CHAPTER 8

Sampling in Two Stages

8.1 Introduction

For the single-stage cluster sampling discussed in the last chapter, observations are obtained from all the elements of each selected cluster. When the clusters contain a large number of elements, estimates for the population means, totals, and proportions are usually obtained from subsamples of elements in the clusters. Mahalanobis (1946) describes this procedure as the **two-stage sampling**. For this approach, clusters are the first-stage units, which are also known as **primary sampling units** (PSUs). The elements are the subunits or **secondary sampling units** (SSUs).

For an agricultural survey, districts or enumeration areas can be the PSUs, and the villages can be considered to be the SSUs. For estimation of employment rates and related characteristics in a geographical region, the counties and city blocks in the region are usually considered to be the PSUs and SSUs, respectively. For health-related surveys, counties or a group of counties can be considered as the PSUs and the households or hospitals as SSUs. For large-scale surveys, stratification may precede selection of the sample at any stage.

This chapter presents some of the frequently used procedures for selecting the first- and second-stage units with equal or unequal probabilities. For some surveys, more than two stages of selection is considered.

8.2 Equal size first-stage units

The notation for this case is similar to that in Sections 7.2 through 7.6 except that the cluster totals are denoted by Y_i. Let y_{ij}, $i = 1,\ldots,N$ and $j = 1,\ldots,M$, denote the observations of the M second-stage units on each

of the N first-stage units. The totals of the ith first-stage unit and the population are $Y_i = \Sigma^M y_{ij}$ and $Y = \Sigma^N Y_i$. The mean of the ith first-stage unit is $\bar{Y}_i = Y_i/M$. The population mean per primary unit and per subunit or element are $\bar{\bar{Y}} = Y/N$ and $\bar{\bar{Y}} = Y/M_o = (\Sigma^N Y_i)/N$, where $M_o = NM$. The variances among the means of the first-stage units and within the elements of the second-stage units, respectively, are

$$S_1^2 = \sum_1^N (\bar{Y}_i - \bar{\bar{Y}})^2/(N-1) \qquad \text{and} \qquad S_2^2 = \sum_1^N S_{2i}^2/N, \tag{8.1}$$

where $S_{2i}^2 = \Sigma_1^M (\bar{y}_{ij} - \bar{Y}_i)^2/(M-1)$. Note that S_{2i}^2 is the variance within the ith primary unit with $(M-1)$ d.f. and S_2^2 is the pooled within variance with $N(M-1)$ d.f.

At the first stage, a sample of n primary units is selected, and at the second stage a sample of m secondary units is selected from each of the selected primary units. The mean of the m subsampled units of the ith primary unit and the overall mean of the nm subsampled units, respectively, are

$$\bar{y}_i = \sum_1^m y_{ij}/m \qquad \text{and} \qquad \bar{\bar{y}} = \sum_1^n \sum_1^m y_{ij}/nm = \sum_1^n \bar{y}_i/n. \tag{8.2}$$

The sample variances among the primary units and the secondary units, respectively, are

$$s_1^2 = \sum_1^n (\bar{y}_i - \bar{\bar{y}})^2/(n-1) \qquad \text{and} \qquad s_2^2 = \sum_1^n s_{2i}^2/n, \tag{8.3}$$

where $s_{2i}^2 = \Sigma_1^m (y_{ij} - \bar{y}_i)^2/(m-1)$ is the variance within the ith primary unit.

8.3 Estimating the mean

Estimator for the mean and its variance

For each of the primary units selected at the first stage, $\bar{y}_i$ is unbiased for the mean $\bar{Y}_i$. As shown in Appendix A8, $\bar{\bar{y}}$ is unbiased for $\bar{\bar{Y}}$ and

its variance is given by

$$V(\bar{\bar{y}}) = (1 - f_1)S_1^2/n + (1 - f_2)S_2^2/nm, \tag{8.4}$$

where $f_1 = n/N$ and $f_2 = m/M$ are the sampling fractions at the first and second stages.

Since only m of the M secondary units are sampled, the variance increased from single-stage sampling by the amount in the second term. In practice, the optimum values of n and m for minimizing the variance or cost of sampling are found.

Example 8.1. Establishments per county: For the number of establishments in Table 7.1 for the nine clusters with six counties in each, $\bar{\bar{Y}} = 3.02$, $S_1^2 = 2.63$, and $S_2^2 = 15.96$. If $n = 3$ and $m = 3$, from (8.4), $V(\bar{\bar{y}}) = (6/27)(2.63) + 15.96/18 = 1.47$.

Since the population variance S^2 is equal to 15.94, for the alternative procedure of selecting a simple random sample of size $nm = 9$ from the 54 counties, the variance is $V(\bar{\bar{y}}_{srs}) = (5/54)(15.94) = 1.48$. In this illustration, there is very little difference in the precision of $\bar{\bar{y}}$ and $\bar{\bar{y}}_{srs}$. However, there can be saving in costs for cluster sampling.

Estimation of the variance

The sample variance s_{2i}^2 is unbiased for S_{2i}^2, and hence s_2^2 is unbiased for S_2^2. This result can be used for estimating the second term of (8.4). As shown in Appendix A8, $s_1^2 - (1-f_2)s_2^2/m$ is unbiased for S_1^2. Using this estimator for S_1^2 and s_2^2 for S_2^2, an unbiased estimator for the variance in (8.4) is

$$v(\bar{\bar{y}}) = (1 - f_1)s_1^2/n + f_1(1 - f_2)s_2^2/nm. \tag{8.5}$$

Example 8.2. Standard error for estimating employment per county: In a random sample of size three from the nine clusters described in Table 7.1, clusters 3, 5, and 9 appeared. In random samples of size three from each of these clusters, counties (2, 4, 5), (1, 4, 6), and (2, 3, 6), respectively, were drawn. The numbers of establishments and the persons employed in them are presented in Table 8.1.

Table 8.1. A sample of three clusters of equal sizes ($M = 6$) and samples of three counties from the selected clusters.

Clusters (PSUs)	Counties (SSUs)	Establishments	Employments
	2	1.9	341.2
3	4	11.2	14.4
	5	5.1	7.8
	1	2.0	20.7
5	4	0.5	3.8
	6	1.6	17.8
	2	5.5	90.0
9	3	6.2	69.9
	6	13.1	170.5

For the establishments, the sample means $\bar{y}_i$ are 6.07, 1.37, and 8.27. The sample variances s_{2i}^2 are 22.32, 0.60, and 17.64. From these figures, $\bar{\bar{y}} = (6.07 + 1.37 + 8.27)/3 = 5.24$. Note that the actual mean is $\bar{Y} = 3.02$, as seen in Example 8.1. Further, $s_1^2 = 2.42$ and $s_2^2 = (22.32 + 0.60 + 17.64)/3 = 13.52$. Now, from (8.5), $v(\bar{\bar{y}}) = (2/9)(2.42) + 13.52/54 = 0.79$. The actual variance of $\bar{\bar{y}}$ as seen in Example 8.1 is 1.47.

8.4 Sample size determination

Continuing with Example 8.1, if $n = 4$ and $m = 2$, from (8.4), $V(\bar{\bar{y}}) = 5(2.63)/36 + 15.96/12 = 1.03$. On the other hand, if $n = 2$ and $m = 4$, $V(\bar{\bar{y}}) = 7(2.63)/18 + 15.96/24 = 1.6878$. In both cases, $V(\bar{\bar{y}}_{\text{srs}}) = 46(15.94)/432 = 1.6973$. These illustrations show that proper choice of n and m can result in a high precision for cluster sampling.

The cost of sampling the nm elements may be of the form $E = enm$, where e is the expense for collecting information from a selected unit. The sample sizes can be determined for a specified cost and the variance in (8.4).

As an illustration, if $e = \$100$ and $E = \$1200$, $nm = 12$. Now, if $S_1^2 = 2.63$ and $S_2^2 = 15.96$ as in Example 8.1, from (8.4), the variance of $\bar{\bar{y}}$ equals $2.63/n - 2.63/9 + 15.96/12 - 15.96/(6n) = 1.0378 - 0.03/n$. Thus, if $(n, m) = (4, 3)$ or $(3, 4)$, the variance will be close to 1.03. Other types of cost functions suitable to the application can be considered to examine the required sample sizes.

8.5 Unequal size primary units

The notation for this case is the same as in Section 7.7, except that the total of the ith cluster is denoted by Y_i. When the ith primary unit contains M_i secondary units, its total and mean are $Y_i = \Sigma^{Mi} y_{ij}$ and $\bar{Y}_i = Y_i/M_i$. The population mean per primary and secondary unit, respectively, are $\bar{\bar{Y}} = \Sigma^N \bar{Y}_i/N$ and $\bar{\bar{Y}} = \Sigma^N Y_i/M_0 = \Sigma^N M_i \bar{Y}_i/M_0$, where $M_0 = \Sigma^N M_i$ is the total number of secondary units. The variance within the ith primary unit is $S_{2i}^2 = \Sigma^{Mi}(y_{ij} - \bar{Y}_i)^2/(M_i - 1)$.

As before, at the first stage, a sample of n primary units is selected without replacement. At the second stage, a sample of size m_i is selected randomly without replacement from the M_i secondary units of the selected primary unit. The mean of this sample is $\bar{y}_i = \Sigma^{mi} y_{ij}/m_i$, which is unbiased for $\bar{Y}_i$. Its variance $s_{2i}^2 = \Sigma^{mi}(y_{ij} - \bar{y}_i)^2/(m_i - 1)$ is unbiased for S_{2i}^2.

Estimators for the population total and mean

An unbiased estimator of Y_i is $\hat{Y}_i = M_i \bar{y}_i$. As shown in Appendix A8, $\hat{Y} = (N/n)\Sigma^n \hat{Y}_i$ is unbiased for Y and its variance is given by

$$V(\hat{Y}) = [N^2(1 - f_1)/n)] \sum^N (Y_i - \bar{\bar{Y}})^2/(N - 1) + (N/n) \sum^N M_i^2 (1 - f_{2i}) S_{2i}^2 / m_i. \quad (8.6)$$

An unbiased estimator of $\bar{\bar{Y}}$ is $\hat{\bar{\bar{Y}}} = \hat{Y}/N = (\Sigma^n \hat{Y}_i)/n$. Its variance is obtained by dividing (8.6) by N^2. Similarly, $\hat{Y}/M_0$ is unbiased for $\bar{\bar{Y}}$ and its variance is obtained by dividing (8.6) by M_0^2. If the totals Y_i vary with the sizes, the first term of (8.6) becomes large. The ratio method to adjust the estimator for the sizes is described in Section 8.6.

Example 8.3. Establishments: Consider a random sample of size three from the clusters described in Table 7.3. For second-stage sampling, consider $m_i = 3$ counties for the first two and last two clusters selected at the first stage and $m_i = 2$ counties for the remaining five clusters selected at the first stage.

For the establishments, the first term of (8.6) equals $9(6)(16.42)^2/3 = 4853.1$ and the second term equals 5065.1. Combining these figures, $V(\hat{Y}) = 9918.1$ and $V\hat{\bar{\bar{Y}}} = 9918.1/(54)^2 = 3.4$.

Estimation of the variance

From the second and third expressions in Appendix A8, an unbiased estimator of (8.6) is given by

$$\begin{aligned} v(\hat{Y}) &= N^2[(1-f_1)/n]\sum^{n}(\hat{Y}_i - \hat{\bar{Y}})^2/(n-1) \\ &\quad -(N/n)^2(1-f_1)\left[\sum^{n} M_i^2(1-f_{2i})s_{2i}^2/m_i\right] \\ &\quad +(N/n)^2\sum^{n} M_i^2(1-f_{2i})s_{2i}^2/m_i \\ &= N^2[(1-f_1)/n]\sum^{n}(\hat{Y}_i - \hat{\bar{Y}})^2/(n-1) \\ &\quad +(N/n)\sum^{n} M_i^2(1-f_{2i})s_{2i}^2/m_i. \end{aligned} \tag{8.7}$$

To estimate the variance of $\hat{\bar{\bar{Y}}} = \hat{Y}/M_0$, divide (8.7) by M_0^2.

Example 8.4. Standard error from the sample: In a random sample of size three from the clusters of unequal size described in Table 7.3, clusters 1, 4, and 8 were drawn. In random subsamples of sizes 3, 2, and 3 from the selected clusters, counties (3, 4, 10), (2, 4), and (2, 4, 5), respectively, were drawn. Figures for the establishments and employments for these samples are presented in Table 8.2. For the establishments, the subsample means $\bar{y}_i$ and variances s_i^2 for the sample clusters are (7.77, 2.75, 4.57) and (137.4, 6.125, 8.9), respectively. Estimates $\hat{Y}_i$ for the three selected clusters are 10(7.77) = 77.7, 4(2.75) = 11, and 7(4.57) = 31.99. Hence, $\hat{Y}$ = (9/3)(77.7 + 11 + 31.99) = 362.1, $\hat{\bar{Y}}$ = 362.1/9 = 40.23, and $\hat{\bar{\bar{Y}}}$ = 362.1/54 = 6.7. Note that Y = 162.9 and $\bar{\bar{Y}}$ = 162.9/54 = 3.02.

The above estimates for the total and mean are about twice their actual values. This has happened because the two largest clusters, 1 and 8, appeared in the sample. Estimators with smaller variances can be obtained if the PSUs are stratified according to their sizes. If supplementary information is available, the ratio and regression procedures presented in Chapters 9 and 10 can be beneficial.

Since $\Sigma_1^3(\hat{\bar{Y}}_i - \hat{\bar{Y}})^2/2$ = 1163.1, the first term of (8.7) equals 20,936.6. With the second-stage variances, the second term equals 9940.7. Thus, $v(\hat{Y})$ = 30,877.3 and S.E.$(\hat{Y})$ = 175.7. For the estimation of $\bar{\bar{Y}}$, S.E. $(\bar{\bar{Y}})$ = 175.7/54 = 3.25.

Table 8.2. Sample of clusters of unequal size and samples of counties.

Clusters (PSUs)	Cluster Sizes M_i	Counties (SSUs)	Establishments	Employments
1	10	3	21.3	341.2
		4	1.2	14.4
		10	0.8	10.1
4	4	2	1.0	13.7
		4	4.5	84.7
8	7	2	3.1	52.5
		4	8.0	132.0
		5	2.6	34.2

Hansen and Hurwitz (1949) and Cochran (1977, p. 313) determine the sample sizes for the two stages with a suitable cost function.

8.6 Ratio adjustments for the sizes

The first term of the variance in (8.6) becomes large if Y_i vary much from each other. In many applications Y_i depends on the size M_i. In such cases, the ratio method can be used for estimating Y with a small MSE. The ratio-type estimator for Y is

$$\begin{aligned}\hat{Y}_R &= M_0\left(\sum^n \hat{Y}_i\right)\Big/\left(\sum^n M_i\right) = N\overline{M}\left(\sum_1^n \hat{Y}_i/n\right)\Big/\left(\sum^n M_i/n\right) \\ &= N\overline{M}\left(\sum^n M_i\bar{y}_i/n\right)\Big/\overline{m},\end{aligned} \tag{8.8}$$

where $\overline{M} = M_0/N$ is the average size of the primary units and $\overline{m} = \Sigma^n M_i/n$ is the mean of the sizes of the cluster selected at the first stage. The estimator of the population mean is given by $\hat{\bar{Y}}_R = \hat{Y}_R/M_0$. For large n, the bias of $\hat{Y}_R$ in (8.8) becomes small, and from Appendix A8 its variance is approximately given by

$$\begin{aligned}V(\hat{Y}_R) &= N^2[(1-f_1)/n]\sum^N M_i^2(\bar{Y}_i - \bar{\bar{Y}})^2/(N-1) \\ &\quad + (N/n)\left[\sum^N M_i^2(1-f_{2i})S_{2i}^2/m_i\right].\end{aligned} \tag{8.9}$$

The first term of this expression will be small if Y_i is close to $\bar{\bar{Y}}M_i$, that is, if Y_i is correlated with M_i. Notice that the second terms of (8.6) and (8.9) are the same.

Starting with the expectation of $\Sigma^n M_i^2(\bar{y}_i - \hat{\bar{\bar{Y}}}_R)^2/(n-1)$ and adopting the procedure in Appendix A8, an estimator for $V(\hat{Y}_R)$ is given by

$$v(\hat{Y}_R) = N^2[(1-f_1)/n]\sum^n M_i^2(\bar{y}_i - \hat{\bar{\bar{Y}}}_R)^2/(n-1) + (N/n)\sum^n M_i^2(1-f_{2i})s_{2i}^2/m_i. \quad (8.10)$$

Note that the second terms in (8.7) and (8.10) are the same.

Example 8.5. Ratio estimation for the establishments: With the samples drawn in Example 8.4, for the number of establishments, $\hat{Y}_R = 54(120.69)/21 = 310.35$ and $\hat{\bar{\bar{Y}}}_R = 310.35/54 = 5.75$. Since $\Sigma^n M_i^2(\bar{y}_i - \hat{\bar{\bar{Y}}}_R)^2 = 592.5$, the first term of (8.10) equals 5332.3. The second term as in Example 8.4 equals 9940.7. Thus, $v(\hat{Y}_R) = 15{,}273$ and S.E.$(\hat{Y}_R) = 123.6$. Hence, S.E.$(\hat{\bar{\bar{Y}}}_R) = 2.29$.

The above ratio estimators for the total and mean differ only slightly from $\hat{Y}$ and $\bar{\bar{Y}}$ found in Example 8.4. For the estimation of these quantities in this illustration, the ratio method has helped only by a small amount.

8.7 Proportions and totals

Let C_i and P_i denote the total number and proportion of the M_i secondary units having the characteristic of interest. The total number and proportion for the population are $C = \Sigma^N C_i$ and $P = C/M_0$. With the one-zero notation for y_{ij}, the totals and means can be expressed as $Y_i = C_i$, $\bar{Y}_i = P_i$, and $\bar{\bar{Y}} = P$. Similarly, let c_i and $p_i = c_i/m_i$ denote the total number and proportion of the m_i secondary units of the sample having the attribute.

Equal size PSUs

When $M_i = M$ and $m_i = m$, from (8.2), an unbiased estimator for P is

$$p = \sum_1^n p_i/n = \sum^n c_i/nm. \quad (8.11)$$

The variance of this estimator can be obtained from (8.4). From (8.5), the estimator of variance is

$$v(p) = \frac{(1-f_1)}{n}\frac{\sum^n (p_i - p)^2}{n-1} + \frac{f_1(1-f_2)}{n^2(m-1)}\sum_1^n p_i(1-p_i). \quad (8.12)$$

An unbiased estimator of C is given by $M_0 p$. For the sample variance of this estimator, multiply (8.12) by M_0^2.

Example 8.6. Employment in the counties: Observations from a sample of $n = 3$ clusters and subsamples of $m = 3$ counties from the selected clusters are presented in Table 8.1. Numbers of counties with employment of at least 20,000 persons are $c_1 = 1$, $c_2 = 1$, and $c_3 = 3$. Thus, $p_1 = 1/3$, $p_2 = 1/3$, and $p_3 = 1$. Estimates for the proportion and total number of counties with this characteristic are $p = (5/9) = 0.56$ and $\hat{C} = 54(5/9) = 30$. From Table 7.1, the actual total and proportion are $C = 25$ and $P = 25/54 = 0.46$.

The first and second terms of (8.12) equal (16/729) and 3/729. Hence $v(p) = 19/729 = 0.0261$, S.E.$(p) = 0.16$, and S.E. $(\hat{C}) = 54(0.16) = 9$.

Unequal size PSUs

When M_i are unequal, from Section 8.5, an unbiased estimator for the total is

$$\hat{C} = \frac{N}{n}\sum_1^n M_i p_i. \quad (8.13)$$

Its variance can be obtained from (8.6). This variance will be small if C_i do not differ much from each other. From (8.7), an unbiased estimator of the variance is

$$v(\hat{C}) = \frac{N^2(1-f_1)}{n}\frac{\sum^n \left[M_i p_i - \sum^n M_i p_i/n\right]^2}{n-1} + \frac{N}{n}\sum_1^n M_i^2 \frac{(1-f_{2i})}{m_i - 1} p_i(1-p_i). \quad (8.14)$$

An alternative estimator

Adjusting for the sizes through the ratio method, from (8.8), an alternative estimator is given by

$$\hat{C}_R = M_0 \frac{\sum^n M_i p_i}{\sum^n M_i} = M_0 \bar{p}, \tag{8.15}$$

where $\bar{p}$ is the weighted average of the p_i. Its variance, which can be obtained from (8.9), is small if P_i are close to each other. From (8.10), an estimator for this variance is approximately given by

$$v(\hat{C}_R) = \frac{N^2(1-f_1)}{n} \frac{\sum^n M_i^2 (p_i - \bar{p})^2}{n-1} + \frac{N}{n} \sum_1^n M_i^2 \frac{(1-f_{2i})}{m_i - 1} p_i(1-p_i). \tag{8.16}$$

To estimate P when M_i are unequal, divide (8.13) or (8.15) by M_0. The corresponding estimates of variances are obtained by dividing (8.14) and (8.16) by M_0^2.

Example 8.7. S.E. for estimating employment: Sample observations from the clusters of unequal size in Table 7.3 and subsamples from the selected clusters are presented in Table 8.2. The observed numbers of counties with at least 20,000 employed persons are $c_1 = 1$, $c_2 = 1$, $c_3 = 3$. The sample proportions are $p_1 = 1/3$, $p_2 = 1/2$, $p_3 = 1$.

Since $(\Sigma_1^3 M_i p_i)/3 = 74/18$, from (8.13), $\hat{C} = 9(74/18) = 37$. The first and second terms of (8.14) equal 120.7 and 29.3. Hence $v(\hat{C}) = 150$ and S.E.$(\hat{C}) = 12.25 = 12$. The estimate of the proportion for the above attribute is $37/54 = 0.69$. The values of 37 and 0.69 for this estimation procedure are rather large relative to the actual values 25 and 0.46.

From (8.15), with the ratio adjustment for the sizes, $\hat{C}_R = (54/21)(74/6) = 32$. The first term of (8.16) equals 134.3 and as seen above, the second term equals 29.3. Thus, $V(\hat{C}_R) = 163.6$ and S.E.$(\hat{C}_R) = 13$. The corresponding estimate of the proportion of the counties with the above attribute is $32/54 = 0.59$, which has an S.E. of 0.24. The estimates 32 and 0.59 for the total and proportion are closer to the actual values than the above estimates without the ratio adjustment for the sizes.

8.8 Unequal probability selection

When the sizes of the primary units are not the same, they can be selected with unequal probabilities through the procedures of the type described in Sections 7.11 through 7.13. From each of the selected first-stage units, samples can be selected at the second stage with equal or unequal probabilities. Selection of the first-stage units with unequal probabilities and without replacement, followed by the Horvitz-Thompson estimator, and selection of the first-stage units with unequal probabilities and with replacement are described in this section. Des Raj (1968, Chapter 6) and Cochran (1977, Chapter 11), for instance, describe additional procedures for selecting the first-stage units.

Horvitz-Thompson estimator

As described in Section 7.11, when the primary units are selected with unequal probabilities and **without** replacement, denote the inclusion probabilities for a unit and a pair of units by ϕ_i and ϕ_{ij}. The second-stage units can be selected with equal or unequal probabilities and without or with replacement. Let $\hat{Y}_i$ denote an unbiased estimator for the total Y_i of the selected primary unit obtained from the sample at the second stage, and denote its variance $V_2(\hat{Y}_i)$ by V_{2i}.

The Horvitz-Thompson estimator for the total Y now is

$$\hat{Y}_{HT} = \sum_{1}^{n} \frac{\hat{Y}_i}{\phi_i}. \tag{8.17}$$

As shown in Appendix A8, this estimator is unbiased for Y and its variance is given by

$$V(\hat{Y}_{HT}) = \sum_{i}^{N} \sum_{i<j}^{N} (\phi_i \phi_j - \phi_{ij}) \left(\frac{Y_i}{\phi_i} - \frac{Y_j}{\phi_j} \right)^2 + \sum_{1}^{N} \frac{V_{2i}}{\phi_i}. \tag{8.18}$$

For the procedure in Section 8.2, where the samples at both the stages are drawn randomly without replacement, $\phi_i = \phi_j = n/N$, $\phi_{ij} = n(n-1)/N(N-1)$ and $V_{2i} = M^2(1-f_2)S_{2i}^2/m$. The inclusion probabilities for the procedure in Section 8.5 for the unequal size clusters also have these values, but V_{2i} in this case is equal to $M_i^2(1-f_{2i})S_{2i}^2/m_i$.

As shown in Appendix A8, an unbiased estimator for $V(\hat{Y}_{HT})$ in (8.18) is

$$v(\hat{Y}_{HT}) = \sum_{i}^{n} \sum_{i<j}^{n} \left(\frac{(\phi_i \phi_j - \phi_{ij})}{\phi_{ij}} \right) \left(\frac{\hat{Y}_i}{\phi_i} - \frac{\hat{Y}_j}{\phi_j} \right)^2 + \sum_{1}^{n} \frac{\hat{V}_{2i}}{\phi_i}. \tag{8.19}$$

Selection with unequal probabilities and replacement

As described in Section 7.13 for the *pps* method, consider selecting the n first-stage units with probabilities u_i and replacement. From each of the units selected at the first stage, consider selecting the second-stage units independently. Note that each time a first-stage unit is selected, the second-stage sample is selected from it independently.

With Y_i denoting the total of the first-stage unit, let $r_i = Y_i/u_i$. For Y_i, consider an unbiased estimator $\hat{Y}_i$ obtained from the sample at the second stage, and let $\hat{r}_i = \hat{Y}_i/u_i$. Now, an estimator for Y is given by

$$\hat{Y}_{pps} = \sum_{i}^{n} \hat{r}_i/n. \tag{8.20}$$

Since $E_2(\hat{r}_i) = r_i$, $E(\hat{Y}_{pps}) = Y$. Thus, $\hat{Y}_{pps}$ is unbiased for Y. As shown in Appendix A8, the variance of $\hat{Y}_{pps}$ and its estimator are given by

$$V(\hat{Y}_{pps}) = \sum_{i}^{N} u_i (r_i - Y)^2/n + (1/n) \sum_{i}^{N} V_2(\hat{Y}_i)/u_i. \tag{8.21}$$

and

$$v(\hat{Y}_{pps}) = \sum_{i}^{n} (\hat{r}_i - \hat{Y}_{pps})^2/n(n-1). \tag{8.22}$$

Exercises

8.1. For a sample of three from the nine clusters of Table 7.1 and a subsample of three counties from each of the selected clusters, find the variance for estimating

the employment per county. Compare this with the variance of the mean of a simple random sample of nine counties. Note that the means and the variances of the clusters are presented in Table 7.2, and S^2 for the employment equals 5057.79.

8.2. With the results of the two-stage sampling considered in Example 8.2 and presented in Table 8.1, (a) estimate the average amount of employment per county and find its standard error. (b) Use these figures to estimate the total employment and its standard error.

8.3. Consider a sample of three clusters at the first stage from the nine clusters of Table 7.3. As in Example 8.3, at the second stage consider samples of size three for the first two and the last two clusters, and samples of size two for the rest of the five clusters. Find the standard errors for estimating the total and average employment for the 54 counties.

8.4. With the samples drawn in Example 8.4 at the two stages presented in Table 8.2, find (a) estimates for the total and the average employment and compare them with the actual values. (b) Find the standard errors from the samples.

8.5. With the sample sizes in Exercise 8.3, (a) find the standard errors for estimating the total and average employment through ratio adjustment with the cluster sizes. (b) Compare these standard errors with those in Exercise 8.3.

8.6. With the sample observations in Table 8.1, estimate the proportion and total number of counties that have at least 2000 establishments. Find the standard errors of the estimates.

8.7. Using the sample observations in Table 8.2 for the unequal size clusters, estimate the total and proportion of counties that have at least 2000 establishments. Obtain the estimates with and without ratio adjustment for sizes. Find the standard errors of the estimates.

8.8. *Project.* Consider the first three, next three, and the remaining four states in Table T7 in the Appendix as three clusters. Select all the samples of size two from each of the three clusters. From each selected cluster, select all the samples of two states. Select the clusters and the states randomly without replacement. For the

total number of physicians, from each of the samples selected above in two stages, find the sample estimate in Section 8.5 and the ratio estimate in (8.8). From these results, (a) find the expectation and variance of the sample estimate and show that they coincide with the actual total Y and the variance in (8.6). (b) Find the expectation, bias, and MSE of the ratio estimator, and compare this exact MSE with the approximation in (8.8).

Appendix A8

Expected value and variance of the estimator for the mean

The expected value and variance of $\bar{\bar{y}}$ can be obtained from the general results in Appendix A3. From these expressions,

$$E(\bar{\bar{y}}) = E_1 E_2(\bar{\bar{y}})$$

and

$$V(\bar{\bar{y}}) = V_1 E_2(\bar{\bar{y}}) + E_1 V_2(\bar{\bar{y}}),$$

where the subscript 2 refers to the expectation conditional on the units selected at the first stage and 1 to the unconditional expectation.

Since $\bar{y}_i$ is unbiased for the mean $\bar{Y}_i$ of the selected primary unit,

$$E_2(\bar{\bar{y}}) = E_2\left(\frac{1}{n}\sum_1^n \bar{y}_i\right) = \frac{1}{n}\sum_1^n \bar{Y}_i$$

and

$$E(\bar{\bar{y}}) = E_1 E_2(\bar{\bar{y}}) = \frac{1}{N}\sum_1^N \bar{Y}_i = \bar{\bar{Y}}.$$

Thus, $\bar{\bar{y}}$ is unbiased for $\bar{\bar{Y}}$.

Further,

$$V_1 E_2(\bar{\bar{y}}) = (1 - f_1) S_1^2 / n,$$

where $f_1 = n/N$ is the sampling fraction at the first stage.

Samples at the second stage are drawn independently from the selected primary units. Consequently, the covariance between $\bar{y}_i$ and $\bar{y}_j$ for $(i \neq j)$ vanishes. Hence,

$$V_2(\bar{\bar{y}}) = V_2\left(\sum_1^n \bar{y}_i\right) / n^2 = [(1 - f_2) / m] \sum_1^n S_{2i}^2 / n^2,$$

where $f_2 = m/M$ is the sampling fraction at the second stage. From this expression,

$$E_1 V_2(\bar{\bar{y}}) = (1 - f_2) S_2^2 / nm.$$

Finally,

$$V(\bar{\bar{y}}) = V_1 E_2(\bar{\bar{y}}) + E_1 V_2(\bar{\bar{y}}),$$

which is presented in (8.4).

Note that

$$\begin{aligned}(n-1)s_1^2 = \sum_1^n (\bar{y}_i - \bar{\bar{y}})^2 &= \sum_1^n [(\bar{y}_i - \bar{\bar{Y}}) - (\bar{\bar{y}} - \bar{\bar{Y}})]^2 \\ &= \sum_1^n (\bar{y}_i - \bar{Y})^2 - n(\bar{\bar{y}} - \bar{\bar{Y}})^2.\end{aligned}$$

Hence,

$$(n-1)E(s_1^2) = E\sum_1^n (\bar{y}_i - \bar{\bar{Y}})^2 - n V(\bar{\bar{y}}).$$

Now,

$$\sum_1^n (\bar{y}_i - \bar{\bar{Y}})^2 = \sum_1^n [(\bar{y}_i - \bar{Y}_i) + (\bar{Y}_i - \bar{\bar{Y}})]^2$$

$$= \sum_1^n (\bar{y}_i - \bar{Y}_i)^2 + \sum_1^n (\bar{Y}_i - \bar{\bar{Y}})^2 + 2\sum_1^n (\bar{y}_i - \bar{Y}_i)(\bar{Y}_i - \bar{\bar{Y}}).$$

Since $E_2(\bar{y}_i) = \bar{Y}_i$, expectation of the last term vanishes. Further, $E_2(\bar{y}_i - \bar{Y}_i)^2 = V_2(\bar{y}_i) = (1 - f_2)S_{2i}^2/m$, and hence $E\sum_1^n (\bar{y}_i - \bar{Y}_i)^2 = n[(1 - f_2)/m]\sum_1^N S_{2i}^2/N = n(1 - f_2)S_2^2/m$. Expectation of the second term is equal to $(n/N)\sum_1^N (\bar{Y}_i - \bar{\bar{Y}})^2 = n(N - 1)S_1^2/N$.

From these results,

$$(n - 1)E(s_1^2) = n(1 - f_2)S_2^2/m + n(N - 1)S_1^2/N - nV(\bar{\bar{y}})$$

$$= (n - 1)(1 - f_2)S_2^2/m + (n - 1)S_1^2.$$

Thus,

$$s_1^2 - (1 - f_2)s_2^2/m \text{ is unbiased for } S_1^2.$$

Expectation and variance of the estimator for the total

Since

$$E_2(\hat{Y}) = N\sum^n \hat{Y}_i/n$$

$$E(\hat{Y}) = E_1[E_2(\hat{Y})] = \sum^N Y_i = Y.$$

Hence, $\hat{Y}$ is unbiased for the population total. Now,

$$V_1[E_2(\hat{Y})] = N^2[(1 - f_1)/n]\sum^N (Y_i - \bar{Y})^2/(N - 1).$$

Noting that the covariance between $\hat{Y}_i$ and $\hat{Y}_j$ for $(i \neq j)$ vanishes, from (8.10),

$$V_2(\hat{Y}) = (N/n)^2 \sum^{n} M_i^2 (1 - f_{2i}) S_{2i}^2 / m_i,$$

where $f_{2i} = m_i/M_i$. Hence,

$$E_1[V_2(\hat{Y})] = (N/n) \sum^{N} M_i^2 (1 - f_{2i}) S_{2i}^2 / m_i.$$

Finally, $V(\hat{Y}) = V_1[E_2(\hat{Y})] + E_1[V_2(\hat{Y})]$, which is presented in (8.6).

Estimation of the variance — unequal cluster sizes

Note that

$$E\left[\sum_1^n (\hat{Y}_i - \hat{\bar{Y}})^2\right] = [(n-1)/N] \sum_1^N M_i^2 (1 - f_{2i}) S_{2i}^2 / m_i + [(n-1)/(N-1)]\left[\sum_1^N (Y_i - \bar{Y})^2\right].$$

Thus,

$$N^2[(1 - f_1)/n] E\left[\sum_1^n (\hat{Y}_i - \hat{\bar{Y}})^2/(n-1)\right] =$$

$$N[(1 - f_1)/n] \sum_1^N M_i^2 (1 - f_{2i}) S_{2i}^2 / m_i$$

$$+ N^2[(1 - f_1)/n] \sum_1^N (Y_i - \bar{Y})^2/(N-1).$$

Further,

$$E\left[(1/n)\sum_1^n M_i^2\frac{1-f_{2i}}{m_i}s_{2i}^2\right] = (1/N)\sum_1^N M_i^2\frac{1-f_{2i}}{m_i}S_{2i}^2.$$

$$E\left[(1/n)\sum_1^n M_i^2(1-f_{2i})s_{2i}^2/m_i\right] = (1/N)\sum_1^N M_i^2(1-f_{2i})S_{2i}^2/m_i.$$

Variance of the ratio estimator for the total

From (8.8),

$$E_2(\hat{Y}_R) = N\overline{M}\left(\sum^n M_i(\overline{Y}_i/n)\right)/\overline{m}.$$

Hence, for large n,

$$V_1E_2(\hat{Y}_R) = N^2[(1-f_1)/n]\sum^N M_i^2(\overline{Y}_i-\overline{\overline{Y}})^2/(N-1).$$

Now,

$$V_2(\hat{Y}_R) = (N/n)^2\left[\sum^n M_i^2(1-f_{2i})S_{2i}^2/m_i\right]$$

$$E_1V_2(\hat{Y}_R) = (N/n)\left[\sum^N M_i^2(1-f_{2i})S_{2i}^2/m_i\right].$$

The variance of $\hat{Y}_R$ is obtained by combining the second and last terms.

The conditional expectation of (8.17) is

$$E_2(\hat{Y}_{HT}) = \sum_1^n\frac{1}{\phi_i}E_2(\hat{Y}_i) = \sum_1^n\frac{Y_i}{\phi_i}$$

and hence

$$E(\hat{Y}_{HT}) = E_1\left(\sum_1^n \frac{Y_i}{\phi_i}\right) = \sum_1^N Y_i = Y.$$

Now, with the procedure in Section 7.12,

$$V_1[E_2(\hat{Y}_{HT})] = \sum_i^N \sum_{i<j}^N (\phi_i\phi_j - \phi_{ij})\left(\frac{Y_i}{\phi_i} - \frac{Y_j}{\phi_j}\right)^2.$$

Noting that $\hat{Y}_i$ and $\hat{Y}_j$ are uncorrelated, from (8.17),

$$V_2(\hat{Y}_{HT}) = \sum_1^n \frac{V_{2i}}{\phi_i^2}$$

and hence

$$E_1[V_2(\hat{Y}_{HT})] = \sum_1^N \frac{V_{2i}}{\phi_i}.$$

Finally,

$$\begin{aligned} V(\hat{Y}_{HT}) &= V_1[E_2(\hat{Y}_{HT})] + E_1[V_2(\hat{Y}_{HT})] \\ &= \sum_i^N \sum_{i<j}^N (\phi_i\phi_j - \phi_{ij})\left(\frac{Y_i}{\phi_i} - \frac{Y_j}{\phi_j}\right)^2 + \sum_1^N \frac{V_{2i}}{\phi_i}. \end{aligned}$$

Estimator for $V(\hat{Y}_{HT})$

As suggested by Des Raj (1968, p. 118),

$$E_2\left(\frac{\hat{Y}_i}{\phi_i} - \frac{\hat{Y}_j}{\phi_j}\right)^2 = V_2\left(\frac{\hat{Y}_i}{\phi_i} - \frac{\hat{Y}_j}{\phi_j}\right) + \left(\frac{Y_i}{\phi_i} - \frac{Y_j}{\phi_j}\right)^2.$$

Since the samples at the second stage are selected independently, the covariance between $\hat{Y}_i$ and $\hat{Y}_j$ vanishes, and this equation can be

expressed as

$$E_2\left(\frac{\hat{Y}_i}{\phi_i} - \frac{\hat{Y}_j}{\phi_j}\right)^2 = \left(\frac{V_{2i}}{\phi_i^2} + \frac{V_{2j}}{\phi_j^2}\right) + \left(\frac{Y_i}{\phi_i} - \frac{Y_j}{\phi_j}\right)^2.$$

Now,

$$E\left[\sum_i^n \sum_{i<j}^n \left(\frac{(\phi_i\phi_j - \phi_{ij})}{\phi_{ij}}\right)\left(\frac{\hat{Y}_i}{\phi_i} - \frac{\hat{Y}_j}{\phi_j}\right)^2\right]$$

$$= \sum_i^N \sum_{i<j}^N (\phi_i\phi_j - \phi_{ij})\left[\left(\frac{V_{2i}}{\phi_i^2} + \frac{V_{2j}}{\phi_j^2}\right) + \left(\frac{Y_i}{\phi_i} - \frac{Y_j}{\phi_j}\right)^2\right]$$

$$= \sum_i^N \sum_{i<j}^N (\phi_i\phi_j - \phi_{ij})\left(\frac{V_{2i}}{\phi_i^2} + \frac{V_{2j}}{\phi_j^2}\right) + \sum_i^N \sum_{i<j}^N (\phi_i\phi_j - \phi_{ij})\left(\frac{Y_i}{\phi_i} - \frac{Y_j}{\phi_j}\right)^2.$$

Following (7.33c), the first term of this expression is the same as $\sum_i^N (V_{2i}/\phi_i^2)\phi_i(1 - \phi_i) = \sum_i^N V_{2i}/\phi_i - \sum_i^N V_{2i}$. Thus,

$$E\left[\sum_i^n \sum_{i<j}^n \left(\frac{(\phi_i\phi_j - \phi_{ij})}{\phi_{ij}}\right)\left(\frac{\hat{Y}_i}{\phi_i} - \frac{\hat{Y}_j}{\phi_j}\right)^2\right] = V(\hat{Y}_{HT}) - \sum_i^N V_{2i}.$$

At the second stage, let $\hat{V}_{2i}$ denote an unbiased estimator for V_{2i}, that is, $E_2(\hat{V}_{2i}) = V_{2i}$. Now, $\sum_i^n (\hat{V}_{2i}/\phi_i)$ is unbiased for $\sum_i^N V_{2i}$, and from the above expression an unbiased estimator for $V(\hat{Y}_{HT})$ is given by

$$v(\hat{Y}_{HT}) = \sum_i^n \sum_{i<j}^n \left(\frac{\phi_i\phi_j - \phi_{ij}}{\phi_{ij}}\right)\left(\frac{\hat{Y}_i}{\phi_i} - \frac{\hat{Y}_j}{\phi_j}\right)^2 + \sum_1^n \frac{\hat{V}_{2i}}{\phi_i}.$$

Variance of $\hat{Y}_{pps}$ and its estimator

First note that since $\hat{Y}_i$ are independent, $V(\hat{Y}_{pps}) = V(\hat{r}_i)/n$. Now, as found in Section 7.13, $V_1E_2(\hat{r}_i) = V_1(r_i) = \sum_i^N u_i(r_i - Y)^2$. Further, $V_2(\hat{r}_i) =$

$V_2(\hat{Y}_i)/u_i^2$, and $E_1 V_2(\hat{r}_i) = \sum_i^N V_2(\hat{Y}_i)/u_i$. Thus,

$$V(\hat{Y}_{pps}) = \sum_i^N u_i(r_i - Y)^2/n + (1/n)\sum_i^N V_2(\hat{Y}_i)/u_i.$$

As in section 7.13, an unbiased estimator of $V(\hat{r}_i)$ is given by $v(\hat{r}_i) = \sum_i^n (\hat{r}_i - \hat{Y}_{pps})^2/(n-1)$. Hence an unbiased estimator of $V(\hat{Y}_{pps})$ is given by $v(\hat{Y}_{pps}) = \sum_i^n (\hat{r}_i - \hat{Y}_{pps})^2/n(n-1)$.