CHAPTER 6

Subpopulations

6.1 Introduction

From the observations of a sample drawn from a population or region, estimates for the totals, means, and proportions for specified groups, **subpopulations**, **domains**, and areas can also be obtained. Households classified according to their sizes, income levels, and age or educational level of the head of the household are examples of subpopulations. Industries classified according to the employee sizes, production levels, and profits provide other examples of these groups. Certain number of units of each of the subpopulations or groups will appear in a sufficiently large sample selected from the entire population.

Political polls provide several illustrations of estimating percentages for subpopulations. For instance, in the local and national elections for public offices, including those of the president, governors, senators, representatives, and mayors, polls conducted before and on election day are classified into several categories. The classifications are usually based on age, education, income, sex, marital status, political affiliation, race, religion, and other demographic and socioeconomic characteristics of the voters. Percentages of the voters favoring each of the candidates are estimated for the entire nation as well as for each of the classifications.

In the studies related to specified subpopulations, importance is given to the estimation for each subpopulation and differences between the subpopulations. If the population is initially stratified, estimates for a subpopulation of interest are obtained from its units appearing in the samples of each of the strata. These estimates can also be obtained from the poststratification described in Section 5.16.

Estimation procedures for the totals and means of subpopulations from a simple random or stratified sample are presented in the following sections. For **small areas**, subpopulations, and regions of small sizes, procedures for increasing the precision of the estimators are described in Chapter 12.

6.2 Totals, means, and variances

Consider a population with k groups with N_i observations, $N = \Sigma_1^k N_i$, in the ith group. Let y_{ij}, $j = 1,\ldots, N_i$, represent the jth observation in the ith group. The total, mean, and variance of the ith group are

$$Y_i = \sum_1^{N_i} y_{ij} = y_{i1} + y_{i2} + \cdots + y_{iN_i}, \tag{6.1}$$

$$\bar{Y}_i = \frac{Y_i}{N_i} \tag{6.2}$$

and

$$S_i^2 = \frac{\sum_1^{N_i} (y_{ij} - \bar{Y}_i)^2}{N_i - 1}. \tag{6.3}$$

6.3 Estimation of the means and their differences

When a sample of size n is selected without replacement from the N population units, the probability of observing $(n_1, n_2,\ldots, n_k)$ units, $n = \Sigma_1^k n_i$, in the groups is

$$P(n_1, n_2, \ldots, n_k) = \frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\binom{N_k}{n_k}}{\binom{N}{n}}. \tag{6.4}$$

Note that conditional on n_i, the probability for a unit from the ith group to appear in the sample is n_i/N_i and the probability for a specified pair of units from the ith group to be selected into the sample is $n_i(n_i - 1)/N_i(N_i - 1)$.

Means

The sample mean and variance for the ith group are

$$\bar{y}_i = \frac{1}{n_i}\sum^{n_i} y_{ij}. \tag{6.5}$$

and

$$s_i^2 = \frac{\sum_1^{n_i}(y_{ij} - \overline{Y}_i)^2}{n_i - 1}. \qquad (6.6)$$

For an observed n_i, $\bar{y}_i$ is unbiased for $\overline{Y}_i$ and has the variance

$$V(\bar{y}_i) = \frac{N_i - n_i}{N_i n_i} S_i^2. \qquad (6.7)$$

From the probability distribution in (6.4), $E(n_i) = nW_i$, where $W_i = N_i/N$ is the proportion of the units in the ith domain. Substitution of this expectation for n_i provides an approximation to the average of the above variance.

An estimator of the above variance is given by

$$v(\bar{y}_i) = \frac{N_i - n_i}{N_i n_i} s_i^2. \qquad (6.8)$$

If N_i is not known, it may be replaced by $\hat{N}_i = Nw_i = N(n_i/n)$, where w_i is the sample proportion of the units observed in the ith group. In this case (6.8) becomes $(1 - f)s_i^2/n_i$.

Example 6.1. Hospital expenditures: In the American Hospital Association *Guide to the Health Care Field* (1988), complete information on the following characteristics is available for 217 short-term hospitals (average length of stay of the patients is less than 30 days) in New York State with general medical and surgical services: (1) Number of beds, (2) average number of patients treated per day, (3) rate of occupancy of the beds, and (4) total annual expenses. Among the 217 hospitals, 150 were owned by individuals, partnerships, or churches. The remaining 67 were operated by federal, state, local or city governments.

A sample of 40 from the 217 hospitals has been selected randomly without replacement. In this sample, 29 hospitals of the first type and 11 of the second type appeared. The means and standard deviations of all the 40 hospitals and the samples observed in the two groups are presented in Table 6.1.

As can be seen, the estimate of the average amount of expenses for the first group is $\bar{y}_1 = \$61.1$ million. With the known value of $N_1 = 150$ and the sample size $n_1 = 29$, from (6.8), $v(\bar{y}_1) = 106.6$ and S.E.$(\bar{y}_1) = 10.3$.

Table 6.1. Means (top) and standard deviations (bottom) for a sample of 40 hospitals.

	Beds	Patients	Occupancy Rates	Expenses, (millions $)
Entire				
sample	339.7	289.7	81.1	69.2
$n = 40$	227.2	201.5	11.6	65.0
Group 1	325.9	276.5	81.7	61.1
$n_1 = 29$	234.8	206.9	11.7	61.9
Group 2	376.1	324.5	79.6	90.7
$n_2 = 11$	211.9	191.5	11.9	71.2

Source: American Hospital Association, *Guide to the Health Care Field* (1988).

If N_1 is not known, its estimate is $\hat{N}_1 = 217(29/40) = 157$. Substituting this value for N_1 in (6.8), $v(\bar{y}_1) = 112.7$ and $\text{S.E.}(\bar{y}_1) = 10.6$.

Difference between two groups

An unbiased estimator of the difference $(\bar{Y}_i - \bar{Y}_j)$ of the means of two groups is $(\bar{y}_i - \bar{y}_j)$ and its variance is given by

$$V(\bar{y}_i - \bar{y}_j) = (1 - f_i)\frac{S_i^2}{n_i} + (1 - f_j)\frac{S_i^2}{n_j}, \tag{6.9}$$

where $f_i = n_i/N_i$ and $f_j = n_j/N_j$. For an estimate of this variance, replace the group variances S_i^2 and S_j^2 by their sample estimates s_i^2 and s_j^2. If the group sizes N_i and N_j are not known, they are replaced by the estimates $n_i(N/n)$ and $n_j(N/n)$, that is, f_i and f_j are replaced by the common sampling fraction $f = n/N$.

From Table 6.1, we find that an estimate for the increase $(\bar{Y}_2 - \bar{Y}_1)$ in the average expenditure for the second group from that of the first is $(\bar{y}_2 - \bar{y}_1) = 90.7 - 61.1 = \$29.6$ million. From (6.9), an estimate of $V(\bar{y}_2 - \bar{y}_1)$ is given by $v(\bar{y}_2 - \bar{y}_1) = (67 - 11)(71.2)^2/(67 \times 11) + (150 - 29)(61.9)^2/(150 \times 29) = 491.78$, and hence S.E. $(\bar{y}_2 - \bar{y}_1) = 22.18$.

As seen in Example 6.1, $\hat{N}_1 = 157$ and hence $\hat{N}_2 = 60$. With these estimated sizes, $v(\bar{y}_2 - \bar{y}_1) = 484$ and hence $\text{S.E.}(\bar{y}_2 - \bar{y}_1) = 22$.

6.4 Totals of subpopulations

Known sizes

In several applications, it is often of interest to estimate the totals of subpopulations. If N_i is known, an unbiased estimator of Y_i is

$$\hat{Y}_i = N_i \bar{y}_i = \frac{N_i}{n_i} \sum_1^{n_i} y_{ij}. \tag{6.10}$$

For a given n_i, its variance is obtained from

$$V(\hat{Y}_i) = N_i^2 V(\bar{y}_i) = \frac{N_i(N_i - n_i)}{n_i} S_i^2. \tag{6.11}$$

For estimating this variance, substitute s_i^2 for S_i^2.

Estimated sizes

If N_i is not known, it can be estimated from $\hat{N}_i = N(n_i/n)$. With this estimator, from (6.10), an estimator for the total is

$$\hat{\hat{Y}}_i = \hat{N}_i \bar{y}_i = \frac{N}{n} \sum_1^{n_i} y_{ij}. \tag{6.12}$$

Since

$$E(\hat{\hat{Y}}_i \,|\, n_i) = \hat{N}_i \bar{Y}_i, \tag{6.13}$$

$$E(\hat{\hat{Y}}_i) = E(\hat{N}_i)\bar{Y}_i = N_i \bar{Y}_i = Y_i. \tag{6.14}$$

For a given n_i, the variance of this unbiased estimator is

$$V(\hat{\hat{Y}}_i) = \hat{N}_i^2 V(\bar{y}_i) = \frac{N^2 n_i(N_i - n_i)}{N_i n^2} S_i^2. \tag{6.15}$$

If n_i is replaced by its expected value nW_i, approximations to the averages of both (6.11) and (6.15) will be equal to $N(N-n)W_i S_i^2/n$.

Notice that when N_i is known, the estimator in (6.10) is obtained by dividing the sample total by n_i/N_i, which is the probability of a unit of the ith group appearing in a sample of size n_i from the N_i units of the ith group. If N_i is not known, as in (6.12), the estimator is obtained by dividing the sample total by n/N, which is the probability of selecting a unit in the sample of size n from the entire population.

Example 6.2. Hospital expenditures: From the data in Table 6.1, with the known value of $N_1 = 150$, the estimate for the total expenses of the first group is $\overline{Y}_1 = 150(61.1) = \$9165$ million, that is, close to \$9.2 billion. From Example 6.1, S.E. $(\bar{y}_1) = 10.3$, and hence S.E.$(\hat{\overline{Y}}_1) = 150(10.3) =$ \$1545 million.

If N_1 is not known, with $\hat{N}_1 = 157$, $(\hat{\hat{Y}}_1) = 157(61.1) = \$9592.7$ milion, that is, about \$9.6 billion. Since S.E.$(\bar{y}_1) = 10.6$ for this case, S.E.$(\hat{\hat{Y}}_1) =$ $157(10.6) = \$1664.2$ million.

6.5 Sample sizes for estimating the means and totals

The overall sample size n for a survey can be determined for estimating the means or totals of subpopulations with required precision. If it is required that the variance of $\bar{y}_i$ should not exceed V_i, $i = 1,\ldots,k$, substituting $E(n_i) = nW_i$ for n_i in (6.7),

$$\left(\frac{1}{n} - \frac{1}{N}\right)\frac{S_i^2}{W_i} \leq V_i\,. \tag{6.16}$$

Solving this equation with the equality sign, n is obtained from $n_m/(1 + n_m/N)$, where n_m is the maximum value of S_i^2/V_iW_i.

Example 6.3. Sample size for hospital expenditures: The proportions of the units in the two groups of the previous section are $W_1 = 150/217 = 0.69$ and $W_2 = 0.31$. If it is of interest to estimate the average expenses of these groups with standard errors not exceeding 15 and 30, using the sample estimates 61.9 and 71.2 for S_1 and S_2.

$$n_m = \text{Max}\left[\frac{(61.9)^2}{(15)^2(0.69)}, \frac{(71.2)^2}{(30)^2(0.31)}\right] = 25.$$

The required sample size n is equal to $25/(1 + 25/217) = 23$.

Similar methods can be used to find the overall sample size with prescriptions on the errors of estimation, coefficients of variations, and confidence widths of the group means. The same type of approaches can also be used to estimate the differences of group means. Sedransk (1965, 1967), Booth and Sedransk (1969), and Liao and Sedransk (1983) suggest procedures for determining the sample sizes with the above type of constraints.

6.6 Proportions and counts

Consider an attribute with two classes and k subpopulations. Among the N_i units of the ith subpopulation, $i = 1, 2, \ldots, k$, let C_i and $(N_i - C_i)$ of the units belong to the two classes. In a random sample of size n from the population, n_i units will be observed in the ith subpopulation, with c_i and $(n_i - c_i)$ units from the two classes. Table 6.2 describes the compositions of the population and sample.

The sample proportion $p_i = c_i/n_i$ is unbiased for $P_i = C_i/N_i$. Its variance and estimator of variance are given by

$$V(p_i) = \frac{N_i - n_i}{N_i - 1} \frac{P_i(1 - P_i)}{n_i} \tag{6.17}$$

and

$$v(p_i) = \frac{N_i - n_i}{N_i} \frac{p_i(1 - p_i)}{n_i - 1}. \tag{6.18}$$

An unbiased estimator of C_i is $\hat{C}_i = N_i p_i$. If the size of the subpopulation N_i is not known, an unbiased estimator of the total is given by $\hat{N}_i p_i$ where $\hat{N}_i = N(n_i/n)$.

For the case of more than two classes, estimates and their standard errors for the proportions and totals in the different subpopulations

Table 6.2. Household composition.

	Residences with Number of Persons						
	1	2	3	4	5	6+	Total
Owners	58	89	46	29	15	5	242
Renters	42	58	33	18	4	3	158

and their differences can be found through suitable modifications of the expressions in Section 4.7.

Example 6.4. Household composition: The figures in Table 6.2 refer to the numbers of persons in a random sample of 400 from the 14,706 residences in a town in New York State. During 1 year, the town assessor reported that 8549 of the houses were occupied by owners and the remaining 6157 were rented.

An estimate of the proportion of the four-member households in the town is $(29 + 18)/400 = 0.1175$. From (6.18), the variance of this estimate is $(1{,}4706 - 400)(0.1175)(0.8825)/(1{,}4706 \times 399) = 2.53 \times 10^{-4}$ and hence its S.E. is equal to 0.016. An estimate for the total number of houses of this type is $14{,}706(0.1175) = 1728$ and its S.E. is $14{,}706(0.016) = 235$.

For the owners, the estimates of the above proportion is $29/242 = 0.1198$. With the known size of 8549 for this group, from (6.18), the variance of this estimate is $(8549 - 242)(0.1198)(0.8802)/(8549 \times 241) = 4.25 \times 10^{-4}$ and hence its S.E. is equal to 0.021. If the size is unknown, its estimate is $14{,}706(242/400) = 8897$. Substituting this figure in (6.18), the S.E. again is close to 0.021.

An estimate for the total number of households of the above type among the owner-occupied residences is $8549(0.1198) = 1024$. The S.E. of this estimate is $8549(0.021) = 180$.

For the owner-occupied residences, an estimate of the difference in the proportions of the two-member and four-member households is $(89/242) - (29/242) = 0.248$. The variance of this estimate is $[(8549 - 242)/8549][0.2325 + 0.1054 + 2(0.0441)]/241 = 0.0017$, and hence its S.E. equals 0.041. An estimate of the difference in the numbers of houses of these types is $8549\ (0.248) = 1624$, and the S.E. of this estimate is $8549\ (0.041) = 351$.

For finding the sample sizes to estimate the proportions of qualitative characteristics in the subpopulations or differences between the proportions, the procedures in Section 6.5 can be used, by replacing S_i^2 with $N_iP_i(1 - P_i)/(N_i - 1)$ or approximately by $P_i(1 - P_i)$.

6.7 Subpopulations of a stratum

Estimates for specified subpopulations can be obtained from the observations of a stratified random sample. Estimates for the subpopulations from a single stratum are described in this section. Section 6.8 presents estimates for the subpopulations obtained from all the strata.

Let N_{gj}, $j = 1,\dots,k, \Sigma_j N_{gj} = N_g$, denote the size of the jth subpopulation in the gth stratum. The observations of these subpopulations can be denoted by y_{gij}, $i = 1,\dots,N_{gj}$. The mean and variance of this group are

$$\bar{Y}_{gj} = \frac{\sum_{1}^{N_{gj}} y_{gij}}{N_{gj}} \qquad \text{and} \qquad S_{gj}^2 = \frac{\sum_{1}^{N_{gj}} (y_{gij} - \bar{Y}_{gj})^2}{N_{gj} - 1}. \tag{6.19}$$

In a random sample of size n_g from the N_g units of the stratum, n_{gj} units, $\Sigma_j N_{gj} = n_g$, of the jth group will be observed. The sample mean and variance for this group are

$$\bar{y}_{gj} = \frac{\sum_{1}^{n_{gj}} y_{gij}}{N_{gj}} \qquad \text{and} \qquad s_{gj}^2 = \frac{\sum_{1}^{n_{gj}} (y_{gij} - \bar{y}_{gj})^2}{n_{gj} - 1}, \tag{6.20}$$

which are unbiased for $\bar{Y}_{gj}$ and S_{gj}^2. The variance of $\bar{y}_{gj}$ and its estimator are

$$V(\bar{y}_{gj}) = \frac{N_{gj} - n_{gj}}{N_{gj} n_{gj}} S_{gj}^2 \tag{6.21}$$

and

$$v(\bar{y}_{gj}) = \frac{N_{gj} - n_{gj}}{N_{gj} n_{gj}} s_{gj}^2. \tag{6.22}$$

An approximation to the average of (6.21) is obtained by replacing n_{gj} by its expectation $n_g(N_{gj}/N_g)$. If N_{gj} is not known, it is replaced by $\hat{N}_{gj} = (n_{gj}/n_g)N_g$ for estimating the variance.

Example 6.5. Athletic facility usage by the males and females of three strata: As an illustration, consider the 4000 undergraduate students, 1000 graduate students, and 200 faculty of a university as the three strata.

The first stratum consists of $N_{11} = 2000$ male students with the total Y_{11}, mean $\bar{Y}_{11}$, and variance S_{11}^2 for a characteristic of interest. This stratum has $N_{12} = 2000$ female students with total Y_{12}, mean $\bar{Y}_{12}$, and variance S_{12}^2. Similarly, for the $N_{21} = 600$ male and $N_{22} = 400$ female graduate students, these quantities are represented by $(Y_{21}, \bar{Y}_{21}, S_{21}^2)$ and $(Y_{22}, \bar{Y}_{22}, S_{22}^2)$, respectively. For the $N_{31} = 150$ and $N_{32} = 50$ female faculty, they can be represented by $(Y_{31}, \bar{Y}_{31}, S_{31}^2)$ and $(Y_{32}, \bar{Y}_{32}, S_{32}^2)$, respectively.

Table 6.3. Usage of athletic facilities.

	Males	Females	Strata and Sample Sizes
	Undergraduates		
N_{1j}	2000	2000	4000
n_{1j}	5	5	10
$\bar{y}_{1j}$	3.4	1.8	
s_{1j}^2	1.3	0.7	
	Graduates		
N_{2j}	600	400	1000
n_{2j}	3	2	5
$\bar{y}_{2j}$	4.67	2.5	
s_{2j}^2	1.29	0.5	
	Faculty		
N_{3j}	150	50	200
n_{2j}	4	1	5
$\bar{y}_{3j}$	4	2	
s_{3j}^2	2.67	0	

For a survey on the athletic facilities on the campus, in a small sample of ten undergraduate students, $n_{11} = 5$ male and $n_{12} = 5$ females were observed. The number of hours a week of utilizing the facilities for the five male students were $y_{11j} = (3, 5, 3, 2, 4)$, and for the five female students, they were $y_{12j} = (3, 1, 2, 2, 1)$. Similarly, in a sample of five graduate students, $n_{21} = 3$ males and $n_{22} = 2$ females were observed. The number of hours for these groups were $y_{21j} = (4, 6, 4)$ and $y_{22j} = (3, 2)$. For a sample of five from the faculty, $n_{31} = 4$ and $n_{32} = 1$, with the number of hours $y_{31j} = (4, 4, 6, 2)$ and $y_{32j} = 2$. The sizes of these groups along with their sample means and variances are presented in Table 6.3 and Figure 6.1.

From the above observations, an estimate of the average for the undergraduate male students is $\bar{y}_{11} = 3.4$ hours. From (6.22), $v(\bar{y}_{11}) = [(2000 - 5)/2000](1.3/5) = 0.2594$, and hence S.E.$(\bar{y}_{11}) = 0.51$. Similarly, for the male graduate students, $\bar{y}_{21} = 4.67$, $v(\bar{y}_{21}) = [(600 - 3)/600](1.29/3) = 0.4279$ and S.E.$(\bar{y}_{21})$. For the male faculty, $\bar{y}_{31} = 4$, $v(\bar{y}_{31}) = [(150 - 4)/150](2.67/4) = 0.6497$ and S.E.$(\bar{y}_{31}) = 0.81$.

An unbiased estimator of the total $Y_{gj} = N_{gj}\bar{Y}_{gj}$ is $\hat{Y}_{gj} = N_{gj}\bar{y}_{gj}$. If N_{gj} is not known, it is replaced by $\hat{N}_{gj}$ and the procedure in Section 6.4 can be used to find its average variance.

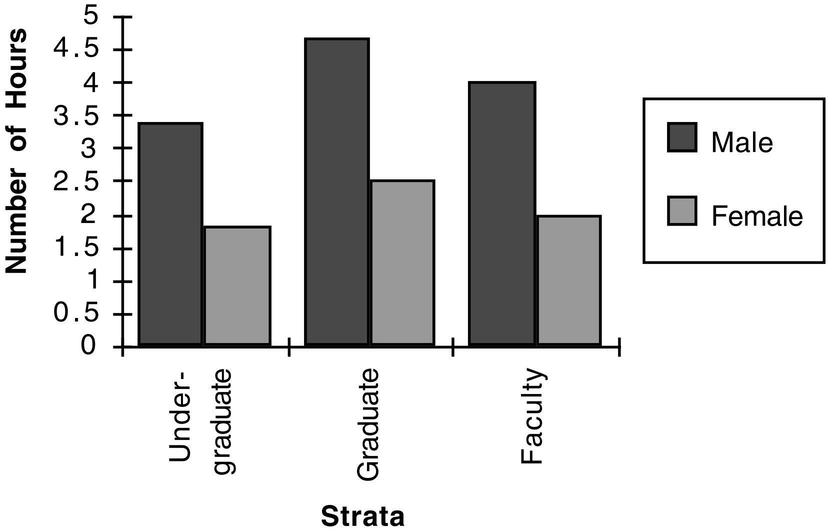

Figure 6.1. Usage of the athletic facility by the male and female groups.

6.8 Totals and means of subpopulations in strata

In the population, the jth group consists of $N_j = \Sigma_g^G N_{gj}$ units with total, mean, and variance:

$$Y_j = \sum_g^G Y_{gj}$$
$$\bar{Y}_j = Y_j/N_j,$$

and

$$S_j^2 = \sum_g \sum_i (y_{gij} - \bar{Y}_j)^2/(N_j - 1). \qquad (6.23)$$

If the sizes N_{gj} of the jth group in the strata are known, an unbiased estimator of Y_j is

$$\hat{Y}_j = \sum_g^G \hat{Y}_{gj} = \sum_g^G N_{gj}\, \bar{y}_{gj}. \qquad (6.24)$$

Its variance $\Sigma_g N_{gj}^2\ V(\bar{y}_{gj})$ and estimator of variance $\Sigma_h N_{gj}^2\, v(\bar{y}_{gj})$ can be obtained from (6.21) and (6.22).

For the above case of known N_{gj}, an unbiased estimator of $\bar{Y}_j$ is $\hat{Y}_j/N_j$. Its variance and estimator of variance are given by $\Sigma_g N_{gj}^2\ V(\bar{y}_{gj})/N_j^2$ and $\Sigma_g N_{gj}^2\ v(\bar{y}_{gj})/N_j^2$, respectively.

From Example 6.5, estimates for the total number of hours for the males in the three strata are $\hat{Y}_{11} = 2000(3.4) = 6800$, $\hat{Y}_{21} = 600(4.67) = 2802$, and $\hat{Y}_{31} = 150(4) = 600$. Hence, estimates of the total and mean for the 2750 males are $\hat{Y}_1 = 6800 + 2802 + 600 = 10{,}202$ and $\hat{\bar{Y}}_1 = 10{,}202/2750 = 3.71$ hours per week. The sample variance of this estimator is $v(\hat{\bar{Y}}_1) = (2000/2750)^2(0.2594) + (600/2750)^2(0.4279) + (150/2750)^2(0.6497) = 0.1595$ and hence S.E.$(\hat{\bar{Y}}_1) = 0.40$.

The sizes N_{gj} of the groups are usually not known. Using their estimates, an estimator for $\bar{Y}_j$ is

$$\hat{\bar{Y}}_j = \frac{\hat{Y}_j}{\hat{N}_j} = \frac{\sum_1^G \hat{N}_{gj}\bar{y}_{gj}}{\sum_1^G \hat{N}_{gj}} = \frac{\sum_1^G N_g \frac{n_{gj}}{n_g}\bar{y}_{gj}}{\sum_1^G N_g \frac{n_{gj}}{n_g}}. \tag{6.25}$$

This estimator resembles the ratio of two sample means and it is biased since the denominator is a random variable; see Chapter 9. For large samples, its variance is approximately given by

$$V(\hat{\bar{Y}}_j) = \frac{1}{N_j^2}\sum_1^G \frac{(N_g - n_g)N_{gj}}{n_g} S_{gj}^2. \tag{6.26}$$

An estimator of this variance can be obtained by replacing N_{gj}, N_j, and S_{gj}^2 by their unbiased estimators. Durbin (1958) and Cochran (1977) present an alternative estimator for (6.26).

An estimator for the difference $(\bar{Y}_j - \bar{Y}_{j'})$ of the means of two domains is $(\hat{\bar{Y}}_j - \hat{\bar{Y}}_{j'})$. The variance of this difference depends on the sample sizes n_g in the strata and the observed sizes n_{gj} and $n_{gj'}$ in the domains.

6.9 Stratification: Proportions and counts of subpopulations

Proportions and totals of qualitative characteristics for subpopulations in strata can be estimated through the notation in Section 6.6 and by suitably modifying the notation in Section 6.8. The following example provides an illustration.

Example 6.6. Off-campus sports: In the samples of Example 6.5, one each of the undergraduate and graduate male students and two of the male faculty participated in off-campus sports. Thus, $p_{11} = 1/5$, $p_{21} = 1/3$, and $p_{31} = 2/4$ are the estimates for the proportions in the three strata who participate in these sports. The sample variances of these estimates are $v(p_{11}) = [(2000 - 5)/2000 \times 4](1/5)(4/5) = 0.0399$, $v(p_{21}) = [(600 - 3)/600 \times 2](1/3)(2/3) = 0.1106$, and $v(p_{31}) = [(150 - 4)/150 \times 3](1/2)(1/2) = 0.0811$.

From these observations, an estimate for the proportion of the 2750 males participating in off-campus sports is $\hat{P}_1 = [2000(1/5) + 600(1/3) + 150(1/2)]/2750 = 0.2455$ or close to 25%. The sample variance of this estimator is $v(\hat{P}_1) = [2000^2(0.0399) + 600^2(0.1106) + 150^2(0.0811)]/(2750)^2 = 0.0266$ and hence $\text{S.E.}(\hat{P}_1) = 0.1632$.

Exercises

6.1. From the data in Table 6.1, (a) estimate the total number of beds in the second group and (b) find its S.E. when its size is known and when it is estimated.

6.2. (a) Using the figures in Table 6.1, estimate the total number of patients in the second group and find the S.E. of the estimate when the actual size is known. (b) Find the estimate and its S.E. when the actual size is estimated.

6.3. Using the sample data in Table 6.2 and the additional information in Example 6.4, estimate the following quantities and find their S.E.: (a) Proportions and numbers of residences with at least four persons, separately for the owner-occupied and rented households. (b) Differences of the above proportions and numbers. (c) Differences of the proportions of the single-member and two-member households in the rented households. In (a) and (b), estimate the numbers of households using the known sizes and by estimating the sizes.

6.4. With the sample data in Table 6.2 and Example 6.4, (a) estimate the proportion and total number of residences in that town with six or more persons and find the S.E. of the estimates. (b) Find the 90% confidence limits for this proportion and total.

6.5. Using the sample data in Table 6.2, estimate the average number of persons per household and find the 95% confidence limits for the average.

6.6. With the sample information presented in Table 6.3, estimate the average number of hours of utilizing the athletic facilities by the 2450 females and find the S.E. of the estimate.

6.7. With the information presented in Table 6.3 estimate the differences of the average number of hours for the 2750 males and the 2450 females, and find the S.E. of the estimate.

6.8. In the samples of Example 6.5, two of the undergraduate and one of the graduate female students but none of the female faculty participated in the off-campus sports. Estimate the proportion of the 2450 females participating in such sports and find the S.E. of the estimate.

6.9. *The New York Times* of February 25, 1994, summarized the results of a survey conducted by Klein Associates, Inc. on 2000 lawyers on sexual advances in the office. Between 85 and 98% responded to the questions in the survey; 49% of the responding women and 9% of the responding men agreed that some sorts of harassment exist in the offices. Assume that the population of lawyers is large and there are equal number of female and male lawyers, and ignore the nonresponse; that is, consider the respondents to be a random sample of the 2000 lawyers. (a) Find the standard errors for the above percentages. (b) Estimate the difference of the percentages for women and men, and find the S.E. of the estimate.

6.10. For the survey in Exercise 6.9, find the sample sizes needed for the following requirements: (a) The standard errors for women and men should not exceed 4 and 1%, respectively, and (b) the S.E. for the difference of the percentages for women and men should not exceed 3%.

6.11. Without the assumption in Exercise 6.9, find the standard errors for the women, men, and their difference.

6.12. *Project*. Consider the two groups of states in Table T5 in the Appendix with the total enrollments below 40,000 and 40,000 or more, and find their means and variances. For each of the 210 samples of size six from the ten states, find the means, variances, and the differences of the means. (a) For the first group, show that for each observed sample size, the average of the sample means coincides with the actual mean of that group and their variance coincides with (6.7). (b) Show that the average of the variances in (a) over the sample sizes is larger than that obtained by replacing n_i by $n(N_i/N)$, as described in Section 6.3.