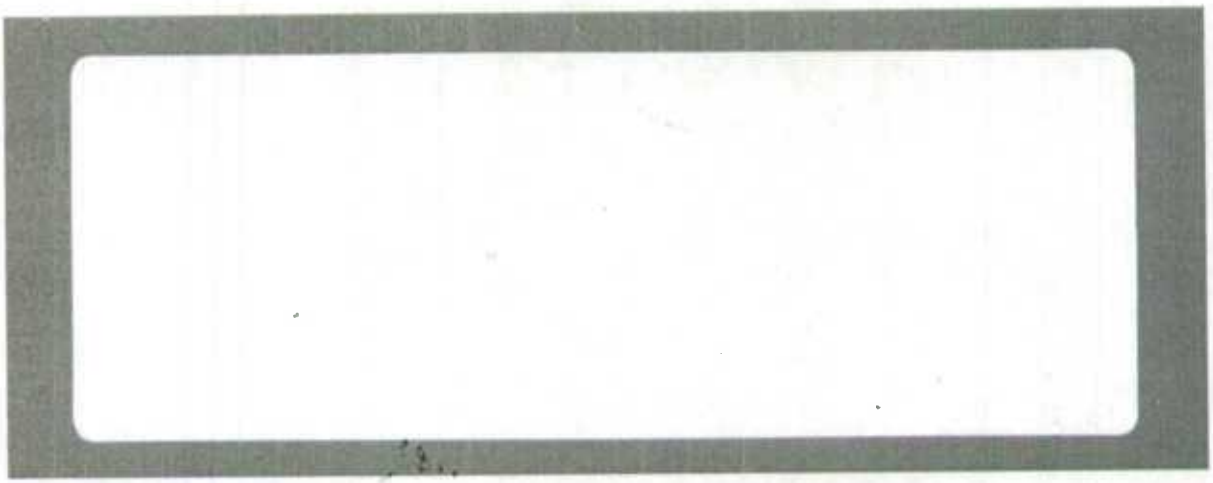

Methodology Branch

Household Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes des ménages

Canada

WORKING PAPER
METHODOLOGY BRANCH

A NOTE ON VARIANCE ESTIMATION IN MULTI-STAGE SAMPLING

HSMD-2004-003E

Paddison Wong

Household Survey Methods Division
Statistics Canada

The work presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada.

A NOTE ON VARIANCE ESTIMATION IN MULTI-STAGE SAMPLING

Paddison Wong[1]

ABSTRACT

In multi-stage sampling surveys, the variance of an estimator has contributions from all stages of the survey design. The expressions of the total variance and its estimator are usually complicated. However, with certain choice of sampling scheme, they can be simplified. In the case of sample survey with r stages, if the first stage is with replacement, the total variance of an estimator has a simpler expression because the combined variance of the last r-1 stages has been integrated into the other terms in the equation. This is one advantage of sampling with replacement in the first stage over sampling without replacement. However, it should be noted that the variance of an estimator for sampling with replacement in the first stage is always greater than that for sampling without replacement in the first stage.

[1] Paddison Wong, Expenditure Survey Analysis Methods Section, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

NOTE CONCERNANT L'ESTIMATION DE LA VARIANCE DANS L'ÉCHANTILLONNAGE À PLUSIEURS DEGRÉS

Paddison Wong[2]

RÉSUMÉ

Dans des enquêtes à plusieurs degrés, toutes les étapes de l'enquête contribuent à la variance d'un estimateur. Les expressions de la variance totale et de son estimateur sont habituellement complexes, mais elles peuvent être simplifiées si l'on choisit certains plans de sondage. Dans le cas d'un plan de sondage de degré *r* avec remise au premier degré, la variance totale d'un estimateur n'a qu'une expression simple, puisque la variance combinée des derniers *r*-1 degrés a été incorporée dans les autres termes de l'équation. C'est un avantage de l'échantillonnage avec remise au premier degré sur l'échantillonnage sans remise. Il faut cependant préciser que la variance d'un estimateur pour l'échantillonnage avec remise au premier degré est toujours plus grande que celle pour l'échantillonnage sans remise au premier degré.

[2] Paddison Wong, Méthodes d'enquêtes sur les dépenses et d'analyse, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Ottawa, Ontario, K1A 0T6.

1 Introduction

In elementary sampling courses, sample surveys are usually assumed to have only one stage. Estimators, as well as their variance estimators, of population parameters are constructed and studied under such a simple system. However, many surveys, especially the large scale ones, are carried out with multi-stages. The estimation of the parameters and of the variance of the estimators, in multi-stage sample surveys, could be very complicated due to the complexity of the survey designs. It is reasonable to think that the variance estimation of an estimator should have contributions from the variation in all stages of the design. However, it can be shown that it may not be true in all cases. The estimation of the variance could be simplified dramatically with some smart choice of sampling schemes. This surprising and important result has a crucial impact on the variance estimation of the estimators because it saves time and cost in the estimation procedure.

In this paper, we study the variance estimators in the case of multi-stage sampling following the same approach given by Stuart [4]. It is shown that the first stage of sampling always plays an important role regarding the variance of an estimator $\hat{\mu}$ of a population parameter which has a linear form as in equation (1). In general, the true sampling variance of $V(\hat{\mu})$ consists of two parts: the variance due to the first-stage sampling and the variance due to the last $r-1$ stages of sampling. If the first stage of sampling is without replacement, the total variance involves the variance of the last $r-1$ stages explicitly as in equations (10) and (12). However, if the first stage of sampling is with replacement, the total variance is less complicated and it is not necessary to calculate the variance of the last $r-1$ stages explicitly as in equations (14) and (15) because it has been integrated into the other terms. As a result, the estimated variance has a simpler form than the one for sampling without replacement in the first stage.

2 Total Sampling Variance in Multi-Stage Surveys

Suppose in a multi-stage sampling, there are r stages, and μ is a population parameter of interest with an estimator of the form

$$\hat{\mu} = \sum_{i=1}^{n} t_i \tag{1}$$

where n first-stage units are selected from a population of size N and t_i is a sample statistics from the ith selected PSU based on the last $r-1$ stages. The Horvitz-Thompson estimator is a special case of $\hat{\mu}$. The Horvitz-Thompson estimator of the population total as defined in equation (6.12) in Lohr [2] is given by

$$\hat{t}_{HT} = \sum_{i=1}^{n} \frac{\hat{t}_i}{\pi_i} \tag{2}$$

where $\hat{t}_i$ is an estimator for the ith unit from the last $r-1$ stage. It is easy to see that t_i in equation (1) is a weighted statistic corresponding to $\hat{t}_i/\pi_i$ in equation (2). The statistics t_i and t_j are independent for i not equal to j. An unconditional variance of $\hat{\mu}$ can be expressed as a sum of the variance of conditional mean and the mean of conditional variance, namely,

$$V\big[\hat{\mu}\big] = V_1 E_L\big[\hat{\mu}\big] + E_1 V_L\big[\hat{\mu}\big] \tag{3}$$

where L means the last $r-1$ stages. Furthermore,

$$E_L\big[\hat{\mu}\big] = \sum_{i=1}^{n} E_L\big[t_i\big] \tag{4}$$

and because t_i and t_j are independent for $i \neq j$,

$$V_L\big[\hat{\mu}\big] = \sum_{i=1}^{n} V_L\big[t_i\big] \tag{5}$$

and hence expression (3) can be written as

$$V\big[\hat{\mu}\big] = V_1 \sum_{i=1}^{n} E_L\big[t_i\big] + E_1 \sum_{i=1}^{n} V_L\big[t_i\big] \ . \tag{6}$$

Let τ_i represent a population value of the ith selected PSU with an unbiased estimator t_i with respect to the last $r-1$ sampling stages, that is $E_L[t_i] = \tau_i$, the first term in (6) can be written as

$$V_1\left[\sum_{i=1}^{n} E_L[t_i]\right] = E_1\left[\sum_{i=1}^{n} \tau_i^2\right] + E_1\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n} \tau_i\tau_j\right] - \left(E_1 \sum_{i=1}^{n} \tau_i\right)^2 . \qquad (7)$$

Now let T_j be a population value of the jth PSU for $j = 1, \ldots, N$ which has the form of the τ_i's such that when the jth PSU is selected in the ith draw, $\tau_i = T_j$. For instance, τ_i could be the total of the ith selected PSU and T_j the total of the jth PSU in the population frame. Also let γ_i be a population value of the ith selected PSU with an unbiased estimator t_i^2 with respect to the $r-1$ sampling stages, that is $\gamma_i = E_L[t_i^2]$ and let Γ_j be defined similarly as T_j for all PSUs in the sampling frame and has the same function as γ_i's. That is, when the jth PSU is selected in the ith draw, $\gamma_i = \Gamma_j$. Note that both τ_i and γ_i are parameters with respect to the last $r-1$ stages but they are random variables with respect to the first sampling stage. Suppose π_i is defined as in Appendix (A.1), the second term of (6) can be expressed as

$$E_1\left[\sum_{i=1}^{n} V_L[t_i]\right] = E_1\left[\sum_{i=1}^{n} \left(E_L[t_i^2] - \tau_i^2\right)\right] = E_1\left[\sum_{i=1}^{n} \left(\gamma_i - \tau_i^2\right)\right] = \sum_{i=1}^{N} \pi_i \left(\Gamma_i - T_i^2\right) \qquad (8)$$

since $E_1\left[\sum_{i=1}^{n} \tau_i^2\right] = \sum_{i=1}^{N} \pi_i T_i^2$, and $E_1\left[\sum_{i=1}^{n} \gamma_i\right] = \sum_{i=1}^{N} \pi_i \Gamma_i$ as shown in Appendix (A.1). Equation (6) with expression (7) and (8) is always true whether the first stage is with or without replacement.

3 Sampling Without Replacement At The First Stage

For sampling without replacement at the first stage, it has been shown in Appendix (A.1) that $E_1\Big[\sum_{i=1}^{n}\sum_{j\neq i}^{n}\tau_i\tau_j\Big] = \sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_{ij}\,T_iT_j$ and (7) becomes

$$\begin{aligned} V_1\left[\sum_{i=1}^{n}E_L\big[t_i\big]\right] &= \sum_{i=1}^{N}\pi_i\,T_i^2+\sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_{ij}\,T_iT_j-\left(\sum_{i=1}^{N}\pi_i\,T_i\right)^2 \\ &= \sum_{i=1}^{N}\pi_i(1-\pi_i)T_i^2+\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_{ij}-\pi_i\pi_j)\,T_iT_j\;. \end{aligned} \tag{9}$$

Therefore, the total variance after the substitution of (8) and (9) in (6) is

$$V_{WOR}\big[\hat{\mu}\big] = \sum_{i=1}^{N}\pi_i(1-\pi_i)T_i^2+\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_{ij}-\pi_i\pi_j)\,T_iT_j+\sum_{i=1}^{N}\pi_i\left(\Gamma_i-T_i^2\right) \tag{10}$$

Note that this expression is equivalent to (4.4.2) in Särndal et al. [3] and equation (6.20) in Lohr [2] where the first two terms here equal to V_{psu} and the last term is V_{ssu} in (6.20). After some arrangement of the terms, the above equation can be written as:

$$\begin{aligned} V_{WOR}\big[\hat{\mu}\big] &= \sum_{i=1}^{N}\left(\pi_i\Gamma_i-\pi_i^2\Gamma_i\right)+\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_{ij}-\pi_i\pi_j)\,T_iT_j+\sum_{i=1}^{N}\left(\pi_i^2\Gamma_i-\pi_i^2T_i^2\right) \\ &= \sum_{i=1}^{N}\pi_i(1-\pi_i)\Gamma_i+\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_{ij}-\pi_i\pi_j)\,T_iT_j+\sum_{i=1}^{N}\pi_i^2\left(\Gamma_i-T_i^2\right) \end{aligned} \tag{11}$$

with an unbiased estimator

$$\hat{V}_{WOR}\big[\hat{\mu}\big] = \sum_{i=1}^{n}(1-\pi_i)t_i^2+\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}t_it_j+\sum_{i=1}^{n}\pi_i\hat{V}_L\big[t_i\big] \tag{12}$$

provided that $\hat{V}_L\big[t_i\big]$ is unbiased for $V_L\big[t_i\big]$ with respect to the last $r-1$ stages of sampling. Note that equation (11) and (12) are the same as equation (6.13) and (6.14) for the Horvitz-Thompson estimator respectively in Lohr [2]. To show that (12) is unbiased, note that

$$E_1\,E_L\Big[\sum_{i=1}^{n}(1-\pi_i)t_i^2\Big] = E_1\Big[\sum_{i=1}^{n}(1-\pi_i)\gamma_i\Big] = \sum_{i=1}^{N}\pi_i(1-\pi_i)\Gamma_i,$$

$$E_1\,E_L\Big[\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}t_it_j\Big]=E_1\Big[\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\tau_i\tau_j\Big]=\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_{ij}-\pi_i\pi_j)T_iT_j\,,$$

and

$$E_1\,E_L\Big[\sum_{i=1}^{n}\pi_i\hat{V}_L[t_i]\Big]=E_1\Big[\sum_{i=1}^{n}\pi_iV_L[t_i]\Big]=E_1\Big[\sum_{i=1}^{n}\pi_i(\gamma_i-\tau_i^2)\Big]=\sum_{i=1}^{N}\pi_i^2\left(\Gamma_i-T_i^2\right)\,.$$

The first two terms in (12) is the contribution from the first stage of sampling and the last term is from the last $r-1$ stages.

4 Sampling With Replacement At The First Stage

If the sampling is with replacement at the first stage, $E_1\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n}\tau_i\tau_j\right]=\sum_{k=1}^{N}\sum_{l=1}^{N}\pi_{kl}\,T_kT_l$. Note that the sum is over all k and l. The variance in (7) becomes

$$\begin{aligned}V_1\left[\sum_{i=1}^{n}E_L[t_i]\right] &= \sum_{k=1}^{N}\pi_k\,T_k^2+\sum_{k=1}^{N}\sum_{l=1}^{N}\pi_{kl}\,T_kT_l-\left(\sum_{k=1}^{N}\pi_k\,T_k\right)^2\\ &= \sum_{k=1}^{N}\pi_k(1-\pi_k)T_k^2+\sum_{k=1}^{N}\sum_{l=1}^{N}\pi_{kl}\,T_kT_l-\sum_{k=1}^{N}\sum_{l\neq k}^{N}\pi_k\pi_lT_kT_l\,. \qquad (13)\end{aligned}$$

Substituting (13) and (8) into (6), we obtained

$$\begin{aligned}V_{WR}[\hat{\mu}] &= \sum_{k=1}^{N}\pi_k(1-\pi_k)T_k^2+\sum_{k=1}^{N}\sum_{l=1}^{N}\pi_{kl}\,T_kT_l-\sum_{k=1}^{N}\sum_{l\neq k}^{N}\pi_k\pi_lT_kT_l+\sum_{k=1}^{N}\pi_k\left(\Gamma_k-T_k^2\right)\\ &= \sum_{k=1}^{N}\pi_k\Gamma_k+\sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl}-\pi_k\pi_l)T_kT_l \qquad (14)\end{aligned}$$

which has an unbiased estimator given by

$$\hat{V}_{WR}[\hat{\mu}]=\sum_{i=1}^{n}t_i^2+\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}t_it_j \qquad (15)$$

which shows that, when the first stage is with replacement, the overall variance and its unbiased estimator are less complicated because the variance from the $r-1$ stages has

and let $g(y_i, y_j)$ and $g(Y_i, Y_j)$ be functions of sample values $(y_i,\ y_j)$ and population values $(Y_i,\ Y_j)$ respectively,

$$\begin{aligned} E\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n} g(y_i, y_j)\right] &= E\left[\sum_{i=1}^{N}\sum_{j\neq i}^{N} I_i\, I_j\, g(Y_i, Y_j)\right] \\ &= \sum_{i=1}^{N}\sum_{j\neq i}^{N} E\big[I_i\, I_j\big] g(Y_i, Y_j) \\ &= \sum_{i=1}^{N}\sum_{j\neq i}^{N} \pi_{ij}\, g(Y_i, Y_j)\ . \end{aligned}$$

A.2 Example: Simple random sample without replacement

For simple random sample, the probability of inclusion is $\pi_i = n/N$ and hence

$$E\big[\overline{y}\big] = \sum_{i=1}^{N} \pi_i \frac{Y_i}{n} = \sum_{i=1}^{N} \frac{Y_i}{N} = \overline{Y}$$

which completes the proof of unbiasedness. To show that the variance of the sample mean is $(1-f)S^2/n$, consider the following

$$\begin{aligned} V\left[\sum_{i=1}^{n} \frac{y_i}{n}\right] &= E\left[\left(\sum_{i=1}^{n} \frac{y_i}{n} - \overline{Y}\right)^2\right] \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n}\sum_{j\neq i}^{n} y_i\, y_j\right] - \overline{Y}^2 \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{N} E\big[I_i\big] Y_i^2 + \sum_{i=1}^{N}\sum_{j\neq i}^{N} E\big[I_i\, I_j\big] Y_i\, Y_j\right) - \overline{Y}^2 \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{N} \frac{n}{N} Y_i^2 + \sum_{i=1}^{N}\sum_{j\neq i}^{N} \frac{n(n-1)}{N(N-1)} Y_i\, Y_j\right) - \overline{Y}^2 \\ &= \frac{1}{n^2}\left\{\sum_{i=1}^{N}\left(\frac{n}{N} - \frac{n(n-1)}{N(N-1)}\right) Y_i^2 + \frac{n(n-1)}{N(N-1)}\left(\sum_{i=1}^{N} Y_i\right)^2\right\} - \overline{Y}^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{N} \frac{n(N-n)}{N(N-1)} Y_i^2 + \frac{N(n-1)}{n(N-1)}\overline{Y}^2 - \overline{Y}^2 \end{aligned}$$

$$E_1\,E_L\Big[\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}t_it_j\Big]=E_1\Big[\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\tau_i\tau_j\Big]=\sum_{i=1}^{N}\sum_{j\neq i}^{N}(\pi_{ij}-\pi_i\pi_j)T_iT_j\,,$$

and

$$E_1\,E_L\Big[\sum_{i=1}^{n}\pi_i\hat{V}_L\big[t_i\big]\Big]=E_1\Big[\sum_{i=1}^{n}\pi_iV_L\big[t_i\big]\Big]=E_1\Big[\sum_{i=1}^{n}\pi_i(\gamma_i-\tau_i^2)\Big]=\sum_{i=1}^{N}\pi_i^2\left(\Gamma_i-T_i^2\right)\,.$$

The first two terms in (12) is the contribution from the first stage of sampling and the last term is from the last $r-1$ stages.

4 Sampling With Replacement At The First Stage

If the sampling is with replacement at the first stage, $E_1\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n}\tau_i\tau_j\right]=\sum_{k=1}^{N}\sum_{l=1}^{N}\pi_{kl}\,T_kT_l$. Note that the sum is over all k and l. The variance in (7) becomes

$$\begin{aligned}V_1\left[\sum_{i=1}^{n}E_L\big[t_i\big]\right] &= \sum_{k=1}^{N}\pi_k\,T_k^2+\sum_{k=1}^{N}\sum_{l=1}^{N}\pi_{kl}\,T_kT_l-\left(\sum_{k=1}^{N}\pi_k\,T_k\right)^2\\ &= \sum_{k=1}^{N}\pi_k(1-\pi_k)T_k^2+\sum_{k=1}^{N}\sum_{l=1}^{N}\pi_{kl}\,T_kT_l-\sum_{k=1}^{N}\sum_{l\neq k}^{N}\pi_k\pi_lT_kT_l\,. \qquad (13)\end{aligned}$$

Substituting (13) and (8) into (6), we obtained

$$\begin{aligned}V_{WR}\big[\hat{\mu}\big] &= \sum_{k=1}^{N}\pi_k(1-\pi_k)T_k^2+\sum_{k=1}^{N}\sum_{l=1}^{N}\pi_{kl}\,T_kT_l-\sum_{k=1}^{N}\sum_{l\neq k}^{N}\pi_k\pi_lT_kT_l+\sum_{k=1}^{N}\pi_k\left(\Gamma_k-T_k^2\right)\\ &= \sum_{k=1}^{N}\pi_k\Gamma_k+\sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl}-\pi_k\pi_l)T_kT_l \qquad (14)\end{aligned}$$

which has an unbiased estimator given by

$$\hat{V}_{WR}\big[\hat{\mu}\big]=\sum_{i=1}^{n}t_i^2+\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}t_it_j \qquad (15)$$

which shows that, when the first stage is with replacement, the overall variance and its unbiased estimator are less complicated because the variance from the $r-1$ stages has

already been integrated into the other terms. To prove the unbiasedness of the estimator in (15), note that

$$E_1 E_L\Big[\sum_{i=1}^{n} t_i^2\Big] = E_1\Big[\sum_{i=1}^{n}\gamma_i\Big] = \sum_{k=1}^{N}\pi_i\Gamma_k \tag{16}$$

and

$$E_1 E_L\Big[\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}t_i t_j\Big] = E_1\Big[\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\tau_i\tau_j\Big] = \sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl}-\pi_k\pi_l)T_k T_l\ . \tag{17}$$

Compare (12) and (15), equation (15) has a simpler form which makes the estimation much easier.

5 Conclusion

In sampling with replacement, the same sampling unit may be selected more than once. It may not be a favourable sampling scheme to everyone because it is less efficient than sampling without replacement. If we compare equation (10) for sampling without replacement and equation (14) for sampling with replacement, we obtain the following relationship:

$$V_{WOR}(\hat{\mu}) = V_{WR}(\hat{\mu}) - \sum_{i=1}^{N}\pi_i T_i^2\ .$$

If we make the similar comparison on the unbiased estimators of the variances in (12) and (15), we have

$$\hat{V}_{WOR}(\hat{\mu}) = \hat{V}_{WR}(\hat{\mu}) - \sum_{i=1}^{n}\pi_i\left(\hat{V}_L[t_i] - t_i^2\right)\ .$$

Whether we are interested in the variance or its unbiased estimator, they always have a simpler form when the first stage of sampling is with replacement. This is a major advantage of sampling with replacement over sampling without replacement in the first stage of selection in multi-stage sample surveys.

APPENDIX A

A.1 Expectations for Sampling Without Replacement

Suppose $Y_1, Y_2, \ldots, Y_N$ are measurements of individuals in a population of size N and $y_1, y_2, \ldots, y_n$ are measurements of individuals in a sample of fixed size n from this population. Let $I_i = 1$ ($I_i = 0$) if the ith individual is selected (not) in the sample. Define π_i to be the probability that the ith individual is selected in the sample, that is, $\pi_i = P(I_i = 1)$, and define π_{ij} to be the probability that both individuals i and j are selected in the sample, that is $\pi_{ij} = P(I_i = 1, I_j = 1)$. It is easy to show that $E[I_i] = \pi_i$ and $E[I_i\, I_j] = \pi_{ij}$. Consider the sample mean

$$\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n} = \sum_{i=1}^{N} I_i \frac{Y_i}{n}$$

which has an expectation

$$E[\overline{y}] = \sum_{i=1}^{N} E[I_i] \frac{Y_i}{n} = \sum_{i=1}^{N} \pi_i \frac{Y_i}{n} \; .$$

Similarly, the expectation of $\sum_{i=1}^{n} \sum_{j \neq i}^{n} y_i\, y_j$ is given by

$$\begin{aligned} E\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n} y_i\, y_j\right] &= E\left[\sum_{i=1}^{N}\sum_{j\neq i}^{N} I_i\, I_j\, Y_i\, Y_j\right] \\ &= \sum_{i=1}^{N}\sum_{j\neq i}^{N} E[I_i\, I_j] Y_i\, Y_j \\ &= \sum_{i=1}^{N}\sum_{j\neq i}^{N} \pi_{ij}\, Y_i\, Y_j \; . \end{aligned}$$

More generally, let $g(y_i)$ and $g(Y_i)$ be functions of sample value y_i of the ith draw and population value Y_i of the ith unit respectively, the expectation of $\sum_{i=1}^{n} g(y_i)$ is given by

$$E\left[\sum_{i=1}^{n} g(y_i)\right] = E\left[\sum_{i=1}^{N} I_i g(Y_i)\right] = \sum_{i=1}^{N} E[I_i] g(Y_i) = \sum_{i=1}^{N} \pi_i\, g(Y_i)$$

and let $g(y_i, y_j)$ and $g(Y_i, Y_j)$ be functions of sample values (y_i, y_j) and population values (Y_i, Y_j) respectively,

$$\begin{aligned} E\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n} g(y_i, y_j)\right] &= E\left[\sum_{i=1}^{N}\sum_{j\neq i}^{N} I_i\, I_j\, g(Y_i, Y_j)\right] \\ &= \sum_{i=1}^{N}\sum_{j\neq i}^{N} E\big[I_i\, I_j\big] g(Y_i, Y_j) \\ &= \sum_{i=1}^{N}\sum_{j\neq i}^{N} \pi_{ij}\, g(Y_i, Y_j)\ . \end{aligned}$$

A.2 Example: Simple random sample without replacement

For simple random sample, the probability of inclusion is $\pi_i = n/N$ and hence

$$E\big[\overline{y}\big] = \sum_{i=1}^{N} \pi_i \frac{Y_i}{n} = \sum_{i=1}^{N} \frac{Y_i}{N} = \overline{Y}$$

which completes the proof of unbiasedness. To show that the variance of the sample mean is $(1-f)S^2/n$, consider the following

$$\begin{aligned} V\left[\sum_{i=1}^{n} \frac{y_i}{n}\right] &= E\left[\left(\sum_{i=1}^{n} \frac{y_i}{n} - \overline{Y}\right)^2\right] \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n}\sum_{j\neq i}^{n} y_i\, y_j\right] - \overline{Y}^2 \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{N} E\big[I_i\big] Y_i^2 + \sum_{i=1}^{N}\sum_{j\neq i}^{N} E\big[I_i\, I_j\big] Y_i\, Y_j\right) - \overline{Y}^2 \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{N} \frac{n}{N} Y_i^2 + \sum_{i=1}^{N}\sum_{j\neq i}^{N} \frac{n(n-1)}{N(N-1)} Y_i\, Y_j\right) - \overline{Y}^2 \\ &= \frac{1}{n^2}\left\{\sum_{i=1}^{N}\left(\frac{n}{N} - \frac{n(n-1)}{N(N-1)}\right) Y_i^2 + \frac{n(n-1)}{N(N-1)}\left(\sum_{i=1}^{N} Y_i\right)^2\right\} - \overline{Y}^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{N} \frac{n(N-n)}{N(N-1)} Y_i^2 + \frac{N(n-1)}{n(N-1)}\overline{Y}^2 - \overline{Y}^2 \end{aligned}$$

$$
\begin{aligned}
&= \frac{N-n}{nN(N-1)} \sum_{i=1}^{N} Y_i^2 - \frac{N-n}{n(N-1)} \overline{Y}^2 \\
&= \frac{N-n}{nN(N-1)} \left(\sum_{i=1}^{N} Y_i^2 - N\overline{Y}^2 \right) \\
&= \left(1 - \frac{n}{N}\right) \frac{S^2}{n} \, .
\end{aligned}
$$

A.3 Some properties of π_i and π_{ij} without replacement

For sampling without replacement, we have the following relations:

$$
\sum_{i=1}^{N} \pi_i = n; \qquad \sum_{j \neq i}^{N} \pi_{ij} = (n-1)\pi_i; \qquad \sum_{i=1}^{N} \sum_{j \neq i}^{N} \pi_{ij} = n(n-1) \, .
$$

Note that I_i has a value of 1 if the ith individual is selected in the sample, otherwise 0. To prove the first relation, note that $\pi_i = E[I_i]$ and we can write $\sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N} E[I_i]$. Because the sum of an expectation is the expectation of the sum, we can write $\sum_{i=1}^{N} \pi_i = E[\sum_{i=1}^{N} I_i]$ and the sum of I_i over the total population is equal to the sample size which is assumed to be fixed. Therefore, the sum of π_i over N is equal to n. To prove the second relation, note that π_{ij} is the inclusion probability that both individuals i and j are in the sample and $E[I_i \, I_j] = \pi_{ij}$. We use the similar argument that $\sum_{j \neq i}^{N} \pi_{ij} = \sum_{j \neq i}^{N} E[I_i I_j]$, and therefore

$$
\begin{aligned}
\sum_{j \neq i}^{N} \pi_{ij} &= \sum_{j \neq i}^{N} E\big[I_i I_j\big] \\
&= E\Big[\sum_{j \neq i}^{N} I_i I_j \Big] \\
&= E\Big[I_i \sum_{j \neq i}^{N} I_j \Big] \\
&= E\Big[\{I_i = 1\} \, E\Big[\sum_{j \neq i}^{N} I_j \Big| I_i \Big] \Big] \\
&= E\Big[\{I_i = 1\}(n-1) \Big] \\
&= (n-1)\pi_i \, .
\end{aligned}
$$

Therefore, the second relation is true. The third relation is easy to deduce from the first two.

APPENDIX B

B.1 Expectations for Sampling With Replacement

For sampling with replacement, the fixed sample size n can be considered as the number of independent draws. At each draw, an individual k has a probability p_k of being selected. Measurements y_i and Y_i are similarly defined as in the sampling with replacement context. Let I_{ik} be an indicator variable for the kth individual in the ith draw. If the kth individual is selected in the ith draw, $I_{ik} = 1$ with probability p_k, and if it is not selected in the ith draw, $I_{ik} = 0$ with probability $(1 - p_k)$. Consider the ith individual in the sample,

$$y_i = \sum_{k=1}^{N} I_{ik} Y_k \ .$$

The sum of y_i has an expectation

$$E\left[\sum_{i=1}^{n} y_i\right] = E\left[\sum_{i=1}^{n}\sum_{k=1}^{N} I_{ik} Y_k\right] = \sum_{k=1}^{N}\sum_{i=1}^{n} p_k Y_k = \sum_{k=1}^{N} \pi_k Y_k$$

where $\pi_k = \sum_{i=1}^{n} p_k = n p_k$. Also consider the expectation of $\sum_{i=1}^{n}\sum_{j\neq i}^{n} y_i\, y_j$

$$\begin{aligned} E\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n} y_i\, y_j\right] &= E\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k=1}^{N}\sum_{l=1}^{N} I_{ik} I_{jl} Y_k\, Y_l\right] \\ &= \sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k=1}^{N}\sum_{l=1}^{N} E\big[I_{ik} I_{jl}\big] Y_k\, Y_l \ . \end{aligned}$$

Because the draws i and j are independent $E\big[I_{ik} I_{jl}\big] = E\big[I_{ik}\big] E\big[I_{jl}\big]$ and hence

$$\begin{aligned} E\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n} y_i\, y_j\right] &= \sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k=1}^{N}\sum_{l=1}^{N} E\big[I_{ik}\big] E\big[I_{jl}\big] Y_k\, Y_l \\ &= \sum_{k=1}^{N}\sum_{l=1}^{N}\sum_{i=1}^{n}\sum_{j\neq i}^{n} p_k\, p_l\, Y_k\, Y_l \ . \end{aligned}$$

Let $\pi_{kl} = \sum_{i=1}^{n} p_k \sum_{j \neq i}^{n} p_l = n(n-1)p_k p_l$, we can re-write the above equation as

$$E\left[\sum_{i=1}^{n}\sum_{j\neq i}^{n} y_i\, y_j\right] = \sum_{k=1}^{N}\sum_{l=1}^{N} \pi_{kl}\, Y_k\, Y_l \ .$$

Another approach for the above results is by looking at the multiplicity of an individual k in n draws, that is, $J_k^n = \sum_{i=1}^{n} I_{ik}$ and $\sum_{i=1}^{n} y_i = \sum_{k=1}^{N} J_k^n Y_k$. Let the expectation of J_k^n be $\pi_k = E[J_k^n]$, it is easy to justify that $\pi_k = np_k$ which is the expected number of times the kth individual is selected in n draws. Now let $J_{kl}^n = \sum_{i=1}^{n} \sum_{j\neq i}^{n} I_{ik}\, I_{jl}$ which is the number of times that the individuals k and l are selected in n draws and we can write $\sum_{i=1}^{n} \sum_{j\neq i}^{n} y_i\, y_j = \sum_{k=1}^{N} \sum_{l=1}^{N} J_{kl}^n Y_k\, Y_l$ with $\pi_{kl} = E\big[J_{kl}^n\big] = E\big[\sum_{i=1}^{n} \sum_{j\neq i}^{n} I_{ik}\, I_{jl}\big]$. We can rewrite this expectation as

$$\begin{aligned}
E\big[J_{kl}^n\big] &= E\Big[\sum_{i=1}^{n}\sum_{j\neq i}^{n} I_{ik}\, I_{jl}\Big] \\
&= \sum_{i=1}^{n}\sum_{j\neq i}^{n} E\big[I_{ik}\, I_{jl}\big] \\
&= \sum_{i=1}^{n}\sum_{j\neq i}^{n} E\Big[\{I_{ik}=1\}\, E\Big[I_{jl}\Big|I_{ik}\Big]\Big] \\
&= \sum_{i=1}^{n} E\Big[\{I_{ik}=1\}\sum_{j\neq i}^{n} E\Big[I_{jl}\Big|I_{ik}\Big]\Big] \ .
\end{aligned}$$

Note that the probability of selecting individual k at each random draw is p_k which is the same for all i. The conditional expectation, $E\Big[I_{jl}\Big|I_{ik}\Big]$, is the same as $E\Big[\sum_{j=1}^{n-1} I_{jl}\Big] = E[J_l^{n-1}]$ because the draws are independent. Notice that $E[J_l^{n-1}]$ is not a function of i and hence,

$$\begin{aligned}
E\big[J_{kl}^n\big] &= E\Big[\sum_{i=1}^{n}\{I_{ik}=1\}\Big]\, E\Big[J_l^{n-1}\Big] \\
&= E\big[J_k^n\big] E\big[J_l^{n-1}\big] \\
&= n(n-1)p_k p_l
\end{aligned}$$

which can be interpreted as a product of the expected number of times the kth individual is selected in n draws and the expected number of times the lth individual is selected in $n-1$

draws.

As in sampling without replacement, let $g(y_i)$ and $g(Y_k)$ be functions of sample value y_i of the ith draw and population value Y_k of the kth unit respectively, the expectation of $\sum_{i=1}^{n} g(y_i)$ is given by

$$\begin{aligned} E\left[\sum_{i=1}^{n} g(y_i)\right] &= E\left[\sum_{i=1}^{N} J_k^n g(Y_k)\right] \\ &= \sum_{i=1}^{N} \pi_k \, g(Y_k) \end{aligned}$$

and let $g(y_i, y_j)$ and $g(Y_k, Y_l)$ be functions of sample values (y_i, y_j) and population value (Y_k, Y_l) respectively, the expectation of $\sum_{i=1}^{n} \sum_{j \neq i}^{n} g(y_i, y_j)$ is given by

$$\begin{aligned} E\left[\sum_{i=1}^{n} \sum_{j \neq i}^{n} g(y_i, y_j)\right] &= E\left[\sum_{k=1}^{N} \sum_{j=1}^{N} J_{kl'}^n \, g(Y_k, Y_l)\right] \\ &= \sum_{k=1}^{N} \sum_{j=1}^{N} \pi_{kl} \, g(Y_k, Y_l) \ . \end{aligned}$$

B.2 Example: Simple random sample with replacement

For simple random sample with replacement, the probability that an individual k in the population is selected at random in any single draw i is $p_k = 1/N$ for all i and $\pi_k = np_k$. The expectation of the sample mean is given by

$$E[\overline{y}] = E\left[\sum_{i=1}^{n} \frac{y_i}{n}\right] = \sum_{k=1}^{N} np_k \frac{Y_k}{n} = \sum_{i=1}^{N} \frac{Y_i}{N} = \overline{Y}$$

which completes the proof of unbiasedness. Note that $\pi_{kl} = n(n-1)\, p_k p_l = n(n-1)/N^2$. For the variance of the sample mean, we have

$$
\begin{aligned}
V\left[\sum_{i=1}^{n}\frac{y_i}{n}\right] &= E\left[\left(\sum_{i=1}^{n}\frac{y_i}{n}-\overline{Y}\right)^2\right] \\
&= \frac{1}{n^2}E\left[\left(\sum_{i=1}^{n}y_i\right)^2\right]-\overline{Y}^2 \\
&= \frac{1}{n^2}E\left[\sum_{i=1}^{n}y_i^2+\sum_{i=1}^{n}\sum_{j\neq i}^{n}y_i\,y_j\right]-\overline{Y}^2 \\
&= \frac{1}{n^2}\left(\sum_{k=1}^{N}\pi_k Y_k^2+\sum_{k=1}^{N}\sum_{l=1}^{N}\pi_{ij}Y_k\,Y_l\right)-\overline{Y}^2 \\
&= \frac{1}{n^2}\left(\sum_{k=1}^{N}np_k\,Y_k^2+\sum_{k=1}^{N}\sum_{l=1}^{N}n(n-1)p_kp_l\,Y_k\,Y_l\right)-\overline{Y}^2 \\
&= \frac{1}{n^2}\left(\sum_{k=1}^{N}\frac{n}{N}Y_k^2+\sum_{k=1}^{N}\sum_{l=1}^{N}\frac{n(n-1)}{N^2}Y_k\,Y_l\right)-\overline{Y}^2 \\
&= \frac{1}{nN}\left(\sum_{k=1}^{N}Y_k^2+\frac{n-1}{N}\left(\sum_{k=1}^{N}Y_k\right)^2\right)-\overline{Y}^2 \\
&= \frac{1}{nN}\left(\sum_{k=1}^{N}Y_k^2+N(n-1)\overline{Y}^2-nN\overline{Y}^2\right) \\
&= \frac{1}{nN}\left(\sum_{k=1}^{N}Y_k^2-\overline{Y}^2\right) \\
&= \left(1-\frac{1}{N}\right)\frac{S^2}{n}\ .
\end{aligned}
$$

B.3 Some properties of π_i and π_{ij} with replacement

For sampling with replacement, we have the following relations very similar to those in samplings without replacement:

$$
\sum_{k=1}^{N}\pi_k=n;\qquad \sum_{l=1}^{N}\pi_{kl}=(n-1)\pi_k;\qquad \sum_{k=1}^{N}\sum_{l=1}^{N}\pi_{kl}=n(n-1)\ .
$$

Note that $\sum_{k=1}^{N} p_k = 1$ for all draws and the first relation is proved. Since $\sum_{j \neq i}^{n} \sum_{l=1}^{N} p_l = (n-1)$, the second relation is then obvious. The third relation is a consequence of the first two. Also note that π_{kl} is not equal to $\pi_k \, \pi_l$ but

$$\pi_{kl} = n(n-1)p_k p_l = \frac{(n-1)}{n} \pi_k \pi_l$$

for all k and l.

References

[1] Cochran, W. G. (1977). *Sampling Technique.* John Wiley & Sons, 3rd ed.

[2] Lohr, S. L. (1999). *Sampling: Design and Analysis.* Duxbury Press.

[3] Särndal, C. E., Swensson, B. & Wretman, J. H. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

[4] Stuart, A. (1963). Some remarks on sampling with unequal probabilites. *Bulletin of International Statistical Institute* **40**, 773–779.