CHAPTER 7

Cluster Sampling

7.1 Introduction

A cluster is usually a large unit consisting of smaller units, also known as the **elements**. Cities, city blocks, colleges, dormitories, residence halls, shopping malls, and apartment complexes are examples of clusters. In agricultural, demographic, economic, and political surveys, the geographical area of interest is divided into clusters with counties, provinces, or enumeration districts as elements. For some surveys, proximity of the elements is used as a guide for cluster formation.

Interest in estimates for individual clusters, convenience of sampling, and reduction of traveling costs for interviews are some of the reasons for dividing a population into clusters. For a study on elementary schools in a large region, the area can be divided into a few clusters with each of the regions containing a certain number of schools. The needed information can be obtained by visiting the schools in a random sample of the clusters. A random sample of the schools from the entire region would be spread more evenly in the area, but the distances between the selected units may increase the costs of the survey. In such cases, costs should be weighed against precision for preferring cluster sampling.

To estimate the population totals, means, and proportions, it becomes convenient in some situations to consider clusters of **equal sizes**, each containing the same number of elements. To increase the precisions in the case of **unequal size** clusters, the estimators can be adjusted for their sizes through what is known as the ratio method and similar procedures. Alternatively, precisions of the estimators can also be increased by selecting the cluster units with probabilities related to their sizes, instead of selecting them through simple random sampling.

In large-scale multistage surveys, the population is first divided into strata, and samples from the clusters in each stratum are selected with equal or unequal probabilities.

7.2 Clusters of equal sizes

For the sake of illustration, data on the number of establishments and employment in 54 counties of New York State, divided into nine clusters each containing six counties are presented in Table 7.1.

When the population consists of N clusters with M elements in each, let y_{ij}, $i = 1, \ldots, N$ and $j = 1, \ldots, M$ denote the values of a characteristic of interest. The total and mean of the ith

Table 7.1. Employment in 54 counties of New York State, divided into nine clusters of size six each; number of establishments (1000s) and number of persons employed (1000s).

	Establishments	Persons Employed	Establishments	Persons Employed	Establishments	Persons Employed
	1		4		7	
	4.2	63.8	0.7	11.7	1.0	13.7
	21.3	341.2	3.2	41.9	1.1	15.1
	0.6	6.1	1.7	22.2	1.2	13.8
	1.2	14.4	0.8	10.1	0.5	4.1
	0.7	7.8	1.8	25.9	3.1	52.5
	1.0	11.6	1.9	29.0	8.0	132.0
Total	29.0	444.9	10.1	140.8	14.9	231.2
Mean	4.8	74.2	1.7	23.5	2.5	38.5
	2		5		8	
	15.1	341.0	2.0	20.7	2.6	34.2
	2.0	25.0	1.7	21.6	4.5	84.7
	1.4	15.8	2.0	24.2	1.1	11.7
	0.4	4.1	0.5	3.8	1.1	8.4
	0.5	8.6	0.9	9.1	1.3	12.3
	1.5	17.5	1.6	17.8	1.9	17.7
Total	20.9	396.0	8.7	105.9	12.5	169.0
Mean	3.5	66.0	1.5	17.7	2.1	28.2
	3		6		9	
	0.3	3.1	1.1	14.5	3.6	41.8
	1.9	34.9	1.3	13.7	5.5	90.0
	0.9	14.0	1.1	10.3	6.2	69.9
	11.2	203.9	1.9	21.3	1.7	12.4
	5.1	74.9	1.0	10.2	6.8	73.4
	1.2	11.5	2.9	29.7	13.1	170.5
Total	20.6	342.3	9.3	99.7	36.9	458.0
Mean	3.4	57.1	1.6	16.6	6.2	76.3

Source: U.S. Bureau of the Census; 1985 figures.

Table 7.2. Means and variances of the nine clusters of Table 7.1.

	No. of Establishments		No. of Persons Employed	
Cluster	Mean	Variance	Mean	Variance
1	4.83	66.89	74.15	17,587.70
2	3.48	32.77	68.67	17,852.40
3	3.43	17.33	57.05	5,842.58
4	1.68	0.82	23.46	139.01
5	1.45	0.38	16.03	67.03
6	1.55	0.54	16.62	57.43
7	2.48	8.10	38.53	2,378.86
8	2.08	1.73	28.17	851.12
9	6.15	15.09	76.33	2,876.83

cluster are

$$y_i = \sum_{j=1}^{M} y_{ij} = y_{i1} + y_{i2} + \cdots + y_{iM} \tag{7.1}$$

and

$$\bar{y}_i = \frac{y_i}{M}. \tag{7.2}$$

The totals and means for the above nine clusters of New York State are presented in Table 7.1. The means of the clusters, along with their variances are presented in Table 7.2.

The population **mean per unit** is obtained by summing the N totals in (7.1) and dividing by N; that is,

$$\bar{\bar{Y}} = \frac{1}{N} \sum_{1}^{N} y_i. \tag{7.3}$$

The **mean per element** is obtained by summing the observations of the NM elements and dividing by NM; that is,

$$\bar{\bar{\bar{Y}}} = \frac{1}{M_o} \sum_{1}^{N} \sum_{1}^{M} y_{ij} = \frac{\bar{\bar{Y}}}{M}, \tag{7.4}$$

where $M_o = NM$ is the total number of elements.

For the above illustration, $\bar{Y}$ can denote the average number of establishments per cluster and $\bar{\bar{Y}}$ the average number of the establishments per county. Similarly, $\bar{Y}$ can denote the employment per cluster and $\bar{\bar{Y}}$ the employment per county. For another illustration, if an agricultural region, for example, is divided into N fields with M farms in each, $\bar{Y}$ can denote the production per field and $\bar{\bar{Y}}$ the production per farm. Both types of means are important in practice.

The population variance is

$$S^2 = \frac{\sum_1^N \sum_1^M (y_{ij} - \bar{\bar{Y}})^2}{NM - 1}. \tag{7.5}$$

The variance of the ith cluster is $s_i^2 = \Sigma_1^M (y_{ij} - \bar{y}_i)^2/(M - 1)$. As shown in Appendix A7, (7.5) can be expressed as

$$S^2 = \frac{N(M-1)}{NM-1} S_w^2 + \frac{N-1}{NM-1} S_b^2 \tag{7.6}$$

$$= \frac{(M-1)S_w^2 + S_b^2}{M}, \tag{7.6a}$$

where $S_w^2 = (\Sigma_1^N s_i^2)/N$ and $S_b^2 = M\Sigma(\bar{y}_i - \bar{\bar{Y}})^2/(N-1)$ are the **within MS** (mean square) and **between MS** with $N(M-1)$ and $(N-1)$ d.f., respectively. The expression in (7.6a) is an approximation of (7.6) for large N.

For the 54 counties of Table 7.1, the average $\bar{\bar{Y}}$ and variance S^2 for the number of employed persons are 44.34 and 5057.79. Corresponding figures for the number of establishments are 3.02 and 15.94, respectively. The correlation between the number of establishments and the number of persons employed is 0.9731.

7.3 Estimation of the means

The population means per unit and per element can be estimated by selecting a sample of size n from the N clusters, and recording the M observations in each of the selected clusters. The sample means on the unit and element basis are

$$\bar{y} = \frac{1}{n} \sum_1^n y_i \tag{7.7}$$

and

$$\bar{\bar{y}} = \frac{1}{nM}\sum_{1}^{n}\sum_{1}^{M} y_{ij} = \frac{\bar{y}}{M}. \tag{7.8}$$

These means are unbiased for $\bar{Y}$ and $\bar{\bar{Y}}$, respectively. Their variances are

$$V(\bar{y}) = \frac{(1-f)}{n}\frac{\sum_{1}^{N}(y_i - \bar{Y})^2}{N-1} \tag{7.9}$$

and

$$\begin{aligned} V(\bar{\bar{y}}) &= \frac{1}{M^2}V(\bar{y}) \\ &= \frac{(1-f)}{nM^2}\frac{\sum(y_i - \bar{Y})^2}{N-1} = \frac{(1-f)}{n}\frac{\sum(\bar{y}_i - \bar{\bar{Y}})^2}{N-1}. \end{aligned} \tag{7.10}$$

From the definition of S_b^2 and from (7.9), this variance can also be expressed as

$$V(\bar{\bar{y}}) = \frac{(1-f)S_b^2}{nM} \tag{7.11}$$

Observe from (7.10) that the variances of $\bar{y}$ and $\bar{\bar{y}}$ will be small if the cluster totals or means do not vary much.

Example 7.1. Estimating employment: Table 7.1 contains the data for private nonfarm establishments. This type of establishment includes manufacturing, retail trade, finance, insurance, and real estate firms and services. One of the counties with only 200 establishments is combined with a neighboring county. New York City alone had 103 thousand establishments with 1988 thousand persons employed, and the adjacent six counties had 177 thousand establishments with a total of 2122 thousand employed persons. Figures for these seven counties are not included in Table 7.1.

If one selects a random sample, for example, of size $n = 3$ from the nine clusters of Table 7.1, any one of the ${}_9C_3 = 84$ combinations of the clusters may be selected.

For the employment figures, from Table 7.1, $\Sigma_1^N(\bar{y}_i - \bar{\bar{Y}})^2/(N-1)$ equals 620.8. If $n = 3$, from (7.10), $V(\bar{\bar{y}}) = 137.96$. Similarly, for the establishments, $\Sigma_1^N(\bar{y}_i - \bar{\bar{Y}})^2/(N-1)$ equals 2.63 and $V(\bar{\bar{y}}) = 0.58$ if $n = 3$.

7.4 Comparison with simple random sampling

As mentioned in Section 7.1, a simple random sample of size nM drawn without replacement from the NM elements is spread more evenly in the population than a sample of n clusters. The mean of these nM elements is unbiased for $\bar{\bar{Y}}$ and has variance:

$$V(\bar{\bar{y}}_{\text{srs}}) = \frac{(1-f)}{nM}S^2. \tag{7.12}$$

Before preferring cluster sampling over simple random sampling, their relative precisions may be compared. By substituting (7.6) in (7.12) and dividing by (7.11), precision of cluster sampling relative to simple random sampling is given by

$$\frac{V(\bar{\bar{y}}_{\text{srs}})}{V(\bar{\bar{y}})} = \frac{S^2}{S_b^2} = \frac{N(M-1)}{(NM-1)F} + \frac{N-1}{NM-1}, \tag{7.13}$$

where $F = S_b^2/S_w^2$ approximately follows the F-distribution with $(N-1)$ and $N(M-1)$ d.f., provided N and M are large. Percentiles of this distribution are available in the standard statistical tables. For large N and M, the expression in (7.13) can also be expressed as $[(M-1)/F + 1]/M$.

Equation 7.13 demonstrates that cluster sampling has higher precision than simple random sampling if S_w^2 is large and S_b^2 is small. This requirement is just the opposite of what is needed for high precision with stratification. Note also that the sample size has no effect on the relative precision of these two procedures of sampling. Following Kish (1965), the ratio $V(\bar{\bar{y}})/V(\bar{\bar{y}}_{\text{srs}})$ is known as the design effect, **deff**. The magnitude of this ratio enables one to examine the different factors affecting $V(\bar{\bar{y}})$.

As has been seen, in the case of stratification, samples are selected from each of the strata. For cluster sampling, a sample of clusters is selected and observations on all the elements of the selected clusters are recorded.

Example 7.2. Gains relative to simple random sampling: Since S^2 for the employment figures in Example 7.1 is 5057.79, the variance of $\bar{y}_{srs}$ from (7.12) for a sample of $nM = 18$ counties from the $NM = 54$ counties is $[(9 - 3)/(9 \times 3)](5057.79) = 187.33$. This variance is larger than the variance of 137.96 for $\bar{y}$ found in Example 7.1. The gain in precision for cluster sampling is $(187.33 - 137.96)/137.6 = 0.36$, that is, 36%.

The estimator for S_b^2 is presented in Section 7.5 and for S^2 in Appendix A7. The relative precision in (7.13) can be found from these estimators.

7.5 Estimation of the standard error

From a sample of n clusters, unbiased estimators of the variances in (7.9) and (7.10) are

$$v(\bar{y}) = \frac{(1-f)}{n} \frac{\sum^n (y_i - \bar{y})^2}{n-1} \tag{7.14}$$

and

$$v(\bar{\bar{y}}) = \frac{(1-f)}{n} \frac{\sum^n (\bar{y}_i - \bar{\bar{y}})^2}{n-1} = \frac{(1-f)}{n} \frac{s_b^2}{M} \tag{7.15}$$

where $s_b^2 = [M\Sigma^n(\bar{y}_i - \bar{\bar{y}})^2]/(n - 1)$ is the sample between MS with $(n - 1)$ d.f.

Example 7.3. Standard error for estimating the employment: If clusters 1, 7, and 9 appear in a sample of three from the nine clusters of Table 7.1, the means for the employments are 74.15, 38.53, and 76.33. Hence, an estimate for the average employment per county is $\bar{\bar{y}} = (74.15 + 38.53 + 76.33)/3 = 63$ (thousands). Since $\Sigma_1^n(\bar{y}_i - \bar{\bar{y}})^2/(n - 1)$ is 450.4, $v(\bar{\bar{y}}) = [(9 - 3)/(9 \times 3)](450.4) = 100.1$. Note that $s_b^2 = 6(450.4) = 2702.4$, and the same value is obtained from the right-hand side expression of (7.15).

The actual mean is $\bar{\bar{Y}} = 44.34$ (thousands). As seen in Example 7.1, the actual variance of $\bar{\bar{y}}$ equals 137.96, and $S_b^2 = 6(620.8) = 3724.8$.

7.6 Optimum cluster and sample sizes

As seen in Sections 7.3 and 7.4, larger values of S_w^2 and smaller values of S_b^2 result in higher precision for cluster sampling. This result can be used for dividing a population into clusters of suitable sizes. For several agricultural surveys, S_w^2 was found to be proportional

to M^h, where h is usually between zero and 2. This observation may also be valid for certain types of demographic and economic surveys. Such information can be used to find the optimum sizes for M and n, for example, by prescribing an upper limit to the variance in (7.11).

Since a sample of n clusters are selected and observations are made on the M elements in each of the selected clusters, the cost of obtaining information from the sample can be a multiple of Mn. In some studies, travel costs are also considered; see Cochran (1977; Chap. 9), for example. The optimum values of M and n are obtained by minimizing $V(\bar{y})$ for the available budget, or minimizing the cost for a required precision. The above type of information regarding S^2 and S_w^2 can be used for this purpose. For telephone and mail surveys, travel costs may not be applicable, and the relevant cost functions should be chosen.

7.7 Clusters of unequal size

Most of the natural clusters of the type described in Section 7.1 usually tend to be of unequal sizes. When the population consists of N clusters with M_i elements in the ith cluster, let y_{ij}, $i = 1,\ldots,N$ and $j = 1,\ldots, M_i$, denote the observations. The cluster totals are $y_i = \Sigma_1^{M_i} y_{ij}$. As before, the means per unit and per element, respectively, are

$$\bar{Y} = \frac{\sum_1^N y_i}{N} \qquad \text{and} \qquad \bar{\bar{Y}} = \frac{\sum_1^N y_i}{M_o} = \frac{\bar{Y}}{\bar{M}}, \tag{7.16}$$

where $M_o = \Sigma_1^N M_i$ is the total number of elements and $\bar{M} = M_o/N$ is the average number of elements per cluster.

One type of grouping of the neighboring counties of New York State, presented in Table 7.1, resulted in nine clusters of unequal size. Their sizes M_i, and the totals x_i and y_i for the number of establishments and the number of employed persons along with their standard deviations and correlations are presented in Table 7.3.

For a sample of n clusters, the estimator $\bar{y}$ for $\bar{Y}$ and its variance $V(\bar{y})$ have the same forms as (7.7) and (7.9), respectively. The unbiased estimator for $V(\bar{y})$ takes the same form as (7.14).

An unbiased estimator of $\bar{\bar{Y}}$ is given by $\hat{\bar{\bar{Y}}} = \bar{y}/\bar{M}$, which has the variance of $V(\bar{y})/\bar{M}^2$.

Table 7.3. Unequal size clusters: totals and variances.

Cluster i	Cluster Size M_i	Number of Establishments		Number Employed	
		Total	Variance	Total	Variance
1	10	49.8	52.04	860.1	18,398.01
2	8	20.5	12.52	322.4	4,427.36
3	3	4.1	0.85	67.0	286.14
4	4	7.1	3.32	124.1	1,281.49
5	4	6.5	0.27	70.8	47.81
6	5	6.4	0.132	70.6	20.66
7	5	9.1	3.45	118.5	837.23
8	7	19.5	6.27	273.2	1,968.01
9	8	39.9	15.46	487.4	2,871.45
Total	54	162.9		2,394.1	
Mean	6	18.1		266.01	
S.D.	2.35	16.42		265.28	

Correlations of employment totals with establishment totals and cluster sizes are 0.98 and 0.92, respectively. Correlation of establishment totals with cluster sizes is 0.92.

Example 7.4. Average employment: Consider the sample size $n = 3$ for the nine clusters in Table 7.3. In this case, from (7.9), $V(\bar{y}) = (9 - 3)(265.28)^2/(9 \times 3) = 15{,}638.55$. Since $\overline{M} = 6$, $V(\hat{\bar{\bar{Y}}}) = 15{,}653.55/36 = 434.4$.

As in the case of equal cluster sizes, the above estimators will have small variances if the cluster totals y_i do not vary much.

7.8 Alternative estimation with unequal sizes

If the sizes are unequal, $\bar{\bar{Y}}$ can also be estimated from

$$\hat{\bar{\bar{Y}}}_R = \frac{\sum_1^n y_i}{\sum_1^n M_i} = \frac{\bar{y}}{\overline{m}}, \tag{7.17}$$

where $\overline{m} = \Sigma_1^n M_i/n$ and the subscript R denotes the ratio.

Although $\bar{y}$ and $\overline{m}$ are unbiased for $\bar{Y}$ and $\overline{M}$, the ratio estimator in (7.17) is not unbiased for $\bar{\bar{Y}} = \bar{Y}/\overline{M}$. For large n, as shown in Appendix A7, the bias of this estimator becomes small and its variance

is approximately given by

$$V(\hat{\bar{Y}}_R) = \frac{(1-f)}{n\bar{M}^2}(S_y^2 + \bar{\bar{Y}}^2 S_m^2 - 2\bar{\bar{Y}}\rho S_y S_m). \tag{7.18}$$

In this expression, S_y^2 and S_m^2 are the variances of (y_i, M_i) of the N clusters, and ρ is the correlation of y_i and M_i. Note that the covariance of y_i with M_i is $S_{ym} = \rho S_y S_m$.

As noted at the end of Section 7.7, for simple random sampling, the variance of $\hat{\bar{\bar{Y}}} = \bar{y}/\bar{M}$ is given by $V(\bar{y})/\bar{M}^2$. The variance of $\hat{\bar{\bar{Y}}}_R$ in (7.18) will be smaller than that of $\hat{\bar{\bar{Y}}}$ if $\rho \geq C_m/2C_y$, where C_y and C_m are the coefficients of variations of y_i and M_i, respectively.

Example 7.5. Ratio estimation adjusting for the sizes: From Table 7.3, the ratio of the totals $\Sigma_1^N y_i$ to $\Sigma_1^N M_i$ is (2394.1/54) = 44.34, which is the same as $\bar{\bar{Y}} = 266.01/6$. For a sample of size three from the nine clusters, the variance of $\hat{\bar{\bar{Y}}}_R$ in (7.18) is

$$V(\hat{\bar{\bar{Y}}}_R) = \frac{9-3}{9(3)(36)}[(265.28)^2 + (44.34)^2(2.35)^2 - 2(44.34)(0.92)(265.28)(2.35)] = 187.47.$$

This variance is only about two fifths of the variance of 434.4 found for the sample mean in Example 7.4.

As can be seen from the comparison of $V(\hat{\bar{\bar{Y}}}_R)$ in (7.18) with $V(\hat{\bar{\bar{Y}}}) = V(\bar{y})/\bar{M}^2$, high positive correlation between the employment and the cluster size has considerably reduced the variance of $\hat{\bar{\bar{Y}}}_R$ relative to that of $\hat{\bar{\bar{Y}}}$.

An estimator for (7.18) is given by

$$V(\hat{\bar{\bar{Y}}}_R) = \frac{(1-f)}{n\bar{M}^2}(S_y^2 + \hat{\bar{\bar{Y}}}_R^2 s_m^2 - 2\hat{\bar{\bar{Y}}}_R s_{ym}), \tag{7.19}$$

where s_{ym} is the sample covariance of y_i and M_i.

For $\bar{\bar{Y}}$, as an alternative to the sample mean $\bar{y}$, one may consider the *ratio estimator:*

$$\hat{\bar{\bar{Y}}}_R = \frac{\sum_1^n y_i}{\sum_1^n M_i}\bar{M}. \tag{7.20}$$

Following the procedure in Appendix A7, the bias of this estimator becomes small as n increases. The approximate variance of (7.20) and

its estimator are given by

$$V(\hat{\bar{Y}}_R) = \frac{(1-f)}{n}(S_y^2 + \bar{\bar{Y}}^2 S_m^2 - 2\bar{\bar{Y}}\rho S_y S_m) \tag{7.21}$$

and

$$v(\hat{\bar{Y}}_R) = \frac{(1-f)}{n}(s_y^2 + \hat{\bar{\bar{Y}}}_R^2 s_m^2 - 2\hat{\bar{\bar{Y}}}_R s_{ym}). \tag{7.22}$$

Note that $V(\hat{\bar{Y}}_R) = \bar{M}^2 V(\hat{\bar{\bar{Y}}}_R)$. Thus, for the above example, $V(\hat{\bar{Y}}_R) = 36(187.47) = 6748.92$. This variance is only about two fifths of the variance of 15,638.55 for $\bar{y}$ found in Example 7.4.

7.9 Proportions and percentages

When a population is divided into clusters, estimation of the total numbers and proportions or percentages related to qualitative characteristics also is usually of interest.

Let c_i and $p_i = (c_i/M_i)$, $i = 1, 2, ..., N$, denote the number and proportion of the M_i elements of the ith cluster having an attribute of interest.

For each of the nine clusters described in Sections 7.7 and 7.8, the numbers c_i and proportions p_i of counties that have more than 1000 establishments are presented in columns 3 and 4 of Table 7.4. The numbers and proportions of the counties that have more than 10,000 employees are presented in columns 5 and 6.

Table 7.4. Composition of the 54 counties.

Cluster	No. of Counties	No. and Proportion of Counties with More Than			
		1000 establishments		10,000 employees	
i	M_i	c_i	p_i	c_i	p_i
1	10	7	0.70	8	0.80
2	8	6	0.75	6	0.75
3	3	2	0.67	2	0.67
4	4	2	0.50	4	1.00
5	4	3	0.75	4	1.00
6	5	5	1.00	5	1.00
7	5	4	0.80	4	0.80
8	7	6	0.86	5	0.71
9	8	8	1.00	8	1.00
Means	6	4.78	0.78	5.11	0.86

Average of the proportions

The average of the N proportions is

$$\bar{P} = \frac{1}{N}\sum_{1}^{N} p_i . \tag{7.23}$$

For either of the two characteristics in Table 7.4, $\bar{P}$ is the average of the nine proportions. For a sample of n clusters, an unbiased estimator of $\bar{P}$ is

$$\hat{\bar{P}} = \frac{1}{n}\sum_{1}^{n} p_i . \tag{7.24}$$

Its variance and estimator of variance are obtained from $V(\bar{y})$ and $v(\bar{y})$ in (2.12) and (2.14) by replacing y_i with p_i.

Example 7.6. Proportions of the counties for the establishments: As shown in Table 7.4, the average of the proportions of counties with more than 1000 establishments is 0.78. If this average is not known, one may estimate it from a random sample from the nine clusters. The variance of the nine proportions is 0.0227. For a sample of three clusters, following (2.12), $V(\hat{\bar{P}}) = (9 - 3)(0.0227)/(9 \times 3) = 0.005$ and S.E.$(\hat{\bar{P}}) = 0.07$.

Proportion of the totals or means

A total of $\Sigma_1^N c_i$ of the $M_o = \Sigma_1^N M_i$ elements have the attribute of interest. Their proportion is

$$P = \frac{\sum_{1}^{N} c_i}{M_o} = \frac{\sum_{1}^{N} M_i p_i}{M_o} = \frac{\left(\sum_{1}^{N} c_i\right)/N}{\bar{M}} . \tag{7.25}$$

In contrast to $\bar{P}$ in (7.23), this is the overall proportion of all the M_0 elements having the attribute of interest. If all the M_i are equal, it coincides with $\bar{P}$.

From a sample of n clusters, an unbiased estimator of P is

$$p = \frac{\left(\sum_{1}^{n} c_i\right)/n}{\bar{M}} = \frac{\bar{c}}{\bar{M}} \tag{7.26}$$

and its variance is given by $V(\bar{c})/\overline{M}^2$. Note that $V(\bar{c})$ and its estimate can be obtained from (2.12) and (2.14) by replacing y_i with c_i.

Another estimator for P is

$$\hat{P} = \frac{\sum_1^n c_i}{\sum_1^n M_i} = \frac{\sum_1^n M_i p_i}{\sum_1^n M_i}, \tag{7.27}$$

which is of the ratio type. Its large sample variance is given by

$$V(\hat{P}) = \frac{(1-f)}{n\overline{M}^2} \frac{\sum_1^N (c_i - PM_i)^2}{N-1}. \tag{7.28}$$

An estimator of this variance is obtained from

$$v(\hat{P}) = \frac{(1-f)}{n\overline{M}^2} \frac{\sum_1^N (c_i - \hat{P}M_i)^2}{n-1}. \tag{7.29}$$

If c_i is highly correlated with M_i, $\hat{P}$ will have smaller variance than p. If $\overline{M}$ is not known, for the variance in (7.28), replace it with the sample mean $\Sigma M_i/n$ of the sizes.

Example 7.7. Establishments and the gain in precision for the ratio estimation: For the number of counties of Table 7.4 that have more than 1000 establishments, $P = (4.78/6) = 0.7967$. The variance of c_i is 4.67 and for a sample of three clusters $V(\bar{c}) = (9-3)(4.67)/(9 \times 3) = 1.0378$. To estimate this proportion with a sample of three clusters, $V(p) = 1.0378/36 = 0.0288$ and hence S.E.$(p) = 0.17$. For the alternative estimator in (7.27), from (7.28), $V(\hat{P}) = 0.005$ and S.E.$(\hat{P}) = 0.07$. The relative gain in precision for the ratio estimation is $(0.029 - 0.005)/0.005 = 4.8$, that is, 480%, which is substantial.

7.10 Stratification

In several large-scale surveys, the population is first divided into strata, with each stratum consisting of a certain number of clusters. For example, the nine clusters of Table 7.3 can be divided into three strata containing the clusters (1, 2, 8, 9), (4, 7), and (3, 5, 6).

With $\hat{Y}_g$ denoting the estimate for the total of the gth stratum obtained through a sample, an estimator for the population total is $\hat{Y} = \Sigma \hat{Y}_g$.

The variance of $\hat{Y}$ and its estimate are given by $\Sigma V(\hat{Y}_g)$ and $\Sigma v(\hat{Y}_g)$. The following example examines the total and average of the employment.

Example 7.8. Total and average of the employment: For the above type of stratification, from Table 7.3, the standard deviations of the employment figures in the three strata equal 265.83, 3.96, and 2.14. If a sample of one cluster is drawn randomly from each of these strata, an unbiased estimator of the total takes the form $\hat{Y} = \Sigma_1^3 N_g y_g$. The variance of this estimator is $V(\hat{Y}) = \Sigma N_g(N_g - 1)S_g^2$.

An unbiased estimator of $\bar{\bar{Y}}$ is $\hat{\bar{\bar{Y}}} = \hat{Y}/M_o$, which has the variance $V(\hat{\bar{\bar{Y}}}) = \Sigma N_g(N_g - 1)S_g^2/M_0^2 = [4(3)(265.83)^2 + 2(1)(3.96)^2 + 3(2)(2.14)^2]/(54)^2 = 290.32$. This variance as expected is smaller than the variance of 434.4 for the sample mean found in Example 7.4.

7.11 Unequal probability selection

Chapters 2 to 6 and the preceding sections of this chapter have considered selection of the units through simple random sampling, that is, with equal probabilities and without replacement. In some applications, the units are selected with unequal probabilities, and unbiased estimators for the population totals and means are obtained. If these probabilities are related to the characteristics of interest, the resulting estimators can have smaller variances than those obtained through simple random sampling.

In cluster sampling, the probabilities for selecting the units can depend on their sizes, the number of elements (M_i). They can also be based on one or more concomitant or supplementary variables highly correlated with the characteristics of interest.

A method of selecting the cluster units

For a simple method of selection with unequal probabilities, consider the nine cluster units in Table 7.3. Their cumulative sizes are 10, 18, 21, 25, 29, 34, 39, 46, and 54. The ranges associated with these figures are 1 to 10, 11 to 18, 19 to 21,…,47 to 54. One number is selected randomly between 1 and 54. If this number is 16, for example, the second cluster is selected into the sample. Since the above ranges correspond to the sizes of the clusters, the ith cluster is selected with probability $u_i = (M_i/M_o)$. If all the M_i are the same, $u_i = (1/N)$ and this procedure is the same as simple random sampling.

Continuing with the above procedure, if the ith unit appears at the first selection, it is removed from the list and the second unit is drawn

with probability $M_j/(M_0 - M_i) = u_j/(1 - u_i)$. The next section examines the probabilities for selecting the units into a sample of size $n = 2$ using the above procedure.

Probabilities of selecting the units

The probabilities of selecting the jth unit after the ith unit and the ith unit after the jth unit, respectively, are

$$P(j|i) = \frac{u_j}{1 - u_i} \quad \text{and} \quad P(i|j) = \frac{u_i}{1 - u_j}. \tag{7.30}$$

The probability of selecting the ith unit into the sample, at the first or the second draw, is

$$\phi_i = u_i + \sum_{j \neq i}^{N} u_j \frac{u_i}{1 - u_j} = u_i\left(1 + \mathrm{T} - \frac{u_i}{1 - u_i}\right) \tag{7.31}$$

where $T = \Sigma_1^N [u_i/(1 - u_i)]$.

The probability that the ith and jth units are drawn into the sample is

$$\phi_{ij} = u_i \frac{u_j}{1 - u_i} + u_j \frac{u_i}{1 - u_j} = \frac{u_i u_j (2 - u_i - u_j)}{(1 - u_i)(1 - u_j)}. \tag{7.32}$$

The above procedure can be continued for selecting more than two units into the sample. If the ith and the jth units appear at the first two draws, they are removed from the population and the third unit is drawn with probability $u_k/(1 - u_i - u_j)$. Continuation of this selection procedure provides a sample of n units.

The expression in (7.31) is the probability that the ith unit appears in the sample—the *inclusion probability*. As can be seen, ϕ_i is not exactly proportional to u_i. For the sake of illustration, consider four units with sizes $M_i = (15, 12, 10, 3)$ and employments $y_i = (105, 80, 45, 10)$ in thousands. The relative sizes of these units are $u_i = (0.375, 0.30, 0.25, 0.075)$. If two units are selected into the sample as above with these probabilities, from (7.31), $\phi_i = (0.6911, 0.6042, 0.5275, 0.1772)$. These inclusion probabilities are in the relative ratios (0.346, 0.302, 0.264, 0.089), which differ slightly from the u_i.

If a sample of n units is selected randomly without replacement from the N population units, as seen in Chapter 2, the inclusion

probability for the ith unit is $\phi_i = n/N = n(1/N)$, which is the same for all N units. Similarly, when $n = 2$ units are selected with unequal probabilities and without replacement, ϕ_i in (7.31) should equal $2u_i = (0.75, 0.6, 0.5, 0.15)$. The above inclusion probabilities $\phi_i = (0.6911, 0.6042, 0.5275, 0.1772)$ are close to these figures but not identical.

To find the initial probabilities p_i to make ϕ_i equal $2u_i$, Narain (1951) and Yates and Grundy (1953) replace u_i on the right-hand side of (7.31) with p_i and solve for the ϕ_i. They also establish the following general results for the selection of a sample of size n with unequal probabilities and *without* replacement.

$$\sum_{i}^{N} \phi_i = n, \tag{7.33a}$$

$$\sum_{j \neq i}^{N} \phi_{ij} = (n-1)\phi_i, \qquad \sum_{i}^{N} \sum_{j \neq i}^{N} \phi_{ij} = n(n-1), \tag{7.33b}$$

and

$$\sum_{j \neq i}^{N} (\phi_i \phi_j - \phi_{ij}) = \phi_i(1 - \phi_i). \tag{7.33c}$$

Note from (7.33a) that $\Sigma_{j \neq i}^{N} \phi_j = n - \phi_i$. With this result, (7.33c) follows from (7.33b).

Brewer (1963), Durbin (1967), and others have suggested procedures for selecting a sample of two units with probabilities proportional to u_i without replacement. Samford (1967) extends Brewer's procedure to the selection of more than two units.

7.12 Horvitz-Thompson estimator

For the estimation of the population total Y by selecting units with probabilities ϕ_i and without replacement, Horvitz and Thompson (1952) consider

$$\hat{Y}_{\mathrm{HT}} = \sum_{1}^{n} \frac{y_i}{\phi_i}. \tag{7.34}$$

As shown in Appendix A7, this estimator is unbiased for Y.

Variance of the estimator

Yates and Grundy (1953) and Sen (1953) derive the variance of $\hat{Y}_{\mathrm{HT}}$ and the two estimators for this variance presented below. As described in Appendix A7,

$$\begin{aligned} V(\hat{Y}_{\mathrm{HT}}) &= E(\hat{Y}_{\mathrm{HT}}^2) - Y^2 \\ &= \sum_i^N \frac{1-\phi_i}{\phi_i} y_i^2 + \sum_{i\neq j}^N \sum^N \frac{\phi_{ij} - \phi_i\phi_j}{\phi_i\phi_j} y_i y_j. \end{aligned} \tag{7.35}$$

By substituting (7.33c) in the first term of (7.35), the variance can also be expressed as

$$\begin{aligned} V(\hat{Y}_{\mathrm{HT}}) &= \sum_i^N \sum_{j\neq i}^N \frac{\phi_i\phi_j - \phi_{ij}}{\phi_i^2} y_i^2 - \sum_{i\neq j}^N \sum^N \frac{\phi_i\phi_j - \phi_{ij}}{\phi_{ij}} y_i y_j \\ &= \sum_i^N \sum_{i<j}^N (\phi_i\phi_j - \phi_{ij}) \left(\frac{y_i}{\phi_i} - \frac{y_j}{\phi_j} \right)^2. \end{aligned} \tag{7.36}$$

For the illustration in Section 7.11 with the four units, the joint probabilities ϕ_{ij} and $(y_i/\phi_i) - (y_j/\phi_j)$ are presented in Table 7.5. From these figures, from (7.36), $V(\hat{Y}_{\mathrm{HT}}) = 445.67$. For these units, $S_y^2 = 1716.67$, and for a simple random sample of $n = 2$ units, $V(\hat{Y}) = 4(4-2)(1716.67)/2 = 6866.67$. Thus, $V(\hat{Y}_{\mathrm{HT}})$ is only 6.5% of $V(\hat{Y})$.

As can be seen from the expressions in (7.34) or (7.36), the variance of the Horvitz–Thompson estimator vanishes when ϕ_i are proportional to y_i. Thus, selecting the population units with probabilities u_i and

Table 7.5. Probabilities of selection.

Clusters (i, j)	ϕ_{ij}	$(y_i/\phi_i) - (y_j/\phi_j)$
1, 2	0.3407	19.53
1, 3	0.1547	66.62
1, 4	0.0754	39.57
2, 3	0.2071	47.10
2, 4	0.0565	20.05
3, 4	0.0453	−27.05

making ϕ_i equal nu_i results in a small variance for $\hat{Y}_{HT}$, provided y_i are proportional to u_i. For cluster sampling, as noted earlier, the u_i can be based on the sizes M_i or the values of a supplementary variable.

The estimator $\hat{Y}_{HT}$ is of a general type obtained by sampling the population units with specified probabilities and without replacement. As seen in Chapter 2, for simple random sampling, all the ϕ_i are equal to (n/N) and $\hat{Y} = N\bar{y}$.

Variance estimators

From (7.35), an unbiased estimator for $V(\hat{Y}_{HT})$ is

$$v_1(\hat{Y}_{HT}) = \sum_i^n \frac{1-\phi_i}{\phi_i^2} y_i^2 + \sum_{i \neq j}^n \sum^n \frac{\phi_{ij} - \phi_i \phi_j}{\phi_i \phi_j \phi_{ij}} y_i y_j. \tag{7.37}$$

As suggested by Yates and Grundy (1953) and Sen (1953), from (7.36), an alternative unbiased estimator is given by

$$v_2(\hat{Y}_{HT}) = \sum_i^n \sum_{i<j}^n \frac{\phi_i \phi_j - \phi_{ij}}{\phi_{ij}} \left(\frac{y_i}{\phi_i} - \frac{y_j}{\phi_j} \right)^2. \tag{7.38}$$

Both these estimators may take negative values.

Estimators for the mean per unit and the population mean

The estimator for the mean per unit $\bar{Y}$ in (7.3) is given by $\hat{\bar{Y}}_{HT} = \hat{Y}_{HT}/N$. Its variance is given by dividing (7.35) or (7.36) by N^2. Similarly, its variance estimator is obtained by dividing (7.37) or (7.38) by N^2. For the above illustration, $V(\hat{\bar{Y}}_{HT}) = 445.67/16 = 27.85$. For a simple random sample of two units, $V(\bar{y}) = 6866.67/16 = 429.17$.

The estimator for the population mean $\bar{\bar{Y}}$ in (7.4) is given by $\hat{\bar{\bar{Y}}}_{HT} = \hat{\bar{\bar{Y}}}_{HT}/M_0$. Its variance is obtained by dividing (7.36) by M_0^2. Similarly, its estimator of variance is obtained by dividing (7.38) by M_0^2. Since $M_0 = 40$ for the above illustration, $V(\hat{\bar{\bar{Y}}}_{HT}) = 445.67/1600 = 0.2785$. For a simple random sample, $\hat{Y}/M_0$ is unbiased for $\bar{\bar{Y}}$, and its variance is given by $V(\hat{Y})/M_0^2$. For a sample of size two, the variance of this estimator is $6866.67/1600 = 4.29$.

7.13 Alternative approaches

Procedures for selecting two or more units with unequal probabilities and without replacement were suggested by Murthy (1957, 1967), J.N.K. Rao (1965), and others. Cochran (1977, pp. 258–270) describes these approaches in detail. Bayless and J.N.K. Rao (1970) empirically compare some of these procedures. Fellegi (1963) considers selection probabilities for rotation sampling. Cochran (1942) describes some of the estimation procedures with cluster units of unequal size. Brewer and Hanif (1983) summarize several methods for selecting units with unequal probabilities.

The following approaches can also be considered to obtain unbiased estimators for the population mean or total, and under suitable conditions they can result in small variances for the estimators.

Probabilities proportional to sum of sizes (PPSS)

Lahiri (1951) suggests the following method for drawing the sample with probability proportional to $\Sigma_1^n M_i$. Let T denote the total of the n largest values of M_i in the population. Draw a simple random sample of n units without replacement, and a random number r between 1 and T. If $\Sigma_1^n M_i \geq r$, retain this sample; otherwise, repeat the procedure. Since the sum of $\Sigma_1^n M_i$ over all the possible ${}_N C_n$ samples equals $(N-1)^C (n-1)M_0$, the probability of drawing a specified sample is

$$P_k = \frac{\sum_1^n M_i}{\binom{N-1}{n-1} M_0} = \frac{\overline{m}}{\binom{N}{n}\overline{M}}, \tag{7.39}$$

where k runs from 1 to ${}_N C_n$.

The estimator for $\overline{Y}$ considered with this procedure is $\hat{\overline{Y}}_{SS} = \overline{M}(\bar{y}/\overline{m})$. As shown in Appendix A7, this estimator is unbiased for $\overline{Y}$ and for large n has the same variance as (7.21) for $\hat{\overline{Y}}_R$. Thus, in addition to being unbiased, $\hat{\overline{Y}}_{SS}$ has smaller variance than $\bar{y}$ provided y_i have a high positive correlation with the cluster sizes M_i. An estimator for the exact variance of $\hat{\overline{Y}}_{SS}$ was obtained by Des Raj (1968), but it may take negative values.

In an alternative procedure suggested by Midzuno (1951), the first unit of the sample is selected with probability proportional to M_i and the remaining $(n-1)$ units are drawn from the $(N-1)$ units randomly without replacement. This approach also results in selecting the sample with probability proportional to $\Sigma_1^n M_i$.

Randomization approach

For the procedure suggested by J.N.K. Rao et al. (1962), the relative size of a unit or a measure of the size is denoted by z_i. In this approach, the population is first divided randomly into n groups of sizes N_g, $g = (1,\ldots,n), N = \Sigma_1^n N_g$. Denoting the total of the gth group by Y_g, the population total becomes $Y = \Sigma_1^g Y_g$. The relative sizes of the population units are denoted by z_i, $i = (1, 2,\ldots,N)$, and the total of the N_g relative sizes of the gth group by Z_g. Now, the probabilities $p_i = z_i/Z_g$ are assigned to the N_g units of the gth group and one unit is selected from each of the n groups with these probabilities.

Let y_g and z_g denote the observation of the sampled unit and its original probability. For a given grouping of the units, let $p_g = z_g/Z_g$. Now, $\hat{Y}_g = (y_g/p_g)$ is unbiased for the total Y_g of the gth group. Hence, $\hat{Y}_{\text{HRC}} = \Sigma_1^n \hat{Y}_g$ is unbiased for the population total Y. The variance of this estimator can be obtained from the approach in Appendix A3. The above authors and Cochran (1977, p. 267), derive this variance and its estimator. When $N_g = N/n$ is an integer, they are given by

$$V(\hat{Y}_{\text{RHC}}) = \frac{N-n}{(N-1)n} \sum_1^N z_i \left(\frac{y_i}{z_i} - Y\right)^2 \tag{7.40}$$

and

$$v(\hat{Y}_{\text{RHC}}) = \frac{N-n}{N(n-1)} \sum_1^n Z_g \left(\frac{y_g}{z_g} - \hat{Y}_{\text{RHC}}\right)^2. \tag{7.41}$$

As an illustration, for the four units described in Sections 7.11 and 7.12, z_i are the same as the relative sizes (0.375, 0.30, 0.25, 0.075), and y_i = (105, 80, 45, 10). A random division of the units into two groups of equal sizes may result, for example, in the groups with units (1, 4) and (2, 3). The z_i for the first group are (0.375, 0.075) and their sum is $Z_1 = 0.45$. Hence, the probabilities of selection for the two units of this group are 0.375/0.45 = 0.83 and 0.075/0.45 = 0.17. One unit is chosen from this group with these probabilities. Similarly, the probabilities of selection for the units of the second group are 0.30/0.55 = 0.545 and 0.25/0.55 = 0.455, and one unit is chosen from this group with these probabilites. In general, since $z_i = M_i/M_0$ for the cluster units, the variance in (7.40) will be small if the correlation between y_i and M_i is positively large.

The estimator for $\bar{Y}$ is obtained from dividing $\hat{Y}_{\text{RHC}}$ by N. The variance of the resulting estimator and its variance estimator are obtained by dividing (7.40) and (7.41) by N^2.

Comparisons of $\hat{Y}_{\text{RHC}}$ with the ratio estimator in Section 7.8 and the PPSS estimator through a suitable model are presented in Chapter 9.

Selection with probabilities proportional to size and replacement (pps)

Following the procedure in Section 7.11 and replacing a unit that has appeared in the sample, one can select a sample of n units with probabilities proportional to the relative sizes u_i. Denoting y_i/u_i by r_i, an estimator for Y is $\hat{Y}_{\text{pps}} = \Sigma_1^n r_i/n$. Since the expectation of r_i is $\Sigma_1^N u_i(y_i/u_i) = Y$, $E(\hat{Y}_{\text{pps}}) = Y$. Thus, this estimator is unbiased for Y.

The variance of $\hat{Y}_{\text{pps}}$ and the estimator of variance are derived in Appendix A7. The variance is given by $V(\hat{Y}_{\text{pps}}) = s_r^2/n$, where $s_r^2 = \sum_1^N u_i\,(r_i - Y)^2$. Thus, the variance of this unbiased estimator also becomes small as y_i is highly correlated with u_i. The estimator of this variance is given by $v(\hat{Y}_{\text{pps}}) = s_r^2/n$, where $s_r^2 = \Sigma_1^n(r_i - \hat{Y}_{\text{pps}})^2/(n-1)$.

Although the procedure is easy to implement, selecting the units with replacement is not practical, since no additional information is obtained by including a unit more than once in the sample. If the sample units, however, are selected with probabilities u_i and replacement, Pathak (1962) shows that the mean $\hat{Y}_d = \Sigma_1^d r_i/d$ of the $d(\leq n)$ distinct sample units obtained by removing the duplications is unbiased for Y and has smaller variance than $\hat{Y}_{\text{pps}}$. As an illustration, when two distinct units appear in a sample of size three selected with replacement, that is, $n = 3$ and $d = 2$, he shows that $r_1 = y_1/u_1$, $r_2 = y_2/u_2$, and $r_{12} = (y_1 + y_2)/(u_1 + u_2)$ are unbiased for Y and have smaller variances than $\hat{Y}_{\text{pps}}$.

Exercises

7.1. For the 54 establishments in Table 7.1, $S^2 = 15.94$. To estimate the average number of establishments per county, compare the precisions of a sample of three clusters and a simple random sample of 18 counties.

7.2. If clusters 2, 3, and 8 are drawn into the sample from the nine clusters of Table 7.1, (a) estimate the average

employment per county and (b) find its standard error. Use the summary figures in Table 7.2.

7.3. With the sample clusters in Example 7.3, estimate the precision of cluster sampling relative to simple random sampling for the estimation of the averages of the (a) employment and (b) number of establishments.

7.4. For estimating the proportion of the 54 counties with more than 10,000 employees, with the data in Table 7.4 and a sample of three clusters, compare the variances of (a) the sample proportion in (7.26) and (b) the ratio type estimator in (7.27).

7.5. (a) From the data in Table 7.4, find the proportion and total number of the 54 counties that have more than 10,000 employees. (b) If clusters 2, 3, and 8 are drawn into the sample from the nine clusters, estimate the above proportion through the sample mean of the proportions and the ratio method, and (c) compare the sample standard errors of the two estimates.

7.6. If clusters 2, 3, and 8 are selected in a sample of three clusters, as in Exercise 7.5, (a) estimate by both approaches the total number of counties that have more than 10,000 employees and (b) find the standard errors of the estimates.

7.7. *Project*. Select all the samples of size three randomly without replacement from the nine clusters of unequal sizes in Table 7.3, and compute the ratio estimate in (7.17) from each sample. (a) From these estimates, find the expectation, bias, and MSE of the ratio estimator, and compare this exact MSE with the approximate expression in (7.18). (b) Compute the variance estimate in (7.19) from each sample, and find the bias of this estimator for the exact variance or MSE of the ratio estimator in (7.17).

7.8. For samples of sizes $n_1 = 2$, $n_2 = 1$, and $n_3 = 2$ selected randomly from the three strata of Section 7.10, find the variance of the estimators for the population total and mean per unit.

7.9. For the stratum with the four clusters 5 through 8 in Table 7.3, the sizes, number of establishments, and employment figures are $M_i = (4, 5, 5, 7)$, $x_i = (6.5, 6.4, 9.1, 19.5)$, and $y_i = (70.8, 70.6, 118.5, 273.2)$. Consider the probabilities $u_i = (0.2, 0.25, 0.25, 0.3)$ for selecting these

units into the sample. For selecting a sample of two units from this stratum as described in Section 7.11, find the probabilities in (7.31) and (7.32).

7.10. Find the variance of the Horvitz–Thomson estimator of $\bar{\bar{Y}}$ for the procedure in Exercise 7.9 and compare it with the estimator obtained from a simple random sample of two units.

7.11. If the second and third units are selected into the sample through the procedure in Exercise 7.9, estimate the S.E. of the Horvitz–Thomson estimators for Y and $\bar{\bar{Y}}$ through both (7.37) and (7.38).

7.12. Following the approach of Des Raj (1968), express $\bar{Y}^2$ as $(\Sigma_i^N y_i^2 + \Sigma_{i\neq j}^N \Sigma^N y_i y_j)/N^2$ and find an expression for the estimator of $V(\hat{\bar{Y}}_{SS})$.

7.13. Show that $V(\hat{\bar{Y}}_{pps})$ and $v(\hat{\bar{Y}}_{pps})$ can be expressed as $\Sigma_i^N \Sigma_{i<j}^N u_i u_j (r_i - r_j)^2/n$ and $\Sigma_i^n \Sigma_{i<j}^n (r_i - r_j)^2/n^2(n-1)$, respectively.

7.14. Find an estimator for $V(\hat{\bar{Y}}_{pps})$ directly from its expression in Section 7.13.

Appendix A7

Alternative expression for S^2

The numerator of (7.5) can be expressed as

$$\begin{aligned}\sum_1^N \sum_1^M (y_{ij} - \bar{\bar{Y}})^2 &= \sum_1^N \sum_1^M [(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{\bar{Y}})]^2 \\ &= \sum_1^N \sum_1^M (y_{ij} - \bar{y}_i)^2 + M \sum_1^N (\bar{y}_i - \bar{\bar{Y}})^2 \\ &= (M-1)\sum_1^N s_i^2 + M\sum_1^N (\bar{y}_i - \bar{\bar{Y}})^2,\end{aligned}$$

where s_i^2 is the variance within the ith cluster. Note that the cross-product term in the square brackets of the first expression vanishes. In the *Analysis of Variance* terminology, the left-hand side is the *total SS* (sum of squares). The two expressions on the right-hand side are the *within SS* and *between SS*, respectively.

Now, from (7.5),

$$S^2 = \frac{N(M-1)}{NM-1}S^2_{\text{wsy}} + \frac{N-1}{NM-1}S^2_b.$$

Estimator for S^2

The variance among the nM sample observations can be expressed as

$$\begin{aligned} s^2 &= \frac{\sum_1^n \sum_1^M (y_{ij} - \bar{\bar{y}})^2}{nM-1} \\ &= \frac{\sum_1^n \sum_1^M (y_{ij} - \bar{y}_i)^2 + M\sum_1^n (\bar{y}_i - \bar{\bar{y}})^2}{nM-1} \\ &= \frac{n(M-1)}{nM-1}s^2_w + \frac{n-1}{nM-1}s^2_b, \end{aligned}$$

where $s^2_w = (\Sigma_1^n s_i^2)/n$ is the sample *within MS* with $n(M-1)$ d.f.

The sample *between MS* and *within MS* are unbiased for S^2_b and S^2_w, respectively. However, s^2 is not unbiased for the population variance S^2 in (7.5). From (7.6), its unbiased estimator is given by

$$\hat{S}^2 = \frac{N(M-1)}{NM-1}s^2_w + \frac{N-1}{NM-1}s^2_b,$$

which becomes $[(M-1)s^2_w + s^2_b]/M$ for large N.

The relative precision in (7.13) can be estimated by substituting $\hat{S}^2$ and s^2_b for S^2 and S^2_b, respectively.

Bias of the ratio estimator for the mean

The bias of $\hat{\bar{\bar{Y}}}_R = \bar{y}/\bar{m}$ is

$$\begin{aligned} B(\hat{\bar{\bar{Y}}}_R) &= E(\hat{\bar{\bar{Y}}}_R) - \bar{\bar{Y}} = E[(\bar{y} - \bar{\bar{Y}}\bar{m})/\bar{m}] \\ &= E[(\bar{y} - \bar{\bar{Y}}\bar{m})/\bar{M}(1+\delta)], \end{aligned}$$

where $\delta = (\bar{m} - \bar{M})/\bar{M}$. By expanding $(1+\delta)^{-1}$ as $1 - \delta + \delta^2 - \delta^3 + \ldots$, and ignoring terms of order $1/n^2$ and smaller, the bias would be of order $1/n$. For large n, ignoring terms of order $1/n$ also, the bias is $E[(\bar{y} - \bar{\bar{Y}}\bar{m})/\bar{M}] = 0$.

Variance of the ratio estimator

From the large sample approximation in Appendix A7, the MSE of $\hat{\bar{\bar{Y}}}_R$ becomes the same as its variance

$$V(\hat{\bar{\bar{Y}}}_R) = E(\bar{y} - \bar{\bar{Y}}\bar{m})^2/\bar{M}^2.$$

The term inside the parenthesis is the mean $\bar{d}$ of $d_i = y_i - \bar{\bar{Y}}m_i$. The population mean of d_i and hence of $\bar{d}$ is equal to zero. The variance $\hat{\bar{\bar{Y}}}_R$ can be expressed as $V(\hat{\bar{\bar{Y}}}_R) = V(\bar{d})/\bar{M}^2 = [(1-f)/n]S_d^2$, where

$$S_d^2 = \frac{1}{N-1}\sum_1^N (y_i - \bar{\bar{Y}}m_i) = \frac{1}{N-1}\sum_1^N [(y_i - \bar{Y}) - \bar{\bar{Y}}(m_i - \bar{M})]^2$$

$$= S_y^2 + \bar{\bar{Y}}^2 S_m^2 - 2\bar{\bar{Y}}S_{my} = S_y^2 + \bar{\bar{Y}}^2 S_m^2 - 2\bar{\bar{Y}}\rho S_m S_y.$$

From the sample,

$$s_d^2 = \frac{1}{n-1}\sum_1^n (y_i - \hat{\bar{\bar{Y}}}_R m_i)^2$$

$$= s_y^2 + \hat{\bar{\bar{Y}}}_R^2 s_m^2 - 2\hat{\bar{\bar{Y}}}_R s_{my} = s_y^2 + \hat{\bar{\bar{Y}}}_R^2 s_m^2 - 2\hat{\bar{\bar{Y}}}_R r s_m s_y,$$

where $r = s_{my}/s_m s_y$.

Hence, $V(\hat{\bar{\bar{Y}}}_R)$ may be approximately estimated from

$$v(\hat{\bar{\bar{Y}}}_R) = \frac{(1-f)}{n\bar{m}^2} s_d^2.$$

Expectation and variance of the Horvitz–Thompson estimator

This estimator $\hat{Y}_{\mathrm{HT}}$ in (7.34) can be expressed as $\Sigma_1^N \delta_i (y_i/\phi_i)$. The random variable δ_i takes the value one when a unit is selected into the sample and zero otherwise. When the ith population unit is selected

into the sample with probability ϕ_i, $P(\delta_i = 1) = \phi_i$ and $P(\delta_i = 0) = 1 - \phi_i$. As a result, $E(\delta_i) = 1(\phi_i) + 0(1 - \phi_i) = \phi_i$. Thus, $E(\hat{Y}_{\mathrm{HT}}) = \Sigma_1^N E(\delta_i)(y_i/\phi_i) = \Sigma_1^N \phi_i(y_i/\phi_i) = Y$.

The variance of $\hat{Y}_{\mathrm{HT}}$ can be expressed as $V(\hat{Y}_{\mathrm{HT}}) = E(\hat{Y}_{\mathrm{HT}} - Y)^2 = E(\hat{Y}_{\mathrm{HT}})^2 - 2YE(\hat{Y}_{\mathrm{HT}}) + Y^2 = E(\hat{Y}_{\mathrm{HT}})^2 - Y^2$. Now,

$$\hat{Y}_{\mathrm{HT}}^2 = \sum_1^n \left(\frac{y_i}{\phi_i}\right)^2 + \sum_{i \neq j}^n \sum^n \frac{y_i}{\phi_i} \frac{y_j}{\phi_j} = \sum_1^N \delta_i \left(\frac{y_i}{\phi_i}\right)^2 + \sum_{i \neq j}^N \sum^N \delta_{ij} \frac{y_i}{\phi_i} \frac{y_j}{\phi_j}.$$

The random variable δ_{ij} takes the value one when the ith and jth units are selected into the sample and zero otherwise, with $P(\delta_{ij} = 1) = \phi_{ij}$ and $P(\delta_{ij} = 0) = 1 - \phi_{ij}$. Thus,

$$E(\hat{Y}_{\mathrm{HT}}^2) = \sum_1^N \phi_i \left(\frac{y_i}{\phi_i}\right)^2 + \sum_{i \neq j}^N \sum^N \phi_{ij} \frac{y_i}{\phi_i} \frac{y_j}{\phi_j}.$$

Now, the variance of $\hat{Y}_{\mathrm{HT}}$ is obtained by subtracting $Y^2 = \Sigma_i y_i^2 + \Sigma_{i \neq j} \Sigma y_i y_j$ from this expression.

Expectation and variance of the PPSS estimator

The expectation of $\hat{\bar{Y}}_{\mathrm{SS}}$ is $\bar{M} \Sigma P_k(\bar{y}/\bar{m})_k = \Sigma \bar{y}_k /_N C_n = \bar{Y}$. Thus, $\hat{\bar{Y}}_{\mathrm{SS}}$ is unbiased for $\bar{Y}$.

Further, the expectation of $\hat{\bar{Y}}_{\mathrm{SS}}^2$ is $\bar{M}^2 \Sigma P_k(\bar{y}/\bar{m})^2 = \bar{M} \Sigma(\bar{y}^2/\bar{m})/_N C_n = \bar{M} E(\bar{y}^2/\bar{m})$, where E now stands for expectation with respect to simple random sampling without replacement. Thus, $V(\hat{\bar{Y}}_{\mathrm{SS}}) = \bar{M} E(\bar{y}^2/\bar{m}) - \bar{Y}^2$. The large sample approach for the ratio estimator leads to

$$V(\hat{\bar{Y}}_{\mathrm{SS}}) = \frac{(1-f)}{n}(S_y^2 + R^2 S_m^2 - 2R\rho S_y S_m),$$

which is the same as (7.21).

Variance of the pps estimator and the variance estimator

Since the sample is selected with replacement, $V(\hat{Y}_{\mathrm{pps}})$ is the same as S_r^2/n, where $S_r^2 = V(r_i)$. Since $E(r_i) = Y$, $V(r_i) = E(r_i^2) - Y^2$. Now, $E(r_i^2) =$

$\Sigma_1^N u_i r_i^2 = \Sigma_1^N (y_i^2/u_i)$. Therefore, $V(r_i) = \Sigma_1^N (y_i^2/u_i) - Y^2 = \Sigma_1^N (y_i - u_i Y)^2/u_i = \Sigma_1^N u_i (r_i - Y)^2$.

To find the estimators for $V(r_i)$ and $V(\hat{Y}_{\text{pps}})$, note that $\Sigma_1^n (r_i - \hat{Y}_{\text{pps}})^2 = \Sigma_1^n [(r_i - Y) - (\hat{Y}_{\text{pps}} - Y)]^2 = \Sigma_1^n (r_i - Y)^2 - n(\hat{Y}_{\text{pps}} - Y)^2$. Thus, $E\Sigma_1^n (r_i - \hat{Y}_{\text{pps}})^2 = n\Sigma_1^N u_i (r_i - Y)^2 - nV(\hat{Y}_{\text{pps}}) = n(n-1)V(\hat{Y}_{\text{pps}})$. Hence, $s_r^2 = \Sigma_1^n (r_i - \hat{Y}_{\text{pps}})^2/(n-1)$ is unbiased for S_r^2 and $v(\hat{Y}_{\text{pps}}) = s_r^2/n$ is unbiased for $V(\hat{Y}_{\text{pps}})$.