CHAPTER 26

Sampling Rare and Mobile Populations

26.1 INTRODUCTION

A rare population is defined as a small subset of a population of interest. Smallness, although subjective, is generally treated as one-tenth, one-hundredth, or even less. For example, surveys on physically disabled persons, child labors, ethnic minority groups, households with very high income, persons with rare diseases, etc. focus on rare populations. Mobile populations include migratory birds, visitors to places of historical interest or shopping malls, pavement dwellers, and hospital outpatients. The main objectives of surveying rare and mobile populations are as follows:

(i) To find the number (M) or the prevalence rate $P = M/N$ where N is the size of the total population. For example, P may be the prevalence of the HIV infection of the total population.

(ii) To estimate characteristics y of the rare or mobile population such as mean income of a child labor. Let y_i be the value of y for the ith unit and let $\delta_i = 1$ if the ith unit belongs to the rare population and $\delta_i = 0$ otherwise, then the size of the rare population and the rare population mean of the character y are, respectively,

$$M = \sum_{i=1}^{N} \delta_i \text{ and } \overline{Y} = \sum_{i=1}^{N} y_i \delta_i / M$$

If there is a separate frame for the rare population, then the sample of rare population can be selected very easily by following any standard sampling technique. But in reality such a frame is rarely available. Furthermore, most surveys are multicharacter surveys where information on several characteristics are collected at the same time with some of the characteristics being rare and others not. For example, in a household survey, we collect information on the income of the household as well as if they have been a victim of theft, among others. In this case, being victim of theft is a rare event but inquiring about household income is not. A number of methods have been proposed

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00026-1

Copyright © 2017 Raghunath Arnab.
Published by Elsevier Ltd.
All rights reserved.

for sampling of rare and mobile populations when a separate list is not available. Good details are given by Kish (1965, 1991), Sudman (1976), Kalton and Anderson (1986), Sudman and Kalton (1986), Thompson and Seber (1996), and Kalton (2001). Some of the methods are as follows:

(i) screening, (ii) disproportionate sampling, (iii) multiplicity or network sampling, (iv) multiframe sampling, (v) snowball sampling, (vi) location sampling, (vii) sequential sampling, (viii) adaptive sampling, and (ix) capture—recapture method.

26.2 SCREENING

In the case that members of the rare population can be identified from the information given in the sampling frame, the frame should be cleaned by deleting the nonrare population from the list and the sample should be selected from the cleaned frame. If any nonrare member is selected in the sample, he/she may be eliminated at the time of data collection. However, the sampling frame generally does not contain enough information to identify whether a unit belongs to the rare population or not. In this case a relatively large sample may be selected from the population and then the members of the rare population may be identified at the stage of data collection. The cost of screening for rare population will be high if the degree of rarity is high. In this case the following routes may be used.

26.2.1 Telephonic Interview

If the majority of households (units) have telephone facilities, then the population can be screened by telephone calls, which will incur a small cost. The sampled people can then be interviewed face to face. Obviously, telephone interviewing underrepresents households that cannot be accessible through telephone calls, e.g., households with low income, the deaf, those with limited activity due to illness, single or divorced heads. The telephonic interview is widely used in the United States for various surveys.

26.2.2 Mail Questionnaire

The mail questionnaire method can be used if full addresses of the households are available. The questionnaires should include items of information that can identify whether the household is a member of rare population or not. The mail questionnaire method was found very useful in the United Kingdom. The response was very high at 86% (Kalton and Anderson, 1986) and even more in certain areas.

26.2.3 Cluster Sampling

The cluster sampling method, although less efficient, from the efficiency point of view is useful for reducing the cost of data collection, as the travel cost from one unit to the other is minimum. If the rare population is underrepresented in many clusters and cannot be identified in advance, then the sampling of the rare population becomes unproductive and expensive because of the high screening cost. Sudman (1972, 1978, 1985) and Wakesberg (1978) provided a method of selection of clusters that prevents selection of clusters that do not contain a rare population. In this method clusters are selected with probability proportional to the number of rare population of the cluster.

26.2.3.1 Sudman–Waksberg Method

Suppose a population consists of N primary stage units (psu's) and each of them consists of M subunits. Let X_i be the number of a rare population that belongs to the ith psu, $i = 1,\ldots,N$. The number X_i is unknown and no listing of the subunits is available. Here, we assume that a cluster either contains no member of rare population or contains at least $n(k+1)$ members of rare population. Here, n and $k+1$ denote the desired number of psu's and the number of members of rare population from each of psu's to be selected in the sample, respectively. The values of n and k are determined optimally from the consideration of efficiency or cost of the survey.

In the Sudman–Waksberg method a psu is chosen at random and one unit (subunit) is selected from it randomly. If the selected subunit is a member of the rare population, then the psu is retained in the sample and additional k subunits are selected from this psu by simple random sampling without replacement (SRSWOR) method. If the selected subunit does not contain a member of the rare population, then the psu is not selected. The procedure is continued until a sample of n psu's is selected. The Sudman–Waksberg method selects a sample of n psu's by probability proportional to the measure of size with replacement (PPSSWR) method where the measure of size is the unknown number of the rare population of the psu. This follows from the fact that the probability of selection of ith psu in an effective draw (where a psu is selected)

$$\begin{aligned}
&= \text{Probability of the selection of } i\text{th psu in the first draw and it is retained} \\
&\quad + \text{Probability that the first draw does not select a psu and it is selected in} \\
&\quad \text{the second draw } + \cdots \\
&= X_i/(NM) + (1-\pi)X_i/(NM) + (1-\pi)^2 X_i/(NM) + \cdots
\end{aligned}$$

$$\left(\text{where } \pi = \frac{1}{N}\sum_{i=1}^{N}\frac{X_i}{M}\right)$$
$$= X_i \bigg/ \sum_{i=1}^{N} X_i.$$

26.2.4 Two-Phase Sampling

Identification of the rare population in some situations may be expensive, e.g., a medical diagnosis may be required for the prevalence of a neurological disorder. In this approach, a large sample is selected from the entire population. It is then classified into two or more strata according to the likelihood of rarity of the units by some relatively cheap but imperfect screening procedure. Subsamples from each of the strata are selected with numbers proportional to degree of likelihood of being a member of the rare population. Finally, an expensive method is used to detect whether the selected units are a member of the rare population or not. The two-phase method is useful if the cost of the first-phase method of screening is much less than that of the second phase. Deming (1977) recommended that the two-phase method should be used if the ratio of cost of screening for the second to first phase is at least 6:1.

26.3 DISPROPORTIONATE SAMPLING

Sometimes, rare populations are more heavily concentrated in certain pockets of the populations. In this situation, we treat various pockets as strata and samples are selected with higher sampling fractions for those strata with higher concentrations of the rare population. This procedure is cost-efficient because less screening is needed to identify the rare population in the strata with the higher concentration of the rare population.

Let $\overline{Y}_h$ and S_h^2 denote the population mean and population variance of the variable under study y for the stratum h, respectively, and let c_{sh} and c_{Rh} denote the cost of screening a member of the nonrare population and collecting data for a member of a rare population for the stratum h, respectively. Under the assumptions (i) $\overline{Y}_h = \overline{Y}$, (ii) $S_h^2 = S^2$, (iii) $c_{sh} = c_s$, and (iv) $c_{Rh} = c_R$ for all h, Kalton (2001) derived the optimum sampling fraction for the stratum h under simple random sampling with replacement (SRSWR) sampling as

$$f_h \propto \sqrt{\frac{P_h}{1+(r-1)P_h}} \tag{26.3.1}$$

where $P_h = M_h/N_h =$ prevalence of the rare population in the stratum h, $M_h =$ number of the rare population, and N_h is the size of the corresponding stratum h and $r = c_R/c_s$.

The value of r generally exceeds 1. For $r = 1$, the optimum sampling fraction reduces to

$$f_h \propto \sqrt{P_h}$$

The gain in precision with the use of disproportionate sampling with the optimum sampling fraction over the proportionate stratification with $r = 1$ is provided by Kalton (2001) as

$$R = \left(\sum \sqrt{A_h W_h}\right)^2 \tag{26.3.2}$$

where A_h and W_h are the proportion of rare and total population in the stratum h, respectively.

It is observed that the disproportionate stratification produces appreciable gain in efficiency only when the following two conditions are realized: (i) the strata to be oversampled should have high concentration of the rare population and (ii) the strata need to contain a substantial proportion of the rare population. For further details readers are referred to Kalton and Anderson (1986) and Kalton (2001).

26.4 MULTIPLICITY OR NETWORK SAMPLING

Multiplicity or network sampling was originally proposed by Birnbaum and Sirken (1965) and was later developed by Sirken (1970, 1972), Sirken and Levy (1974), and Nathan (1976), among others. Sampling of a rare population requires many contacts to identify sampling units possessing the rare trait. The objective of the multiplicity sampling is to spread the identification of the rare population more broadly over the total population and hence it reduces the number of contacts needed. In a conventional household survey, a number of households are selected in the initial sample, and information of the rare trait is collected from the members of the sampled households if they possess the required characteristic. In multiplicity sampling design, information regarding the membership of the rare trait is collected not only from each of the selected households but also from other households that are linked to that household. The linkage should be clearly defined. For example, the linkage may include any of the family relationship such as children, brothers, parents, or siblings; coworkers; members of the same church or social organization; and adjacent neighbors of the selected households. An individual living in an institution has no chance of selection in a conventional household survey, but he/she may have a chance of selection in the multiplicity sampling. The linkage provides necessary information of the existence of the rare trait. Sirken et al. (1978) and Czaja et al. (1884) used close relatives as linkage for survey of rare illnesses, Nathan (1976) for survey of births and deaths, whereas Brown and Ritchie (1981) for the survey of ethnic minorities.

Multiplicity sampling may be useful in estimating prevalence of the rare population or simply to identify members of the rare population.

The selection probabilities of the sampled members of the rare trait are determined by the defined linkage. For example, in an equal probability selection of households, all members of a household with a rare trait are given equal weight, but in a multiplicity sampling, the weight associated with an individual with a rare trait is inversely proportional to the number of households in which an individual member has a link including the household.

The main advantage of a multiplicity sample is that it needs a smaller sample to yield the required sample size of the rare population. The disadvantage is that the sampling design yields unequal probability of selection of members of the rare population and hence requires adjustment of weights. In addition, in the case of nonresponse caused by failure to trace the identified members of the rare population, the weight adjustments become tedious. Furthermore, the cost tracing and contacting all the individuals in a multiplicity sampling may be considerably high. The main demerit of a multiplicity sampling is that it may jeopardize the privacy of the individuals possessing the rare trait, e.g., an HIV infected person may not wish to be identified.

26.5 MULTIFRAME SAMPLING

A complete list of the rare population may not be available, but different partial (incomplete) lists of the rare population may be available from different sources. The lists are not generally disjoint, but their union may cover the entire rare population under study. For example, a complete list of antiretroviral (ARV) treatment recipients is difficult to obtain, but several lists such as hospital records, and records from government and nongovernment clinics and pharmacies may cover almost all ARV recipients. The lists may not be disjoint because a person may visit different clinics or hospitals.

One can prepare a complete frame by combining the incomplete frames and omitting duplicates. But in practice, removing duplication is not easy and the process may be highly error-prone because the spelling of the name or address of the same person may be different in the two lists. On the contrary, an incomplete list but containing a high proportion of rare individuals may provide a better result and be cheaper to sample. Selection of the ARV treatment recipients from the sampled clinics is easy and less expensive than surveying households because the hospital lists comprise of a high proportion of ARV treatment recipients than the complete household lists. Hartley (1962, 1974) pointed out that sampling from both the incomplete and complete frame may produce better estimate than sampling from the complete frame only.

The methodology of multiframe survey was introduced by Hartley (1962). The theory was extended by Lund (1968), Fuller and Burmeister (1972), Bankier (1986), Kalton and Anderson (1986), Singh and Wu (1996), and Skinner and Rao (1996), among others. In this section we will consider estimation of population characteristics from dual frame survey only. The theory can be extended for multiframe surveys.

26.5.1 Methods of Estimation

Let U_A and U_B be two incomplete frames that cover the population U under consideration, i.e., $U = U_A \cup U_B$. The frames U_A and U_B have overlap $U_{ab} = U_A \cap U_B = U_{ba}$. Let $U_a = U_A \cap U_B^c =$ consisting units of U_A only while $U_b = U_B \cap U_A^c$ consisting of units of U_B only. Let N, N_A, N_B, N_a, N_b, and N_{ab} ($=N_{ba}$) be sizes of U, U_A, U_B, U_a, U_b, and $U_{ab} = (U_{ba})$, respectively. Clearly, $N = N_a + N_b + N_{ab}$, and $N_A = N_a + N_{ab}$ and $N_B = N_b + N_{ab}$. Let y_i be the value of the variable under study for the ith unit of the population U. Let us define $Y = \sum_{i \in U} y_i$, $Y_A = \sum_{i \in U_A} y_i$, $Y_B = \sum_{i \in U_B} y_i$, $Y_a = \sum_{i \in U_a} y_i$, $Y_b = \sum_{i \in U_b} y_i$, and $Y_{ab} = \sum_{i \in U_{ab}} y_i = Y_{ba}$. Then we have

$$\begin{aligned} Y &= Y_a + Y_b + Y_{ab} \\ Y_A &= Y_a + Y_{ab} \\ Y_B &= Y_b + Y_{ab} \end{aligned}$$

Let samples s_A and s_B of sizes n_A and n_B be selected independently from the frames A and B, respectively, using suitable sampling designs p_A and p_B. Let samples s_a and s_{ab} ($s_A = s_a \cup s_{ab}$) of sizes n_a and $n_{ab}(n_A = n_a + n_{ab})$ fall in the population U_a and U_{ab}, respectively. Similarly, samples s_b and $s_{ba}(s_B = s_b \cup s_{ba})$ of sizes n_b and $n_{ba}(n_B = n_b + n_{ba})$ fall in the population U_b and U_{ab}, respectively. It should be noted that $s_{ab} \neq s_{ba}$ and $n_{ab} \neq n_{ba}$ in general. The estimation problem depends on the knowledge of the size of the population sizes N_A, N_B, and N_{ab}. There are three scenarios (i) all N_A, N_B, and N_{ab} are known, (ii) N_A, and N_B are known but N_{ab} unknown, and (iii) all N_A, N_B, and N_{ab} are unknown.

Let $\widehat{Y}_a$ and $\widehat{Y}_{ab}$ be suitable estimators of the totals Y_a and Y_{ab} obtained from the samples s_a and s_{ab}, respectively. Similarly, $\widehat{Y}_b$ and $\widehat{Y}_{ba}$ denote the estimators of Y_b and $Y_{ba}(=Y_{ab})$ obtained from the samples s_b and s_{ba}, respectively.

26.5.2 Simple Random Sampling Without Replacement

Let us assume that the samples s_A and s_B are selected by SRSWOR method independently. Here we consider the following cases:

Case 1: N_a, N_b, and N_{ab} are known

Hartley (1962) proposed the following result relating to estimation of the population total Y.

Theorem 26.5.1

(i) $\widehat{Y}_H = N_a\overline{y}_a + N_{ab}\left\{p\overline{y}_{ab} + (1-p)\overline{y}_{ba}\right\} + N_b\overline{y}_b$ is an unbiased estimator of Y (26.5.1)

where $\overline{y}_a = \sum_{i \in s_a} y_i/n_a, \overline{y}_{ab} = \sum_{i \in s_{ab}} y_i/n_{ab}, \overline{y}_b = \sum_{i \in s_b} y_i/n_b, \overline{y}_{ba} = \sum_{i \in s_{ba}} y_i/n_{ba},$ and $p(0 \leq p \leq 1)$ is a suitably chosen constant.

$$\text{(ii) } V\left(\widehat{Y}_H\right) \cong \frac{N_A^2}{n_A}\left\{(1-\alpha)S_a^2 + \alpha p^2 S_{ab}^2\right\} + \frac{N_B^2}{n_B}\left\{(1-\beta)S_b^2 + \beta(1-p)^2 S_{ab}^2\right\} \tag{26.5.2}$$

where

$$S_a^2 = \frac{1}{N_a - 1}\sum_{i \in U_a}\left(y_i - \overline{Y}_a\right)^2, S_b^2 = \frac{1}{N_b - 1}\sum_{i \in U_b}\left(y_i - \overline{Y}_b\right)^2,$$

$$S_{ab}^2 = \frac{1}{N_{ab} - 1}\sum_{i \in U_{ab}}\left(y_i - \overline{Y}_{ab}\right)^2, \overline{Y}_a = Y_a/N_a, \overline{Y}_b = Y_b/N_b,$$

$\overline{Y}_{ab} = Y_{ab}\big/N_{ab} = \overline{Y}_{ba}$, $\alpha = N_{ab}/N_A$ and $\beta = N_{ab}/N_B$.

Proof

(i) $E\left(\widehat{Y}_H\right) = E\left\{E\left(\widehat{Y}_H \middle| n_a, n_b, n_{ab}, n_{ba}\right)\right\} = E(Y) = Y$

(ii) $V\left(\widehat{Y}_H\right) = E\left\{V\left(\widehat{Y}_H \middle| n_a, n_b, n_{ab}, n_{ba}\right)\right\} + V\left\{E\left(\widehat{Y}_H \middle| n_a, n_b, n_{ab}, n_{ba}\right)\right\}$

Now noting $E\left(\widehat{Y}_H \middle| n_a, n_b, n_{ab}, n_{ba}\right) = Y$, we find

$$\begin{aligned} V\left(\widehat{Y}_H\right) = E\Bigg[& N_a^2\left(\frac{1}{n_a} - \frac{1}{N_a}\right)S_a^2 + p^2 N_{ab}^2\left(\frac{1}{n_{ab}} - \frac{1}{N_{ab}}\right)S_{ab}^2 \\ & + N_b^2\left(\frac{1}{n_b} - \frac{1}{N_b}\right)S_b^2 + (1-p)^2 N_{ab}^2\left(\frac{1}{n_{ba}} - \frac{1}{N_{ab}}\right)S_{ab}^2 \Bigg| n_a, n_b, n_{ab}, n_{ba}\Bigg] \end{aligned}$$

Now neglecting the sampling fractions $f_a = n_a/N_a, f_b = n_b/N_b, f_{ab} = n_{ab}/N_{ab}$, and $f_{ba} = n_{ba}/N_{ab}$, we get

$$\begin{aligned} V\left(\widehat{Y}_H\right) &= E\Big[N_a^2 S_a^2/n_a + p^2 N_{ab}^2 S_{ab}^2/n_{ab} \\ &\qquad + N_b^2 S_b^2/n_b + (1-p)^2 N_{ab}^2 S_{ab}^2/n_{ba} \Big| n_a, n_b, n_{ab}, n_{ba}\Big] \\ &\cong \frac{N_a^2}{E(n_a)}S_a^2 + p^2\frac{N_{ab}^2}{E(n_{ab})}S_{ab}^2 + \frac{N_b^2}{E(n_b)}S_b^2 + (1-p)^2\frac{N_{ab}^2}{E(n_{ba})}S_{ab}^2 \end{aligned} \tag{26.5.3}$$

Now, noting $E(n_a) = n_A N_a/N_A$, $E(n_b) = n_B N_b/N_B$, $E(n_{ab}) = n_A N_{ab}/N_A$, and $E(n_{ba}) = n_B N_{ab}/N_B$, we find

$$\begin{aligned} V\left(\widehat{Y}_H\right) &= \frac{N_A^2}{n_A}\left\{(1-\alpha)S_a^2 + \alpha p^2 S_{ab}^2\right\} + \frac{N_B^2}{n_B}\left\{(1-\beta)S_b^2 + \beta(1-p)^2 S_{ab}^2\right\} \\ &= \frac{N_A^2}{n_A}\left\{(1-\alpha)S_a^2 + \alpha p^2 S_{ab}^2\right\} + \frac{N_B^2}{n_B}\left\{(1-\beta)S_b^2 + \beta(1-p)^2 S_{ab}^2\right\} \end{aligned}$$

Lund (1968) obtained the optimum value of p by minimizing the expression (Eq. 26.5.3) as a function of p as

$$p_0 = \frac{n_{ab}}{n_{ab} + n_{ba}} \tag{26.5.4}$$

Substituting the optimum value of $p = p_0$ in expression (Eq. 26.5.1), Lund (1968) derived an improved estimator of Y as

$$\widehat{Y}_L = N_a\overline{y}_a + N_{ab}\overline{y}_{ab}^{(w)} + N_b\overline{y}_b \tag{26.5.5}$$

where

$$\overline{y}_{ab}^{(w)} = p_0\overline{y}_{ab} + (1-p_0)\overline{y}_{ba} \tag{26.5.6}$$

Ignoring finite population correction terms, Lund (1968) derived the approximate expression of the variance of $\widehat{Y}_L$ as

$$V\left(\widehat{Y}_L\right) = \frac{N_A^2}{n_A}(1-\alpha)S_a^2 + \frac{N_A N_B \alpha\beta}{\alpha n_A + \beta n_B}S_{ab}^2 + \frac{N_B^2}{n_B}(1-\beta)S_b^2 \qquad (26.5.7)$$

Although the variance of $\widehat{Y}_L$ is less than or equal to the variance of $\widehat{Y}_H$, substituting Eq. (26.5.4) in Eq. (26.5.2) provides the same approximate variance expressions.

Case 2: N_a, N_b, and N_{ab} are unknown

In this case the proportions $w_a = \dfrac{n_a}{n_A}$, $w_{ab} = \dfrac{n_{ab}}{n_A}$, $w_b = \dfrac{n_b}{n_B}$, and $w_{ba} = \dfrac{n_{ba}}{n_B}$ are unbiased estimators of $\dfrac{N_a}{N_A}, \dfrac{N_{ab}}{N_A}, \dfrac{N_b}{N_B}$, and $\dfrac{N_{ab}}{N_B}$, respectively. Here, N_{ab} possesses two unbiased estimators: $N_A w_{ab}$ and $N_B w_{ba}$. Now, replacing N_a, N_b, and N_{ab} in Eq. (26.5.1) by their estimates, Hartley obtained the following unbiased estimator for Y.

$$\widehat{Y}_H^* = N_A\left\{w_a\bar{y}_a + p w_{ab}\bar{y}_{ab}\right\} + N_B\left\{w_b\bar{y}_b + (1-p)w_{ba}\bar{y}_{ba}\right\} \qquad (26.5.8)$$

Theorem 26.5.2

(i) $E\left(\widehat{Y}_H^*\right) = Y$

(ii) $$V\left(\widehat{Y}_H^*\right) \cong N_a\left(\frac{1}{f_A}-1\right)S_a^2 + N_{ab}\left\{\left(\frac{1}{f_A}-1\right)p^2 + (1-p)^2\left(\frac{1}{f_B}-1\right)\right\}S_{ab}^2 + N_b\left(\frac{1}{f_B}-1\right)S_b^2 + \frac{g_A N_a N_{ab}}{n_A}\left(\overline{Y}_a - p\overline{Y}_{ab}\right)^2 + \frac{g_B N_b N_{ab}}{n_B}\left\{\overline{Y}_b - (1-p)\overline{Y}_{ab}\right\}^2$$

where $f_A = \dfrac{n_A}{N_A}$, $f_B = \dfrac{n_B}{N_B}$, $g_A = \dfrac{N_A - n_A}{N_A - 1}$ and $g_B = \dfrac{N_B - n_B}{N_B - 1}$

Proof

(i) $$E\left(\widehat{Y}_H^*\right) = E\left[E\left(\widehat{Y}_H^* \middle| n_a, n_{ab}, n_b, n_{ba}\right)\right] = E\left[N_A\left\{w_a\overline{Y}_a + p w_{ab}\overline{Y}_{ab}\right\} + N_B\left\{w_b\overline{Y}_b + (1-p)w_{ba}\overline{Y}_{ba}\right\}\middle| n_a, n_{ab}, n_b, n_{ba}\right] = \overline{Y}$$

$$\text{(ii)}\;\; V\left(\widehat{Y}_H^*\right) = V\left\{E\left(\widehat{Y}_H^*\middle| n_a, n_{ab}, n_b, n_{ba}\right)\right\} + E\left\{V\left(\widehat{Y}_H^*\middle| n_a, n_{ab}, n_b, n_{ba}\right)\right\}$$

$$\begin{aligned} &= V\left[N_A\left(w_a\overline{Y}_a + pw_{ab}\overline{Y}_{ab}\right)\right] + V\left[N_B\left\{w_b\overline{Y}_b + (1-p)w_{ba}\overline{Y}_{ba}\right\}\right] \\ &\quad + E\left[V\left(\widehat{Y}_H^*\middle| n_a, n_{ab}, n_b, n_{ba}\right)\right] \end{aligned} \tag{26.5.9}$$

The first component of Eq. (26.5.9) is

$$\begin{aligned} &V\left[N_A\left(w_a\overline{Y}_a + pw_{ab}\overline{Y}_{ab}\right)\right] \\ &= N_A^2\left\{\overline{Y}_a^2 V(w_a) + p^2\overline{Y}_{ab}^2 V(w_{ab}) + 2p\overline{Y}_a\overline{Y}_{ab}Cov(w_a, w_{ab})\right\} \\ &= N_A^2\left(\frac{1}{n_A} - \frac{1}{N_A}\right)\frac{N_A}{N_A - 1}\left[W_a(1-W_a)\overline{Y}_a^2 + p^2\overline{Y}_{ab}^2 W_{ab}(1-W_{ab})\right. \\ &\qquad\qquad \left. -2p\overline{Y}_a\overline{Y}_{ab}W_aW_{ab}\right] \end{aligned}$$

(where $W_a = N_a/N_A$ and $W_{ab} = N_{ab}/N_A$)

$$= N_A^2\left(\frac{1}{n_A} - \frac{1}{N_A}\right)\frac{N_A}{N_A - 1}W_aW_{ab}\left(\overline{Y}_a - p\overline{Y}_{ab}\right)^2$$

(noting $W_a = 1 - W_{ab}$)

$$= \frac{g_A N_a N_{ab}}{n_A}\left(\overline{Y}_a - p\overline{Y}_{ab}\right)^2 \tag{26.5.10}$$

Similarly, the second component of Eq. (26.5.9) is

$$\begin{aligned} &V\left[N_B\left\{w_b\overline{Y}_b + (1-p)w_{ba}\overline{Y}_{ba}\right\}\right] \\ &= \frac{g_B N_b N_{ab}}{n_B}\left\{\overline{Y}_b - (1-p)\overline{Y}_{ab}\right\}^2. \end{aligned} \tag{26.5.11}$$

The third component of Eq. (26.5.9) is

$$\begin{aligned} &E\left\{V\left(\widehat{Y}_H^*\middle| n_a, n_{ab}, n_b, n_{ba}\right)\right\} \\ &= N_A^2 E\left\{w_a^2\left(\frac{1}{n_a} - \frac{1}{N_a}\right)S_a^2 + p^2 w_{ab}^2\left(\frac{1}{n_{ab}} - \frac{1}{N_{ab}}\right)S_{ab}^2\right\} \\ &\quad + N_B^2 E\left\{w_b^2\left(\frac{1}{n_b} - \frac{1}{N_b}\right)S_b^2 + (1-p)^2 w_{ba}^2\left(\frac{1}{n_{ba}} - \frac{1}{N_{ab}}\right)S_{ab}^2\right\} \end{aligned} \tag{26.5.12}$$

Now, noting
$E(w_a) = \frac{N_a}{N_A} = W_a,\ V(w_a) = \left(\frac{1}{n_A} - \frac{1}{N_A}\right)\ \frac{N_A}{N_A - 1} W_a(1 - W_a),$
we find

$$\begin{aligned}
& N_A^2 E\left\{w_a^2\left(\frac{1}{n_a} - \frac{1}{N_a}\right)\right\} S_a^2 \\
&= N_A^2 \left\{\frac{W_a}{n_A} - \frac{W_a^2 + V(w_a)}{N_a}\right\} S_a^2 \\
&= N_A^2 \left[\frac{W_a}{n_A} - \frac{W_a^2}{N_a} - \frac{1}{N_a}\left(\frac{1}{n_A} - \frac{1}{N_A}\right)\ \frac{N_A W_a(1 - W_a)}{N_A - 1}\right] S_a^2 \\
&= N_A^2 W_a \left[\frac{1}{n_A} - \frac{W_a}{N_a} - \frac{1}{N_a}\left(\frac{1}{n_A} - \frac{1}{N_A}\right)\ \frac{N_A(1 - W_a)}{N_A - 1}\right] S_a^2 \\
&= N_a \left[\frac{N_A}{n_A} - 1 - \frac{1}{N_a}\left(\frac{N_A}{n_A} - 1\right)\ \frac{N_A(1 - W_a)}{N_A - 1}\right] S_a^2 \\
&= N_a \left(\frac{1}{f_A} - 1\right)\left\{1 - \frac{1}{N_a}\,\frac{N_A(1 - W_a)}{N_A - 1}\right\} S_a^2 \\
&\cong N_a \left(\frac{1}{f_A} - 1\right)\left\{1 - \frac{(1 - W_a)}{N_a}\right\} S_a^2 \\
&\cong N_a \left(\frac{1}{f_A} - 1\right) S_a^2 \left\{\text{neglecting the term } \left(\frac{1}{N_a} - \frac{1}{N_A}\right)\right\}
\end{aligned} \tag{26.5.13}$$

Similarly,

$$\begin{aligned}
& N_A^2 E\left\{w_{ab}^2\left(\frac{1}{n_{ab}} - \frac{1}{N_{ab}}\right) S_{ab}^2\right\} \cong N_{ab}\left(\frac{1}{f_A} - 1\right) S_{ab}^2, \\
& N_B^2 E\left\{w_b^2\left(\frac{1}{n_b} - \frac{1}{N_b}\right)\right\} S_{ab}^2 \cong N_b\left(\frac{1}{f_B} - 1\right) S_b^2 \ \text{ and} \\
& N_B^2 E\left\{w_{ba}^2\left(\frac{1}{n_{ba}} - \frac{1}{N_{ab}}\right) S_{ab}^2\right\} \cong N_{ab}\left(\frac{1}{f_B} - 1\right) S_{ab}^2
\end{aligned} \tag{26.5.14}$$

Expressions (Eqs. 26.5.12–26.5.14) yield

$$\begin{aligned}
E\left\{V\left(\widehat{Y}_H^* \middle| n_a, n_{ab}, n_b, n_{ba}\right)\right\} &\cong N_a\left(\frac{1}{f_A} - 1\right) S_a^2 \\
&+ N_{ab}\left\{\left(\frac{1}{f_A} - 1\right) p^2 + \left(\frac{1}{f_B} - 1\right)(1 - p)^2\right\} S_{ab}^2 + N_b\left(\frac{1}{f_B} - 1\right) S_b^2
\end{aligned} \tag{26.5.15}$$

The proof of the theorem follows from Eqs. (26.5.9)–(26.5.11) and (26.5.15).

Corollary 26.5.1

If sampling fraction f_A and f_B are ignored we get

$$V\left(\widehat{Y}_H^*\right) \cong \frac{N_A^2}{n_A}\left\{(1-\alpha)S_a^2 + \alpha p^2 S_{ab}^2\right\} + \frac{N_B^2}{n_B}\left\{(1-\beta)S_b^2 + \beta(1-p)^2 S_{ab}^2\right\}$$
$$+ \frac{N_A^2\alpha(1-\alpha)}{n_A}\left(\overline{Y}_a - p\overline{Y}_{ab}\right)^2 + \frac{N_B^2\beta(1-\beta)}{n_B}\left\{\overline{Y}_b - (1-p)\overline{Y}_{ab}\right\}^2$$

Lund (1968) modified Hartley estimator $\widehat{Y}_H^*$ using the combined mean of the overlapping frame (including repetition of units) $\overline{y}_w = \dfrac{n_{ab}\overline{y}_{ab} + n_{ba}\overline{y}_{ba}}{n_{ab} + n_{ba}}$ as an estimator of $\overline{Y}_{ab}$. Lund estimator is as follows:

$$\widehat{Y}_L^* = N_A w_a \overline{y}_a + \{N_A p w_{ab} + N_B(1-p)w_{ba}\}\overline{y}_w + N_B w_b \overline{y}_b$$

Using Theorem 26.5.2, we can derive the following:

Theorem 26.5.3

(i) $E\left(\widehat{Y}_L^*\right) = Y$

(ii) $V\left(\widehat{Y}_L^*\right) = N_a\left(\dfrac{1}{f_A} - 1\right)S_a^2 + \dfrac{N_{ab}}{(f_A + f_B)^2}\{f_A(1-f_A) + f_B(1-f_B)\}S_{ab}^2$
$$+ N_B\left(\frac{1}{f_B} - 1\right)S_b^2 + \frac{g_A}{n_A}N_{ab}N_a\left(\overline{Y}_a - p\overline{Y}_{ab}\right)^2$$
$$+ \frac{g_B}{n_B}N_{ab}N_b\left\{\overline{Y}_b - (1-p)\overline{Y}_{ab}\right\}^2$$

Lund (1968) derived the optimum value of p as

$$p_{0L} = \frac{(1-\alpha)\overline{y}_a/f_A + (1-\beta)\left(\overline{y}_w - \overline{y}_b\right)/f_B}{[(1-\alpha)/f_A + (1-\beta)/f_B]\overline{y}_w} \tag{26.5.16}$$

Lund proposed the following estimator for p_{0L}

$$\widehat{p}_{0L} = \frac{n_a n_B \overline{y}_a/f_A + n_A n_b\left(\overline{y}_w - \overline{y}_b\right)/f_B}{[n_B n_a/f_A + n_A n_b/f_B]\overline{y}_w} \tag{26.5.17}$$

The expression of the variance $V\left(\widehat{Y}_L^*\right)$ given in Theorem 26.5.3 was provided by Fuller and Burmeister (1972). The estimator $\widehat{Y}_L^*$ possesses a much lower variance than that of Hartley estimator $\widehat{Y}_H^*$.

Fuller and Burmeister (1972) proposed an alternative estimator for Y as

$$\widehat{Y}_{FB} = \left(N_A - \widehat{N}_{ab}\right)\overline{y}_a + \widehat{N}_{ab}\overline{y}_{w^*} + \left(N_B - \widehat{N}_{ab}\right)\overline{y}_b \tag{26.5.18}$$

where $\overline{y}_{w^*} = w^*\overline{y}_{ab} + \left(1 - w^*\right)\overline{y}_{ba}$, $w^* = \dfrac{n_{ab}(1 - f_B)}{n_{ab}(1 - f_B) + n_{ba}(1 - f_A)}$, and $\widehat{N}_{ab}$ is the smallest root of the quadratic,

$$\begin{aligned}(n_A g_B + n_B g_A)\widehat{N}_{ab}^2 - (n_A N_B g_B + n_B N_A g_A + n_{ab} N_A g_B + n_{ba} N_B g_A)\widehat{N}_{ab} \\ + (n_{ab} g_B + n_{ba} g_A) N_A N_B = 0\end{aligned} \tag{26.5.19}$$

We state the following theorem (without derivation) that is related to the mean and variance of $\widehat{Y}_{FB}$ derived by Fuller and Burmeister (1972).

Theorem 26.5.4

(i) $E\left(\widehat{Y}_{FB}\right) = Y + o\left(\dfrac{1}{n}\right)$

(ii) $$\begin{aligned}V\left(\widehat{Y}_{FB}\right) = {} & N_a\left(\frac{1}{f_A} - 1\right)S_a^2 + \frac{(1 - f_A)(1 - f_B)N_{ab}}{(1 - f_B)f_A + (1 - f_A)f_B}S_{ab}^2 + N_b\left(\frac{1}{f_B} - 1\right)S_b^2 \\ & + \frac{N_{ab}N_a N_b g_A g_B}{n_A N_b g_B + n_B N_a g_A}\left(\overline{Y}_{ab} - \overline{Y}_a - \overline{Y}_b\right)^2 + o(1)\end{aligned}$$

Fuller and Burmeister (1972) pointed out that the bias in the Lund estimator $\widehat{Y}_L^*$ is $o(1)$ while that of $\widehat{Y}_{FB}$ is $o\left(\dfrac{1}{n}\right)$. Furthermore, $\widehat{Y}_{FB}$ is advantageous over $\widehat{Y}_L^*$ in the computational point of view. In cases where the units in samples s_{ab} and s_{ba} are identifiable, the following estimator $\widehat{Y}_{FB}^*$ has smaller variance than $\widehat{Y}_{FB}$.

$$\widehat{Y}_{FB}^* = \left(N_A - \widehat{N}_{ab,m}\right)\overline{y}_a + \widehat{N}_{ab,m}\overline{y}_d + \left(N_B - \widehat{N}_{ab,m}\right)\overline{y}_b \tag{26.5.20}$$

where $\overline{y}_d = \dfrac{1}{n_d}\displaystyle\sum_{i \in s_d} y_i$, s_d is the set of distinct units in $s_{ab} \cup s_{ba}$, n_d is the number of distinct units in s_d, and $\widehat{N}_{ab,m}$ is the smallest root of the quadratic equation

$$\begin{aligned}(n_a + n_d + n_b)\widehat{N}_{ab,m}^2 - \left\{n_a N_B + n_d(N_A + N_B) + n_b N_A - n_a n_b\right\}\widehat{N}_{ab,m} \\ + n_d N_A N_B = 0\end{aligned} \tag{26.5.21}$$

Fuller and Burmeister (1972) recommended that $\widehat{N}_{ab,m}$ should be considered as a maximum likelihood estimate (MLE) of N_{ab} with score (Eq. 26.5.21), whereas Skinner (1991) provided reasons for considering $\widehat{Y}^*_{FB}$ as an MLE of Y.

26.5.3 General Sampling Procedures

Let π_{Ai} and π_{Bi} be the inclusion probabilities of selection of ith unit of the population U_A based on sampling design p_A and U_B based on sampling design p_B, respectively. In case N_A is known, Y_a and Y_{ab} can be estimated by

$$\widehat{Y}_a = \sum_{i\in s_a} w_{Ai}y_i \text{ and } \widehat{Y}_{ab} = \sum_{i\in s_{ab}} w_{Ai}y_i \tag{26.5.22}$$

where $w_{Ai} = N_A\left(\frac{1}{\pi_{Ai}}\Big/\sum_{i\in s_A}\frac{1}{\pi_{Ai}}\right)$.

Similarly, if N_B is known, estimators for Y_b and Y_{ab} are, respectively,

$$\widehat{Y}_b = \sum_{i\in s_b} w_{Bi}y_i \text{ and } \widehat{Y}_{ba} = \sum_{i\in s_{ba}} w_{Bi}y_i \tag{26.5.23}$$

where $w_{Bi} = N_B\left(\frac{1}{\pi_{Bi}}\Big/\sum_{i\in s_B}\frac{1}{\pi_{Bi}}\right)$.

The Hartley (1974) estimator for Y is

$$\widetilde{Y}_H = \widehat{Y}_a + \widehat{Y}_b + \varphi\widehat{Y}_{ab} + (1-\varphi)\widehat{Y}_{ba} \tag{26.5.24}$$

where the constant φ is chosen to minimize $V\left(\widetilde{Y}_H\right)$. Clearly, the optimum φ will involve unknown parameters, which should be estimated from the selected sample.

Fuller and Burmeister (1972) proposed the following estimator for Y

$$\widetilde{Y}_{FB} = \widehat{Y}_a + \widehat{Y}_b + \varphi\widehat{Y}_{ab} + (1-\varphi)\widehat{Y}_{ba} + \phi\left(\widehat{N}_{ab} - \widehat{N}_{ba}\right) \tag{26.5.25}$$

where $\widehat{N}_{ab} = \sum_{i\in s_{ab}} w_{Ai}$ and $\widehat{N}_{ba} = \sum_{i\in s_{ba}} w_{Bi}$.

Here also φ and ϕ are chosen to minimize the variance of $\widetilde{Y}_{FB}$. The optimum values of φ and ϕ will involve unknown parameters, which need to be estimated from the data.

The pseudo-MLE for Y proposed by Skinner and Rao (1996) is given by

$$\widehat{Y}_{PML} = \left(N_A - \widehat{N}_{ab,PML}\right)\widehat{\mu}_a + \left(N_B - \widehat{N}_{ab,PML}\right)\widehat{\mu}_b + \widehat{N}_{ab,PML}\widehat{\mu}_w \tag{26.5.26}$$

where

$$\widehat{\mu}_a = \widehat{Y}_a \Big/ \widehat{N}_a,\ \widehat{N}_a = \sum_{i\in s_a} w_{Ai},\ \widehat{\mu}_b = \widehat{Y}_b \Big/ \widehat{N}_b,\ \widehat{N}_b = \sum_{i\in s_b} w_{Bi},$$

$$\widehat{\mu}_w = \left(\frac{n_A}{N_A}\widehat{N}_{ab}\widehat{\mu}_{ab} + \frac{n_B}{N_B}\widehat{N}_{ba}\widehat{\mu}_{ba}\right) \Big/ \left(\frac{n_A}{N_A}\widehat{N}_{ab} + \frac{n_B}{N_B}\widehat{N}_{ba}\right),$$

$\widehat{\mu}_{ab} = \widehat{Y}_{ab} \Big/ \widehat{N}_{ab}$, and $\widehat{\mu}_{ba} = \widehat{Y}_{ba} \Big/ \widehat{N}_{ba}$.

Furthermore, $\widehat{N}_{ab,PML}$ is the smallest root of the quadratic equation

$$(n_A + n_B)x^2 - \left(n_A N_B + n_B N_A + n_A \widehat{N}_{ab} + n_B \widehat{N}_{ba}\right)x + n_A \widehat{N}_{ab} N_B + n_B \widehat{N}_{ba} N_A = 0 \tag{26.5.27}$$

Skinner and Rao (1996) showed that the estimators $\widetilde{Y}_H$ and $\widetilde{Y}_{FB}$ are all consistent for Y and are asymptotically normally distributed with mean Y. They also provided expressions of the asymptotic variances of the estimators.

In case N_A, N_B, and N_{ab} are known, we may write $\widehat{N}_{ab} = \widehat{N}_{ba} = \widehat{N}_{ab,PML} = N_{ab}$ in the expression of $\widehat{Y}_{PML}$ given in Eq. (26.5.26) and obtain the following modified estimator

$$\widehat{Y}_{PML} = N_a\widehat{\mu}_a + N_b\widehat{\mu}_b + N_{ab}\widehat{\mu}_w$$

where $\widehat{\mu}_w = w\widehat{\mu}_{ab} + (1-w)\widehat{\mu}_{ba}$ and $w = \dfrac{(n_A/N_A)}{(n_A/N_A) + (n_B/N_B)}$.

In case none of N_A, N_B, or N_{ab} is known, the estimators $\widetilde{Y}_H$ and $\widetilde{Y}_{FB}$ can be modified by simply writing $w_{Ai} = 1/\pi_{Ai}$, $w_{Bi} = 1/\pi_{Bi}$ in expressions (Eqs. 26.5.24 and 26.5.25), respectively. The estimator $\widehat{Y}_{PML}$ can also be modified by replacing N_A and N_B with their estimates $\sum_{i\in s_A} 1/\pi_{Ai}$ and $\sum_{i\in s_B} 1/\pi_{Bi}$, respectively.

26.5.4 Horvitz–Thompson-Based Estimators

If the units in the overlapping population U_{ab} can be identified, the sample $s = s_A \cup s_B$ can be regarded as a sample from the population U with inclusion probability for the ith unit

$$\pi_i = \begin{cases} \pi_{Ai} & \text{for } i\in s_a \\ \pi_{Bi} & \text{for } i\in s_b \\ \pi_{ABi} & \text{for } i\in s_{AB} \end{cases} \tag{26.5.28}$$

where $\pi_{ABi} = \pi_{Ai} + \pi_{Bi} - \pi_{Ai}\pi_{Bi}$ and $s_{AB} = s_A \cap s_B$.

In this case the Horvitz–Thompson estimator for the population total Y is given by

$$\widehat{Y}_{ht} = \sum_{i \in s_a} \frac{y_i}{\pi_{Ai}} + \sum_{i \in s_b} \frac{y_i}{\pi_{Bi}} + \sum_{i \in s_{AB}} \frac{y_i}{\pi_{ABi}} \qquad (26.5.29)$$

It is very difficult to compare the performance of $\widehat{Y}_{ht}$ with the alternatives $\widetilde{Y}_H$, $\widetilde{Y}_{FB}$, and $\widehat{Y}_{PML}$. However, $\widehat{Y}_{ht}$ is expected to perform better as it is based on distinct units only.

26.5.5 Concluding Remarks

The estimation problem from the dual frame survey depends on the knowledge of the size of the population sizes N_A, N_B, and N_{ab}. Estimation of the population total and the mean were considered by Hartley (1962), Lund (1968), and Fuller and Burmeister (1972) for SRSWOR sampling while Hartley (1974), Fuller and Burmeister (1972), Skinner (1991), Skinner and Rao (1996), and Singh and Wu (1996) studied for the general sampling designs. Fuller and Burmeister considered "pseudo"-MLEs under SRSWR and complex survey designs. The "pseudo"-MLE uses the same survey weights for all the variables, unlike Hartley (1974) and Fuller and Burmeister (1972). Hence "pseudo"-MLEs possess a computational advantage. The asymptotic properties of the proposed estimators are studied by Skinner and Rao (1996). They conducted limited simulation studies relating to the performances of these estimators and recommended the use of "pseudo"-MLEs, as they perform better than the others and bring significant gain in efficiency over the single frame estimators. Multiframe surveys for multistage sampling design were studied by Saxena et al. (1984). Singh and Wu (1996) used multiauxiliary variables for multiframe complex surveys. More comprehensive researches are required to obtain conclusive results relating to the performances of the proposed estimators used for multiframe surveys.

26.6 SNOWBALL SAMPLING

Snowball sampling is used for surveying an extremely rare population where the cost of the survey using one of the methods discussed earlier remains prohibitively high. In snowball sampling, members of the rare population are assumed to be known to each other. This condition is very restrictive, but may hold true for some rare populations such as ethnic minorities and religious groups. Initially, a few members of the rare population are identified and each of them is asked to identify the other members. Then each of these members are contacted and asked to identify other members and so on. After a certain stage, no new member is found, i.e., the list of the rare population is completed. Finally, from the completed list of the rare population a sample is selected using a suitable sampling design. For example, snowball sampling can be used to select doctors in a locality by identifying a few doctors and asking them if they know other doctors and so on. After a certain stage no new identification of doctors could be found.

Some applications of snowball sampling avoid construction of the sampling frame for the rare population. Instead, the snowball sampling process is continued till a sizable number of the rare population is identified and information is collected from these identified individuals. Because this sampling is not a probability sampling, unbiased estimation of population characteristics is not possible.

Snowball sampling was used by Snow et al. (1981) to select a sample of Hispanics in Atlanta, by Welch (1975) for selection of samples of Mexican Americans in Omaha, and Biernacki and Waldrof (1981) for sampling ex—heroin addicts.

26.7 LOCATION SAMPLING

Location sampling is widely used to sample populations that have no fixed abode for both census and surveys: nomads may be sampled at water points when they take their animals for water, and homeless persons may be sampled at soup kitchens when they go for food.

Location sampling is used to sample rare mobile populations at the time of their visits to specific locations such as airports, game parks, churches, shopping malls, and playgrounds. Location sampling fails to cover those who do not visit such places during the survey period. Here, the unit of analysis may be visits or visitors (Kalton, 1991).

Location sampling can readily produce a probability sample of visits, with known probabilities, and hence visits are easily analyzed. Visits may be the appropriate unit of analysis for, say, a survey about satisfaction with visits to a museum. In this case no issues of multiplicity arise.

However, for many surveys, the visitor is the appropriate unit of analysis, for example, in a survey of visitors to soup kitchens over a week to estimate the number of homeless, a survey of nomads visiting watering holes to estimate the size of the nomadic population. In this situation issues of multiplicity arise because a visitor can visit more than once during the survey reference period. To avoid this problem one may treat the first visits during the time period as an eligible visit. Otherwise, multiplicity adjustments will be required. For example, each sampled person is asked whether this visit is the first since the start of the survey or not. If the answer is "yes" the person is selected and if the answer is "no" the person is not selected. In this procedure most visits near the start of the survey will be accepted, as they are the first visits, whereas at the end of the survey most of the visits will not be accepted because these are not the first visits.

Kalton (1991, 2009) recommended a two-stage sampling procedure where psu's are combination of locations (entrances or exits) and time segments when the locations are open (e.g., on week days from 10:00 a.m. to 5:00 p.m. and on holidays from 8:00 a.m. to 4:00 p.m.). The psu's may be selected with probability proportional to size, with careful stratification by location and time. Then systematic sampling may be used to select visitors entering or exiting the location.

Location sampling has been used to sample men, with locations being gay bars, bathhouses, and bookstores (Kalton, 1993; MacKellar et al., 1996), who have sex with men. The Young Men's Survey conducted in seven cities in 1994—98 in 194 public locations is a major survey of this type (Vallerory et al., 2000). Mckenzie and Mistiaen (2009) carried out studies to compare location sampling with area sampling and snowball sampling for sampling on Nikkei (Brazilians of Japanese descents) in Sao Paulo and Panama. The locations were places where Nikkei often visit such as the metro station, ethnic grocery stores, sports clubs, and other locations where family members of Nikkei community congregate. The studies reveal that the location and snowball sampling are unlikely to provide a representative sample.

26.8 SEQUENTIAL SAMPLING

Because the size of the rare population is unknown, it is difficult to determine the sample of the desired size needed for estimation of the rare population characteristics (for example, prevalence) with a specified degree of accuracy. The problem is acute, especially when there are no past survey data of the rare population available. In this case, sample size may be determined sequentially. At the first stage, a preliminary estimate is made on the basis of expert judgment or small pilot survey, and on the basis of this estimate the desired sample size is determined. At the second stage a survey is conducted with half of the required sample size and a revised estimate of the population characteristic is obtained. Using the estimates in the second step, the optimal sample size is determined and the additional required sample is selected at the third step. The sequential sampling procedure generally increases time and cost of the study but is worth the expense to obtain an efficient estimate of the rare population.

26.9 ADAPTIVE SAMPLING

In adaptive sampling, the selection of the sample depends on the values of the characteristic under study y. Adaptive sampling was motivated for sampling rare and mobile populations such as rare contagious diseases, drug use, rare species of animals or birds, and density of animals in a forest (Thompson, 1990). In sampling animals or rare species of trees in a forest, the entire forest may be divided into number of square plots of equal size. A sample of n_1 plots can be selected by some probability sampling such as SRSWOR method. If a sampled unit (plot) satisfies a certain condition C, for example, the number of animals y_i of the plot i exceeds a certain prespecified number c, then the neighborhood units are added in the sample. Here, neighborhoods are defined in a certain manner such as adjacent four plots viz. north, east, south, and west. If the other units in the neighborhood satisfy the condition C, then their neighborhoods are also included in the sample. This process is continued until a cluster of units is obtained that contains a boundary of units (called edge units) that do not satisfy the condition C. Thus from the selection of initial n_1 plots, we select n_1 clusters of plots. The selected clusters may not be distinct. It should be noted that if no animal is selected in the initial selected plot, the neighboring plots are not sampled. In this case, it becomes a cluster of one unit plot. The details of the selection procedures are given by Thompson and Seber (1996).

The neighborhood relationship can be defined in various ways and neighborhoods need not be contiguous. For example, the neighborhood relation may be brothers, sisters, siblings, or the same ethnic group. This type of sampling is known as adaptive cluster sampling. Here, the total number of sampled units (plots) is a random variable. Thus adaptive sampling includes sequential sampling where sampling is continued until some condition is satisfied.

26.9.1 Unbiased Estimation of Population Mean

Let a region U be divided into a finite number of N square plots and let a sample s of n_1 plots be selected by SRSWOR method. We define a network A_i for the unit i as a cluster generated by the unit i, but its edge units are removed. Thus selection of any unit in A_i will lead the selection of all the units of A_i. If a unit i is the only one unit in a cluster satisfying C, then A_i is itself a network of size 1. We also define a unit that does not satisfy C as a network of size 1. Note that any edge unit is also a network of size 1. Thus all the clusters of size 1 are also a network of size 1. Hence according to the definition any two different networks are disjoint and networks form the partition of the population U. Let $B_1,\ldots,B_K$ be the distinct networks that form the partition of the population U. Following Thompson and Seber (1996) we propose the following unbiased estimators of the population mean.

26.9.1.1 Use of Intersection Probabilities

Here we define the population total as

$$Y = \sum_{i=1}^{N} y_i = \sum_{j=1}^{K} y_j^* \tag{26.9.1}$$

where $y_j^* = \sum\limits_{k\in B_j} y_k$.

Let $I_j(s) = 1$ if $s \cap B_j \neq \phi$, i.e., some of the selected units in s also belong to B_j, $I_j(s) = 0$ if $s \cap B_j = \phi$, and x_j is the number of units in the jth network.

Then, we have

$$\begin{aligned}\alpha_j &= \Pr ob\{I_j(s) = 1\}\\ &= \text{Probability that the network } B_j \text{ will have non} - \text{null intersectionwith } s\\ &= \text{Probability that at least one of the } x_j \text{ units of the network}\\ &\qquad B_j \text{ is included in the sample } s\\ &= 1 - \text{ Probability that none of the } x_j \text{ units of the network } B_j \text{ is selected in } s\\ &= 1 - \binom{N-x_j}{n_1}\Big/\binom{N}{n_1}\end{aligned} \tag{26.9.2}$$

and

$$\begin{aligned}\alpha_{jk} &= \Pr ob\{I_j(s) = 1, I_k(s) = 1\} \text{ for } j \neq k\\ &= \text{Probability that both the network } B_j \text{ and } B_k \text{ will have non-}\\ &\qquad \text{null intersections with } s\\ &= 1 - \left[\binom{N-x_j}{n_1} + \binom{N-x_k}{n_1} - \binom{N-x_j-x_k}{n_1}\right]\Big/\binom{N}{n_1}\end{aligned} \tag{26.9.3}$$

Theorem 26.9.1

(i) $\widehat{\overline{Y}}_1 = \dfrac{1}{N}\sum_{j=1}^{K}\dfrac{y_j^*}{\alpha_j}I_j(s)$ is an unbiased for $\overline{Y} = \sum_{j=1}^{K} y_j^* \Big/ N$

(ii) $$V\left(\widehat{\overline{Y}}_1\right) = \frac{1}{N^2}\left[\sum_{j=1}^{K} y_j^{*2}\left(\frac{1}{\alpha_j} - 1\right) + \sum_{j\neq}^{K}\sum_{k=1}^{K}\frac{y_j^*}{\alpha_j}\,\frac{y_k^*}{\alpha_k}\left(\frac{\alpha_{jk}}{\alpha_j\alpha_k} - 1\right)\right]$$

(iii) $$\widehat{V}\left(\widehat{\overline{Y}}_1\right) = \frac{1}{N^2}\left[\sum_{j=1}^{K} y_j^{*2}\left(\frac{1}{\alpha_j} - 1\right)\,\frac{I_j(s)}{\alpha_j}\right.$$
$$\left. + \sum_{j\neq}^{K}\sum_{k=1}^{K}\frac{y_j^*}{\alpha_j}\,\frac{y_k^*}{\alpha_k}\left(\frac{\alpha_{jk}}{\alpha_j\alpha_k} - 1\right)\,\frac{I_j(s)I_k(s)}{\alpha_{jk}}\right]$$

Proof

(i) $$E\left(\widehat{\overline{Y}}_1\right) = \frac{1}{N}\sum_{j=1}^{K}\frac{y_j^*}{\alpha_j}E\{I_j(s)\} = \overline{Y}$$

(ii) $$V\left(\widehat{\overline{Y}}_1\right) = \frac{1}{N^2}E\left\{\sum_{j=1}^{K}\frac{y_j^*}{\alpha_j}I_j(s)\right\}^2 - \overline{Y}^2$$

$$= \frac{1}{N^2}\left[\sum_{j=1}^{K}\frac{y_j^{*2}}{\alpha_j^2}E\{I_j(s)\} + \sum_{j\neq}^{K}\sum_{k=1}^{K}\frac{y_j^*}{\alpha_j}\,\frac{y_k^*}{\alpha_k}E\{I_j(s)I_k(s)\}\right] - \overline{Y}^2$$

$$= \frac{1}{N^2}\left[\sum_{j=1}^{K} y_j^{*2}\left(\frac{1}{\alpha_j} - 1\right) + \sum_{j\neq}^{K}\sum_{k=1}^{K}\frac{y_j^*}{\alpha_j}\,\frac{y_k^*}{\alpha_k}\left(\frac{\alpha_{jk}}{\alpha_j\alpha_k} - 1\right)\right]$$

(iii) Noting $E\{I_j(s)\} = \alpha_j$ and $E\{I_j(s)I_k(s)\} = \alpha_{jk}$, we can prove $E\left[\widehat{V}\left(\widehat{\overline{Y}}_1\right)\right] = V\left(\widehat{\overline{Y}}_1\right)$.

26.9.1.2 Use of the Number of Intersections

Let A_i be the network containing the ith unit, $i = 1,\ldots, N$. The networks A_i's need not be distinct. The networks A_1 and A_2 are identical if both of them contain the units 1 and 2. Let $f_i(s)$ be the number of units in the initial sample s that belongs to the network A_i (ignoring the edge units of the clusters). Clearly, $f_i(s) = 0$ if no unit in the initial sample intersects A_i. Hence $f_i(s)$ follows hypergeometric distribution and

$$v_i = E[\,f_i(s)] = n_1\frac{m_i}{N} \tag{26.9.4}$$

where m_i is the number of units in the ith network.

Hence

$$\widehat{\overline{Y}}_2 = \frac{1}{N}\sum_{i=1}^{N} y_i \frac{f_i(s)}{\nu_i}$$

is an unbiased estimator of the population mean $\overline{Y}$. We can write $\widehat{\overline{Y}}_2$ as follows:

$$\begin{aligned}\widehat{\overline{Y}}_2 &= \frac{1}{n_1}\sum_{i=1}^{N} y_i \frac{f_i(s)}{m_i}\\ &= \frac{1}{n_1}\sum_{i\in s} w_i \\ &= \overline{w}_s\end{aligned} \tag{26.9.5}$$

where $w_i = \frac{1}{m_i}\sum_{j\in A_i} y_j$.

From the expression of $\widehat{\overline{Y}}_2$ given in Eq. (26.9.5), we get the following theorem.

Theorem 26.9.6

(i) $E(\overline{w}_s) = \overline{Y}$

(ii) $V(\overline{w}_s) = \frac{N-n_1}{n_1 N} S_w^2$

(iii) $\widehat{V}(\overline{w}_s) = \frac{N-n_1}{n_1 N} s_w^2$

where $S_w^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(w_i - \overline{Y}\right)^2$ and $s_w^2 = \frac{1}{n_1 - 1}\sum_{i\in s}(w_i - w_s)^2$.

Proof

(i) Noting $\overline{w}_s$ is the sample mean of n_1 units selected by SRSWOR method, we have

$$E(\overline{w}_s) = \frac{1}{N}\sum_{i=1}^{N} w_i = \frac{1}{N}\sum_{i=1}^{K}\sum_{i\in B_k} y_i = \overline{Y}$$

Proofs of (ii) and (iii) follow straight from the properties of SRSWOR sampling.

Details of adaptive sampling for unequal sampling schemes are given by Thompson and Seber (1996). But we have omitted them here because of their highly technical nature.

26.10 CAPTURE—RECAPTURE METHOD

The capture—recapture method was developed at least in the 100 years ago in the field of ecology and wildlife studies for estimating the unknown population size (N). The application of this method in epidemiological problems (e.g., incidence of disease) and demography (e.g., population size, survival, recruitment, emigration, and migration) came relatively late to the literature. Other names of the method include capture—mark—recapture, mark—recapture, mark—release—recapture, and band recovery. The basic method involves capturing a sample of animals and then marking and releasing them into the population. A second sample is then selected from the population and the number of marked animals discovered. From the data thus collected, the population size is estimated.

A population that remains unchanged during the period of investigation will be called a closed population. In this case there is neither addition in the population due to birth or immigration nor removal due to migration or death. Here, the only parameter of interest is the population size N. A population is called an open population if it may be subject to change because of processes such as immigration, migration, births, deaths, etc. In the open population interest is given to the population dynamics such as birth and death rates, survival rates, and population change. Capture—recapture method does not correspond to the sampling of a finite population paradigm because the frame does not exist. The capture—recapture method was considered by Peterson (1896) and Lincoln (1930) and great details are provided by Seber (1982).

26.10.1 Closed Population

26.10.1.1 Peterson and Lincoln Method

Let N be the total number of the unknown population (birds or animals). From the population a sample s_1 of n_1 animals are caught, marked (or tagged), and released into the population. After sometime when the animals have resettled with the unmarked animals, another sample s_2 of size n_2 is selected from the population. Let $s_{2m}(\subset s_2)$ of size m_2 be the set of marked animals and $s_{2u}(s_2 - s_{2m})$ be the sample of unmarked animals of size $u_2 = n_2 - m_2$. Now, we make the following assumptions:

(a) Each animal is equally likely to be caught in the two samples, i.e., healthy, unhealthy, and animals once caught have the same chance of being captured again.

(b) Animals do not lose their marks

Under these assumptions, we find that the proportion of marked individuals that are caught (m_2/n_1) should be equal to the proportion of total animals caught (n_2/N), i.e.,

$$\frac{m_2}{n_1} = \frac{n_2}{N}$$
$$\text{i.e.,} \quad \widehat{N} = \widehat{N}_P = \frac{n_1 n_2}{m_2} \tag{26.10.1}$$

Eq. (26.10.1) is called the Peterson (1896) estimate or Lincoln (1930) Index.

26.10.1.2 Hypergeometric Model

Under the assumptions (a) and (b), the conditional probability distribution of m_2 given n_1, n_2, and N is the hypergeometric distribution

$$f(m_2|n_1, n_2, N) = \frac{\binom{n_1}{m_2}\binom{N-n_1}{n_2-m_2}}{\binom{N}{n_2}}; m_2 = 0, 1, \ldots, \min(n_1, n_2) \tag{26.10.2}$$

Here, $E(m_2) = n_2\frac{n_1}{N}$, hence we can choose an estimator of N to be $\widehat{N} = \widehat{N}_P = \frac{n_1 n_2}{m_2}$. Clearly, $\widehat{N}_P$ is a biased estimator. Chapman (1951) showed that $\widehat{N}_P$ is asymptotically normal as $N \to \infty$, but the bias of $\widehat{N}_P$ cannot be negligible for a small sample. For $n_1 + n_2 \geq N$, Chapman proposed the following exactly unbiased estimator for N.

$$\widehat{N}_c = \frac{(n_1+1)(n_2+1)}{(m_2+1)} - 1 \tag{26.10.3}$$

However, for $n_1 + n_2 < N$, the estimator $\widehat{N}_c$ is biased. The expression of bias was obtained by Robson and Regier (1964).

Assuming Poisson approximation to Eq. (26.10.2), Chapman (1951) computed the variance of $\widehat{N}_c$ as

$$V\left(\widehat{N}_c\middle|n_1, n_2\right) = N^2\left(\frac{1}{\mu} + \frac{2}{\mu^2} + \frac{6}{\mu^3}\right) \tag{26.10.4}$$

where $\mu = E(m_2|n_1, n_2) = n_1 n_2/N$.

Seber (1970) proposed an approximate unbiased estimator of $V\left(\widehat{N}_c\right)$ as

$$\widehat{V}\left(\widehat{N}_c\right) = \frac{(n_1+1)(n_2+1)(n_1-m_2)(n_2-m_2)}{(m_2+1)^2(m_2+2)} \tag{26.10.5}$$

26.10.1.3 Bailey's Binomial Model

Using Binomial approximation to Eq. (26.10.2), Bailey (1951) obtained the probability distribution of m_2 as

$$f(m_2|n_1,n_2)=\binom{n_2}{m_2}\left(\frac{n_1}{N}\right)^{m_2}\left(1-\frac{n_1}{N}\right)^{n_2-m_2};\ m_2=0,1,\ldots,n_2 \quad (26.10.6)$$

Because $E(m_2)=n_2\dfrac{n_1}{N}$, the MLE of N comes out as $\widehat{N}_P$. Because $\widehat{N}_P$ is biased, Bailey proposed the following alternative estimator of N as

$$\widehat{N}_b=\frac{n_1(n_2+1)}{(m_2+1)} \quad (26.10.7)$$

The proposed estimator of the variance of $\widehat{N}_b$ is

$$\widehat{V}\left(\widehat{N}_b\right)=\frac{n_1^2(n_2+1)(n_2-m_2)}{(m_2+1)^2(m_2+2)} \quad (26.10.8)$$

26.10.1.4 Ratio Method

Let us attach a variable y_i for the ith member of the population such that $y_i=1$. Then $Y=\sum_{i=1}^{N}y_i=N=$ population size. Similarly, we can attach a variable x_i to the ith individual such that $x_i=1$ if the ith individual is marked and 0 if it is unmarked. In this case $X=\sum_{i=1}^{N}x_i=n_1,\overline{y}(s_{2m})=1$ and $\overline{x}(s_2)=m_2/n_2$. Hence the ratio estimator of N is given by

$$\widehat{N}_R=\frac{\overline{y}}{\overline{x}}X=\frac{n_2}{m_2}n_1=\widehat{N}_P \quad (26.10.9)$$

The ratio estimator $\widehat{N}_R$ is biased as usual and an approximate estimator of the variance of $\widehat{N}_R$ was obtained by Lohr (1999) as

$$\widehat{V}\left(\widehat{N}_R\right)\cong\frac{n_1^2n_2(n_2-m_2)}{m_2^3} \quad (26.10.10)$$

26.10.1.5 Inverse Sampling Methods

26.10.1.5.1 Without Replacement Method

Here, the sample s_1 of size n_1 animals is selected as before. The selection of sample s_2 is continued until m_2 of the tagged n_1 animals are recaptured again. Here, n_2, the sample size of s_2, is a random variable. The condition probability distribution of n_2 given n_1 and m_2 is the negative hypergeometric distribution

$$f(n_2|n_1,m_2,N)=\frac{\begin{pmatrix} n_1 \\ m_2-1 \end{pmatrix}\begin{pmatrix} N-n_1 \\ n_2-m_2 \end{pmatrix}}{\begin{pmatrix} N \\ n_2-1 \end{pmatrix}}\frac{n_1-m_2+1}{N-n_2+1}; \tag{26.10.11}$$

$$n_2=m_2,m_2+1,\ldots,N+m_2-n_1$$

Here the MLE of N is

$$\widehat{N}=\widehat{N}_I=\frac{n_2(n_1+1)}{m_2}-1 \tag{26.10.12}$$

The estimator $\widehat{N}_I$ is unbiased for N with variance

$$\begin{aligned} V\left(\widehat{N}_I\middle|n_1,m_2\right) &= \frac{(N+1)(N-n_1)(n_1-m_2+1)}{m_2(n_1+2)} \\ &\cong \frac{N^2}{m_2} \end{aligned} \tag{26.10.13}$$

The coefficient of variation of $\widehat{N}_I$ is approximately equal to

$$C\left(\widehat{N}_I\right)=\left(\frac{n_1-m_2+1}{m_2(n_1+2)}\right)^{1/2} \tag{26.10.14}$$

Because $C\left(\widehat{N}_I\right)$ is independent of N, one can choose m_2 for prescribed values of $C\left(\widehat{N}_I\right)$ and n_1.

26.10.1.5.2 With Replacement Method

Here, animals in the second sample s_2 are caught one by one and released into the original population. The catch and release procedure is continued until m_2 (a prespecified) marked animals are caught. In this case the probability distribution of n_2 given n_1 and m_2 is the negative binomial distribution

$$f(n_2|n_1,m_2,N)=\begin{pmatrix} n_2-1 \\ m_2-1 \end{pmatrix}\left(\frac{n_1}{N}\right)^{m_2}\left(1-\frac{n_1}{N}\right)^{n_2-m_2}; n_2=m_2,m_2+1,\ldots \tag{26.10.15}$$

Because $E(n_2) = N\dfrac{m_2}{n_1}$, an unbiased estimator of N is

$$\widehat{N}_{IW} = \widehat{N}_P = n_1 n_2/m_2 \tag{26.10.16}$$

The variance of $\widehat{N}_{IW}$ and its unbiased estimator are given, respectively, as follows:

$$V\left(\widehat{N}_{IW}\right) = \left(N^2 - Nn_1\right)/m_2 \text{ and } \widehat{V}\left(\widehat{N}_{IW}\right) = \frac{n_2 n_1^2 (n_2 - m_2)}{m_2^2 (m_2 + 1)} \tag{26.10.17}$$

A serious disadvantage of the use of inverse sampling is that the variance of n_2 becomes large for inappropriate choices of n_1 and m_2. The coefficient of variation of $\widehat{N}_{IW}$ under this sampling scheme is

$$C\left(\widehat{N}_{IW}\right) \cong \frac{1}{\sqrt{m_2}} \tag{26.10.18}$$

So, one can easily find the value of m_2 for a given value of $C\left(\widehat{N}_{IW}\right)$.

26.10.1.6 Interval Estimation

In determining the confidence interval of N based on capture—recapture method, one should be certain about the normality of the distribution of $\widehat{N}$. For a small sample the distribution of $\widehat{N}$ is skewed, which results in poor coverage of the probability of the confidence interval. Several methods of determining the confidence interval of N have been proposed by Seber (1970). Some of them are reported below as follows.

As $N \to \infty$, $\widehat{N}_c$ is asymptotically normally distributed with mean N and estimated variance $\widehat{V}\left(\widehat{N}_c\right)$. Hence $100(1-\alpha)\%$ confidence interval of N is given by

$$\widehat{N}_c \pm z_{\alpha/2}\sqrt{\widehat{V}\left(\widehat{N}_c\right)} \tag{26.10.19}$$

where $z_{\alpha/2}$ is the upper $100\alpha/2$ percent point of standard normal distribution.

For $\widehat{p} = m_2/n_2 < 0.1$ and $m_2 > 50$, normal approximation of $\widehat{p}$ is valid and the $100(1-\alpha)\%$ confidence interval of $p = n_1/N$ is

$$\widehat{p} \pm \left\{ z_{\alpha/2}\sqrt{(1-f)\widehat{p}(1-\widehat{p})/(n_2-1)} + 1/(2n_2) \right\} \tag{26.10.20}$$

The confidence interval of N is obtained from the inversion of Eq. (26.10.20). Here, the unknown sampling fraction $f(=n_2/N)$ can be ignored if its estimate m_2/n_1 is assumed to be less than 0.1.

In case $N > 150$, $n_1 > 50$ and $n_2 > 50$, the distribution of m_2 is asymptotically normal (Seber, 1973). In this case the two largest roots of the equation

$$\frac{\left(m_2 - \dfrac{n_1 n_2}{N}\right)^2}{n_2\dfrac{n_1}{N}\left(1 - \dfrac{n_1}{N}\right)\dfrac{N - n_2}{N - 1}} = \left(z_{\alpha/2}\right)^2 \qquad (26.10.21)$$

provide the confidence interval of N.

26.10.1.7 Multiple Marking

Consider the Schnabel census (1938) where a series of samples $s_1, s_2, \ldots, s_t$ of sizes $n_1, n_2, \ldots, n_t$ were selected independently from the entire population. All the captured animals were tagged or marked and returned to the population. Tags are unique, so that the capture history of each of the individual animal can be followed separately. Let

$$\begin{aligned} m_i &= \text{number of marked animals in the sample } s_i, \\ u_i &= n_i - m_i = \text{number of unmarked animals in the sample } s_i, \\ M_i &= u_1 + u_2 + \cdots + u_{i-1};\, i = 1, \ldots, t+1 \\ &= \text{number of marked animals in the population just before the } i\text{th} \\ &\quad\ \text{sample } s_i \text{ is taken.} \end{aligned}$$

Clearly, $m_1 = M_1 = 0$, $M_2 = u_1 = n_1$, and $M_{t+1}(=r \text{ say})$ is the number of marked animals in the population after performing the experiment.

Let $w = x_1 x_2 \ldots x_t$, with $x_j = 1$ if an animal is caught at the jth sample s_j and $x_j = 0$ if it is not caught at the jth sample. Following Seber (1986) we denote a_w as the number of animals with a capture history, e.g., $t = 2$, a_{10} = number of animals caught in the sample s_1 but not in s_2; a_{01} = number of animals caught in the sample s_2 but not in s_1; a_{11} = number of animals caught in the sample s_1 and also in s_2, and a_{00} = number of animals that were not captured in any of the samples s_1 and $s_2 = N - r$.

Here, we assume that all animals irrespective of capture history will have the same probability of being caught in a particular sample. The caught in the samples $s_1, s_2, \ldots, s_t$ are independent. Let $p_i(=1 - q_i)$ = probability that an animal will be captured in the ith sample s_i. Then for $t = 2$, the probability of the vector $\{a_w\} = \{a_{00}, a_{01}, a_{10}, a_{11}\}$ follows the multinomial distribution

$$\begin{aligned} f\{a_w\} &= \frac{N!}{a_{10}!a_{01}!a_{11}!(N-r)!}(p_1q_2)^{a_{10}}(q_1p_2)^{a_{01}}(p_1p_2)^{a_{11}}(q_1q_2)^{N-r} \\ &\propto \frac{N!}{(N-r)!}p_1^{n_1}p_2^{n_2}q_1^{N-n_1}q_2^{N-n_2} \end{aligned} \tag{26.10.22}$$

For general t, the probability distribution of $\{a_w\}$ is given by

$$f\{a_w\} \propto \frac{N!}{(N-r)!}\prod_{i=1}^{t}\left(p_i^{n_i}q_i^{N-n_i}\right) \tag{26.10.23}$$

Eq. (26.10.23) indicates that n_i's are independent binomial variable with distribution

$$f(n_i) = \binom{N}{n_i}p_i^{n_i}q_i^{N-n_i} \tag{26.10.24}$$

The conditional distribution of $\{a_w\}$ for fixed sample sizes $\{n_i\}$ is given by

$$f(\{a_w\}|\{n_i\}) \propto \frac{N!}{(N-r)!}\prod_{t=1}^{t}\binom{N}{n_i}^{-1} \tag{26.10.25}$$

The MLE of N is the unique root greater than r of the following $(t-1)$ degree polynomial (Seber, 1970)

$$1-\frac{r}{N} = \prod_{i=1}^{t}\left(1-\frac{n_i}{N}\right) \tag{26.10.26}$$

For $t=2$, MLE of N is $\widehat{N}_P = n_1n_2/m_2$, the Peterson estimate.

Chapman (1952) derived Eq. (26.10.26) by considering the conditional distribution of $\{m_2,\ldots, m_t\}$ given $\{n_1, n_2,\ldots, n_t\}$ as a product of hypergeometric distributions

$$f(m_2,\ldots,m_t|n_1,\ldots,n_t) = \prod_{i=2}^{t}\frac{\binom{M_i}{m_i}\binom{N-M_i}{u_i}}{\binom{N}{n_i}} \propto \frac{N!}{(N-r)!}\prod_{i=2}^{t}\binom{N}{n_i}^{-1} \tag{26.10.27}$$

For further details readers are referred to Seber (1986).

26.10.2 Open Model

In the open model, provision is made for death, birth, immigration, and migration. The models are broadly classified into two categories: The first one is based on the bird-banding and fish-tagging studies where a number of animals are caught and banded for several (k) periods. Each band or tag carries a unique identification number. Data were recorded of bands or tags collected from dead animals for each of the k periods. The recovery data are analyzed for estimation of the annual survival rate and the annual band recovery rate for each of the k periods. A compressive review is given by Brownie et al. (1985). The second one, known as the Jolly—Seber model, deals with multiple recaptures of alive marked animals. The links between the two models were established by Brownie et al. (1985). The details are given by Seber (1973). Because there is not much difference in the collection and analysis of Brand recovery data and live animal capture data, our present discussions are limited to the Jolly—Seber model only.

26.10.2.1 Jolly—Seber Model

Jolly (1965) and Seber (1965) independently provided the most important stochastic model for capture—recapture sampling in the open population setup. The model allows estimates of survival, capture probability, and population size for each sampling time, and recruitment between sampling times.

Experimental protocol: Here, we capture and recapture animals over $k(>1)$ successive occasions. In each occasion, animals are captured, tagged uniquely, and then released. The capture or sighting history of each individual is recorded. On the first occasion a sample s_1 of n_1 animals is captured, all the animals are tagged, a few d_1 of them died in the capturing process, and the remaining $R_1(=n_1 - d_1)$ animals are released. On the second occasion, a sample s_2 of n_2 animals is captured of which u_2 are untagged and the remaining $n_2 - u_2$ are tagged (recaptured). All the recaptured and newly captured animals are uniquely marked and capture history is recorded. If d_2 of them die during the capture process, the remaining $R_2(=n_2 - d_2)$ are released and treated as cohort of the sample s_2. In general, at the ith occasion, a sample s_i of size n_i is captured of which u_i are untagged, $n_i - m_i$ are tagged, and d_i dies. All the $R_i(=n_i - d_i)$ animals are tagged uniquely, released, and treated as cohort of ith sample s_i. Let m_{ij} be the number of animals recaptured the first time from the released R_i animals in the jth sample, $i = 1,\ldots, k - 1; j = i + 1,\ldots, k$.

26.10.2.1.1 Summary Data

The summary of the data is presented as follows:

					s_1	s_2	s_3	$\ldots$	s_{i+1}	$\ldots$	s_k		
Samples	n_i	u_i	d_i	R_i		m_{ij}						Total r_i	z_{i+1}
s_1	n_1	u_1	d_1	R_1	0	m_{12}	m_{13}	$\ldots$	m_{1i+1}	$\ldots$	m_{1k}	r_1	$z_2 = r_1 - m_{12}$
s_2	n_2	u_2	d_2	R_2			m_{23}	$\ldots$	m_{2i+1}	$\ldots$	m_{2k}	r_2	$z_3 = r_2 - m_{23}$
$\ldots$								$\ldots$	$\ldots$	$\ldots$	$\ldots$		
s_i	n_i	u_i	d_i	R_i					m_{ii+1}	$\ldots$	m_{ik}	r_i	$z_{i+1} = r_i - m_{i,i+1}$
$\ldots$													
s_{k-1}	n_{k-1}	u_{k-1}	d_{k-1}	R_{k-1}							m_{k-1k}	r_{k-1}	$z_k = r_{k-1} - m_{k-1,k}$
Total					$m_1 = 0$	m_2	m_3	$\ldots$	m_{i+1}	$\ldots$	m_k		

Descriptions of the statistics:

$r_i =$ number of marked animal released in the sample s_i, which were ever captured again = ith row total.

$m_i =$ number of animals captured in the sample $s_i =$ ith column total.

$z_i =$ number of animals released before the ith sample, not caught in the sample s_i but are captured in the subsequent samples $= r_{i-1} - m_{i-1,i}$.

26.10.2.1.2 Assumptions of the Model

(i) Capture probability p_i: Each animal (marked or unmarked) in the population has the same probability p_i of capture at the time of ith sample ($i = 1,\ldots, k$).

(ii) Survival probability ϕ_i: Each marked animal has the same survival probability ϕ_i for the period of taking *i*th sample to $i+1$th sample ($i = 1,\ldots, k-1$).

(iii) Tags are not lost or damaged.

(iv) All samples are instantaneous: each release is made immediately after the selection of sample.

26.10.2.1.3 Estimation of Parameters

The conditional distribution of $m_{ii+1}, m_{ii+2},\ldots, m_{ii+k}, R_i - \sum_{j=i+1}^{k} m_{ij} = R_i - r_i$ given R_i follows multinomial distribution with cell probabilities:

$$\begin{aligned}\pi_{ii+1} &= \{\phi_i p_{i+1}\}, \pi_{ii+2} = \{\phi_i(1-p_{i+1})\}\{\phi_{i+1}p_{i+2}\}, \ldots,\\ \pi_{ij} &= \left[\{\phi_i(1-p_{i+1})\}\{\phi_{i+1}(1-p_{i+2})\}\cdots\{\phi_{j-1}p_j\}\right], \ldots,\\ \pi_{ik} &= [\{\phi_i(1-p_{i+1})\}\{\phi_{i+1}(1-p_{i+2})\}\cdots\{\phi_{k-1}p_k\}].\end{aligned}$$

Hence the likelihood function of the recaptured conditional to $R_1,\ldots, R_{k-1}$ is

$$L(\mathbf{m}) \propto \prod_{i=1}^{k-1} \lambda_i^{R_i - r_i} \prod_{j=i+1}^{k} \{\pi_{ij}\}^{m_{ij}} \qquad (26.10.28)$$

Let U_i be the number of unmarked animals just before the ith sample is taken, then the likelihood function of $u_1,\ldots, u_k$ is

$$L(\mathbf{u}) \propto \prod_{i=1}^{k-1} \binom{U_i}{u_i} p_i^{u_i}(1-p_i)^{U_i - u_i} \qquad (26.10.29)$$

Let M_i be the number of marked animal just before the ith sample is taken. $M_1 = 0$ but M_i's for $i = 2,\ldots, k$ are unknown. The M_i's are estimated by Seber (1973) as follows.

Among the marked M_i animals, m_i are caught in the sample s_i and the remaining $M_i - m_i$ are not caught. Out of $M_i - m_i$ marked animals, z_i are subsequently caught. R_i animals are realized in the sample s_i of which r_i are caught. Hence,

$$\frac{z_i}{M_i - m_i} = \frac{r_i}{R_i} \tag{26.10.30}$$

Eq. (26.10.30) yields an estimate of M_i as

$$\widehat{M}_i = m_i + \frac{R_i z_i}{r_i} \tag{26.10.31}$$

Population size N_i:

The number of the marked animal in the sample s_i of size n_i is m_i and M_i is the number of the marked sample in the population of size N_i when the sample s_i was taken. Hence

$$\frac{m_i}{n_i} \cong \frac{M_i}{N_i} \tag{26.10.32}$$

Eq. (26.10.32) leads

$$\widehat{N}_i = \widehat{M}_i \frac{n_i}{m_i} \tag{26.10.33}$$

Capture probability p_i:

Proportion of marked animals captured in ith sample s_i to the total number of animals at that time

$$\text{i.e.,} \quad \widehat{p}_i = \frac{n_i}{\widehat{N}_i} = \frac{m_i}{\widehat{M}_i} \tag{26.10.34}$$

Survival rate ϕ_i:

The number of marked animals in the population immediately after sample s_i is $M_i - m_i + R_i$. The number of the marked animals in the population just before sample s_{i+1} is selected is M_{i+1}. Hence

$$\widehat{\phi}_i = \frac{\widehat{M}_{i+1}}{\widehat{M}_i - m_i + \widehat{R}_i} \tag{26.10.35}$$

Recruitment B_i:

Estimated number of survival between the sampling period i to $i+1$ is $(N_i - n_i + R_i)\phi_i$. Hence recruitment between the sampling period i to $i+1$ is

$$\widehat{B}_i = \widehat{N}_{i+1} - \left(\widehat{N}_i - n_i + R_i\right)\widehat{\phi}_i \tag{26.10.36}$$

Out of the five estimators only $\widehat{\phi}_i$ and $\widehat{p}_i$ are MLEs and the remaining $\widehat{M}_i, N_i$, and $\widehat{B}_i$ are intuitive estimators. These estimators are not unbiased. Seber (1982) proposed approximate unbiased estimators of these five estimators and their variances.

Example 26.10.1

Consider the following artificial data where animals are captured and tagged for 7 consecutive months.

Month	Sample size	Released	(m_{ij})						
(i)	(n_i)	(R_i)	1	2	3	4	5	6	7
1	80	80		15	10	6	2	1	0
2	125	120			30	10	3	0	1
3	150	146				35	18	5	4
4	180	178					35	20	10
5	140	135						30	15
6	100	98							20

Estimates of parameters:

i	n_i	m_i	r_i	R_i	z_i	$\widehat{M}_i$	$\widehat{N}_i$	$\widehat{\phi}_i$	$\widehat{p}_i$	$\widehat{B}_i$
1	80	0	34	80						
2	125	15	44	120	19	66.9	556.8	0.835	0.225	39.3
3	150	40	62	146	14	73.0	273.6	0.425	0.548	252.7
4	180	51	65	178	27	125.0	441.0	0.698	0.484	99.4
5	140	58	45	135	30	148.0	357.2	0.875	0.392	28.5
6	100	56	20	98	15	129.5	231.5	0.576	0.432	
7		50			0					

The Jolly–Seber model assumes that the marked and unmarked animals have equal capture and survival probabilities, which may not hold in practice. This model estimates time-specific apparent survival rates and is restricted to specific localities only. The estimators may be highly biased in the presence of unequal capture and survival probabilities. Several extensions of the Jolly—Seber model are available in the literature. Pollock (1975) allowed behavioral effects in the captured and survivors; Brownie et al. (1986) considered survival and captured parameters unchanged over time; while Cormack (1981) used log-linear models on open population. A comprehensive review of the recent development in the capture—recapture model for the open population was given by Seber (1986), Boswell et al. (1988), Pollock et al. (1990), and Barker (1995).

26.11 EXERCISES

26.11.1 A sample of 800 fish were captured from a lake, then marked uniquely, and realized alive. One month later, a sample of 1000 fish was captured from the same lake and 80 marked fish captured earlier were found. Estimate the total number of fish of the lake and find 95% confidence interval of the number of fishes using (i) Peterson—Lincoln, (ii) Chapman, (iii) Bailey, and (iv) ratio methods.

26.11.2 A sample of 1000 people is selected from a locality and 20 of them were found to be illegal immigrants. One month later another sample of 825 people was selected by the inverse sampling method without replacement from the same locality till 10 of them were found to be illegal immigrants selected in the earlier sample. Estimate the number of illegal immigrant in that locality and compute its standard error.

26.11.3 Samples of animals are captured from a certain game park for 8 consecutive years. All the animals that were captured are uniquely marked and released unless they die in the capturing process. Estimate the total number of animals, capture probability rates, survival rates, and new recruitments for each of the sampling years. The following table gives the capture record:

Year	***i***		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
i	n_i	R_i					m_{ij}			
1	100	98		20	10	8	6	4	0	2
2	125	125			30	20	7	5	4	3
3	160	158				30	25	10	2	0
4	180	175					40	15	10	5
5	150	150						25	10	2
6	130	130							30	15
7	100	100								20

26.11.4 Consider two incomplete frames U_A and U_B, which cover the population $U(=U_A \cup U_B)$. Let us define $U_{ab} = U_A \cap U_B$, $U_a = U_A \cap U_B^c$, and $U_b = U_B \cap U_A^c$. Two independent samples s_A and s_B of sizes n_A and n_B are selected from U_A and U_B by the PPSWR method using normed size measures p_{i1} and p_{j2} of the ith unit of U_A and the jth unit of U_B, respectively. Let $t_a = \frac{1}{n_a}\sum_{s_a}\frac{y_i}{p_{1i}}$, $t_b = \frac{1}{n_b}\sum_{s_b}\frac{y_i}{p_{2i}}$, $t_{ab} = \frac{1}{n_{ab}}\sum_{s_{ab}}\frac{y_i}{p_{1i}}$, and $t_{ba} = \frac{1}{n_{ba}}\sum_{s_{ba}}\frac{y_i}{p_{2i}}$, where $s_a = s_A \cap U_a$, $s_{ab} = s_A \cap U_{ab}$, $s_b = s_B \cap U_b$, $s_{ba} = s_B \cap U_{ab}$; $\sum_{s_t}$ and n_t denote respectively the sum over the units and its size in s_t with repletion, $t = a, b, ab, ba$. Show that $T = w_a t_a + t_w + w_b t_b$ is an unbiased estimator of the population total of $Y = \sum_{i \in U} y_i$, where $t_w = \alpha w_{ab} t_{ab} + (1-\alpha) w_{ba} t_{ba}$, $w_a = \frac{n_a}{n_A}$, $w_b = \frac{n_b}{n_B}$, $w_{ab} = \frac{n_{ab}}{n_A}$, $w_{ba} = \frac{n_{ba}}{n_B}$ and α is a known constant. Find the variance of T and an unbiased estimator of the variance.

26.11.5 Consider a population consisting of N units, which are classified into K overlapping clusters. The ith cluster consists of N_i primary units on unknown size and $\sum_{i=1}^{K} N_i = M \geq N$. A primary unit may be found in more than one cluster. Let y_{ij} be the value of y for the jth unit of the ith cluster and f_{ij} its frequency of occurring in K clusters. Let a sample s of size k clusters be selected by the PPSWR method with selection probability $p_i = N_i/M$ for the ith cluster. If the ith cluster is selected in s, a subsample s_i of n_i primary units are selected from the ith cluster by the SRSWOR method. Define $Z_{ij} = y_{ij}/f_{ij}$, $w_{ij} = 1/f_{ij}$, $\overline{z}_i = \frac{1}{n_i}\sum_{i \in s_i} z_{ij}$, $\overline{Z}_i = \frac{1}{N_i}\sum_{i=1}^{N_i} z_{ij}$, $\overline{w}_i = \frac{1}{n_i}\sum_{j \in s_i} w_{ij}$, and $\overline{W}_i = \frac{1}{N_i}\sum_{i=1}^{N_i} w_{ij}$. Show that the ratio estimator $\widehat{Y}_R = \sum_{i=1}^{k}\overline{z}_i \Big/ \sum_{i=1}^{k}\overline{w}_i$ is a biased estimator of the population mean $\overline{Y}$ and the mean square error of $\widehat{Y}_R$ is

$$MSE\left(\widehat{Y}_R\right) \cong \frac{M}{kN^2}\sum_{i=1}^{K} N_i\left[\left(\overline{Z}_i - \overline{YW}_i\right)^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)D_i^2\right],$$

where $D_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left\{\left(z_{ij} - \overline{Z}_i\right) - \overline{Y}\left(w_{ij} - W_i\right)\right\}^2$ (Tracy and Osahan, 1994).