetermined proportion r of the $n_{1(2)}$ units in Sample $s_{1(2)} = s_1 \cap U_2$ is selected by SRSWOR. That is, selection is from the $n_{1(2)}$ units which were selected for Sample s_1 and remain in Population U_2. A further $(1 - r)n_{1(2)}$ units are selected by SRSWOR from the common Population U_c excluding those units in Sample s_1.
>
> *Sampling Plan B*: a predetermined proportion r of the $n_{1(2)}$ units in Sample $s_{1(2)} = s_1 \cap U_2$ is selected by SRSWOR. A further $(1 - r)n_{1(2)}$ units are selected randomly from the common Population U_c excluding only the previously selected $rn_{1(2)}$ units. That is, the $(1 - r)n_{1(2)}$ units rejected from Sample s_1 are eligible for reselection in Sample s_2, which is not the case in *Sampling Plan A*.

It is of interest to note that $rn_{1(2)}$ is equivalent to n_c in Tam (1984) and Tam's ratio n_c/n_1 is equivalent to r. This is because the sampling plans in Laniel (1987) are extensions of those in Tam (1984) to cover overlapping but different populations. For a constant population, as considered in Tam (1984), $n_{1(2)} = n_1$. For a changing population, as considered in Laniel (1987), $n_{1(2)}$ is random. This is relevant to the applicability of the results of Section 2, as considered at the end of this subsection.

A feature of Laniel's sampling plans is that, for Period 2, births are sampled separately from units in the common population. In addition, the number of units in Sample s_2 which belong to the common Population U_c is set equal to the (random) number of units in Sample s_1 which belong to U_c. That is, $n_{2(1)} = n_{1(2)}$. A consequence of this, acknowledged implicitly in Equation 10 of Laniel (1987), is that, for units in the common population, $\pi_2 = \pi_1$.

To compare Equation (14) with the results from Laniel (1987), we can rewrite it as

$$\mathrm{Cov}_{\mathrm{SRS}}(\hat{T}_1, \hat{T}_2) = N_1\left(\frac{\pi_{2|1}}{\pi_2} - 1\right)\frac{\sum_{k\in U_c}(y_{1k} - \bar{y}_{1c})(y_{2k} - \bar{y}_{2c}) + (\bar{y}_{1c} - \bar{y}_1)\sum_{k\in U_c} y_{2k}}{(N_1 - 1)} \tag{20}$$

where $\bar{y}_{1c}$ and $\bar{y}_{2c}$ are the means over the common Population U_c of the $\{y_{1k}\}$ and $\{y_{2k}\}$, respectively.

The equivalent expressions from Laniel (1987), modified to use the notation of this article, may be written as

$$\mathrm{Cov}_{\mathrm{SRS}}(\hat{T}_1, \hat{T}_2) = N_c\left(\frac{R}{\pi_1} - 1\right)\frac{\sum\limits_{k\in U_c}(y_{1k} - \bar{y}_{1c})(y_{2k} - \bar{y}_{2c})}{(N_c - 1)} \tag{21}$$

where R depends on the sampling plan.

Consider first the weighting factors $((\pi_{2|1}/\pi_2) - 1)$ and $((R/\pi_1) - 1)$.

Under *Sampling Plan A*, $R = r$ and r is the predetermined proportion of the $n_{1(2)}$ units in Sample $s_{1(2)}$ which is selected by SRSWOR for inclusion in Sample s_2. Since, for this sampling plan, these are the only units from Sample s_1 which are included in Sample s_2, r is therefore, in principle, the conditional probability that a unit in the first sample is selected for the second sample. If $rn_{1(2)}$ is an integer, this result is exact. Laniel (1987) does not discuss how to apply SRSWOR in the much more likely event that $rn_{1(2)}$ is not an integer. One possibility, denoting the integer part of $rn_{1(2)}$ as $[rn_{1(2)}]$, is to select either $[rn_{1(2)}]$ or $[rn_{1(2)}] + 1$ units with probabilities $[rn_{1(2)}] + 1 - rn_{1(2)}$ and $rn_{1(2)} - [rn_{1(2)}]$, respectively. The resultant, conditional inclusion probability for each unit is then

$$\{[rn_{1(2)}] + 1 - rn_{1(2)}\}\frac{[rn_{1(2)}]}{n_{1(2)}} + \{rn_{1(2)} - [rn_{1(2)}]\}\frac{[rn_{1(2)}] + 1}{n_{1(2)}} = \frac{rn_{1(2)}}{n_{1(2)}} = r$$

On this basis, we therefore have $\pi_{2|1} = r$. It is clear from Equation 14 of Laniel (1987) that this is the intention underlying Laniel's analysis, whatever modification of SRSWOR is applied. In addition, as discussed above, we have $\pi_2 = \pi_1$ for units in the common population. So $((r/\pi_1) - 1) = ((\pi_{2|1}/\pi_2) - 1)$ and Laniel's weighting factor accords with Equation (20).

Under *Sampling Plan B*, $R = r + (1 - r)^2\mathrm{E}[n_{1(2)}/(N_c - rn_{1(2)})]$.

Under this sampling plan, using the same argument as under *Sampling Plan A* and with reference to Equation 14 of Laniel (1987),

$$\pi_{2|1} = r + (1 - r)\mathrm{E}[(1 - r)n_{1(2)}/(N_c - rn_{1(2)})] = R$$

and, again, Laniel's weighting factor accords with Equation (20).

[In fact, Laniel (1987) presents the expression equivalent to

$$R = r + (1 - r)\mathrm{E}[n_{1(2)}/(N_c - rn_{1(2)})]$$

but this inadvertently omits the squaring of the term $(1 - r)$. The squared term $(1 - r)^2$ is evidently necessary from the version for *Sampling Plan B* of Equations 14 and 15 in Laniel (1987).]

Although the weighting factors are equivalent, there are two differences between Equations (20) and (21). The bias correction factors $N_1/(N_1 - 1)$ and $N_c/(N_c - 1)$ are different and Equation (21) omits the term $(\bar{y}_{1c} - \bar{y}_1)\sum\limits_{k\in U_c} y_{2k}$. These differences relate, respectively, to differences in the sizes and mean values in Period 1 of Populations U_1 and U_c. These differences are negligible if both populations are large and overlap to a large extent (that is, there are few deaths). Both these conditions are necessary to justify Laniel's simplifying approximation that the probability of there being no units from the common population in the first sample is negligible. In essence, Laniel (1987) has assumed that U_1 and U_c are almost identical.

Considering now the applicability of Equation (14) to Laniel's *sampling plans*, note first that the sampling schemes of Laniel (1987) and Nordberg (2000) have the common features that n_1 is fixed and sampling is by SRSWOR. So Expression (17) is zero and there should be negligible correlation between δ_{kl} and $y_{1k}y_{2l}/(\pi_{1k}\pi_{2l})$. The residual covariance should therefore also be close to zero and, on first inspection, the results of Section 2 may be applied as approximations. We now consider Laniel's *sampling plans* in more detail.

For *Sampling Plan A*, $\pi_{2|1} = r$, a predetermined constant, as demonstrated above, and Condition (3) applies. However, Condition (4) does not apply universally. It does apply for births, because they are sampled independently of units in Population U_1. For unit

$$l \in U_1 \text{ with } l \notin s_1,\ \Pr(l \in s_2|s_1) = \frac{(1-r)n_{1(2)}}{N_c - n_{1(2)}}$$

which is not a constant but a random variable. The conditional probability

$\Pr(l \in s_2|k \in s_1\ \ \&\ \ l \notin s_1)$ therefore depends on whether unit k is a death or survives into the common population. Although the presence of unit k in Sample s_1 affects the value of $n_{1(2)}$, this value does not depend on which particular unit has died or survived because sampling is by SRSWOR. The set $\{\delta_{kl} : l \in U_c\}$ therefore includes only two distinct numerical values, which we denote as δ_c for units k in the common population and δ_d for deaths.

Algebraic development presented in Appendix D leads to the following approximate expression for the relative error in Expression (14) (defining the relative error as the residual covariance divided by Expression (14))

$$\frac{(1-r)}{(\pi_{2|1}-\pi_2)}\frac{(N_1-N_c)\{N_c(\bar{y}_{1c}-\bar{y}_{1d})\bar{y}_{2c}-C_{12}\}}{\{N_1(N_c-1)C_{12}+N_c(N_1-N_c)(\bar{y}_{1c}-\bar{y}_{1d})\bar{y}_{2c}\}}\frac{(N_c-1)}{(N_c-2)}\left\{1-\frac{(N_1-N_c)}{(N_1-n_1)(N_c-2)}\right\} \tag{22}$$

where $\bar{y}_{1d}=\sum_{\substack{k\in U_1\\ k\notin U_c}} y_{1k}/(N_1-N_c)$ is the mean response in Period 1 for deaths and: $C_{12}=\sum_{k\in U_c}(y_{1k}-\bar{y}_{1c})y_{2k}/(N_c-1)$ is the finite population covariance between the response in Period 1 and the response in Period 2 for units in the common population.

If $N_c(\bar{y}_{1c}-\bar{y}_{1d})\bar{y}_{2c}$ is negligible compared to C_{12}, this ratio is of order $(N_1-N_c)/(N_1N_c)$. If $N_c \approx N_1$, as required by Laniel's *sampling plans*, the ratio is of order N_1^{-2}, so the residual covariance is immaterial even for relatively small values of N_1.

If $N_c(\bar{y}_{1c}-\bar{y}_{1d})\bar{y}_{2c}$ is dominant, the ratio approaches the value $(1-r)/(\pi_{2|1}-\pi_2)$ and is determined by the rotation and inclusion probabilities. For *Sampling Plan A*, $\pi_{2|1} = r$ and the ratio is less than one if $r > (1+\pi_2)/2$. Note that, for Laniel's *sampling plans*, the term $N_c(\bar{y}_{1c}-\bar{y}_{1d})\bar{y}_{2c}$ is the main mechanism through which correlation between δ_{kl} and $y_{1k}y_{2l}/(\pi_{1k}\pi_{2l})$ gives rise to potentially nontrivial values for the residual covariance.

For relatively small common population sizes, the rightmost adjustment term $\{1-((N_1-N_c)/(N_1-n_1)(N_c-2))\}$ acts to reduce the magnitude of the residual covariance, counteracting the tendency of the other terms in Expression (22) to increase the residual covariance when the common population size is small. This gives further support to the claim that the residual covariance is usually small. However, the approximate relative error above is valid only for relatively large values of N_c. For small values of N_c, a more detailed and rigorous analysis would be required than is presented here.

Under *Sampling Plan B*, $\pi_{2|1} = R = r + (1 - r)\mathrm{E}[(1 - r)n_{1(2)}/(N_c - rn_{1(2)})$, so even Condition (3) is not met because for units $l \in U_1$ with $l \in s_1$

$$\Pr(l \in s_2 | s_1) = r + \frac{(1 - r)^2 n_{1(2)}}{N_c - n_{1(2)}}$$

which is a random variable. However, Equation (28) in Appendix D still applies and further algebraic development in Appendix D presents derivations for the required conditional inclusion probabilities. These are more numerous and more complicated than for *Sampling Plan A* but they have the same general form and the general conclusions from the discussion for *Sampling Plan A* also apply for *Sampling Plan B*.

4.3. *Comparison with Nordberg (2000)*

Comparison with Nordberg (2000) is difficult because he does not present explicit formulae for the covariance or its estimator. The nearest equivalent is an expression for the covariance estimator, conditioned on the sample sizes (his Equation 3.9). The unconditional estimator is obtained by taking the mean of the conditioned estimators over a set of computerised simulations of sample sizes.

The context is one of stratified simple random sampling, where the two populations are subject to different stratifications. The conditional covariance estimator is expressed as a sum of a standard formula over all possible combinations of strata for Population U_1 with strata for Population U_2. For the present purpose, we may regard any one such combination as representing two overlapping populations under SRSWOR, allowing comparison of Nordberg's Equation 3.9 with Equation (15) above.

Using the notation of this article, each term in the double summation for Nordberg's Equation 3.9 is equivalent to:

$$\frac{n_c\left(1 - \dfrac{n_{1(2)}n_{2(1)}}{n_c N_c}\right)}{\pi_1 \pi_2\left(1 - \dfrac{n_c}{n_{1(2)}n_{2(1)}}\right)}\left\{\left(\overline{y_1 y_2}\right)_{s_c} - \bar{y}_{s_{1(2)}}\bar{y}_{s_{2(1)}}\right\} \tag{23}$$

This is similar in form to Equation (15) but the finite population correction factor and the bias correction factor both contain random sample sizes (Formula (23) is, of course, conditioned on the sample sizes). The premultiplying factor for the product of sample means is exactly 1. We have already noted in Section 2 that the expected value of the corresponding factor in Equation (15) is approximately 1. Finally, the mean response for Sample s_1 is taken over those units in the common population only, whereas the corresponding mean in Equation (15) is for the whole of Sample s_1.

Since Formula (23) is conditioned on the sample sizes, we may take the expected values of these sample sizes in the finite population correction factor and the bias correction factor to obtain an approximate direct comparison with Equation (15). This approximation is

$$\frac{n_c\left(1-\dfrac{\pi_2}{\pi_{2|1}}\right)}{\pi_1\pi_2\left(1-\dfrac{\pi_{2|1}}{\pi_2 N_c}\right)}\left\{\left(\overline{y_1 y_2}\right)_{s_c}-\bar{y}_{s_{1(2)}}\bar{y}_{s_{2(1)}}\right\}$$

For this approximation, the finite population correction factor is equivalent to that in Equation (15) but the bias correction factor contains the term N_c, rather than the term $n_1/\pi_1 = N_1$ of Equation (15). This reflects the reduced number of degrees of freedom for estimating the sample covariance because Nordberg's sample mean for the first sample is only taken over units in the common population. This comparison is similar to the comparison of Equation (14) against the covariance formula from Laniel (1987), which also uses N_c instead of N_1. This shared difference arises because both Laniel (1987) and Nordberg (2000) consider the decomposition of the covariance into the expected value of conditional covariances and the covariance of conditional expected values

$$\mathrm{Cov}(\hat{T}_1, \hat{T}_2) = \mathrm{E}_\Omega\left[\mathrm{Cov}(\hat{T}_1, \hat{T}_2|\Omega)\right] + \mathrm{Cov}_\Omega\left[\mathrm{E}\left(\hat{T}_1|\Omega\right), \mathrm{E}\left(\hat{T}_2|\Omega\right)\right]$$

where Ω is a random set of sample sizes.

Laniel (1987) assumes that the second component, the covariance of the conditional expected values, is negligible. Nordberg (2000) estimates this component by computer simulations, a consequence of which is that a straightforward comparison with the formulae presented in this article is not possible.

The contribution from this second component arises because the random selection of relatively small or large numbers of deaths in the sample has an effect on the size of the common sample and thus a potential effect on the covariance. The covariance is affected only if the mean response for deaths is different from the mean response for the other units in Population U_1. The final term in Equation (20) specifies the magnitude and direction of this effect. For the usual occurrence that $\bar{y}_{2c} > 0$, a relatively low mean response for deaths will produce an increase in the covariance and a relatively high mean response for deaths will produce a reduction in the covariance.

To assess the applicability of Equation (15) to Nordberg's sampling scheme, we need to understand how Nordberg's rotation scheme affects the conditional inclusion probabilities. Nordberg (2000) applies rotation based on the use of permanent random numbers (PRNs), a common occurrence in National Statistical Institutes, but does not describe the rotation method applied in any detail. In principle, the start of the PRN range is moved forward a predefined amount and thereby defines the constant, conditional inclusion probability $\Pr(l \in s_2|l \in s_1)$. The end of the PRN range is then moved forward to select sufficient units to produce the intended sample size for Period 2. In practice, it can happen that, if there is a large excess of births over deaths in this PRN range, more units need to be removed from rather than added to the sample to achieve the intended sample size for Period 2. Since the probability of this happening depends on whether unit $k \in s_1$ is a death or survives, it follows that Conditions (3) and (4) are violated when this occurrence can arise.

To assess this effect, it is helpful to redefine the PRN rotation scheme in a manner similar to the descriptions in Laniel (1987), using the following steps.

1 We select by SRSWOR a sample of size n_1 from Population U_1.

2 For each of the $n_{1(2)}$ units which remain in Population U_2, we randomly decide whether to retain that unit in the sample, for each unit independently and with probability r of retention and probability $1 - r$ of rejection. This process corresponds to shifting the start of the PRN range forward. Independence arises because the PRNs are assigned to units independently. The effect is Bernoulli sampling from Sample $s_{1(2)}$. Denote the resultant random number of retained units in the common sample as n_c^*.

3 For each of the $N_2 - N_c$ births in Population U_2, we randomly decide whether to add that unit to the sample, for each unit independently and with probability $r\pi_1$. This process corresponds to the appearance of births in the shifted PRN range. Again, independence arises because the PRNs are assigned to units independently and the effect is Bernoulli sampling, this time from the population of births. Denote the resultant number of selected births as n_b^* and the total number selected as $n_2^* = n_b^* + n_c^*$.

4 The final action depends on the value of n_2^*.

a If $n_2^* < n_2$, select $n_2 - n_2^*$ units by SRSWOR from the $N_2 - n_{1(2)} - n_b^*$ units in Population U_2 not already selected for Sample s_1 or s_2. This corresponds to the process of extending the end of the PRN range until the desired sample size n_2 is achieved. I ignore the implausible possibility that the end of the PRN range circles round to pass the start of the original PRN range. That is, I assume that $n_2 \le N_2 - (n_{1(2)} - n_c^*)$ for all possible values of $n_{1(2)}$ and n_c^*. This condition is met if $n_2 \le N_2 - n_1$.

b If $n_2^* = n_2$, no further action is required because the desired sample size n_2 has been achieved. In the analysis below, for convenience, I shall treat this action as a special case of action 4a above, with $n_2 - n_2^* = 0$.

c If $n_2^* > n_2$, select n_2 units by SRSWOR from the n_2^* units selected in Steps 2 and 3. This corresponds to the process of reducing the length of the PRN range until the desired sample size n_2 is achieved.

We thus have

$$\begin{aligned}\pi_{2|1} &= \mathrm{E}\left[\frac{n_c^* n_2}{n_{1(2)} n_2^*}\,|n_2^* > n_2\right]\Pr(n_2^* > n_2) + \mathrm{E}\left[\frac{n_c^*}{n_{1(2)}}\,|n_2^* \le n_2\right]\left\{1 - \Pr(n_2^* > n_2)\right\} \\ &= r - \mathrm{E}\left[\frac{n_c^*(n_2^* - n_2)}{n_{1(2)} n_2^*}\,|n_2^* > n_2\right]\Pr(n_2^* > n_2) \qquad (24)\end{aligned}$$

So Condition (3) is met only if $\Pr(n_2^* > n_2) = 0$.

We also have

$$\Pr(l \in s_2 | l \in U_1 \ \& \ l \notin s_1) = \mathrm{E}\left[\frac{n_2 - n_2^*}{N_2 - n_{1(2)} - n_b^*}\,|n_2^* \le n_2\right]\left\{1 - \Pr(n_2^* > n_2)\right\} \qquad (25)$$

and

$$\begin{aligned}
\Pr(l \in s_2 | l \notin U_1) =& \mathrm{E}\left[r\pi_1 \frac{n_2}{n_2^*} | n_2^* > n_2 \right] \Pr(n_2^* > n_2) \\
&+ \mathrm{E}\left[r\pi_1 + \frac{n_2 - n_2^*}{N_2 - n_{1(2)} - n_b^*} | n_2^* \leq n_2 \right] \left\{ 1 - \Pr(n_2^* > n_2) \right\} \\
=& r\pi_1 - r\pi_1 \mathrm{E}\left[\frac{n_2^* - n_2}{n_2^*} | n_2^* > n_2 \right] \Pr(n_2^* > n_2) \\
&+ \mathrm{E}\left[\frac{n_2 - n_2^*}{N_2 - n_{1(2)} - n_b^*} | n_2^* \leq n_2 \right] \{ 1 - \Pr(n_2^* > n_2) \}
\end{aligned} \tag{26}$$

So Condition (4) is not met, even if $\Pr(n_2^* > n_2) = 0$. We thus have a situation analogous to that for Laniel's *sampling plans*. If $\Pr(n_2^* > n_2) = 0$, Condition (3) applies but not Condition (4). This is similar to the situation under Laniel's *Sampling Plan A*, and the same arguments used there may be applied here.

If $\Pr(n_2^* > n_2) > 0$, neither Condition (3) nor Condition (4) applies and a more complicated analysis, similar to that under Laniel's *Sampling Plan B*, is required. In fact, the analysis for Nordberg's sampling scheme is even more complicated than for Laniel's because, as is evident from Equation (26), $\delta_{kl} \neq 0$ also for births. However, this difference is essentially one of scale because it is clear from Equations (24) to (26) that the $\{\delta_{kl}\}$ for births are almost exactly equal to linear combinations of the $\{\delta_{kl}\}$ for units in the common population. The general arguments made above in this and the previous subsection therefore still apply.

5. Discussion and Conclusion

This article has presented general formulae for the covariance and an unbiased estimator of the covariance between Horvitz-Thompson estimators obtained in different periods of a repeating survey. These formulae apply exactly or almost exactly for a large class of rotation schemes, which should encompass all but the most unusual and exceptional circumstances. The formulae are also consistent, insofar as this can be determined, with previously published formulae for special cases, as presented in Tam (1984), Laniel (1987) and Nordberg (2000).

The fundamental Result 2.1 expresses the important joint inclusion probability π_{1k2l} in terms of single and joint inclusion probabilities within each sample and the conditional inclusion probabilities $\pi_{2|1,l}$. This result is exact when Conditions (3) and (4) apply. When Conditions (3) and (4) do not apply, Equation (6) provides a means of determining π_{1k2l} from the joint inclusion probabilities for Sample s_1 and conditional inclusion probabilities for Sample s_2.

Result 2.2 provides an expression, with unequal inclusion probabilities, for the population covariance when the populations are known. Result 2.3 provides a corresponding unbiased covariance estimator. Results 2.4 and 2.5 provide, respectively, simplified versions of these for the special case of simple random sampling without replacement.

All these results are exact when Conditions (3) and (4) apply. The analyses in Sections 3 and 4 demonstrate that they are also approximately correct for a much wider range of circumstances. Further empirical work to identify the circumstances for which these results are unreliable would be useful. As the examples in Section 4 demonstrate, calculating exact covariance formulae when Conditions (3) and (4) do not apply is excessively complicated and it would be helpful to be able to use the relatively simple formulae of Section 2.

The analysis at the end of Subsection 4.2 on the residual covariance for Laniel's *Sampling Plan A* suggests that, for this sampling plan and possibly more generally, the conditions which lead to relatively large values of the residual covariance are, singly or in combination:

- a relatively small common population
- a large difference between units in the common population and deaths for the mean value of the response variable in Period 1
- large inclusion probabilities for Period 2
- small probabilities of retaining units selected for Sample s_1.

These conditions are most likely to arise when Periods 1 and 2 are far apart, in which case the applicable rotation scheme is most likely to be the net effect of a series of defined rotation schemes over intervening periods. In such circumstances, the relative simplicity of the results in Section 2 is even more advantageous. These circumstances also give rise to low values of the covariance between estimates for Periods 1 and 2, so that even large relative errors in the covariance formulae of Section 2 may be unimportant in the context of estimating the variance of changes in estimates.

The covariance estimators in Results 2.3 and 2.5 require the combination of data from samples and populations in different periods. Such estimators may be difficult to calculate in practice. Further work, to determine under what conditions simpler, biased estimators may be preferable, is now possible. The covariance formulae in Results 2.2 and 2.4 enable the application of simulation studies to estimate the biases and mean squared errors of alternative estimators. Such studies would ideally be based on sample designs and rotation schemes which satisfy Conditions (3) and (4), to ensure exact validity of the covariance formulae.

Appendix A. Derivation of the Covariance and Its Estimator

From Equations (2) and (8), we have:

$$\mathrm{Cov}(\hat{T}_1,\hat{T}_2)=\sum_{k\in U_1}\sum_{l\in U_2}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\left\{\begin{array}{l}\Pr(l\in s_2|l\in s_1)\Pr(k\in s_1\,\&\,l\in s_1)\\[2ex]+\Pr(l\in s_2|l\notin s_1)\Pr(k\in s_1\,\&\,l\notin s_1)\\[2ex]+\delta_{kl}-\pi_{1k}\pi_{2l}\end{array}\right\}$$

$$=\sum_{k\in U_1}\sum_{l\in U_2}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\left\{\begin{array}{l}\Pr(l\in s_2|l\in s_1)\Pr(k\in s_1\,\&\,l\in s_1)\\[2ex]+\Pr(l\in s_2|l\notin s_1)\Pr(k\in s_1\,\&\,l\notin s_1)\\[2ex]-\pi_{1k}\pi_{2l}\end{array}\right\}+\sum_{k\in U_1}\sum_{l\in U_2}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\delta_{kl}$$

Ignoring the residual covariance $\sum\limits_{k\in U_1}\sum\limits_{l\in U_2}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\delta_{kl}$ and using Result 2.1, we have

$$\mathrm{Cov}(\hat{T}_1,\hat{T}_2)=\sum_{k\in U_1}\sum_{\substack{l\in U_c\\(\pi_{1l}\neq 1)}}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\left\{\pi_{1kl}\pi_{2|1,l}+\frac{(\pi_{1k}-\pi_{1kl})(\pi_{2l}-\pi_{1l}\pi_{2|1,l})}{(1-\pi_{1l})}-\pi_{1k}\pi_{2l}\right\}$$

$$=\sum_{k\in U_1}\sum_{\substack{l\in U_c\\(\pi_{1l}\neq 1)}}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\left\{\begin{matrix}\pi_{1kl}\pi_{2|1,l}-\pi_{1kl}\pi_{2|1,l}\pi_{1l}+\pi_{1k}\pi_{2l}-\pi_{1k}\pi_{1l}\pi_{2|1,l}\\ \dfrac{-\pi_{1kl}\pi_{2l}+\pi_{1kl}\pi_{1l}\pi_{2|1,l}-\pi_{1k}\pi_{2l}+\pi_{1k}\pi_{2l}\pi_{1l}}{(1-\pi_{1l})}\end{matrix}\right\}$$

$$=\sum_{k\in U_1}\sum_{\substack{l\in U_c\\(\pi_{1l}\neq 1)}}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\left\{\frac{\pi_{1kl}\pi_{2|1,l}-\pi_{1k}\pi_{1l}\pi_{2|1,l}-\pi_{1kl}\pi_{2l}+\pi_{1k}\pi_{2l}\pi_{1l}}{(1-\pi_{1l})}\right\}$$

$$=\sum_{k\in U_1}\sum_{\substack{l\in U_c\\(\pi_{1l}\neq 1)}}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\left\{\frac{(\pi_{1kl}-\pi_{1k}\pi_{1l})(\pi_{2|1,l}-\pi_{2l})}{1-\pi_{1l}}\right\}$$

To obtain an unbiased estimator for this covariance, we multiply each term in the double summation by the random, inclusion indicator variables I_{1k} and I_{2l} and divide by the expected value of their product, π_{1k2l}. This gives

$$\hat{\mathrm{Cov}}(\hat{T}_1,\hat{T}_2)=\sum_{k\in U_1}\sum_{\substack{l\in U_c\\(\pi_{1l}\neq 1)}}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\left\{\frac{(\pi_{1kl}-\pi_{1k}\pi_{1l})(\pi_{2|1,l}-\pi_{2l})}{1-\pi_{1l}}\right\}\frac{I_{1k}I_{2l}}{\pi_{1k2l}}$$

$$=\sum_{k\in s_1}\sum_{\substack{l\in s_{2(1)}\\(\pi_{1l}\neq 1)}}\frac{y_{1k}y_{2l}}{\pi_{1k}\pi_{2l}}\left\{\frac{(\pi_{1kl}-\pi_{1k}\pi_{1l})(\pi_{2|1,l}-\pi_{2l})}{(1-\pi_{1l})\pi_{1kl}\pi_{2|1,l}+(\pi_{1k}-\pi_{1kl})(\pi_{2l}-\pi_{1l}\pi_{2|1,l})}\right\}$$

$$=\sum_{k\in s_1}\sum_{\substack{l\in s_{2(1)}\\(\pi_{1l}\neq 1)}}\frac{y_{1k}y_{2l}(\pi_{1kl}-\pi_{1k}\pi_{1l})(\pi_{2|1,l}-\pi_{2l})}{\pi_{1k}\pi_{2l}\left[\pi_{1kl}(\pi_{2|1,l}-\pi_{2l})+\pi_{1k}(\pi_{2l}-\pi_{1l}\pi_{2|1,l})\right]}$$

Appendix B. Derivation of the Covariance and Its Estimator Under SRSWOR

Under simple random sampling, we have $\pi_{1k}=\pi_{1l}=\pi_1$, $\pi_{2l}=\pi_2$, $\pi_{2|1,l}=\pi_{2|1}$, $\pi_{1kk}=\pi_{1k}=\pi_1$ and $\pi_{1kl}=\pi_1(N_1\pi_1-1)/(N_1-1)\quad$ for $k\neq l$.

Substituting these values into Equation (12) and assuming $\pi_1 < 1$ gives

$$\mathrm{Cov}_{\mathrm{SRS}}(\hat{T}_1,\hat{T}_2)=\sum_{k\in U_c}\frac{y_{1k}y_{2k}}{\pi_1\pi_2}\frac{\left(\pi_1-\pi_1^2\right)(\pi_{2|1}-\pi_2)}{(1-\pi_1)}$$

$$+\sum_{k\in U_1}\sum_{\substack{l\in U_c\\ l\neq k}}\frac{y_{1k}y_{2l}}{\pi_1\pi_2}\frac{\left\{\dfrac{\pi_1(N_1\pi_1-1)}{N_1-1}-\pi_1^2\right\}(\pi_{2|1}-\pi_2)}{(1-\pi_1)}$$

$$=\sum_{k\in U_c}y_{1k}y_{2k}\left(\frac{\pi_{2|1}}{\pi_2}-1\right)+\sum_{k\in U_1}\sum_{\substack{l\in U_c\\ l\neq k}}\frac{y_{1k}y_{2l}}{\pi_1\pi_2}\frac{\left(N_1\pi_1^2-\pi_1-N_1\pi_1^2+\pi_1^2\right)(\pi_{2|1}-\pi_2)}{(N_1-1)(1-\pi_1)}$$

$$=\sum_{k\in U_c}y_{1k}y_{2k}\left(\frac{\pi_{2|1}}{\pi_2}-1\right)-\frac{1}{N_1-1}\sum_{k\in U_1}\sum_{\substack{l\in U_c\\ l\neq k}}y_{1k}y_{2l}\left(\frac{\pi_{2|1}}{\pi_2}-1\right)$$

$$=\left(\frac{\pi_{2|1}}{\pi_2}-1\right)\frac{\left\{N_1\sum_{k\in U_c}y_{1k}y_{2k}-\sum_{k\in U_1}\sum_{l\in U_c}y_{1k}y_{2l}\right\}}{N_1-1}$$

$$=N_1\left(\frac{\pi_{2|1}}{\pi_2}-1\right)\frac{\left\{\sum_{k\in U_c}y_{1k}y_{2k}-\bar{y}_1\sum_{l\in U_c}y_{2l}\right\}}{N_1-1}$$

$$=N_1\left(\frac{\pi_{2|1}}{\pi_2}-1\right)\frac{\sum_{k\in U_c}(y_{1k}-\bar{y}_1)y_{2k}}{N_1-1}$$

Substituting into Equation (13) gives

$$\hat{\mathrm{Cov}}_{\mathrm{SRS}}(\hat{T}_1,\hat{T}_2)=\sum_{k\in s_c}\frac{y_{1k}y_{2k}\left(\pi_1-\pi_1^2\right)(\pi_{2|1}-\pi_2)}{\pi_1\pi_2\left[\pi_1(\pi_{2|1}-\pi_2)+\pi_1(\pi_2-\pi_1\pi_{2|1})\right]}$$

$$+\sum_{k\in s_1}\sum_{\substack{l\in s_{2(1)}\\ l\neq k}}\frac{y_{1k}y_{2l}\left\{\dfrac{\pi_1(N_1\pi_1-1)}{N_1-1}-\pi_1^2\right\}(\pi_{2|1}-\pi_2)}{\pi_1\pi_2\left\{\dfrac{\pi_1(N_1\pi_1-1)}{N_1-1}(\pi_{2|1}-\pi_2)+\pi_1(\pi_2-\pi_1\pi_{2|1})\right\}}$$

$$= \sum_{k\in s_c} \frac{y_{1k}y_{2k}}{\pi_1\pi_2}\left(1-\frac{\pi_2}{\pi_{2|1}}\right)$$

$$+\sum_{k\in s_1}\sum_{\substack{l\in s_{2(1)}\\ l\neq k}} \frac{y_{1k}y_{2l}\left\{N_1\pi_1^2-\pi_1-N_1\pi_1^2+\pi_1^2\right\}(\pi_{2|1}-\pi_2)}{\pi_1\pi_2\left\{\left(N_1\pi_1^2-\pi_1\right)(\pi_{2|1}-\pi_2)+N_1\pi_1\pi_2-\pi_1\pi_2-N_1\pi_1^2\pi_{2|1}+\pi_1^2\pi_{2|1}\right\}}$$

$$= \sum_{k\in s_c} \frac{y_{1k}y_{2k}}{\pi_1\pi_2}\left(1-\frac{\pi_2}{\pi_{2|1}}\right)+\sum_{k\in s_1}\sum_{\substack{l\in s_{2(1)}\\ l\neq k}} \frac{-y_{1k}y_{2l}\left(\pi_1-\pi_1^2\right)(\pi_{2|1}-\pi_2)}{\pi_1\pi_2\left(-N_1\pi_1^2\pi_2-\pi_1\pi_{2|1}+N_1\pi_1\pi_2+\pi_1^2\pi_{2|1}\right)}$$

$$= \sum_{k\in s_c} \frac{y_{1k}y_{2k}}{\pi_1\pi_2}\left(1-\frac{\pi_2}{\pi_{2|1}}\right)-\sum_{k\in s_1}\sum_{\substack{l\in s_{2(1)}\\ l\neq k}} \frac{y_{1k}y_{2l}(\pi_{2|1}-\pi_2)}{\pi_1\pi_2(N_1\pi_2-\pi_{2|1})}$$

$$= \sum_{k\in s_c} \frac{y_{1k}y_{2k}}{\pi_1\pi_2}\left(1-\frac{\pi_2}{\pi_{2|1}}\right)-\sum_{k\in s_1}\sum_{\substack{l\in s_{2(1)}\\ l\neq k}} \frac{y_{1k}y_{2l}\left(1-\dfrac{\pi_2}{\pi_{2|1}}\right)}{\pi_1\pi_2\left(N_1\dfrac{\pi_2}{\pi_{2|1}}-1\right)}$$

$$= \frac{\left(1-\dfrac{\pi_2}{\pi_{2|1}}\right)}{\pi_1\pi_2\left(N_1\dfrac{\pi_2}{\pi_{2|1}}-1\right)}\left\{\frac{N_1\pi_2}{\pi_{2|1}}\sum_{k\in s_c}y_{1k}y_{2k}-\sum_{k\in s_1}\sum_{l\in s_{2(1)}}y_{1k}y_{2l}\right\}$$

$$= \frac{\left(1-\dfrac{\pi_2}{\pi_{2|1}}\right)}{\pi_1\pi_2\left(1-\dfrac{\pi_{2|1}}{\pi_2 N_1}\right)}\left\{\sum_{k\in s_c}y_{1k}y_{2k}-\frac{\pi_{2|1}}{N_1\pi_2}\sum_{k\in s_1}y_{1k}\sum_{l\in s_{2(1)}}y_{2l}\right\}$$

$$= \frac{n_c\left(1-\dfrac{\pi_2}{\pi_{2|1}}\right)}{\pi_1\pi_2\left(1-\dfrac{\pi_1\pi_{2|1}}{\pi_2 n_1}\right)}\left\{\left(\overline{y_1y_2}\right)_{s_c}-\frac{\pi_1\pi_{2|1}n_{2(1)}}{\pi_2 n_c}\bar{y}_{s_1}\bar{y}_{s_{2(1)}}\right\}$$

Appendix C. Deviations from Conditions (3) and (4)

To demonstrate Expression (17), we have, using the random indicator variables I_{1k}, I_{1l} and I_{2l} introduced in Section 2

$$\delta_{kl}=\pi_{1k2l}-\Pr(l\in s_2|l\in s_1)\Pr(k\in s_1\,\&\,l\in s_1)-\Pr(l\in s_2|l\notin s_1)\Pr(k\in s_1\,\&\,l\notin s_1)$$

$$=\mathrm{E}[I_{1k}I_{2l}]-\Pr(l\in s_2|l\in s_1)\mathrm{E}[I_{1k}I_{1l}]-\Pr(l\in s_2|l\notin s_1)\mathrm{E}[I_{1k}(1-I_{1l})]$$

$$=\mathrm{E}[I_{1k}]\mathrm{E}[I_{2l}]+\mathrm{Cov}[I_{1k},I_{2l}]\ -\Pr(l\in s_2|l\in s_1)\{\mathrm{E}[I_{1k}]\mathrm{E}[I_{1l}]+\mathrm{Cov}[I_{1k},I_{1l}]\}$$

$$-\Pr(l\in s_2|l\notin s_1)\{\mathrm{E}[I_{1k}]\mathrm{E}[(1-I_{1l})]+\mathrm{Cov}[I_{1k},1-I_{1l}]\}$$

$$=\pi_{1k}\pi_{2l}+\mathrm{Cov}[I_{1k},I_{2l}]-\Pr(l\in s_2|l\in s_1)\Pr(l\in s_1)\pi_{1k}$$

$$-\Pr(l\in s_2|l\in s_1)\mathrm{Cov}[I_{1k},I_{1l}]-\Pr(l\in s_2|l\notin s_1)\Pr(l\notin s_1)\pi_{1k}+\Pr(l\in s_2|l\notin s_1)\mathrm{Cov}[I_{1k},I_{1l}]$$

$$=\mathrm{Cov}[I_{1k},I_{2l}-I_{1l}\{\Pr(l\in s_2|l\in s_1)-\Pr(l\in s_2|l\notin s_1)\}]$$

We thus have

$$\sum_{k\in U_1}\sum_{\substack{k\in U_2\\ l\neq k}}\delta_{kl}=\sum_{k\in U_1}\sum_{l\in U_2}\mathrm{Cov}[I_{1k},I_{2l}-I_{1l}\{\Pr(l\in s_2|l\in s_1)-\Pr(l\in s_2|l\notin s_1)\}]$$

$$=\mathrm{Cov}\left[\sum_{k\in U_1}I_{1k},\sum_{l\in U_2}I_{2l}-\sum_{l\in U_2}I_{1l}\{\Pr(l\in s_2|l\in s_1)-\Pr(l\in s_2|l\notin s_1)\}\right] \tag{27}$$

$$=\mathrm{Cov}\left[n_1,n_2-\sum_{l\in U_c}I_{1l}\left\{\pi_{2|1,l}-\frac{\pi_{2l}-\pi_{2|1,l}\pi_{1l}}{1-\pi_{1l}}\right\}\right]$$

because $I_{1l}=0$ for $I_{1l}\notin U_1$

$$=\mathrm{Cov}\left[n_1,\left(n_2-\sum_{l\in s_{1(2)}}\frac{\pi_{2|1,l}-\pi_{2l}}{1-\pi_{1l}}\right)\right]$$

Appendix D. Derivation of Residual Covariances for Laniel (1987)

Under Laniel's *sampling plans*

$$\delta_{kl}=\begin{cases}0: l\notin U_c\\ \delta_c: k,l\in U_c\ \ \&\ \ k\neq l\\ \delta_d: k\notin U_c, l\in U_c\end{cases}$$

Because n_1 is fixed we have, from Expression (17)

$$\sum_{k\in U_1}\sum_{\substack{l\in U_2\\ l\neq k}}\delta_{kl}=0\quad=\sum_{\substack{k\in U_1\\ k\notin U_c}}\sum_{l\in U_c}\delta_d+\sum_{k\in U_c}\sum_{\substack{k\in U_c\\ k\notin U_c}}\delta_c\quad=(N_1-N_c)N_c\delta_d+N_c(N_c-1)\delta_c$$

Hence $(N_1 - N_c)\delta_d = -(N_c - 1)\delta_c$ and the residual covariance is

$$\begin{aligned}\sum_{k\in U_1}\sum_{\substack{l\in U_2\\ l\neq k}} \frac{y_{1k}y_{2l}}{\pi_1\pi_2}\delta_{kl} &= \sum_{\substack{k\in U_1\\ k\notin U_c}}\sum_{l\in U_c}\frac{y_{1k}y_{2l}}{\pi_1\pi_2}\delta_d + \sum_{k\in U_c}\sum_{\substack{l\in U_c\\ l\neq k}}\frac{y_{1k}y_{2l}}{\pi_1\pi_2}\delta_c \\ &= -\frac{N_c - 1}{(N_1 - N_c)}\sum_{\substack{k\in U_1\\ k\notin U_c}}\sum_{l\in U_c}\frac{y_{1k}y_{2l}}{\pi_1\pi_2}\delta_c + \sum_{k\in U_c}\sum_{l\in U_c}\frac{y_{1k}y_{2l}}{\pi_1\pi_2}\delta_c - \sum_{k\in U_c}\frac{y_{1k}y_{2k}}{\pi_1\pi_2}\delta_c \qquad (28) \\ &= \frac{(N_c - 1)N_c(\bar{y}_{1c} - \bar{y}_{1d})\bar{y}_{2c} - \sum_{k\in U_c}(y_{1k} - \bar{y}_{1c})y_{2k}}{\pi_1\pi_2}\delta_c\end{aligned}$$

where $\bar{y}_{1d} = \sum_{\substack{k\in U_1\\ k\notin U_c}} y_{1k}/(N_1 - N_c)$ is the mean response in Period 1 for deaths.

Note that this general result applies to both *Sampling Plan A* and *Sampling Plan B*.

Under *Sampling Plan A*, $\Pr(l \in s_2 | l \notin s_1) = \mathrm{E}[(1 - r)n_{1(2)}/(N_c - n_{1(2)})]$. As Laniel (1987) notes, $n_{1(2)}$ has a hypergeometric distribution. His Equation 19 presents a second order approximation to $\mathrm{E}[(1 - r)n_{1(2)}/(N_c - n_{1(2)})]$ but the second order term is of order $1/N_c$. Assuming that $N_c \approx N_1$, as required by Laniel's sampling specification, we confine our attention to the first order approximation

$\Pr(l \in s_2 | l \notin s_1) \approx (1 - r)\mathrm{E}[n_{1(2)}]/(N_c - \mathrm{E}[n_{1(2)}])$ in order to simplify the analysis below.

In the current context, the expectation in the previous paragraph is applied only to those samples for which $l \notin s_1$. That is, we require

$$\Pr(l \in s_2 | l \notin s_1) = \mathrm{E}\left[\frac{(1 - r)n_{1(2)}}{N_c - n_{1(2)}} | l \notin s_1\right]$$

Since the condition $l \notin s_1$ reduces by one the number of units available for the random selection process in both Populations U_1 and U_2, $n_{1(2)}$ has a hypergeometric distribution with parameters $N_1 - 1$, $N_c - 1$ and n_1. This gives the following first order approximation

$$\begin{aligned}\Pr(l \in s_2 | l \notin s_1) &\approx \frac{(1 - r)n_1(N_c - 1)/(N_1 - 1)}{N_c - (n_1(N_c - 1)/(N_1 - 1))} = \frac{(1 - r)n_1}{N_1 - n_1 + ((N_1 - N_c)/(N_c - 1))} \\ &\approx \frac{(1 - r)n_1}{N_1 - n_1}\left\{1 - \frac{(N_1 - N_c)}{(N_1 - n_1)(N_c - 1)}\right\}\end{aligned}$$

To obtain a similar expression for $\Pr(l \in s_2 | k \in s_{1(2)} \ \& \ l \notin s_1)$ we need to assess the effect of the additional condition $k \in s_{1(2)}$ on the probability distribution of $n_{1(2)}$. Since the additional condition reduces by one more unit the number of units available for the random selection process in both populations and in the sample for Period 1, $n_{1(2)} - 1$ has a hypergeometric distribution, with parameters $N_1 - 2$, $N_c - 2$ and $n_1 - 1$. So:

$$\Pr(l \in s_2 | k \in s_{1(2)} \ \& \ l \notin s_1) \approx \frac{(1-r)\left\{1 + ((n_1 - 1)(N_c - 2)/(N_1 - 2))\right\}}{N_c - 1 - ((n_1 - 1)(N_c - 2)/(N_1 - 2))}$$

$$= \frac{(1-r)\{((N_1 - 2)/(N_c - 2)) - 1 + n_1)\}}{N_1 - 1 - n_1 + ((N_1 - 2)/(N_c - 2))}$$

$$= \frac{(1-r)\{((N_1 - N_c)/(N_c - 2)) + n_1)\}}{N_1 - n_1 + ((N_1 - N_c)/(N_c - 2))}$$

$$\approx \frac{(1-r)\{((N_1 - N_c)/(N_c - 2)) + n_1\}}{N_1 - n_1}\left\{1 - \frac{(N_1 - N_c)}{(N_1 - n_1)(N_c - 2)}\right\}$$

Hence

$$\Pr(l \in s_2 | k \in s_{1(2)} \ \& \ l \notin s_1) - \Pr(l \in s_2 | l \notin s_1)$$
$$\approx \frac{(1-r)(N_1 - N_c)}{(N_1 - n_1)(N_c - 2)}\left\{1 - \frac{(N_1 - N_c)}{(N_1 - n_1)(N_c - 2)}\right\}$$

and

$$\delta_c = \{\Pr(l \in s_2 | k \in s_{1(2)} \ \& \ l \notin s_1) - \Pr(l \in s_2 | l \notin s_1)\}\Pr(k \in s_1 \ \& \ l \notin s_1)$$
$$\approx \frac{(1-r)(N_1 - N_c)}{(N_1 - n_1)(N_c - 2)}\left\{1 - \frac{(N_1 - N_c)}{(N_1 - n_1)(N_c - 2)}\right\}\frac{n_1(N_1 - n_1)}{N_1(N_1 - 1)} \tag{29}$$
$$= \frac{\pi_1(1-r)(N_1 - N_c)}{(N_1 - 1)(N_c - 2)}\left\{1 - \frac{(N_1 - N_c)}{(N_1 - n_1)(N_c - 2)}\right\}$$

Substituting Equation (29) into Equation (28) thus gives

$$\sum_{k \in U_1}\sum_{\substack{l \in U_2 \\ l \neq k}} \frac{y_{1k} y_{2l}}{\pi_1 \pi_2}\delta_{kl} \approx \frac{(1-r)(N_1 - N_c)(N_c - 1)}{\pi_2(N_1 - 1)(N_c - 2)}\{N_c(\bar{y}_{1c} - \bar{y}_{1d})\bar{y}_{2c} - C_{12}\}$$
$$\times\left\{1 - \frac{(N_1 - N_c)}{(N_1 - n_1)(N_c - 2)}\right\} \tag{30}$$

where: $C_{12} = \sum_{k \in U_c}(y_{1k} - \bar{y}_{1c})y_{2k}/(N_c - 1)$ is the finite population covariance between the response in Period 1 and the response in Period 2 for units in the common population.

To compare this with Equations (14) and (20), we may rewrite Equation (20) as

$$\mathrm{Cov}_{\mathrm{SRS}}(\hat{T}_1, \hat{T}_2) \approx \frac{(\pi_{2|1} - \pi_2)N_1}{\pi_2(N_1 - 1)}\left\{(N_c - 1)C_{12} + \frac{N_c}{N_1}(N_1 - N_c)(\bar{y}_{1c} - \bar{y}_{1d})\bar{y}_{2c}\right\} \tag{31}$$

The relative error in the approximate covariance of Equation (14) is therefore approximately the ratio of Expression (30) to Expression (31)

$$\frac{(1-r)}{(\pi_{2|1}-\pi_2)}\frac{(N_1-N_c)\{N_c(\bar{y}_{1c}-\bar{y}_{1d})\bar{y}_{2c}-C_{12}\}}{\{N_1(N_c-1)C_{12}+N_c(N_1-N_c)(\bar{y}_{1c}-\bar{y}_{1d})\bar{y}_{2c}\}}\frac{(N_c-1)}{(N_c-2)}$$

$$\times\left\{1-\frac{(N_1-N_c)}{(N_1-n_1)(N_c-2)}\right\}$$

For *Sampling Plan B*, Conditions (3) and (4) are not met. For units $l \in U_1$ and $l \notin s_1$, we have

$$\Pr(l\in s_2|l\notin s_1)=\mathrm{E}\left[\frac{(1-r)n_{1(2)}}{N_c-rn_{1(2)}}\middle| l\notin s_1\right]$$

and, using similar arguments as for *Sampling Plan A*

$$\Pr(l\in s_2|l\notin s_1)\approx\frac{(1-r)\{n_1(N_c-1)/(N_1-1)\}}{N_c-(rn_1(N_c-1)/(N_1-1))}\approx\frac{(1-r)n_1}{N_1-rn_1}\left\{1-\frac{(N_1-N_c)}{(N_1-rn_1)(N_c-1)}\right\}$$

$$\Pr(l\in s_2|k\in s_{1(2)}\,\&\,l\notin s_1)\approx\frac{(1-r)\{1+((n_1-1)(N_c-2))/(N_1-2)\}}{N_c-r-r((n_1-1)(N_c-2))/(N_1-2)}$$

$$=\frac{(1-r)\{((N_1-2)/(N_c-2))-1+n_1\}}{N_1-2+((2-r)(N_1-2))/(N_c-2)-r(n_1-1)}$$

$$=\frac{(1-r)\{(N_1-N_c)/(N_c-1)+n_1\}}{N_1-rn_1+(2-r)((N_1-N_c)/(N_c-1))}$$

$$\approx\frac{(1-r)\{((N_1-N_c)/(N_c-1))+n_1\}}{N_1-rn_1}\left\{1-\frac{(2-r)(N_1-N_c)}{(N_1-rn_1)(N_c-1)}\right\}$$

Similarly

$$\Pr(l\in s_2|l\in s_1)=\mathrm{E}\left[r+\frac{(1-r)^2n_{1(2)}}{N_c-rn_{1(2)}}\middle| l\in s_1\right]$$

$$\approx r+\frac{(1-r)^2\{1+((n_1-1)(N_c-1))/(N_1-1)\}}{N_c-r-r((n_1-1)(N_c-1))/(N_1-1)}$$

$$=r+\frac{(1-r)^2\{((N_1-1)/(N_c-1))-1+n_1\}}{N_1-1+((1-r)(N_1-1))/(N_c-1)-r(n_1-1)}$$

$$=r+\frac{(1-r)^2\{((N_1-N_c)/(N_c-1))+n_1\}}{N_1-rn_1+(1-r)((N_1-N_c)/(N_c-1))}$$

$$\approx r+\frac{(1-r)^2\{((N_1-N_c)/(N_c-1))+n_1\}}{N_1-rn_1}\left\{1-\frac{(1-r)(N_1-N_c)}{(N_1-rn_1)(N_c-1)}\right\}$$

$$\begin{aligned}
\Pr(l \in s_2 | k \in s_{1(2)} \ \& \ l \in s_1) &= \mathrm{E}\left[r + \frac{(1-r)^2 n_{1(2)}}{N_c - r n_{1(2)}} \middle| k \in s_{1(2)} \ \& \ l \in s_1 \right] \\
&\approx r + \frac{(1-r)^2 \left\{ 2 + ((n_1 - 2)(N_c - 2))/(N_1 - 2) \right\}}{N_c - 2r - r((n_1 - 2)(N_c - 2))/(N_1 - 2)} \\
&= r + \frac{(1-r)^2 \{((2(N_1 - 2)/(N_c - 2)) - 2 + n_1\}}{N_1 - 2 + 2((1-r)(N_1 - 2))/(N_c - 2) - r(n_1 - 2)} \\
&= r + \frac{(1-r)^2 \{(2(N_1 - N_c)/(N_c - 2)) + n_1\}}{N_1 - r n_1 + (2(1-r)(N_1 - N_c))/(N_c - 2)} \\
&\approx r + \frac{(1-r)^2 \{2(N_1 - N_c)/(N_c - 2) + n_1\}}{N_1 - r n_1} \left\{ 1 - \frac{2(1-r)(N_1 - N_c)}{(N_1 - r n_1)(N_c - 2)} \right\}
\end{aligned}$$

Substituting these expressions into the relevant equations, in the same way as for *Sampling Plan A*, produces corresponding expressions for the residual covariance and the relative error for *Sampling Plan B*. Because these expressions have the same form as with *Sampling Plan A*, the considerations and resultant conclusions are also the same.

6. References

Berger, Y. (2004a). Variance Estimation for Measures of Change in Probability Samples. The Canadian Journal of Statistics, 32, 451–467.

Berger, Y. (2004b). Variance Estimation for Change: An Evaluation Based upon the 2000 Finnish Labour Force Survey. http://laplace.wiwi.uni-tuebingen.de/dacseis/RPM/deliverables/DACSEIS-D9-1.pdf

Hansen, M., Hurwitz, W., and Madow, W. (1953). Sample Survey Methods and Theory, Vol. 1 and 2. New York: Wiley.

Hidiroglou, M.G., Särndal, C.-E., and Binder, D. (1995). Weighting and Estimation in Business Surveys. In Business Survey Methods, B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M. Colledge, and P.S. Kott (eds). New York: Wiley, 477–502.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. Journal of the American Statistical Association, 47, 663–685.

Laniel, N. (1987). Variances for a Rotating Sample from a Changing Population. Proceedings of the American Statistical Association, Business and Economic Statistics Section, 496–500.

Nordberg, L. (2000). On Variance Estimation for Measures of Change When Samples Are Coordinated by the Use of Permanent Random Numbers. Journal of Official Statistics, 16, 363–378.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer.

Tam, S.M. (1984). On Covariances from Overlapping Samples. The American Statistician, 38, 288–292.

Valliant, R. (1991). Variance Estimation for Price Indexes from a Two-Stage Sample with Rotating Panels. Journal of Business and Economic Statistics, 9, 409–422.

Received July 2005
Revised July 2006