Bulletin of the Iranian Mathematical Society Vol. 38 *No. 2 (2012), pp 333-347.*

HORVITZ-THOMPSON ESTIMATOR OF POPULATION MEAN UNDER INVERSE SAMPLING DESIGNS

M. MOHAMMADI* AND M. SALEHI M.

Communicated by Ahmad Reza Soltani

Abstract. Inverse sampling design is generally considered to be an appropriate technique when the population is divided into two subpopulations, one of which contains only a few units. Here, we derive the Horvitz-Thompson estimator for the population mean under inverse sampling designs, where subpopulation sizes are known. We then introduce an alternative unbiased estimator, corresponding to post-stratification approach. Both of these are not location-invariant, but this is ignorable for alternative estimator. Using a simulation study, we find that the Horvitz-Thompson estimator is an efficient estimator when the mean of the off-interest subpopulation is close to zero, while the alternative estimator appears to be an efficient estimator in general.

1. Introduction

Inverse sampling design is considered to be an efficient strategy for estimating the population parameters when only few units represent the characteristic of interest. In many situations, the main reason of implementing inverse sampling is not efficiency, in the sense of having smaller variance. Whenever the data from a survey are intended to be used for statistical analysis such as contingency table or logistics regression, we

MSC(2010): Primary: 65F05; Secondary: 11Y50.
Keywords: Finite population, inverse sampling, post-stratification, random sample size.
Received: 25 March 2010, Accepted: 4 December 2010.
*Corresponding author

need to have at least a certain number of units from a rare subpopulation in order to have a valid statistical conclusion. Inverse sampling design has been used to estimate the population proportion (Haldane, [6]). Salehi and Seber [9], using Murthy's estimator [8], obtained an unbiased estimator of the population mean under a simple inverse sampling design. Christman and Lan [3] introduced three inverse sampling designs using stopping rules based on the number of selected units in the rare subpopulation. Salehi and Seber [10] suggested a general version of inverse sampling.
All of the unbiased estimators for the population mean in the mentioned literature are derived based on the assumption of unknown subpopulation sizes. However, if these are known, then we may derive more efficient estimators, similar to post-stratification approach (Cochran, [4]).

Here, we derive the Horvitz-Thompson estimator for the population mean under inverse sampling designs when the size of subpopulations are known. We also consider another unbiased weighted estimator based on post-stratification idea. We will compare the precision of the proposed estimators. The precision of both estimators depend on the coefficient of variation of two subpopulations and the square of the off-interested subpopulation mean. The Horvitz-Thompson estimator is sensitive to distance of the subpopulation mean from zero, while the other estimator is more stable.

2. Horvitz-Thompson estimator under inverse sampling design

Here, we derive the Horvitz-Thompson estimator for the population mean under which the size of subpopulations are known. Such a situation is more likely to occur when the subpopulations are recognized by their domains rather than their y-values.
Suppose that a finite population $U = \{u_1, u_2, \ldots, u_N\}$ of N units is divided into two subpopulations U_C and $U_{\bar{C}}$, with the corresponding sizes M and $N - M$, respectively. With any unit u_k, there is an associated value of the variable of interest y_k, for $k = 1, 2, \ldots, N$. An unbiased estimator of the population mean may be found using the Horvitz-Thompson (HT) estimator. Let s denote the sample set and $\pi_k = P_r(k \in s)$ be the inclusion probability of the kth unit. Hence, the

HT estimator for the mean, say μ, is

$$\hat{\mu}_{HT} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k}. \tag{2.1}$$

The variance of $\hat{\mu}_{HT}$ is

$$\mathrm{Var}(\hat{\mu}_{HT}) = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl}-\pi_k\pi_l)\frac{y_k y_l}{\pi_k\pi_l}, \tag{2.2}$$

with unbiased estimate

$$v(\hat{\mu}_{HT}) = \frac{1}{N^2}\sum_{k\in s}\sum_{l\in s}(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}})\frac{y_k y_l}{\pi_k\pi_l}, \tag{2.3}$$

where π_{kl} is the joint inclusion probability of the kth and the lth units, with $\pi_{kk} = \pi_k$.

2.1. **Simple inverse sampling design.** In Simple Inverse Sampling (SIS) design, units are selected one by one without replacement and equal probabilities until a predetermined number, say r, of $U_{\bar{C}}$ is observed. Let n_s be the random size of s, s_C be the the sample set from U_C, and $s_{\bar{C}} = s - s_C$. In this manner, s_C is equivalent to a simple random sample from U_C, while $s_{\bar{C}}$ conditionally on its size is a simple random sample from $U_{\bar{C}}$.
To find the HT estimator under SIS design we need to obtain the first and second order inclusion probabilities. It can be shown that (see Appendix A):

$$\pi_k = \begin{cases} \frac{r}{M}, & if \quad k \in U_C, \\ \frac{r}{M+1}, & if \quad k \in U_{\bar{C}}. \end{cases} \tag{2.4}$$

The joint inclusion probabilities for the distinct units k and l are given by

$$\pi_{kl} = \begin{cases} \frac{r(r-1)}{M(M-1)}, & if \quad (k,l) \in U_C \\ \frac{r(r+1)}{(M+1)(M+2)}, & if \quad (k,l) \in U_{\bar{C}} \\ \frac{r^2}{M(M+1)}, & if \quad k \in U_C, l \in U_{\bar{C}}. \end{cases} \tag{2.5}$$

Hence, the Horvitz-Thompson estimator is found to be

$$\hat{\mu}_{HT} = \frac{1}{Nr}\left(M\sum_{k\in s_C} y_k + (M+1)\sum_{k\in s_{\bar{C}}} y_k\right). \tag{2.6}$$

By substituting (2.4) and (2.5) into (2.2), and by some algebraic operations, we obtain the variance of $\hat{\mu}_{HT}$ to be

$$\mathrm{Var}(\hat{\mu}_{HT}) = (\frac{M}{N})^2(\frac{M-r}{M-1})\frac{\sigma_C^2}{r} + \frac{(M-r+1)(N-M)}{N^2(M+2)r} \left[(N+1)\bar{y}_{U_{\bar{C}}}^2 + (M+1)\sigma_{\bar{C}}^2\right], \tag{2.7}$$

where $\sigma_{\bar{C}}^2 = (N-M)^{-1}\sum_{k\in U_{\bar{C}}}(y_k - \bar{y}_{U_{\bar{C}}})^2$, $\bar{y}_{U_{\bar{C}}} = (N-M)^{-1}\sum_{k\in U_{\bar{C}}} y_k$ and $\sigma_C^2 = \frac{1}{M}\sum_{k\in U_C}(y_k - \bar{y}_{U_C})^2$, with $\bar{y}_{U_C} = \frac{1}{M}\sum_{k\in U_C} y_k$. The equation (2.7) indicates that for the fixed subpopulation variances σ_C^2 and $\sigma_{\bar{C}}^2$, $\hat{\mu}_{HT}$ has the maximum efficiency when the mean of y-values in the $U_{\bar{C}}$ are zero. An unbiased estimator of $\mathrm{Var}(\hat{\mu}_{HT})$ is obtained using equation (2.3) as:

$$v(\hat{\mu}_{HT}) = (\frac{M}{N})^2(1-\frac{r}{M})\frac{s_C^2}{r} + \frac{(M+1)(M-r+1)}{N^2 r(r+1)} \left[\sum_{k\in s_{\bar{C}}} y_k^2 + \frac{1}{r}\left(\sum_{k\in s_{\bar{C}}} y_k\right)^2\right],$$

where

$$s_C^2 = \frac{1}{r-1}\sum_{k\in s_C}(y_k - \bar{y}_{s_C})^2 \quad ; \quad \bar{y}_{s_C} = \frac{1}{r}\sum_{k\in s_C} y_k.$$

2.2. **Other inverse sampling designs.** Some other versions of inverse sampling including variable stopping rules and general inverse sampling methods are considered here. Christman and Lan [3] defined an inverse sampling procedure as follows.
First, a simple random sample of fixed size n_0 is selected. If at least r units from U_C are selected, then the sampling procedure is stopped. Otherwise, the sampling is continued until r units in U_C are selected.
Salehi and Seber [10] suggested a general inverse sampling (GIS) design

which is useful to have control on the total sample size. In their procedure, a simple random sample of size n_0 is selected. If the selected sample contains at least r units in U_C, then the sampling is stopped. Otherwise, the sampling is continued sequentially until either exactly r units from U_C are selected or the sample size reaches a predetermined value of n_1. In the particular case of $n_1 = N - M + r$, GIS reduces to the procedure of variable stopping rules.
To derive the Horvitz-Thompson estimator and its variance under above designs, we need to find the first and second order inclusion probabilities. Unfortunately, these have not simple forms, specially for general inverse sampling design. However, we present them for the variable stopping rules design in Appendix B.

2.3. **Alternative estimator.** Doss et al. [5] introduced an unbiased estimator for the population mean based on post-stratification under simple random sample. Following their approach, the estimator under general inverse sampling design is giving by

$$\hat{\mu}_{alt} = \frac{M}{N} u_{C,n_1} \bar{y}_{s_C} + \frac{N-M}{N} u_{\bar{C},n_0} \bar{y}_{s_{\bar{C}}}, \tag{2.8}$$

where

$$u_{C,n_1} = \frac{I(n_{s_C} > 0)}{1 - p_C} \qquad and \qquad u_{\bar{C},n_0} = \frac{I(n_{s_{\bar{C}}} > 0)}{1 - p_{\bar{C}}},$$

with $p_C = \binom{N-M}{n_1} / \binom{N}{n_1}$, $p_{\bar{C}} = \binom{M}{n_0} / \binom{N}{n_0}$, and $I(.)$ denoting the indicator function, which has value of 1 if condition (.) holds, and zero otherwise. Let $p^*_{\cdot} = p_{\cdot}/(1 - p_{\cdot})$. Hence, the variance of $\hat{\mu}_{alt}$ is

(2.9)

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_{alt}) = {} & \left(\frac{M}{N}\right)^2 (1 + p^*_C)^2 \left\{ \sum_{x \geq 1} \frac{P_r(n_{s_C} = x)}{x} - \frac{1}{M(1 + p^*_C)} \right\} S^2_C + \\
& \left(\frac{N-M}{N}\right)^2 (1 + p^*_{\bar{C}})^2 \left\{ \sum_{x \geq 1} \frac{P_r(n_{s_{\bar{C}}} = x)}{x} - \frac{1}{(N-M)(1 + p^*_{\bar{C}})} \right\} S^2_{\bar{C}} + \\
& \left(\frac{M}{N}\right)^2 p^*_C \bar{y}^2_{U_C} + \left(\frac{N-M}{N}\right)^2 p^*_{\bar{C}} \bar{y}^2_{U_{\bar{C}}} - 2 \frac{M(N-M)}{N^2} p^*_C p^*_{\bar{C}} \bar{y}_{U_C} \bar{y}_{U_{\bar{C}}},
\end{aligned}$$

where n_{s_C} and $n_{s_{\bar{C}}}$ are the sample sizes of s_C and $s_{\bar{C}}$, respectively, and $S^2_C = M/\sigma^2_C/(M-1)$, $S^2_{\bar{C}} = (N-M)\sigma^2_{\bar{C}}/(N-M-1)$.

The formulas for the variances of the two estimators $\hat{\mu}_{HT}$ and $\hat{\mu}_{alt}$ do not allow analytical comparison of their efficiencies, and for this purpose we only provide the results of a small simulation study in the next section. However, some simplifications are obtained in the case of SIS, where we have $p^*_C = 0$, and $P_r(n_{s_C} = r) = 1$.
A serious drawback for the unbiased estimators $\hat{\mu}_{HT}$ and $\hat{\mu}_{alt}$ is that their variances depend on the origin of the y-values for units in the population subgroups $U_{\bar{C}}$. In the case of SIS, if the mean of y-values in $U_{\bar{C}}$ is zero, then this problem will be removed. However, this problem is ignorable for $\hat{\mu}_{alt}$ in the ordinary rare populations, since p^*_C and $p^*_{\bar{C}}$ are close to zero. For example, if $N = 400$, $M = 20$ and $r = 5$, then under simple inverse sampling design, $p^*_C = 0$ and the value of $p^*_{\bar{C}}$ is less than 1.87×10^{-7}. In this case, $\hat{\mu}_{alt}$ converges to the customary post-stratification estimator (see, Chang et al. ([1, 2])). In other cases, to eliminate the dependence of the variance of mentioned estimators on the origin of the y-values, we can use the ratio estimator suggested by Doss et al. [5]. However, our simulation shows that, in this case, $\hat{\mu}_{alt}$ is more efficient than $\hat{\mu}_{HT}$ and no improvement is provided using the ratio estimator based on $\hat{\mu}_{alt}$.

3. **Simulation study**

In this section, we conduct a small simulation study to investigate the efficiency of the derived unbiased estimators of the mean under inverse sampling designs, $\hat{\mu}_{HT}$ and $\hat{\mu}_{alt}$. We compare the variances of these with the corresponding estimator proposed by Doss et al. [5] under inverse sampling design with post-stratification approach. Define $\hat{\mu}_{y.pst}$ as

$$\hat{\mu}_{y.pst} = \frac{M}{N} u_{C,n} \bar{y}_{s_C} + \frac{N-M}{N} u_{\bar{C},n} \bar{y}_{s_{\bar{C}}}. \tag{3.1}$$

Hence, the ratio estimator of the population mean under simple random sampling is of the form

$$\tilde{\mu}_{pst} = \frac{\hat{\mu}_{y.pst}}{\hat{\mu}_{1.pst}}. \tag{3.2}$$

The variance of $\tilde{\mu}_{pst}$ does not have an explicit form to allow for the analytic comparison of the inverse sampling estimators. However, we give a relative efficiency of $\tilde{\mu}_{pst}$ via a small simulation study. To have a fair comparison, we fix the expected sample size of inverse sampling for

the simple random sampling design, to be $E(n_s) = (N+1)r/(M+1)$. In the case of variable stopping rules, we will compare the results with the corresponding post-stratified estimator based on simple random sample with size $E(n_s)$. It can be shown that

$$E(n_s) = n_0 + \frac{N+1}{M+1} \sum_{r_0=0}^{r-1} \frac{\binom{N-M}{n_0-r_0}\binom{M+1}{r_0}}{\binom{N+1}{n_0}}(r-r_0).$$

We consider two methods of SIS and the variable stopping rules. We consider three underlying models to generate the finite population. These are:

- *Model 1* : $F_{\bar{C}} \sim DG(0)$; $F_C \sim Exp(10, 0.05)$,
- *Model 2* : $F_{\bar{C}} \sim N(0,3)$; $F_C \sim N(40,8)$,
- *Model 3* : $F_{\bar{C}} \sim (|N(0,3)|)$; $F_C \sim Exp(10, 0.05)$,

where DG and Exp are abbreviation of the degenerated and exponential distributions, respectively, and $|N(.,.)|$ is absolute value of the normal distribution.

The population size is $N = 400$. For each model, we consider two types of populations: rare with $M = 20$ and common with $M = 40$. Also, we consider three values for r as $r = 3, 4, 5$. In the case of variable stopping rules, we use two cases $n_0 = E(n_s)/3, 2E(n_s)/3$. The results are shown for SIS in Table 1, and for variable stopping rules in tables 2 and 3.

Table 1. Mean square error of population mean estimators under two simple inverse and simple random sampling designs, with subpopulation size M.

Population	Model	r	$\hat{\mu}_{HT}$	$\hat{\mu}_{alt}$	$\tilde{\mu}_{pst}$
Common	1	3	0.802	0.802	1.873
		4	0.588	0.588	1.204
		5	0.453	0.453	0.839
	2	3	0.497	0.678	2.436
		4	0.374	0.461	1.290
		5	0.293	0.338	0.733
	3	3	2.435	0.978	1.715
		4	1.797	0.694	1.192
		5	1.393	0.535	0.892
Rare	1	3	0.221	0.221	0.477
		4	0.161	0.161	0.322
		5	0.119	0.119	0.225
	2	3	0.226	0.312	0.784
		4	0.154	0.194	0.416
		5	0.116	0.137	0.243
	3	3	1.958	0.288	0.560
		4	1.368	0.194	0.370
		5	0.993	0.144	0.255

Table 2. Mean square error of population mean estimators $\hat{\mu}_{HT}$ and $\hat{\mu}_{alt}$ under variable stopping rules with $n_0 = \frac{E(n_s)}{3}$ and post-stratification estimator $\tilde{\mu}_{pst}$ under simple random sampling design, with subgroup size M.

Population	Model	r	$\hat{\mu}_{HT}$	$\hat{\mu}_{alt}$	$\tilde{\mu}_{pst}$
Common	1	3	0.818	0.806	1.831
		4	0.572	0.580	1.218
		5	0.457	0.450	0.832
	2	3	0.542	0.649	2.305
		4	0.376	0.447	1.216
		5	0.298	0.336	0.725
	3	3	2.394	0.955	1.669
		4	1.816	0.687	1.179
		5	1.464	0.532	0.864
Rare	1	3	0.218	0.213	0.455
		4	0.130	0.127	0.305
		5	0.098	0.098	0.209
	2	3	0.225	0.275	0.702
		4	0.164	0.188	0.389
		5	0.119	0.135	0.220
	3	3	1.809	0.272	0.549
		4	1.468	0.197	0.360
		5	1.115	0.144	0.242

Table 3. Mean square error of population mean estimators $\hat{\mu}_{HT}$ and $\hat{\mu}_{alt}$ under variable stopping rules with $n_0 = \frac{2E(n_s)}{3}$ and post-stratification estimator $\tilde{\mu}_{pst}$ under simple random sampling design, with subpopulation size M.

Population	Model	r	$\hat{\mu}_{HT}$	$\hat{\mu}_{alt}$	$\tilde{\mu}_{pst}$
Common	1	3	0.972	0.774	1.535
		4	0.666	0.560	1.028
		5	0.516	0.439	0.709
	2	3	0.882	0.536	1.614
		4	0.579	0.402	0.876
		5	0.414	0.305	0.512
	3	3	1.849	0.893	1.506
		4	1.468	0.664	1.064
		5	1.143	0.520	0.754
Rare	1	3	0.256	0.210	0.393
		4	0.166	0.136	0.261
		5	0.125	0.105	0.185
	2	3	0.296	0.232	0.516
		4	0.196	0.163	0.283
		5	0.142	0.123	0.174
	3	3	1.240	0.250	0.461
		4	1.003	0.182	0.304
		5	0.800	0.134	0.210

4. **Conclusion**

We derived several estimators for the population mean using inverse sampling deign under the condition that the population groups $(U_C, U_{\bar{C}})$ have known sizes $(M, N-M)$. Our simulation study showed that when M is known, we can achieve more efficient estimates of the population mean with inverse sampling designs as opposed to simple random sampling design using post-stratified estimator. In the case of variable stopping rules, we obtained similar results for simple inverse sampling design if n_0 is chosen as $E(n_s)/3$.

For model 1 representing a population with $y_k = 0$ for any unit in the off-interested subpopulation $U_{\bar{C}}$, $\hat{\mu}_{HT}$ and $\hat{\mu}_{alt}$ are equivalent under

simple random sampling design. The efficiency of $\hat{\mu}_{HT}$ and/or $\hat{\mu}_{alt}$ is almost twice that of $\hat{\mu}_{pst}$.

If $\bar{y}_{U_{\bar{C}}} = 0$, $S^2_{\bar{C}} > 0$, model 2, then $\hat{\mu}_{HT}$ is even more efficient than $\hat{\mu}_{alt}$, when n_0 is not greater than $E(n_s)/3$. As n_0 increases, the precision of $\hat{\mu}_{HT}$ improves over $\hat{\mu}_{alt}$. Contrary to other estimators, when $\bar{y}_{U_{\bar{C}}} = 0$ (model 1 and model 2), the precision of $\hat{\mu}_{HT}$ decreases as the initial sample size n_0 increases from $E(n_s)/3$ to $2E(n_s)/3$. For model 3, having a population with small y-values in the $U_{\bar{C}}$ for which $\hat{\mu}_{HT}$ is very sensitive to distance of the mean from zero in $U_{\bar{C}}$, we may even get less efficiency than $\tilde{\mu}_{pst}$. In the case of variable stopping rules, the estimator $\hat{\mu}_{alt}$ has a better behavior than the others.

Appendix A: Derivations of π_k and π_{kl} under simple inverse sampling

If inverse sampling is without replacement until r units in U_C are observed, then sub-sample s_C is a simple random sample from U_C, and so for any $(k \neq l) \in U_C$, $\pi_k = r/M$ and $\pi_{kl} = r(r-1)/M(M-1)$, (see Särndal et al. [11]). On the other hand, the sub-sample $s_{\bar{C}}$ condition on its size $n_s - r$ is a simple random sample from $U_{\bar{C}}$. The total sample size n_s is a negative hypergeometric random variable (Johnson, et al. [7]) with the probability function

$$P_r(n_s = n) = \frac{\binom{N-n}{M-r}\binom{n-1}{r-1}}{\binom{N}{M}}, \qquad n = r, \ldots, N - M + r,$$

and mathematical expectation and variance respectively given by

$$E(n_s) = \frac{r(N+1)}{M+1} \quad , \quad \mathrm{Var}(n_s) = r\frac{(N+1)(N-M)(M-r+1)}{(M+1)^2(M+2)}.$$

Hence, for any $k \in U_{\bar{C}}$, we have

$$\pi_k = E(I(k)) = E_{n_s}\left[E(I(k))|n_s\right] = E_{n_s}\left[\frac{n_s - r}{N - M}\right]$$

$$= \frac{1}{N-M}\left(\frac{r(N+1)}{M+1} - r\right) = \frac{r}{M+1}.$$

To find the second-order inclusion probabilities, we again use the conditional property of mathematical expectation. For $(k \neq l) \in U_{\bar{C}}$, we

have

$$\begin{aligned}\pi_{kl} = E(I(k)I(l)) &= E_{n_s}\left[E(I(k)I(l))|n_s\right] \\ &= E_{n_s}\left[\frac{(n_s-r)(n_s-r-1)}{(N-M)(N-M-1)}\right] \\ &= \frac{\left(\frac{r^2(N-M)^2}{(M+1)^2}+\frac{r(N+1)(N-M)(M-r+1)}{(M+1)^2(M+2)}-\frac{r(N-M)}{M+1}\right)}{(N-M)(N-M-1)}.\end{aligned}$$

Some algebraic simplifications yield:

$$\pi_{kl} = \frac{r(r+1)}{(M+1)(M+2)}.$$

Finally, for any $k \in U_C$ and $l \in U_{\bar{C}}$, we have

$$\begin{aligned}\pi_{kl} = E(I(k)I(l)) &= E_{n_s}\left[E(I(k)I(l))|n_s\right] \\ &= E(I(k))E_{n_s}\left[I(l)|n_s\right] \\ &= \frac{r^2}{M(M+1)}.\end{aligned}$$

Appendix B: Derivations of π_k and π_{kl} in the case of variable stopping rules

For variable stopping rules, let A_k and B_k be the events of selecting unit k in the initial sample of size n_0 (the first stage), and in the sequential sampling process (the second stage), respectively. So, we have

$$\pi_k = P(A_k) + P(A_k^c)P(B_k|A_k^c),$$

where A^c denotes the complementary of A. It can be shown that

$$\pi_k = \frac{n_0}{N} + \left(1-\frac{n_0}{N}\right)\sum_{r_0=0}^{r-1} P(S_{r_0}|A_k^c)P(B_k|S_{r_0},A_k^c),$$

where S_{r_0} is the event that r_0 units in U_C are selected in the first stage of sampling. Hence,

$$(B.1) \qquad P(S_{r_0}|A_k^c) = \begin{cases} \frac{\binom{N-M}{n_0-r_0}\binom{M-1}{r_0}}{\binom{N-1}{n_0}}, & if \quad k \in U_C \\ \\ \frac{\binom{N-M-1}{n_0-r_0}\binom{M}{r_0}}{\binom{N-1}{n_0}}, & if \quad k \in U_{\bar{C}}. \end{cases}$$

Using the equation (2.4), we have

$$(B.2) \qquad P(B_k|S_{r_0}, A_k^c) = \begin{cases} \frac{r-r_0}{M-r_0}, & if \quad k \in U_C \\ \\ \frac{r-r_0}{M-r_0+1}, & if \quad k \in U_{\bar{C}}. \end{cases}$$

Upon substituting (B.1) and (B.2) into π_k, we get

$$\pi_k = \begin{cases} \frac{n_0}{N} + P_M(r-r_0), & if \quad k \in U_C \\ \\ \frac{n_0}{N} + P_{M+1}(r-r_0), & if \quad k \in U_{\bar{C}}, \end{cases}$$

where

$$P_\alpha(\beta) = \frac{1}{\alpha} \sum_{r_0=0}^{r-1} \frac{\binom{N-\alpha}{n_0-r_0}\binom{\alpha}{r_0}}{\binom{N}{n_0}} \beta.$$

The second order inclusion probability for distinct units k and l is

$$\pi_{kl} = P(A_k \cap A_l) + P(A_k \cap B_l) + P(B_k \cap A_l) + P(B_k \cap B_l).$$

The first term of the above probability is $n_0(n_0-1)/N(N-1)$, for any $(k \neq l) = 1, 2, \ldots, N$, and $P(A_k \cap B_l) = P(B_k \cap A_l)$, for any pair $(k \neq l)$ in the same subpopulation. Again, using the conditional probability, we find

$$\begin{aligned} P(A_k \cap B_l) &= P(A_k \cap A_l^c) P(B_l | A_k \cap A_l^c) \\ &= \frac{n_0(N-n_0)}{N(N-1)} \sum_{r_0=0}^{r-1} P(S_{r_0}|A_k \cap A_l^c) P(B_l|S_{r_0}, A_k \cap A_l^c). \end{aligned}$$

Using the equations (B.1) and (B.2) we get

$$(B.3) \qquad P(S_{r_0}|A_k \cap A_l^c) P(B_l|S_{r_0}, A_k \cap A_l^c) =$$

$$\begin{cases} \frac{\binom{N-M}{n_0-r_0}\binom{M-2}{r_0-1}}{\binom{N-2}{n_0-1}} \left(\frac{r-r_0}{M-r_0}\right), & if \quad (k \neq l) \in U_C \\ \\ \frac{\binom{N-M-1}{n_0-r_0}\binom{M-1}{r_0-1}}{\binom{N-2}{n_0-1}} \left(\frac{r-r_0}{M-r_0+1}\right), & if \quad k \in U_C, l \in U_{\bar{C}} \\ \\ \frac{\binom{N-M-1}{n_0-r_0-1}\binom{M-1}{r_0}}{\binom{N-2}{n_0-1}} \left(\frac{r-r_0}{M-r_0}\right), & if \quad k \in U_{\bar{C}}, l \in U_C \\ \\ \frac{\binom{N-M-2}{n_0-r_0-1}\binom{M}{r_0}}{\binom{N-2}{n_0-1}} \left(\frac{r-r_0}{M-r_0+1}\right), & if \quad (k \neq l) \in U_{\bar{C}}. \end{cases}$$

On the other hand,

$$P(B_k \cap B_l) = P(A_k^c \cap A_l^c) P(B_k \cap B_l | A_k^c \cap A_l^c)$$

$$= \frac{(N-n_0)(N-n_0-1)}{N(N-1)} \sum_{r_0=0}^{r-1} P(S_{r_0} | A_k^c \cap A_l^c) P(B_k \cap B_l | S_{r_0}, A_k^c \cap A_l^c),$$

which results in

$$\text{(B.4)} \qquad P(S_{r_0} | A_k \cap A_l^c) P(B_k \cap B_l | S_{r_0}, A_k \cap A_l^c) =$$

$$\begin{cases} \frac{\binom{N-M}{n_0-r_0}\binom{M-2}{r_0}}{\binom{N-2}{n_0}} \frac{(r-r_0)(r-r_0-1)}{(M-r_0)(M-r_0-1)}, & if \quad (k \neq l) \in U_C \\ \frac{\binom{N-M-1}{n_0-r_0}\binom{M-1}{r_0}}{\binom{N-2}{n_0}} \frac{(r-r_0)^2}{(M-r_0)(M-r_0+1)}, & if \quad k \in U_C, l \in U_{\bar{C}} \\ \frac{\binom{N-M-2}{n_0-r_0}\binom{M}{r_0}}{\binom{N-2}{n_0}} \frac{(r-r_0)(r-r_0+1)}{(M-r_0+1)(M-r_0+2)}, & if \quad (k \neq l) \in U_{\bar{C}}. \end{cases}$$

Substituting (B.3) and (b.4) in π_{kl}, we find after some simplifications,

$$\pi_{kl} = \begin{cases} \frac{n_0(n_0-1)}{N(N-1)} + \frac{P_M((r-r_0)(r+r_0-1))}{M-1}, & if\ (k \neq l) \in U_C \\ \frac{n_0(n_0-1)}{N(N-1)} + \frac{P_M((n_0-r_0)(r-r_0))}{N-M} + \frac{P_{M+1}(r(r-r_0))}{M}, & if\ k \in U_C, l \in U_{\bar{C}} \\ \frac{n_0(n_0-1)}{N(N-1)} + 2\frac{P_{M+1}((n_0-r_0)(r-r_0))}{N-M-1} + \frac{P_{M+1}((r-r_0)(r-r_0+1))}{M}, & if\ (k \neq l) \in U_{\bar{C}}. \end{cases}$$

References

[1] K. C. Chang, J. F. Lio, and C. P. Han, Multiple inverse sampling in post-stratification, *J. Statist. Plann. and Inference* **69** (1998), no. 2, 209-227.

[2] K. C. Chang, C. P. Han, and D. L. Hawkins, Truncated multiple inverse sampling in post-stratification, *J. Statist. Plann. and Inference* **76** (1999), no. 1-2, 215-234.

[3] M. C. Christman and F. Lan, Inverse adaptive cluster sampling, *Biometrics* **57** (2001), no. 4, 1096-1105.

[4] W. G. Cochran, Sampling Techniques, John Wiley & Sons, New York, 1977.

[5] D. C. Doss, H. O. Hartley, and G. R. Somayejulu, An exact small sample theory for post-stratification, *J. Statist. Plann. and Inference* **3** (1979), no. 3, 235-248.

[6] J. B. S. Haldane, On a method of estimating frequencies, *Biometrika* **33** (1945), 222-225.
[7] N. L. Johnson, S. Kotz, and A. W. Kemp, Univariate Discrete Distributions, 2^{nd} ed., Wiley, 1993.
[8] M. N. Murthy, Ordered and unordered estimators in sampling without replacement, *Sankhyã* **18** (1957), 379-390.
[9] M. M. Salehi and A. F. Seber, A new proof of murthy's estimator which applies to sequential sampling, *Aust. N. Z. J. Stat.* **43** (2001), no. 3, 281–286.
[10] M. M. Salehi and A. F. Seber, A general inverse sampling scheme and its application to adaptive cluster sampling, *Aust. N. Z. J. Stat.* **46** (2004), no. 3, 483–494.
[11] C. E. Särndal, B. Swensson, and J. H. Wretman, Model Assisted Survey Sampling, Springer-Verlag, New York, 1992.

Mohammad Mohammadi
Department of Mathematical Science, Isfahan University of Technology, Isfahan, 84156-83111, Iran
Email: m.mohammadi@math.iut.ac.ir

Mohammad Salehi M.
Department of Mathematics, Statistics and Physics, Qatar University, P.O. Box 2713, Doha, Qatar
and
Department of Mathematical Science, Isfahan University of Technology, Isfahan, 84156-83111, Iran
Email: salehi@qu.edu.qa