CHAPTER 5

Stratification

5.1 Introduction

Each layer of a rock is a **stratum**, and the layers are the **strata**. There is usually a good deal of uniformity in the mineral contents and the composition of each of these layers. The Himalayas and Rockys are ranges of mountains. Each mountain is a stratum, and the ranges are the strata. Composition and fertility of the soil along the banks of a river vary as it gallops from its birth place to the ocean. The entire stretch can be divided into strata based on this variation. These types of analogies are used in survey sampling to describe the subdivisions or partitions of a population. The observations of the units of a stratum are closer to each other than to the units of another stratum.

As was seen in Section 1.8, for some types of income and expenditure surveys on households in urban areas, states, provinces, counties, and districts may be considered as the strata. For business surveys on employee size, production, and sales, stratification is usually based on industrial classifications. For agricultural surveys in rural areas, villages and geographical regions compose the strata. In general, stratification of population units depends on the purpose of the survey.

The major advantages of stratification are (1) estimates for each stratum can be obtained separately, (2) differences among the strata can be evaluated, (3) the total, mean, and proportion of the entire population can be estimated with high precision by suitably weighting the estimates obtained from each stratum, and (4) there are frequently savings in time and cost needed for sampling the units. In addition, it is usually convenient to sample separately from the strata rather than from the entire population, especially if the population is large.

The following sections examine the estimation of the mean, total, and proportion of each stratum as well as the entire population, determination of the sample sizes for the strata for specified requirements, and other topics of interest in stratified random sampling.

5.2 Notation

For the sake of illustration, Table 5.1 presents the rice production in 1985 in 43 countries. These figures for 65 of the 191 countries in the world appear in the *Statistical Abstracts of the United States*, 1990. The amounts for the three largest producing countries, China, India, and Indonesia, appear as 171.5, 91.5, and 38.7 million metric tons. Two of the countries produce only 3000 metric tons. Productions for 17 countries are in the NA (not available) category. Based on their production levels, the remaining 43 countries have been divided into two strata (groups) of sizes 32 and 11.

The total, mean, variance, and standard deviation for the population of the $N = 43$ productions in Table 5.1 are $Y = 157.27$, $\bar{Y} = 3.66$, $S^2 = 31.68$, and $S = 5.63$.

When a population consists of G strata of sizes N_g, the observations of the strata can be represented by y_{gi}, $g = 1, 2, \ldots, G$ and $i = 1, 2, \ldots, N_g$. In Table 5.1, the population of $N = 43$ productions is divided into $G = 2$ strata of sizes $N_1 = 32$ and $N_2 = 11$. The observations for the first stratum are $y_{11} = 0.04$, $y_{12} = 0.04$, $y_{13} = 0.06, \ldots, y_{1,32} = 2.80$. Similarly, for the second stratum, $y_{21} = 4.50$, $y_{22} = 5.60$, $y_{23} = 6.17, \ldots, y_{2,11} = 21.90$.

Table 5.1. Rice production (in million metric tons).

First Stratum						Second Stratum	
1	0.04	12	0.30	23	1.10	1	4.50
2	0.04	13	0.40	24	1.40	2	5.60
3	0.06	14	0.42	25	1.76	3	6.17
4	0.08	15	0.46	26	1.90	4	7.86
5	0.09	16	0.47	27	1.90	5	8.30
6	0.11	17	0.48	28	2.18	6	9.02
7	0.11	18	0.52	29	2.31	7	14.58
8	0.15	19	0.86	30	2.60	8	15.40
9	0.16	20	0.97	31	2.63	9	15.60
10	0.20	21	0.99	32	2.80	10	19.52
11	0.27	22	1.06			11	21.90
Total					28.82		128.45
Mean					0.90		11.68
Variance					0.79		35.55
S.D.					0.89		5.96

Source: Statistical Abstracts of the United States (1990), Table 1412.

The total and mean of the observations of the gth stratum are given by

$$Y_g = \sum_1^{N_g} y_{gi} \qquad \text{and} \qquad \bar{Y}_g = Y_g / N_g. \tag{5.1}$$

The variance of this stratum is

$$S_g^2 = \frac{\sum_1^{N_g} (y_{gi} - \bar{Y}_g)^2}{N_g - 1}. \tag{5.2}$$

The expressions in (5.1) and (5.2) are the same as those of the total, mean, and variance presented in Chapter 2, except that the different strata are represented by the subscript g. These summary figures for the two strata appear in Table 5.1.

The total and mean for the entire population are

$$Y = \sum_1^G \sum_1^{N_g} y_{gi} = \sum_1^G Y_g = \sum_1^G N_g \bar{Y}_g \tag{5.3}$$

and

$$\bar{Y} = \frac{Y}{N} = \sum_1^G W_g \bar{Y}_g = W_1 \bar{Y}_1 + W_2 \bar{Y}_2 + \cdots + W_G \bar{Y}_L, \tag{5.4}$$

where $W_g = N_g/N$ is the proportion of the population units in the gth stratum. In the right-hand side of (5.3), the first term is the sum of all the observations in the sample, the second term is the sum of the stratum totals, and in the final term the strata means are weighted by their sizes. In (5.4), the population mean is expressed as a weighted average of the means of the strata. As a verification, the population mean of 3.66 for the 43 countries is the same as (32/43)(0.90) + (11/43)(11.68).

The population variance is

$$S^2 = \frac{\sum_1^G \sum_1^{N_g} (y_{gi} - \bar{Y})^2}{N - 1}. \tag{5.5}$$

The numerator of this expression, $(N - 1)S^2$, is the sum of squares of the deviations from the population mean of all the observations in all the strata, and hence it is known as the **total SS**.

As shown in Appendix A5, (5.5) can also be expressed as

$$S^2 = \frac{\sum_1^G (N_g - 1)S_g^2 + \sum_1^G N_g(\bar{Y}_g - \bar{Y})^2}{N - 1}. \tag{5.6}$$

If N_g and N are large enough that they differ little from $(N_g - 1)$ and $(N - 1)$, this variance can be approximately expressed as $\Sigma W_g S_g^2 + \Sigma W_g(\bar{Y}_g - \bar{Y})^2$. The first term in the numerator of (5.6) is the **within SS** (sum of squares), which is the *pooled* sum of squares of the squared deviations of the observations of the strata from their means. The second term of (5.6) which expresses the variation among the strata means is the **between SS**. Some of the S_g^2 may be larger than S^2, but it is larger than the weighted average of the S_g^2.

5.3 Estimation for a single stratum

Samples of sizes n_g are selected from the N_g units of the strata, randomly without replacement and independently from the strata. The general description of the strata and the samples are presented in Table 5.2.

The mean and variance of a sample of n_g units selected from the gth stratum are

$$\bar{y}_g = \sum_1^{n_g} y_{gi} / n_g \tag{5.7}$$

and

$$s_g^2 = \frac{\sum_1^{n_g} (y_{gi} - \bar{y}_g)^2}{n_g - 1}. \tag{5.8}$$

Table 5.2. Strata and samples.

	1	2	g	G
	Strata			
Sizes	N_1	N_2	N_g	N_G
Means	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_g$	$\bar{Y}_G$
Variances	S_1^2	S_2^2	S_g^2	S_G^2
	Samples			
Sizes	n_1	n_2	n_g	n_G
Means	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_g$	$\bar{y}_G$
Variances	s_1^2	s_2^2	s_g^2	s_G^2

As seen in Chapter 2, this mean and variance are unbiased for $\bar{Y}_g$ and S_g^2, respectively. The variance of $\bar{y}_g$ is

$$V(\bar{y}_g) = \frac{N_g - n_g}{N_g n_g} S_g^2 = \frac{(1 - f_g)}{n_g} S_g^2, \tag{5.9}$$

where $f_g = n_g/N_g$ is the sampling fraction in the gth stratum. The estimator,

$$v(\bar{y}_g) = \frac{N_g - n_g}{N_g n_g} s_g^2 = \frac{(1 - f_g)}{n} s_g^2 \tag{5.10}$$

is unbiased for the above variance.

The expressions in (5.7) through (5.10) with the additional subscript g representing the stratum are the same as those presented in Chapter 2 for the mean of a simple random sample.

For the total Y_g of the gth stratum, an unbiased estimator is given by $N_g \bar{y}_g$. Its variance and estimator of variance are obtained by multiplying (5.9) and (5.10) by N_g^2.

Example 5.1. Rice production: For the sake of illustration, consider a sample of size $n = 10$ from the $N = 43$ countries in Table 5.1, and distribute this sample size proportional to the sizes of the two strata. Thus, the sample sizes are $n_1 = (0.74)n$ and $n_2 = (0.26)n$, which are approximately equal to 7 and 3. For these sample sizes, from (5.9), $V(\bar{y}_1) = [(32 - 7)/32](0.09)/7 = 0.0882$, S.E.$(\bar{y}_1) = 0.297$, $V(\bar{y}_2) = [(11 - 3)/11](35.55)/3 = 8.62$, and S.E. $(\bar{y}_2) = 2.94$.

To examine the estimates from the samples, samples of the above sizes were selected randomly without replacement through the Random Number Table in Appendix T1, independently from the two groups. Through this procedure, countries (4, 6, 9, 12, 17, 22, 31) and (1, 6, 10) appear in the samples of the two groups, respectively. The means and variances of these samples are $\bar{y}_1 = 0.69$, $s_1^2 = 0.85$, $\bar{y}_2 = 11.01$, and $s_2^2 = 59.38$.

Now, from (5.10), $v(\bar{y}_1) = [(32 - 7)/32](0.85/7) = 0.095$, and hence, S.E.$(\bar{y}_1) = 0.31$. Similarly, $v(\bar{y}_2) = [(11 - 3)/11](59.38/3) = 14.395$ and S.E.$(\bar{y}_2) = 3.79$.

5.4 Estimation of the population mean and total

With stratification, an estimator for the population mean $\bar{Y}$ is given by

$$\hat{\bar{Y}}_{\text{st}} = \sum_{1}^{G} W_g \bar{y}_g = W_1\bar{y}_1 + W_2\bar{y}_2 + \cdots + W_G\bar{y}_G, \tag{5.11}$$

where the subscript st denotes stratification. This estimator is obtained by substituting the sample means for the stratum means in (5.4). Note that the sample means are multiplied by the strata weights. Since $E(\bar{y}_g) = \bar{Y}_g$, this weighted mean is unbiased for $\bar{Y}$. Since the samples are selected independently from the strata, as shown in Appendix A5, the covariances between the sample means vanish and the variance of $\hat{\bar{Y}}_{\text{st}}$ becomes

$$V(\hat{\bar{Y}}_{\text{st}}) = \sum_{1}^{G} W_g^2 V(\bar{y}_g) = W_1^2 V(\bar{y}_1) + W_2^2 V(\bar{y}_2) + \cdots + W_G^2 V(\bar{y}_G). \tag{5.12}$$

Substituting the variances of the strata $V(\bar{y}_g)$ from (5.9), this variance can be expressed as

$$V(\hat{\bar{Y}}_{\text{st}}) = \sum_{1}^{G} W_g^2 \frac{(1-f_g)}{n_g} S_g^2 = \sum \frac{W_g^2 S_g^2}{n_g} - \frac{\sum W_g S_g^2}{N}. \tag{5.13}$$

An unbiased estimator of this variance is obtained by replacing $V(\bar{y}_g)$ in (5.12) by $v(\bar{y}_g)$ in (5.10), that is, replacing S_g^2 in (5.13) by s_g^2. Thus,

$$v(\hat{\bar{Y}}_{\text{st}}) = \sum_{1}^{G} W_g^2 v(\bar{y}_g) = \sum_{1}^{G} W_g^2 \frac{(1-f_g)}{n_g} s_g^2. \tag{5.14}$$

An unbiased estimator of the population total Y is $N\hat{\bar{Y}}_{\text{st}}$. Notice that this estimator is the same as $N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots + N_G\bar{y}_G$. Thus, the population total is estimated by adding the estimates for the totals of all the strata. Its variance and estimator of variance are obtained by multiplying (5.13) and (5.14) by N^2.

Example 5.2. Rice production in the world: With the sample means in Example 5.1, the estimate for the mean of the 43 countries is $\hat{\bar{Y}}_{st}$ = (32/43)(0.69) + (11/43)(11.01) = 3.33. From the first term on the right-hand side of (5.13), $V(\hat{\bar{Y}}_{st})$ = $(32/43)^2(0.088)$ + $(11/43)^2(8.62)$ = 0.613, and S.E.$(\hat{\bar{Y}}_{st})$ = 0.78. For the estimate of the variance, from (5.14), $v(\hat{\bar{Y}}_{st})$ = $(32/43)^2(0.095)$ + $(11/43)^2(14.395)$ = 0.9946, and hence S.E. $(\hat{\bar{Y}}_{st})$ = 0.9973.

Now, an estimate for the total rice production for the 43 countries is $\hat{Y}_{st}$ = 43(3.33) = 143.19 and S.E.$(\hat{Y}_{st})$ = 43(0.9973) = 42.88. From the sample, the estimate for S.E.$(\hat{Y}_{st})$ is 43(0.78) = 33.54.

By adding the 301.7 million pounds of the three largest producing countries to this estimate, the estimate for the world rice production is close to 445 million metric tons a year. The actual production, as appears in *Statistical Abstracts*, is 466 million metric tons.

Note that without stratification, for selecting a simple random sample of 10 from the 43 units, from (2.12), $V(\bar{y})$ = [(43 − 10)/43] (31.68/10) = 2.431. Thus, the precision of $\hat{\bar{Y}}_{st}$ relative to $\bar{y}$ is 2.431/ 0.613 = 3.97. This result shows that there is a 297% gain in precision for stratification.

As noted in Section 3.6, Cochran (1977, pp. 215–216) shows that if there is a linear trend in the population units, systematic sampling results in smaller variance for the mean than for simple random sampling. He also shows that for this case, the variance of $\hat{\bar{Y}}_{st}$ is smaller than the means of both simple random and systematic sampling. Further comparisons of these three procedures appear in Cochran (1946).

5.5 Confidence limits

As in Section 2.9, (1 − α)% confidence limits for $\bar{Y}_g$ are obtained from $\bar{y}_g \pm Z$. S.E.$(\bar{y}_g)$. Similarly, one can obtain the confidence limits for $\bar{Y}$ from $\hat{\bar{Y}}_{st} \pm Z$. S.E.$(\hat{\bar{Y}}_{st})$. Multiply these limits by N to obtain the limits for the population total Y.

From the results of Exercise 5.2, the 95% confidence limits for average rice production for the 43 countries are given by 3.33 ± 1.96(0.9973), that is, (1.38, 5.29).

When S_g^2 are estimated, the limits may be obtained from $\bar{y}_g \pm t$ S.E.$(\bar{y}_g)$. Approximation for the d.f. of the t-distribution needed to find these limits is presented, for example, by Cochran (1977, p. 96).

5.6 Proportions and totals

The results of Chapter 4 and Sections 5.3 and 5.4 can be combined to estimate for each stratum and the population the proportions and total numbers of units having a specific qualitative characteristic. All the expressions in this chapter for a single stratum are easily obtained from Chapter 4 with the subscript g for the gth stratum

Let C_g and $P_g = C_g/N_g$, $g = 1,\ldots,G$, denote the number and proportion of units in the gth stratum having the characteristic of interest. If c_g units in a random sample of size n_g are observed to have the characteristic, the sample proportion $p_g = c_g/n_g$ is unbiased for P_g. Its variance and estimator of variance are

$$V(p_g) = \frac{N_g - n_g}{N_g - 1}\frac{P_g Q_g}{n_g} \tag{5.15}$$

and

$$v(p_g) = \frac{N_g - n_g}{N_g}\frac{p_g q_g}{n_g - 1}, \tag{5.16}$$

where $Q_g = 1 - P_g$ and $q_g = 1 - p_g$. These expressions are obtained from (4.7) and (4.8) with the subscript g for the stratum. An unbiased estimator for C_g is given by $\hat{C}_g = N_g p_g$. For the variance of $\hat{C}_g$ and its estimator, multiply (5.15) and (5.16) by N_g^2.

Example 5.3. Energy consumption: For the sake of illustration, energy production and consumption in 1996 for the 25 countries in the world with the highest per capita consumption is presented in Table T9 in the Appendix. For 15 of these countries, that is, 60%, per capita consumption exceeds 200 million BTUs.

To study the production and consumption of energy, one can divide the 25 countries into two strata with the first stratum consisting of countries producing more than 200,000 barrels of petroleum a day, and the second stratum consisting of the rest. From this division, the 10 countries numbered (1, 2, 6, 8, 14, 17, 18, 19, 23, 24) are included in the first stratum, and the remaining 15 in the second stratum.

Among the $N_1 = 10$ countries, for $C_1 = 7$ countries energy consumption exceeds 200 million BTUs. Thus, $P_1 = 7/10$ for this characteristic. Corresponding figures for the second stratum are $N_2 = 15$, $C_2 = 8$, and $P_2 = 8/15$. If these percentages are not known, one can estimate them from samples selected from the two strata. For example, for a sample of $n_1 = 3$

units from the first stratum, from (5.15), $V(p_1) = [(10 - 3)/(9 \times 3)](7/10)(3/10) = 0.0544$. Similarly, for a sample of size $n_2 = 5$ from the second stratum, $V(p_2) = [(15 - 5)/(14 \times 5)](8/15)(7/15) = 0.0356$.

5.7 Population total and proportion

The total and proportion of the population having a specific characteristic, such as the number of countries with per capita energy consumption exceeding 200 million BTUs, are

$$\begin{aligned} C &= \sum_{1}^{G} C_g = C_1 + C_2 + \cdots + C_G \\ &= \sum N_g P_g = N_1 P_1 + N_2 P_2 + \cdots + N_G P_G \end{aligned} \tag{5.17}$$

and

$$P = C/N = \sum W_g P_g. \tag{5.18}$$

An unbiased estimator for P is given by

$$\hat{P}_{\text{st}} = \sum_{1}^{G} W_g p_g \tag{5.19}$$

with variance

$$V(\hat{P}_{\text{st}}) = \sum_{1}^{G} W_g^2 V(p_g) = \sum W_g^2 \frac{N_g - n_g}{N_g - 1} \frac{P_g Q_g}{n_g} \tag{5.20}$$

An unbiased estimator for this variance is given by

$$V(\hat{P}_{\text{st}}) = \sum W_g^2 (1 - f_g) \frac{p_g q_g}{n_g - 1}. \tag{5.21}$$

An unbiased estimator for the total number of units having the characteristic is $\hat{C}_{\text{st}} = N\hat{P}_{\text{st}}$, which has variance $N^2 V(\hat{P}_{\text{st}})$ and estimator of variance $N^2 v(\hat{P}_{\text{st}})$. The S.E. of $\hat{P}_{\text{st}}$ from the sample is given

by $N\sqrt{v(\hat{P}_{st})}$. For large sizes of the population and sample, confidence limits for P can be obtained from $\hat{P}_{st} \pm Z$ S.E.$(\hat{P}_{st})$. For C, multiply these limits by N.

Example 5.4. Energy consumption: For the two strata in Example 5.3, $W_1 = 10/25 = 0.4$ and $W_2 = 15/25 = 0.6$. With samples selected from the two strata, the proportion of the countries with per capita consumption exceeding 200 million BTUs is estimated from $\hat{P}_{st} = W_1 p_1 + W_2 p_2$. For samples of sizes $n_1 = 3$ and $n_2 = 5$ from the strata, we have found in Example 5.3 that $V(p_1) = 0.0544$ and $V(p_2) = 0.0356$. From these figures, $V(\hat{P}_{st}) = (0.4)^2(0.0544) + (0.6)^2(0.0356) = 0.0215$, and hence S.E.$(\hat{P}_{st}) = 0.1466$. For the estimation of the total number of countries with the above characteristic, $V(\hat{C}_{st}) = 25^2(0.0215) = 13.44$. From this variance or directly, S.E.$(\hat{C}_{st}) = 25(0.1466) = 3.67$.

5.8 Proportional and equal division of the sample

If the resources are available for a total sample of n units, they are distributed among the strata. One obvious choice is to divide this sample size proportionate to the stratum sizes; that is, select $n_g = n(N_g/N) = nW_g$ units from the gth stratum. As an alternative, if it is convenient and cost-effective, one may select equal number, $n_g = n/G$, units from the strata. An examination of the precision for these two types of determining the sample sizes follows.

Proportional allocation

For this type of dividing the sample, since $n_g/n = N_g/N = W_g$, the sampling fraction $f_g = n_g/N_g$ is the same for all the strata and is equal to the overall sampling fraction $f = n/N$. The estimator $\bar{Y}_{st}$ in (5.11) now becomes **self-weighting** and can be expressed as

$$\hat{\bar{Y}}_{st} = \frac{\sum^{G} n_g \bar{y}_g}{n} = \frac{\sum^{G} \sum^{n_g} y_{gi}}{n}. \tag{5.22}$$

This estimate is the average of all the n observations. Notice, however, that in general $\hat{\bar{Y}}_{st}$ in (5.11) is the weighted average of the means $\bar{y}_g$ with the strata weights W_g and it is unbiased for $\bar{Y}$ for any type of allocation of the sample.

For proportional allocation, replacing n_g by $n(N_g/N)$, the variance in (5.13) becomes

$$V_P = \frac{(1-f)}{n}\sum_1^G W_g S_g^2. \tag{5.23}$$

Notice that $\Sigma_1^G W_g S_g^2$ is the weighted average of the stratum variances. For the rice production in Example 5.2, from (5.23), $V_P = [(43 - 10)/430][(32/43)(0.79) + (11/43)(35.55)] = 0.743$. However, $V(\hat{\bar{Y}}_{st})$ was found to be 0.613 since the sample sizes were rounded off to 7 and 3 for the strata.

From (2.12), the variance of the sample mean is $V_S = (1-f)S^2/n$. As noted in Section 5.2, S^2 can be approximately expressed as $\Sigma W_g S_g^2 + \Sigma W_g(\bar{Y}_g - \bar{Y})^2$. This approximation is valid when $N_g - 1$ and $N - 1$ do not differ much from N_g and N, respectively. Substituting this expression in V_S, from (5.23), $V_S - V_P = [(1-f)/n]\Sigma W_g(\bar{Y}_g - \bar{Y})^2$, which is non-negative. Thus, proportional allocation results in smaller variance than simple random sampling if the means $\bar{Y}_g$ differ much.

From (2.12) and (5.23) the precision of proportional allocation relative to simple random sampling is

$$\frac{V_S}{V_P} = \frac{S^2}{\sum W_g S_g^2}. \tag{5.24}$$

Since $S^2 > \Sigma W_g S_g^2$ from (5.6), this relative precision is at least 100%. Notice, however, that it does not depend on the sample size.

Equal distribution

As an illustration of assigning equal sample sizes to all the strata, consider a population with five strata and the resources available for a total sample of 100 units. In this case, five interviewers can collect information on 20 sample units of one stratum each or the survey can be completed in 5 days by collecting the information on 20 sample units of a stratum per day. For this type of sampling, the variance of $\hat{\bar{Y}}_{st}$ and its estimator are obtained from (5.13) and (5.14) by substituting

n/G for n_g. The following example compares this approach with proportional allocation.

Example 5.5. Proportional and equal allocation: For the rice production in Example 5.2, with $n_1 = 7$ and $n_2 = 3$ for proportional allocation, $V_P = 0.613$. For equal allocation with $n_1 = n_2 = 5$, from (5.13), $V(\bar{y}_1) = [(32 - 5)/32](0.79)/5 = 0.1333$ and $V(\bar{y}_2) = [(11 - 5)/11](35.55)/5 = 3.88$. Now, from (5.13), $V(\hat{\bar{Y}}_{st}) = (32/43)^2(0.1333) + (11/43)^2(3.88) = 0.328$. By denoting this variance for equal allocation by V_E, $V_E/V_P = 0.328/0.613 = 0.54$. Thus, equal allocation reduces the variance by 46% from proportional allocation.

For the rice production in Example 5.2, as has been seen, $N_1 = 32$, $N_2 = 11$, $S_1 = 0.89$, and $S_2 = 5.96$. Proportional allocation considered more sample units for the larger first stratum. Equal allocation resulted in increasing the sample size by two units for the second stratum, which has larger variance. For the division of the sample size for the strata, the following procedure takes into account both the sizes and the variances of the strata.

5.9 Neyman allocation

For a given n, Neyman (1934) suggested finding n_g by minimizing the variance in (5.13). As shown in Appendix A5, from this optimization,

$$\frac{n_g}{n} = \frac{N_g S_g}{\sum N_g S_g} = \frac{W_g S_g}{\sum W_g S_g}. \tag{5.25}$$

Note that it is enough to know the relative values of S_g for determining the sample sizes for the strata. This allocation suggests that variance for $\hat{\bar{Y}}_{st}$ is reduced when strata with larger sizes N_g and larger variances S_g^2 receive larger sample sizes n_g. With this method, the minimum of the variance in (5.13) becomes

$$V_N = \frac{\left(\sum W_g S_g\right)^2}{n} - \frac{\left(\sum W_g S_g^2\right)}{N}. \tag{5.26}$$

The sample size n_g obtained from (5.25) for one or more strata can be as small as unity in some instances if N_g or S_g are relatively small.

For such strata, $V(\bar{y}_g)$ will be zero. On the other hand, since (5.25) is obtained without the constraint $n_g \leq N_g$, it may result in n_g larger than N_g for some of the strata. In this case, all or a predetermined number of units are sampled from these strata, and the remaining sample size is distributed to the rest of the strata.

5.10 Gains from Neyman allocation

From (5.23) and (5.26),

$$\begin{aligned} V_P - V_N &= \left[\sum W_g S_g^2 - \left(\sum W_g S_g\right)^2\right]/n \\ &= \sum W_g \left(S_g - \sum W_g S_g\right)^2/n. \end{aligned} \tag{5.27}$$

The right-hand side expresses the differences among the standard deviations S_g of the strata. Notice from (5.27) that the variance for Neyman allocation becomes smaller than that for proportional allocation as the differences among the S_g increase. The precision of this allocation relative to proportional allocation is

$$\frac{V_P}{V_N} = \frac{N-n}{N\left(\sum W_g S_g\right)^2/\sum (W_g S_g^2) - n}. \tag{5.28}$$

Since $\Sigma W_g S_g^2 > (\Sigma W_g S_g)^2$, this ratio is larger than unity, and it also increases with n.

With the approximation for S^2 considered in Section 5.8,

$$V_S - V_N = \sum W_g \left(S_g - \sum W_g S_g\right)^2 / n + (1-f)\sum W_g (\bar{Y}_g - \bar{Y})^2 / n. \tag{5.29}$$

The precision of Neyman allocation relative to simple random sampling increases as the differences among the means $\bar{Y}_g$ or the standard deviations S_g increase.

Example 5.6. Neyman allocation: For the rice production strata in Table 5.1, $W_1S_1 = (32/43)(0.89) = 0.66$, $W_2S_2 = (11/43)(5.96) = 1.53$ and $\Sigma W_g S_g = 2.19$. For the Neyman allocation, $n_1/n = 0.66/2.19 = 0.3$ and $n_2/n = 1.53/2.19 = 0.7$. If $n = 10$, the sample sizes for the strata are $n_1 = 3$ and $n_2 = 7$. The second stratum with considerably larger variance than the first receives the larger sample size. Interestingly, in this illustration, the sample sizes are in the opposite direction for proportional allocation. The *equal* allocation is closer to Neyman allocation than proportional allocation.

For Neyman allocation, $V(\bar{y}_1) = [(32 - 3)/32](0.79)/3 = 0.239$ and $V(\bar{y}_2) = [(11 - 7)/11](35.55)/7 = 1.8468$. Now, from (5.13), $V(\bar{Y}_{st}) = (32/43)^2(0.239) + (11/43)^2(1.8468) = 0.253$ and S.E.$(\hat{\bar{Y}}_{st}) = 0.503$. The same values for this variance and S.E. are obtained from (5.26).

As found in Example 5.2, for simple random sampling, the variance of the mean $\bar{y}$ for a sample of ten units is $V_S = 2.431$. Since $V_N/V_S = 0.253/2.431 = 0.104$, Neyman allocation reduces the variance almost by 90%. As found in Example 5.2, for proportional allocation, the variance of $\bar{Y}_{st}$ is $V_P = 0.613$. Now, $V_N/V_P = 0.253/0.613 = 0.413$. Hence, Neyman allocation decreases the variance from proportional allocation by about 58%.

For the $N = 43$ countries in Table 5.1, the precision of Neyman allocation relative to simple random sampling and proportional allocation can be further examined for different sizes of the overall sample. For simple random sampling, one can express the variance of $\bar{y}$ as $V_S = S^2/n - S^2/43 = S^2/n - 31.68/43 = S^2/n - 0.74$. For the variance in (5.23) for proportional allocation, $\Sigma_1^G W_g S_g^2 = 9.68$, and $V_P = 9.68/n - 9.68/43 = 9.68 - 0.23$. Since $\Sigma W_g S_g = 2.19$ and $\Sigma_1^G W_g S_g^2 = 9.68$, for Neyman allocation, from (5.26), $V_N = 4.80/n - 0.23$.

Table 5.3 presents these variances and the relative precisions for Neyman allocation for $n = 5$, 10, 15, and 20. The relative precisions are also presented in Figure 5.1. Notice from (5.24) that the precision of proportional allocation relative to simple random sampling is $V_S/V_P = 31.68/9.68 = 3.27$, and it is the same for any sample size.

Table 5.3. Variances for simple random sampling and proportional and Neyman allocations; relative precision of Neyman allocation.

Sample Size, n	V_S	V_P	V_N	V_S/V_N	V_P/V_N
4	7.180	2.190	0.965	7.4	2.3
6	4.540	1.383	0.567	8.0	2.4
8	3.220	0.980	0.368	8.8	2.7
10	2.428	0.738	0.248	9.8	3.0
12	1.900	0.577	0.168	11.3	3.4
14	1.523	0.461	0.111	13.7	4.1

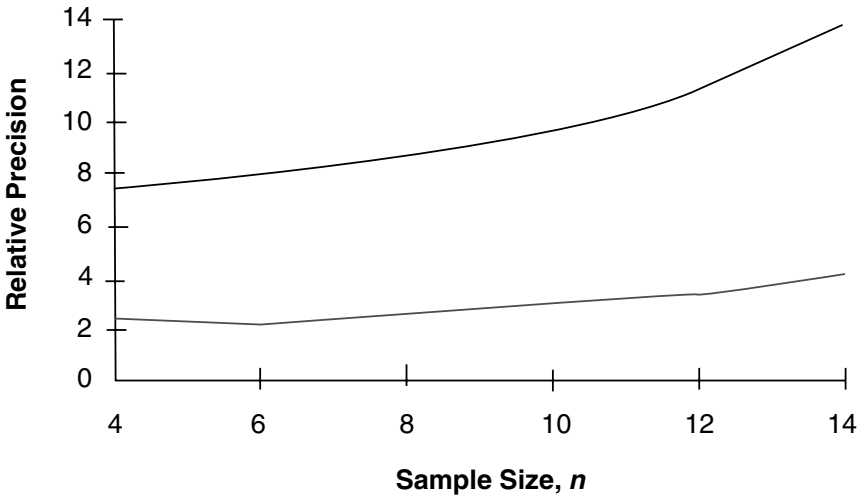

Figure 5.1. Precision of Neyman allocation relative to simple random sampling (top) and proportional allocation (bottom).

5.11 Summary on the precision of the allocations

The results from the comparisons of Sections 5.8 through 5.10 can be summarized as follows;

1. For a given sample size n, proportional and Neyman allocations estimate $\bar{Y}$ and Y with smaller variances than a simple random sample from the entire population.
2. Neyman allocation has much smaller variance than proportional allocation if S_g vary, and much smaller variance than simple random sampling if both $\bar{Y}_g$ and S_g vary. Information on the relative values of S_g is needed for Neyman allocation.
3. As $\bar{Y}_g$ vary, the precision of proportional allocation relative to simple random sampling increases.
4. An increase in n decreases the variances of the estimators for simple random sampling as well as for proportional and Neyman allocations and also for equal division of the sample size for the strata. As n increases, the precision of Neyman allocation relative to proportional allocation and simple random sampling increases, but it has no effect on the precision of proportional allocation relative to simple random sampling.

5. Distributing the sample size equally to the strata can be convenient in some situations. This procedure can have smaller variance than proportional allocation if it results in larger sample sizes for strata with larger sizes and standard deviations, as for Neyman allocation.

5.12 Sample size allocation to estimate proportions

To estimate the population proportion P or the total C, allocation of a given sample can be determined directly from the formulas in Section 5.7 or the results for the mean and total in Sections 5.8 and 5.9.

Proportional allocation

Since $n_g = nW_g$ for this allocation, from (5.20) or (5.23), the variance of $\hat{P}_{\text{st}}$ becomes

$$V_P(\hat{P}_{\text{st}}) = \frac{(1-f)}{n}\sum_1^G W_g \frac{N_g}{N_g - 1} P_g Q_g, \tag{5.30}$$

which is approximately equal to $(1 - f)\Sigma_1^G W_g P_g Q_g / n$.

Following the analysis in Section 5.8, proportional allocation of the sample will have smaller variance for $\hat{P}_{\text{st}}$ than equal distribution if n_g is larger for the strata with larger $S_g^2 = N_g P_g Q_g/(N_g - 1)$, that is, for P_g closer to 0.5 than zero or unity.

Neyman allocation

For this procedure, as in the case of the mean, the variance in (5.20) is minimized for a given $n = \Sigma n_g$. If N_g does not differ much from $(N_g - 1)$, this minimization results in

$$\frac{n_g}{n} = \frac{W_g\sqrt{P_g Q_g}}{\sum W_g\sqrt{P_g Q_g}}. \tag{5.31}$$

From (5.20) or (5.26), the minimum variance is approximately given by

$$V_N(\hat{P}_{\text{st}}) = \frac{\left(\sum W_g\sqrt{P_g Q_g}\right)^2}{n} - \frac{\sum W_g P_g Q_g}{N}. \tag{5.32}$$

For exact expressions when N_g is not large, replace P_gQ_g by $N_gP_gQ_g/(N_g - 1)$ in these equations.

Unlike proportional allocation, this method requires prior information regarding the unknown P_g. The sample sizes obtained from (5.31) do not differ much from proportional allocation even when P_g vary. For example, in the case of a population consisting of two strata, if $P_2 = 1 - P_1$, $n_1 = nW_1$ and $n_2 = nW_2$ from (5.25). Thus, the sample allocation in this case is identically the same as for proportional allocation, and the sample sizes do not depend on P_g. If valid prior information on the P_g is not available, proportional allocation can be recommended for practical situations.

5.13 Sample sizes to estimate means and totals

Sections 5.8 through 5.12, examined the distribution of the available overall sample size n for the strata. For practical situations, the sample size required for a specified criterion can be determined for each type of allocation.

To estimate $\bar{Y}_g$ and Y_g of the individual strata, one can find n_g needed with different criteria from the procedures in Sections 2.12. For proportional and Neyman allocations, procedures to determine the overall sample size n needed to estimate $\bar{Y}$ and Y with each of the following five criteria can be found:

1. The variance of $\hat{\bar{Y}}_{st}$ should not exceed a given value V.
2. The error of estimation $|\hat{\bar{Y}}_{st} - \bar{Y}|$ should not exceed a small positive quantity e except for a probability of α.
3. The relative error $|\hat{\bar{Y}}_{st} - \bar{Y}| / \bar{Y}$ should not exceed $\mathbf{a}$ except with probability α.
4. The confidence width for $\bar{Y}$ for a confidence probability of $(1 - \alpha)$ should not exceed a given value w.
5. The coefficient of variation of $\hat{\bar{Y}}_{st}$ should be smaller than C.

For proportional allocation, the solutions to these requirements are obtained by following the approaches in Section 2.12 with the expression in (5.23) for the variance. Now, for large N, the minimum sample sizes n_l required are obtained by multiplying $\Sigma W_g S_g^2$ by $(1/V)$, $(Z/\varepsilon)^2$, $(Z/\mathbf{a}\bar{Y})^2$, $(2Z/w)^2$, and $(1/C^2\bar{Y}^2)$ for the five criteria, respectively. Note that Z is the $(1 - \alpha)$ percentage point of the standard normal distribution, and if N is not very large, the sample size is given by $n = n_l/(1 + n_l/N)$.

Similarly, for Neyman allocation, the sample sizes are determined from the variance in (5.26). In this case, n_l is obtained by multiplying $(\Sigma W_g S_g)^2$ by the above factors and n is given by from $n_l/[1 + (n_l/N) \Sigma W_g S_g^2 / (\Sigma W_g S_g)^2]$.

Example 5.7. College tuitions: Consider estimating the average tuition for higher education for the colleges and universities in five states: New York, Pennsylvania, Massachusetts, New Jersey, and Connecticut. These states have relatively more institutions of higher education in the Northeast U.S. Baron's *Profiles of American Colleges* (1988) gives data on tuition, room and board expenses, and SAT (Scholastic Aptitude Test) scores for the educational institutions in the entire country.

The above five states together have $N = 329$ colleges and universities and they are classified as *Highly Competitive, Very Competitive*, and *Competitive*. These three categories consist of $N_1 = 55$, $N_2 = 74$, and $N_3 = 200$ institutions. Thus, the relative number of colleges in these three strata are $W_1 = 55/329 = 0.17$, $W_2 = 74/329 = 0.22$, and $W_3 = 200/329 = 0.61$.

From the above data, the averages $\bar{Y}_g$ of the tuitions in the three strata are found to be \$11.58, \$7.57, and \$6.13 thousand, respectively. The standard deviations S_g of the tuitions were \$0.77, \$1.24, and \$1.66 thousand, respectively. From the means of the strata, $\bar{Y} = (0.17)(11.58) + (0.22)(7.57) + (0.61)(6.13) = 7.37$. By using this as a preliminary estimate, sample size needed to estimate the mean for a specified S.E. can be found. For instance, consider the requirement that the S.E. of $\bar{Y}_{st}$ should not exceed 2.5% of the actual mean, that is, $7.37(0.025) = 0.18$ approximately.

From the above figures, $W_1 S_1^2 = 0.10$, $W_2 S_2^2 = 0.34$, and, $W_3 S_3^2 = 1.68$, and hence $\Sigma W_g S_g^2 = 2.12$. Similarly, $W_1 S_1 = 1.13$, $W_2 S_2 = 0.27$, $W_3 S_3 = 1.01$, and hence $\Sigma W_g S_g = 1.41$.

Now, for proportional allocation, $n_l = 2.12/(0.18)^2 = 65$ and $n = 65/(1 + 65/329) = 55$. The sample sizes for the strata are $n_1 = (0.17)(55) = 9$, $n_2 = (0.22)(55) = 12$, and $n_3 = (0.61)(55) = 34$.

Similarly, for Neyman allocation, $n_l = (1.41)^2/(0.18)^2 = 62$, and $n = 62/[1 + (62/329)(2.12)/(1.41)^2] = 52$. The sample sizes for the strata are $n_1 = (0.13/1.41)52 = 5$, $n_2 = (0.27/1.41)52 = 10$, and $n_3 = (1.01/1.41)52 = 37$.

5.14 Sample sizes to estimate proportions

As in the case of the mean and total, sample sizes to estimate P to satisfy the following requirements can be found:

1. The variance of $\hat{P}_{st}$ should not exceed V.
2. The margin of error $|\hat{P}_{st} - P|$ should be smaller than **e** except for a probability of α.

3. The relative error $|\hat{P}_{st} - P|/P$ should not exceed **a** except for a probability of α.
4. Confidence width for P with the confidence probability $(1 - \alpha)$ should not exceed w.
5. The coefficient of variation of $\hat{P}_{st}$ should be smaller than C.

For proportional allocation, from (5.30), the sample size n_l needed for the five prescriptions, respectively, are obtained by multiplying $\Sigma W_g P_g Q_g$ by $(1/V)$, $(Z/\mathbf{e})^2$, $(Z/\mathbf{a}P)^2$, $(2Z/w)^2$ and $(1/CP^2)$, and $n = n_l/(1 + n_l/N)$.

For Neyman allocation, from (5.31), n_l is obtained by multiplying $(\Sigma W_g \sqrt{P_g Q_g})^2$ by the above factors, and $n = n_l / [1 + (n_l / N)(\Sigma W_g P_g Q_g) / (\Sigma W_g \sqrt{P_g Q_g})^2]$.

If the strata sizes N_g are not large, multiply $P_g Q_g$ by $N_g/(N_g - 1)$ in the above solutions. If valid prior information on P_g is not available, all the solutions can be obtained with $P_g = 0.5$.

Note that the same solutions are obtained for the estimation of the total count $C = NP$ with the corresponding prescriptions. For example, the first solution is obtained for estimating C with a variance not exceeding N^2V.

Example 5.8. Prescribed S.E: For the energy consumption in Example 5.3, determine the sample size to estimate the proportion with the S.E. not exceeding 10%. Since $W_1 = 10/25$, $W_2 = 15/25$, $P_1 = 7/10$, and $P_2 = 8/15$.

From these figures, $W_1P_1Q_1 = 0.084$, $W_2P_2Q_2 = 0.1493$, and hence $\Sigma W_g P_g Q_g = 0.2333$. For proportional allocation, $n_l = 0.2333/0.10 = 24$ approximately, and $n = 24/[1 + (24/25)] = 12.2$ or 13 approximately. The sample sizes for the strata are $n_1 = 13(0.4) = 5$ and $n_2 = 13(0.6) = 8$.

Further, $W_1(P_1Q_1)^{1/2} = 0.1833$, $W_2(P_2Q_2)^{1/2} = 0.2993$, and $\Sigma W_g \sqrt{P_g Q_g} = 0.4826$. Now, for Neyman allocation, $n_l = (0.4826)^2/(0.10) = 23.29$ or 24 approximately. Now, $n = 1/[1 = (24/25)(0.2333/0.2329) = 12.23$ or 13 approximately. The sample sizes for the strata are $n_1 = 13(0.1833/0.4826) = 5$ and $13(0.2993/0.4826) = 8$.

5.15 Sample sizes for minimizing variance or cost

In the last two sections, expenses for collecting information from the sample units were not considered. The costs for obtaining information from a unit can be different for the strata. A cost function as suggested in the literature takes the form

$$E = e_0 + e_1 n_1 + \cdots + e_G n_G = e_0 + \sum e_g n_g, \qquad (5.33)$$

see Cochran (1977, Chap. 5), for example. In this linear function, e_0 represents the initial expenses for making arrangements for the survey, which may be ignored for some established surveys, e_g is the cost for interviewing a selected unit in the gth stratum, and E is the total allowable expenditure.

As outlined in Appendix A5, the sample sizes needed to minimize the variance in (5.13) for a given cost E, or for minimizing the cost in (5.33) for a given value V of the variance are obtained from

$$\frac{n_g}{n} = \frac{N_g S_g/\sqrt{e_g}}{\sum(N_g S_g/\sqrt{e_g})} = \frac{W_g S_g/\sqrt{e_g}}{\sum(W_g S_g \sqrt{e_g})}. \tag{5.34}$$

For the variance or the cost to be minimum, this optimum allocation suggests that larger sample sizes are needed for strata with larger sizes and standard deviations, and they can be large if the costs of sampling are not large. If e_g are the same for all the strata, (5.34) coincides with the Neyman allocation in (5.25).

Substituting (5.34) in (5.13), the minimum of the variance is

$$V_{\text{opt}}(\bar{y}_{\text{st}}) = \frac{\sum W_g S_g \sqrt{e_g} \sum (W_g S_g/\sqrt{e_g})}{n} - \frac{\sum W_g S_g^2}{N}. \tag{5.35}$$

For a given V, n is obtained from this expression. Similarly, substituting (5.34) in (5.33), the minimum cost is

$$E = e_0 + \frac{\sum W_g S_g \sqrt{e_g}}{\sum (W_g S_g/\sqrt{e_g})} n. \tag{5.36}$$

The sample size n for the available budget is obtained from this expression.

Example 5.9. Optimum allocation: For the illustration in Example 5.7, consider an initial expense of e_0 = \$400, and the expenses of e_1 = \$50, e_2 = \$50, and e_3 = \$45 to obtain information on a unit on tuition from the institutions in the three strata. With these costs, $W_1 S_1 \sqrt{e_1} = 0.92$, $W_2 S_2 \sqrt{e_2} = 1.93$, and $W_3 S_3 \sqrt{e_3} = 6.80$, and hence $\Sigma W_g S_g \sqrt{e_g} = 9.65$. Similarly, $W_1 S_1/\sqrt{e_1} = 0.0184$, $W_2 S_2/\sqrt{e_2} = 0.0386$, and $W_3 S_3/\sqrt{e_3} = 0.1510$, and hence $\Sigma(W_g S_g/\sqrt{e_g}) = 0.2079$.

If the S.E. of $\hat{\bar{Y}}_{st}$ should not exceed 0.18 as in Example 5.7, from (5.35), $9.65(0.21)/n - 2.13/329 = (0.18)^2$, and hence $n = 53$. The allocation in (5.34) results in $n_1 = (0.0184/0.2079)53 = 5$, $n_2 = (0.0386/0.2079)53 = 10$, and $n_3 = (0.1510/0.2079)53 = 38$. From (5.36), the total cost for the survey is equal to \$2860.

On the other hand, if the total cost should not exceed \$3000, for example, from (5.36), $400 + (9.65/0.2079)n = 3000$, and hence $n = 56$. Now, from (534), $n_1 = (0.0184/0.2079)56 = 5$, $n_2 = (0.0386/0.2079)56 = 10$, and $n_3 = (0.1510/0.2079)56 = 41$. In this case, the variance for $\hat{\bar{Y}}_{st}$ from (5.35) is equal to 0.0293, and hence S.E.$(\hat{\bar{Y}}_{st}) = 0.17$. With the increase of $3000 - 2860 =$ \$140, S.E.$(\hat{\bar{Y}}_{st})$ is reduced by 0.01, that is, by 0.01/0.18, or about 6%.

5.16 Further topics

Poststratification

The observations of a simple random sample of the students of a college can be classified into the freshman, sophomore, junior, and senior classes. The returns of a mail survey on the physicians in a region can be categorized according to their specialties. The responses of the public to a marketing or political survey can be classified into strata defined according to one or more of the characteristics such as age, sex, profession, family size, and income level.

Means, totals, and proportions of the above type of strata can be estimated from the sample units *observed* in the strata in a sample of size n drawn from the N population units. Since the sizes n_g of the observed samples are not fixed and they are random, there will be some loss in precision for the estimators if average variances are considered. This procedure, known as poststratification, is described below for estimating the means of the strata and the population.

The mean $\bar{y}_g$ of the n_g observed sample units can be used to estimate the stratum mean $\bar{Y}_g$. For an observed n_g, the mean $\bar{y}_g$ and variance s_g^2 are unbiased for $\bar{Y}_g$ and S_g^2. Further, $V(\bar{y}_g \mid n_g) = (N_g - n_g)S_g^2/N_g n_g$, which can be estimated from replacing S_g^2 by s_g^2. By replacing n_g by its expectation nW_g, an approximation to the average of this variance becomes $(1 - f)S_g^2/nW_g$.

With poststratification, the estimator of the population mean is

$$\hat{\bar{Y}}_{pst} = \sum_{1}^{G} W_g \bar{y}_g, \tag{5.37}$$

which is unbiased for $\bar{Y}$. Since $V(\hat{\bar{Y}}_{\text{pst}}) = \Sigma W_g^2 V(\bar{y}_g)$, replacing n_g by nW_g as above, the average variance of $\hat{\bar{Y}}_{\text{pst}}$ is approximately given by $(1 - f)\Sigma W_g S_g^2 / n$, which is the same as the variance in (5.23) for proportional allocation. Thus, for large sample sizes, $\hat{\bar{Y}}_{\text{pst}}$ has the same precision as $\hat{\bar{Y}}_{\text{st}}$ with proportional allocation of the sample.

Estimating differences of the stratum means

With samples of sizes n_1 and n_2 selected from two strata, an unbiased estimator of the difference of their means $(\bar{Y}_1 - \bar{Y}_2)$ is given by $(\bar{y}_1 - \bar{y}_2)$, which has the variance of $(N_1 - n_1)S_1^2/n_1 + (N_2 - n_2)S_2^2/n_2$. As shown in Appendix A5, for given $n = n_1 + n_2$, this variance is minimized if n_1 and n_2 are proportional to S_1 and S_2, respectively, that is, $(n_1/n) = S_1/(S_1 + S_2)$ and $(n_2/n) = S_2/(S_1 + S_2)$. In general, this division of the sample is not the same as the proportional or Neyman allocations described in Sections 5.8 and 5.9. However, it becomes the same as Neyman allocation if $N_1 = N_2$ and the same as proportional allocation if $S_1 = S_2$ in addition.

Two-phase sampling for stratification

When the sizes N_g are not known, Neyman (1938) suggests this procedure, which is also known as **double sampling** for stratification. In this procedure, an initial sample of large size n' is selected from the entire population and n_g', $g = (1, 2, \ldots, G)$, $\Sigma n_g' = n'$, units of the gth stratum are identified. Notice that $w_g = n_g'/n'$ is an unbiased estimator for $W_g = N_g/N$. At the second phase, samples of size n_g are selected from the n_g' observed units, the means and variances $(\bar{y}_g, s_g^2)$ are obtained, and the population mean $\bar{Y}$ is estimated from $\Sigma w_g \bar{y}_g$.

For repeated sampling at both phases, Cochran (1977, pp. 327–335) and J.N.K. Rao (1973) present expressions for the variance of $\Sigma w_g \bar{y}_g$ and the estimator for the variance. These authors and Treder and Sedransk (1993) describe procedures for determining the sample sizes n' and n_g.

Sample size determination for two or more characteristics

Several surveys are usually conducted to estimate the population quantities such as the means, totals, and proportions of more than one characteristic. Proportional or equal distribution of the available sample size can be considered for such surveys.

Neyman allocation based on the different characteristics can result in different allocations of the overall sample size. Yates (1960) suggests finding the sample sizes by minimizing the weighted average of the variances in (5.13) for all the characteristics; the weights are determined from the importance of the characteristics. Cochran (1977, pp. 119–123) averages the Neyman allocations of the different characteristics. Chatterjee (1968) and Bethel (1989) suggest alternative procedures for allocating the sample size.

Strata formation

Sections 1.8 and 5.1 presented illustrations of different types of strata considered for different purposes. If the main purpose of stratification is to estimate the population quantities such as the mean with high precision, as noted in Sections 5.8 and 5.9, the units within each stratum should be close to each other but the means of the strata should differ as much as possible. Ekman (1959) and Dalenius and Hodges (1959) suggest procedures for formation of the strata to estimate the mean or total. Cochran (1961) examines the difference procedures available for the division of a population into strata.

For the illustration in Example 5.1 to estimate rice production, the strata based on the levels of rice production have been constructed. To estimate the average tuition in Example 5.7, the different levels of competitiveness were considered to form the strata. In this case, an initial attempt to consider five U.S. states as the strata resulted in almost the same means and variances for all the strata. The reason for this outcome was that each state has some educational institutions with high and some with low tuitions.

Exercises

5.1. With the sample information in Example 5.1, (a) estimate the differences of the means and totals of the rice productions in the two strata, (b) find the S.E. of the estimates, and (c) find the 95% confidence limits for the differences.

5.2. In samples of size three and five selected from the two strata of Example 5.3, respectively, units (1, 6, 24) and (4, 10, 12, 15, 20) appeared. For the average and total consumption of the 25 countries, find (a) the estimates, (b) their S.E., and (c) the 95% confidence limits.

5.3. For the estimators of the average rice production in Example 5.1 that can be obtained through proportional and Neyman allocations of a sample of size $n = 16$, (a) find the variances and (b) compare their precisions relative to the mean of a simple random sample of the same size.

5.4. A total of 209 students responded to the survey conducted by the students in 1999, described in Section 1.8. The number of freshman, sophomore, junior, and senior students responded to the survey were 50, 30, 56, and 73. Among these students, 82, 80, 57, and 60%, respectively, had part-time employment on campus, about 11 hours a week for each group. The university consisted of 807, 798, 966, and 1106 students in the four classes, a total of 3677. For the 1605 freshman-sophomore and the 2072 junior–senior classes, find (a) estimates for the percentages having part-time employment and (b) The standard errors of the estimates. (c) Estimate the difference of the percentages for the two groups and find its S.E.

5.5. Divide the population of the 32 units in Table T2 in the Appendix into two strata with the math scores below 600 and 600 or more. For proportional and Neyman allocations of a sample of size 10, compare the variances of the stratified estimators for the means of the total (verbal + math) scores.

5.6. With the stratification in Exercise 5.5 and a sample of size 10, compare the variances of proportional and Neyman allocations for estimating the proportion of the 32 candidates scoring a total of 1100 or more.

5.7. For the 200 managers and 800 engineers of a corporation, the standard deviations of the number of days a year spent on research were presumed to be 30 and 60 days, respectively. Find the sample size needed for proportional allocation to estimate the population mean with the S.E. of the estimator not exceeding 10 and its allocation for the two groups.

5.8. With the information in Exercise 5.7, (a) find the sample size required if the minimum variance of the estimator for the difference of the means of the two groups should not exceed 200, and (b) find the distribution of the sample size for the groups.

5.9. Using the information in Example 5.7, find the sample sizes needed for proportional and Neyman allocations to estimate the average tuition if the width of the 95% confidence limits should not exceed \$1000.

5.10. To estimate the population proportion in Example 5.4 with an error not exceeding 12% except with a probability of 0.05, find the sample sizes required for proportional and Neyman allocations.

5.11. To estimate the average tuition in Example 5.9 when $e_0 = 500$, $e_1 = 50$, $e_2 = 100$, and $e_3 = 150$ and \$5000 is available for the survey, find the sample sizes for the strata and the minimum variance and S.E. of the estimator that can be obtained with these sample sizes.

5.12. As mentioned in Example 5.7, the standard deviations of the tuitions for the three types of colleges were \$0.77, \$1.24, and \$1.66 thousand. Similarly, for the expenses for room and board (R&B), the standard deviations were \$0.61, \$0.67, and \$0.83 thousand. (a) Find the standard errors of the estimators for the averages of the tuition and R&B expenses for a sample of size 50 allocated proportionally to the three types of colleges. (b) For the tuition, distribute the 50 units among the three types through Neyman allocation and find the S.E. of the estimator. (c) Similarly, distribute the 50 units among the three types for R&B through Neyman allocation and find the S.E. of the estimator.

5.13. *Project*. Consider the first five and next five states of Table T4 in the Appendix as two strata. From all the samples of size three selected independently from each of the strata, estimate (a) average of private enrollments, (b) difference of the averages of public and private enrollments, and (c) difference between the averages of the two strata for private enrollments. From the above estimates, find the expectations and variances of the corresponding estimators, and verify that they coincide with the exact expressions for the expectations and variances. Note that the expectations and variances for (a) and (b) can be obtained from the ${}_5C_3 + {}_5C_3 = 20$ estimates, but they should be obtained from the ${}_5C_3 \times {}_5C_3 = 100$ estimates.

5.14. Divide 47 of the states in Table T3 in the Appendix into two strata consisting of the 25 states with 2000 or more persons over 25 years old and the remaining 22 states.

(a) To estimate the average number of persons over 25 in the 47 states, find the S.E. of the stratified estimator for samples of size 5 from each of the strata. (b) For Neyman allocation of the sample of size 10, find the S.E. of the above estimator. (c) To estimate the above average for all the 51 entries, describe the procedure for combining the figures for the four largest states with the sample estimates in (a) or (b), and find the standard errors of the resulting estimators.

5.15. Express the sample variance s^2 obtained from a sample of n units as in (5.6) for population variance S^2.

5.16. For the observed sample size n_g in the gth stratum from a sample of n units from the entire population, show that $E(1/n_g) \geqslant 1/E(n_g) = nW_g$ (a) from the Cauchy–Schwartz inequality in Appendix A3 and (b) noting that the covariance of n_g and $1/n_g$ is negative. (c) Use this result to show that the approximation to the average variance of the poststratified estimator presented in Section 5.16 is smaller than the actual average.

Appendix A5

An expression for the variance

The numerator of (5.5) can be expressed as

$$\begin{aligned}\sum_1^G \sum_1^{Ng} (y_{gi} - \bar{Y})^2 &= \sum_1^G \sum_1^{Ng} [(y_{gi} - \bar{Y}_g) + (\bar{Y}_g - \bar{Y})]^2 \\ &= \sum_1^G \sum_1^{Ng} (y_{gi} - \bar{Y}_g)^2 + \sum_1^G N_g (\bar{Y}_g - \bar{Y})^2 \\ &= \sum_1^G (N_g - 1) S_g^2 + \sum N_g (\bar{Y}_g - \bar{Y})^2.\end{aligned}$$

The cross-product terms of the first expression on the right-hand side vanish since $\Sigma_1^{Ng}(y_{gi} - \bar{Y}_g) = 0$. The first and second terms of the last two expressions are the *within SS* and *between SS* (sum of squares). These two terms are the expressions for the variations within the strata and among the means of the strata, respectively.

Variance of the estimator for the mean

From (5.11), the variance of $\hat{\bar{Y}}_{\text{st}}$ is

$$V(\hat{\bar{Y}}_{\text{st}}) = \sum_g W_g^2 V(\bar{y}_g) + \sum_{g \neq j} \sum W_g W_j \operatorname{Cov}(\bar{y}_g, \bar{y}_j).$$

Since samples are selected independently from the strata, $E(\bar{y}_g \bar{y}_j) = E(\bar{y}_g)E(\bar{y}_j)$ and hence the covariance on the right side vanishes.

Neyman allocation

To find n_g that minimizes the variance in (5.13) for a given n, let

$$\Delta = \sum W_g^2 \left(\frac{1}{n_g} - \frac{1}{N_g}\right) S_g^2 + \lambda\left(\sum n_g - n\right),$$

where λ is the Lagrangian multiplier. Setting the derivative of Δ with respect to n_g to zero, $n_g = W_g S_g / \sqrt{\lambda}$. Since $n = \Sigma n_g = \Sigma W_g S_g / \sqrt{\lambda}$, the optimum value of n_g is given by (5.25).

This result can also be obtained from the Cauchy–Schwartz inequality in Appendix A3 by writing $a_g = W_g S_g / (n_g)^{1/2}$ and $b_g = (n_g)^{1/2}$.

Optimum allocation

The sample size required for minimizing the variance in (5.13) for a given cost is obtained by setting the first derivative of

$$\Delta_1 = V(\hat{\bar{Y}}_{\text{st}}) + \lambda\left(E - e_0 - \sum e_g n_g\right)$$

with respect to n_g to zero. Similarly, the sample size to minimize the cost for a specified value V of the variance can be found by setting the first derivatives of

$$\Delta_2 = e_0 + \sum e_g n_g + \lambda[V(\hat{\bar{Y}}_{\text{st}}) - V]$$

to zero. The solution for both these procedures is given by (5.34).

The solution can also be obtained from the Cauchy-Schwartz inequality by writing $a_g = W_g S_g / (n_g)^{1/2}$ and $b_g = (e_g n_g)^{1/2}$.