CHAPTER 3

Simple Random Sampling: Related Topics

3.1 Introduction

This chapter contains topics of general interest both in sample surveys and for any type of statistical analysis. The results presented in the appendix to this chapter are also of a general type, and they are widely applicable.

The last chapter demonstrated that the sample mean and variance are unbiased estimators for the population mean and variance. Several estimators in sample surveys take the form of the ratio of two sample means or they may be nonlinear in the observations. The property of unbiasedness may not hold for these types of estimators. In such cases, as described in the following section, in addition to the **variance** of an estimator, one can examine its **bias** and **mean square error** (MSE) for assessing its departure from the population quantity being estimated. The relevance of these concepts to the criteria of **precision**, **accuracy**, and **consistency** is described in Section 3.3.

Covariance and correlation between two characteristics and their sample means are examined in Section 3.4. In the next section, this topic is followed by the estimation of the mean and variance of a linear combination of two characteristics, the sum and difference in particular.

In several types of surveys, for convenience or otherwise, the sample is selected **systematically** from the population. Section 3.6 examines its similarity and difference from simple random sampling. In surveys on human populations, responses usually cannot be obtained from all the sampled units. The bias arising from **nonresponse** is briefly examined in Section 3.7. It is examined further in the exercises of some of the coming chapters, and Chapter 11 is entirely devoted to the various procedures available for counteracting the effects of nonresponse.

Chapter 2 noted that the procedures of confidence limits developed for infinite populations can be used as approximate procedures for finite populations. Section 3.8 considers the adoption of the classical **tests of hypotheses** for inferences regarding the means and totals of finite populations.

3.2 Bias, variance, and mean square error of an estimator

As an illustration, consider the sample standard deviation s to estimate the population standard deviation S. Although s^2 is *unbiased* for S^2, s is biased for S. The bias, variance, and MSE of the sample standard deviation s are

$$B(s) = E(s) - S, \tag{3.1}$$

$$V(s) = E[s - E(s)]^2 \tag{3.2}$$

and

$$\begin{aligned} \mathrm{MSE}(s) = E(s - S)^2 &= E[s - E(s) + E(s) - S]^2 \\ &= E[s - E(s)]^2 + [E(s) - S]^2 \\ &= V(s) + B^2(s). \end{aligned} \tag{3.3}$$

Note that the expectation of the cross-product term in the first expression of (3.3) vanishes.

The bias of an estimator in general can be zero, positive, or negative. In the first case, it is unbiased for the population quantity being estimated. It overestimates in the second case and underestimates in the last.

For the math scores in Table 2.1, the population standard deviation is $S = 60.67$. The expected value of the sample standard deviation obtained from the average of the last column of Table 2.2 is $E(s) = (35.36 + 14.14 + \cdots + 99.00)/15 = 49.03$. Thus the bias of s is $B(s) = 49.03 - 60.7 = -11.64$. This is an illustration of the general result that the sample standard deviation underestimates the population standard deviation.

In some instances, the bias becomes negligible for large sample sizes. For example, consider $\hat{S}^2 = \Sigma_1^n (y_i - \bar{y})^2/n = (n-1)s^2/n$ as an alternative estimator for S^2. Since $E(\hat{S}^2) = (n - 1)S^2/n$, the bias of $\hat{S}^2$ is equal to $-S^2/n$, which becomes small as n increases. Suitable adjustments can be made to reduce or eliminate the bias in some estimators. In this case, note that $n\hat{S}^2/(n - 1)$ is the same as s^2, which is unbiased.

Selection of a sample haphazardly is frequently characterized as biased sampling, and faulty measurements are described as biased observations. In statistics, bias refers to the estimation of a population quantity with its formal definition as in (3.1).

As defined in (3.2) and (3.3), the variance and MSE, respectively, are the averages of the squared deviation of an estimator from its **expected value** and the **actual value**. For the math scores, from Table 2.2, $V(s) = (35.36 - 49.03)^2 + (14.14 - 49.03)^2 + \cdots + (99 - 49.03)^2 = 1276.44$ and $\text{MSE}(s) = (35.36 - 60.67)^2 + (14.14 - 60.67)^2 + \cdots + (99 - 60.67)^2 = 1411.79$. Except for the rounding errors, this value for the MSE can also be obtained by adding the squared bias to the variance. Note also that $\text{S.E.}(s) = (1276.44)^{1/2} = 35.73$, and the square root of the MSE is $(1411.79)^{1/2} = 37.57$.

One can examine the severity of the bias of an estimator by comparing its absolute value with the actual population quantity being estimated. For the sample standard deviation of the math scores, this relative value is 11.63/60.67, which exceeds 17%. The absolute bias of an estimator can also be compared with its S.E. or the square root of its MSE. For the math scores, these relative values, respectively, are $11.73/35.73 = 0.33$ or 33% and $11.73/37.57 = 0.31$ or 31%, both of which are very high.

Since the actual population quantity of such S is not known, one cannot examine the bias as above. As will be seen in some of the coming chapters, it is possible to examine the bias from its expression. Procedures for reducing or completely eliminating the bias of an estimator are presented in Chapter 12.

3.3 Precision, accuracy, and consistency

The **precision** of an estimator is inversely proportional to its variance. As can be seen from Equation 2.12, the precision of the sample mean increases with the sample size.

The **accuracy** of an estimator is inversely proportional to its MSE. It is desirable that any estimator is highly accurate. As can be seen from Equation 3.3, estimators that are highly precise but have a large positive or negative bias cannot have a high accuracy and will not have much practical utility.

For infinite populations, an estimator is defined to be **consistent** if it approaches the population quantity being estimated as the sample size becomes large. From the Tschebycheff inequality in A2.2, note that an unbiased estimator is consistent if its variance becomes small. In the case of a finite population, as noted in Section 2.7, the sample mean, which is unbiased, approaches the population mean if the sample size becomes close to the population size.

3.4 Covariance and correlation

In almost every survey, information is collected on more than one characteristic of the population. Incomes of couples, test scores of students in two or more subjects, and employee sizes, sales, and profits of corporations at several periods of time are some of the examples. Let (x_i, y_i), $i = (1,\dots,N)$, represent two of these characteristics. One can compute the covariance and correlation of these characteristics from the ungrouped or grouped data.

Ungrouped data

Similar to the total and mean of y_i in (2.1) and (2.2), let X and $\bar{X}$ denote the population total and mean of x_i. With the definition in (2.4), let S_x^2 and S_y^2 denote the variances of x_i and y_i. The population covariance and correlation coefficient of these characteristics are defined as

$$S_{xy} = \frac{\sum_1^N (x_i - \bar{X})(y_i - \bar{Y})}{N-1} = \frac{\sum_1^N x_i y_i - N\bar{X}\bar{Y}}{N-1} \tag{3.4}$$

and

$$\rho = \frac{S_{xy}}{S_x S_y}. \tag{3.5}$$

This coefficient ranges from -1 to 1. For high positive and negative correlation between x and y, it will be close to $+1$ and -1, respectively, and for low correlation it will be close to 0. Notice that the covariance can also be expressed as $S_{xy} = \rho\, S_x S_y$.

For the verbal and math scores in Table 2.1, $S_{xy} = [(520 - 550)(670 - 670) + (690 - 550)(720 - 670) + \cdots + (480 - 550)(700 - 670)]/5 = 1920$.

Since $S_x = 76.42$ and $S_y = 60.66$, $\rho = 1920/(76.42 \times 60.66) = 0.41$. Thus, there is a positive correlation between the math and verbal scores, but it is not very high.

For a simple random sample of size n, let $(\bar{x}, \bar{y})$ and (s_x^2, s_y^2) denote the means and variances of (x_i, y_i). The sample covariance and correlation coefficient are

$$s_{xy} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{\sum_1^n x_i y_i - n\bar{x}\bar{y}}{n - 1} \tag{3.6}$$

and

$$r = \frac{s_{xy}}{s_x s_y}. \tag{3.7}$$

As in the case of the population correlation coefficient, r also ranges from -1 to $+1$. Note that the sample covariance can be expressed as $s_{xy} = r s_x s_y$. As will be seen in the next section, s_{xy} is unbiased for S_{xy}. However, r in (3.7) is biased for ρ.

If units (U_1, U_4) appear in a sample of size two from the population in Table 2.1, the sample means of the verbal and math scores are $\bar{x} = (520 + 580)/2 = 550$ and $\bar{y} = (670 + 720)/2 = 695$. The sample variances and covariance are $s_x^2 = [(520 - 550)^2 + (580 - 550)^2]/1 = 1800$, $s_y^2 = [(670 - 695)^2 + (720 - 695)^2]/1 = 1350$, and $s_{xy} = [(520 - 550)(670 - 695) + (580 - 550)(720 - 695)]/1 = 1500$. From (3.7), the correlation is $r = 1500/(1800 \times 1350)^{1/2} = 0.96$. For this sample, the correlation is much larger than the population correlation of 0.41.

Grouped data

For the 50 states plus the District of Columbia in the U.S., the number of persons (y) over 25 years (in millions) and the percent (x) of the population with four or more years of college are grouped in Table 3.1 into a joint distribution and presented in Figure 3.1.

One can denote the mid-values of x by x_i, and its marginal frequencies by f_i, $i = (1, 2,\ldots)$, $\Sigma f_i = N$. The mean and variance of x now are obtained from

$$\bar{X} = \sum x_i f_i / N$$

Table 3.1. Joint distribution of age and education in 1995 in the 50 states and DC.

Percent Graduates	Persons (millions) over 25 Years of Age (y)			
(x)	0–2	2–4	4–20	Total
14–18	3	3	0	6
18–22	9	5	4	18
22–26	6	3	5	14
26–30	5	2	2	9
30–34	1	2	1	4
Total	24	15	12	51

Source: The U.S. Bureau of the Census, Current Population Survey, March 1996, and the *New York Times Almanac,* 1998, p. 374.

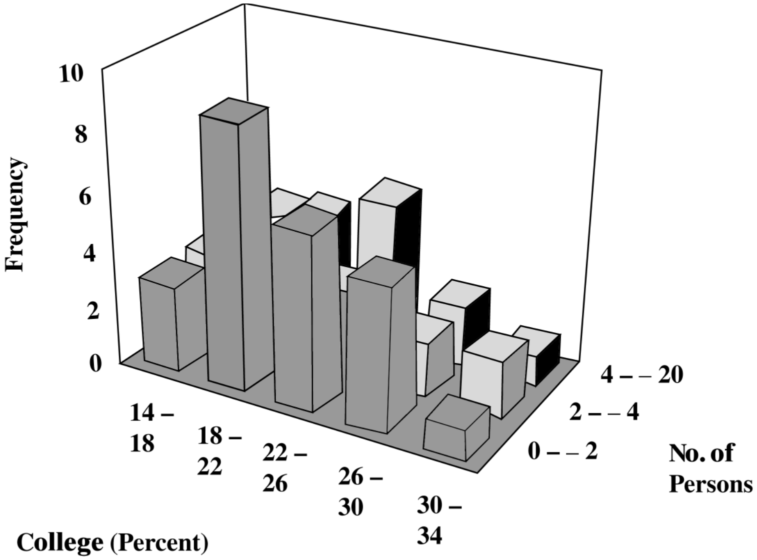

Figure 3.1. Joint distribution of the number of persons over 25 years of age and the percentage of college graduates in the 50 states and DC.

and

$$S_x^2 = \sum (x_i - \bar{X})^2 f_i/(N-1). \qquad (3.8)$$

The numerator of S_x^2 can be expressed as $\Sigma x_i^2 f_i - N\bar{X}^2$. For y, the mean $\bar{Y}$ and variance S_y^2 are obtained similarly. With the joint frequencies f_{ij},

Table 3.2. Arrangement of 40 population units into four systematic samples.

Samples			
1	2	3	4
1	2	3	4
5	6	7	8
⋮	⋮	⋮	⋮
37	38	39	40

the covariance of x and y is obtained from

$$S_{xy} = \sum\sum(x_i - \bar{X})(y_i - \bar{Y})f_{ij}/(N - 1) \tag{3.9}$$

The numerator of this covariance can be expressed as $\Sigma\Sigma x_i y_i f_{ij} - N\bar{X}\bar{Y}$.

For the data in Table 3.1, $\bar{X} = (16 \times 6 + 20 \times 18 + \cdots + 32 \times 4)/51 = 22.98$, $S_x^2 = [(16 - 22.98)^2 \times 6 + (20 - 22.98)^2 \times 18 + \cdots + (32 - 22.98)^2 \times 4]/50 = 20.38$, and $S_x = (20.38)^{1/2} = 4.51$. Similarly, $\bar{Y} = 4.18$, $S_y^2 = 19.95$, and $S_y = 4.47$. From Equation 3.9, $S_{xy} = [(16 - 22.98) \times (1 - 4.18) \times 3 + (16 - 22.98)(3 - 4.18) \times 3 + \cdots + (32 - 22.98)(12 - 4.18) \times 1]/50 = 2.50$. Now, $r = 2.50/(4.51 \times 4.47) = 0.12$.

The joint probability distribution of two random variables and related properties are presented in Appendix A3.

3.5 Linear combination of two characteristics

It is frequently of interest to estimate the mean or total of a sum, difference, or a specified linear combination of two characteristics, for example, the math and verbal scores. Linear combinations of incomes of couples, scores of students on an aptitude test taken on two occasions, and sales or profits in two successive years provide other illustrations.

Linear combination

For each of the N population units, consider a linear combination of x_i and y_i, $g_i = ax_i + by_i + c$, where a, b, and c are chosen constants. This type of linear combination is of importance in several practical situations.

For example, the applicants for admission to a college may be ranked by an index of the SAT verbal and math scores with $a = 5/8$, $b = 3/8$, and $c = -500$.

The population total and mean of g_i are $G = aX + bY + c$ and $\overline{G} = (a\Sigma_1^N x_i + b\Sigma_1^N y_i + Nc)/N = a\overline{X} + b\overline{Y} + c$. As shown in Appendix A3, the variance of g_i is given by

$$S_g^2 = V(ax_i + by_i) = a^2S_x^2 + b^2S_y^2 + 2abS_{xy}. \tag{3.10}$$

From a simple of observations (x_i, y_i), $i = 1, 2, \ldots, n$, the sample mean of g_i is $\bar{g} = a\bar{x} + b\bar{y} + c$, and its variance is given by

$$s_g^2 = a^2s_x^2 + b^2s_y^2 + 2abs_{xy}. \tag{3.11}$$

As in the case of $\bar{x}$ and $\bar{y}$, for simple random sampling, $\bar{g}$ and s_g^2 are unbiased for $\overline{G}$ and S_g^2, respectively. Now, $E(s_g^2) = a^2E(s_x^2) + b^2E(s_y^2) + 2abE(s_{xy})$, which is the same as S_g^2. Equating this expression to the right-hand side of (3.10), yields $E(s_{xy}) = S_{xy}$, that is, the sample covariance s_{xy} is unbiased for the population covariance S_{xy}. This result can also be obtained directly from the expression of s_{xy}.

Further, the variance of $\bar{g}$ is

$$V(\bar{g}) = (1 - f)s_g^2/n, \tag{3.12}$$

and an unbiased estimator of this variance is obtained from $v(\bar{g}) = (1 - f)s_g^2/n$.

For the above index of the SAT scores, from Table 2.1, $\overline{G} = (5/8)(550) + (3/8)(670) - 500 = 95$, and $S_g^2 = (5/8)^2(5840) + (3/8)^2(3680) + 2(5/8) \times (3/8)(1920) = 3698.75$.

If the first and fourth units are selected into a sample of size two, $\bar{g} = (5/8)(550) + (3/8)(695) - 500 = 104.375$, and $s_g^2 = (5/8)^2(1800) + (3/8)^2(1350) + 2(5/8)(3/8)(1500) = 1596.09$.

From (3.12), $V(\bar{g}) = (6 - 2)(3698.75)/12 = 1232.92$ and hence S.E.$(\bar{g}) = 35.11$. The sample estimate of this variance is $(6 - 2)(1596.09)/12 = 532.03$, and hence the sample S.E. of $\bar{g}$ is equal to 23.07.

Note that all the sample observations are not needed to find $\bar{g}$ since it is the same as $a\bar{x} + b\bar{y} + c$. Similarly, the S.E. of $\bar{g}$ can be obtained by first finding s_g^2 from (3.11) as above.

The sum

The population and sample means and variances for the sum $t_i = x_i + y_i$ can be obtained directly or from the corresponding results for g_i in (3.10) with $a = 1$, $b = 1$, and $c = 0$. The total and mean of t_i are $T = X + Y$ and $\bar{T} = \bar{X} + \bar{Y}$. The population variance of t_i is

$$S_t^2 = \sum_1^N (t_i - \bar{T})^2/(N-1) = S_x^2 + S_y^2 + 2S_{xy}. \tag{3.13}$$

The sample mean $\bar{t} = \bar{x} + \bar{y}$ is unbiased for $\bar{T} = \bar{X} + \bar{Y}$. Its variance is given by $V(\bar{t}) = (1-f)S_t^2/n$. An unbiased estimator of this variance is $v(\bar{t}) = (1-f)s_t^2/n$, where

$$s_t^2 = \sum_1^n (t_i - \bar{t})^2/(n-1) = s_x^2 + s_y^2 + 2s_{xy}. \tag{3.14}$$

For the sum of the verbal and math scores in Table 2.1, the mean is 1220. If the first and fourth units are selected in a sample of size two, $\bar{t} = (550 + 695) = 1245$. From (3.13), $S_t^2 = 5840 + 3680 + 2(1920) = 13{,}360$. Hence, $V(\bar{t}) = 13{,}360/3 = 4453.33$ and S.E.$(\bar{t}) = 66.73$.

With the first and fourth units in the sample, the variance in (3.14) is $s_t^2 = 1800 + 1350 + 2(1500) = 6150$. Now, the sample estimate of the variance is $v(\bar{t}) = 6150/3 = 2050$, and hence the sample S.E. of $\bar{t}$ is equal to 45.28. This sample estimate is much smaller than the actual standard error of 66.73.

Note that S_{xy} in (3.13) can be replaced by $\rho S_x S_y$, and s_{xy} in (3.14) by $r s_x s_y$.

The variances of t_i and $\bar{t}$ will be larger when x and y are positively correlated than when they are uncorrelated or negatively correlated.

The difference

Denote the difference of the two characteristics by $d_i = x_i - y_i$, $i = 1, 2, \ldots, N$. An unbiased estimator for the mean $\bar{D} = \bar{X} - \bar{Y}$ is $\bar{d} = \bar{x} - \bar{y}$. The population variance of this difference is

$$S_d^2 = \sum_1^N (d_i - \bar{D})^2/(N-1) = S_x^2 + S_y^2 - 2S_{xy}. \tag{3.15}$$

and its unbiased estimator is

$$s_d^2 = \sum_1^n (d_i - \bar{d})^2/(n-1) = s_x^2 + s_y^2 - 2s_{xy}. \tag{3.16}$$

The variance of $\bar{d}$ is $V(\bar{d}) = (1-f)S_d^2/n$, which is estimated from $v(\bar{d}) = (1-f)s_d^2/n$.

For the population in Table 2.1, $\bar{D} = 550 - 670 = -120$ and the variance in (3.15) is $S_d^2 = 5840 + 3680 - 2(1920) = 5680$. For samples of size two, the variance of $\bar{d}$ is 5680/3 = 1893.33, and hence it has a standard error of 43.51.

If units (U_1, U_4) appear in the sample, $\bar{d} = 550 - 695 = -145$, and $s_d^2 = 1800 + 1350 - 2(1500) = 150$. The sample variance of $\bar{d}$ is 150/3 = 50, and hence the S.E. of $\bar{d}$ becomes 7.1. This sample estimate is considerably smaller than the actual standard error of 43.51.

The variances of d_i and $\bar{d}$ will be larger when x and y are negatively correlated than when they are uncorrelated or positively correlated.

Covariance and correlation of the sample means

Note that

$$V(\bar{t}) = V(\bar{x} + \bar{y}) = V(\bar{x}) + V(\bar{y}) + 2\ \mathrm{Cov}\ (\bar{x}, \bar{y}). \tag{3.17}$$

This variance however is the same as $V(\bar{t}) = (1-f)S_t^2/n = (1-f) \times (S_x^2 + S_y^2 + 2S_{xy})/n$. Thus, $\mathrm{Cov}(\bar{x}, \bar{y}) = (1-f)S_{xy}/n$, which is estimated from $(1-f)s_{xy}/n$. This result can also be obtained from the expressions for the variance of $\bar{g}$ and its estimate described in Section 3.5 or of $\bar{d}$ in Section 3.5. When $n = 2$, for the math and verbal scores in Table 2.1, $\mathrm{Cov}(\bar{x}, \bar{y}) = 1920/3 = 640$. If the first and fourth units appear in the sample, an estimate of this covariance is 1500/3 = 500.

The correlation of the sample means is $\mathrm{Cov}(\bar{x}, \bar{y})/[V(\bar{x})V(\bar{y})]^{1/2}$, which is the same as ρ.

3.6 Systematic sampling

Sample selection

In the common practice for selecting a systematic sample, the population is divided into k groups of size $n = N/k$ in each. One unit is chosen randomly from the first k units and every kth unit following it is included in the sample. If r is the random number drawn from the first group, units numbered $r + uk$, $u = (0, 1,\ldots, n - 1)$ constitute the sample. This procedure results in the **1 in *k* systematic sample**.

Table 3.2 presents $N = 40$ units arranged in $k = 4$ groups of $n = 10$ units in each. If the random number selected from the first four numbers is 2, for example, the sample would consist of the units (2, 6, 10, …, 38). Selecting a systematic sample as above is equivalent to selecting randomly one of the k groups.

If N/k is not an integer, the sizes for the initial and latter samples will be equal to $[N/k] + 1$ and $[N/k]$, where $[N/k]$ is the integer closest to N/k. For instance, if $N = 39$ and $k = 4$, the first three samples will have size 10 and the last one, 9.

It is very convenient to draw systematic samples from telephone and city directories, automobile registries, and electoral rolls. A sample of households from a residential neighborhood can be easily obtained by this method. It also becomes convenient to draw a sample systematically when the population frame is not available. For example, selection of every 10th farm on the spot from the approximately 500 farms in a village will provide a sample of 50 farms. In industrial quality control, a sample of items produced, for example, every 30 minutes or every tenth batch of manufactured items is inspected. In some marketing and political surveys, interviews are attempted from, for example, every tenth person passing a certain location. For surveys supplementing censuses, systematic sampling is frequently employed. In large-scale multistage surveys, samples are selected systematically at the different stages.

Expectation and variance

For the systematic sampling procedure, the observations of the population can be denoted by y_{ij}, $i = 1,\ldots, k$ and $j = 1,\ldots, n$. Since each of the k samples has a $1/k$ chance of being selected, the expectation of the sample mean $\bar{y}_{sy} = \bar{y}_i$ is $\overline{Y}$ and hence it is unbiased. Its variance is given by

$$V(\bar{y}_{sy}) = \frac{1}{k}\sum_{1}^{k}(\bar{y}_i - \overline{Y})^2, \qquad (3.18)$$

which will be small if the variation among $\bar{y}_i$ is not large. Preliminary information on y_{ij} can be used to determine k and n for reducing this variance.

Since only ${}_nC_2 = n(n - 1)/2$ of the ${}_NC_2 = N(N - 1)/2$ pairs of the population units are given an equal chance for being selected, an unbiased

Table 3.3. Systematic samples of the savings ($1000) of 40 employees.

	Samples			
	1	2	3	4
	23	30	19	38
	23	38	19	18
	43	23	14	17
	40	17	40	23
	20	20	24	14
	14	22	24	14
	17	12	30	14
	20	14	15	14
	15	29	10	10
	16	22	16	12
Means	23.1	22.7	21.1	17.4
Variances	104.1	62.01	77.66	65.16

estimator of the variance in (3.18) cannot be found. Approximate procedures for this variance are available in the literature; see Cochran (1977, Chap. 8) and Section 3.9.4 below.

Example 3.1. Systematic sampling to estimate employee savings: The savings ($10,000) of 40 employees are arranged in Table 3.3 into $k = 4$ groups of size $n = 10$ in each. In actual sampling, one of the four samples will be selected randomly.

The population mean and variance of the savings are $\bar{Y} = 21.075$ and $S^2 = 76.48$. From (3.18), $V(\bar{y}_{sy}) = [(23.1 - 21.075)^2 + (22.7 - 21.075)^2 + (21.1 - 21.075)^2 + (17.4 - 21.075)^2]/4 = 5.062$, and hence S.E.$(\bar{y}_{sy}) = 2.25$. For simple random sampling, $V(\bar{y}) = (1 - 0.25)(76.48)/10 = 5.736$, and hence S.E.$(\bar{y}) = 2.4$. The gain in precision for $\bar{y}_{sy}$ relative to $\bar{y}$ is $(5.736 - 5.062)/5.062 = 0.13$ or 13%.

An unbiased estimator for the population total is obtained from $\hat{Y} = N\bar{y}_{sy}$, and S.E.$(\hat{Y}) = N$ S.E.$(\bar{y}_{sy})$. For the above example, the standard errors of $\hat{Y}$ for systematic and simple random sampling, respectively, are $40(2.25) = 90$ and $40(2.4) = 96$.

Comparison with simple random sampling

The variance within the systematic samples is

$$S^2_{wsy} = \frac{\sum_1^k \sum_1^n (y_{ij} - \bar{y}_i)^2}{k(n-1)} = \frac{1}{k}\sum_1^k s_i^2. \qquad (3.19)$$

Note that the **within-variance** s_i^2 of the ith sample has $(n - 1)$ d.f., and (3.19) is obtained by pooling the k sample variances.

From the expression for S^2 in Appendix A3,

$$V(\bar{y}_{sy}) = \frac{N-1}{N}S^2 - \frac{(n-1)}{n}S^2_{wsy}. \tag{3.20}$$

This variance decreases as the within-variance increases. If the units in a systematic sample vary, they represent the population units adequately. If n increases, the multiplier $-(n - 1)/n$ decreases but S^2_{wsy} may increase. As a result, the variance in (3.20) need not decrease with an increase in the sample size; see Exercise 3.8.

From Equation 3.20, the difference in the variances of the mean $\bar{y}$ of a simple random sample and $\bar{y}_{sy}$ is

$$\begin{aligned} V(\bar{y}) - V(\bar{y}_{sy}) &= \left(\frac{N-n}{Nn} - \frac{N-1}{N}\right)S^2 + \frac{(n-1)}{n}S^2_{wsy} \\ &= \frac{n-1}{n}(S^2_{wsy} - S^2). \end{aligned} \tag{3.21}$$

Hence, $\bar{y}_{sy}$ will have higher precision than $\bar{y}$, if the within variance S^2_{wsy} is larger than the population variance S^2.

For the four samples in Example 3.1, the within variances s_i^2 equal 104.1, 62.01, 77.66, and 65.16 as presented in Table 3.3. Hence, $S^2_{wsy} = 77.23$, which is only slightly larger than the population variance $S^2 = 76.48$. This was the reason for the gain in precision for the mean of the systematic sample relative to the mean of the simple random sample to be as small as 13%.

Further properties

As noted in the above two sections, the precision of the systematic sample mean increases if $\bar{y}_i$ do not vary much but s_i^2 differ from each other. These two objectives can be met through any prior information available on the population units.

If the observations of the population units are in an ascending or descending order, that is, if there is a linear trend, Cochran (1977, pp. 215–216) shows that $V(\bar{y}_{sy})$ is smaller than $V(\bar{y})$. If the observations, however, are in a random order, both the procedures can be expected to have almost the same precision. In this case, the variance in (3.18) can be estimated from $v(\bar{y}) = (1 - f)\, s^2/n$ in (2.14). For the random ordering, Madow and Madow (1944) show that the averages of $V(\bar{y}_{sy})$ and $V(\bar{y})$ over the $N!$ permutations of the observations are the same.

3.7 Nonresponse

In almost every survey, some of the units selected in the sample do not respond to the entire survey or to some of the items in the questionnaire. The different reasons for nonresponse and the various procedures for compensating for its effects are described in detail in Chapter 11. This section briefly examines the bias and MSE arising from nonresponse.

To analyze this situation, consider the population of size N to consist of the responding and nonresponding groups of sizes N_1 and $N_2 = N - N_1$, respectively. Denote the totals, means, and variances of these groups by (Y_1, Y_2), $(\bar{Y}_1, \bar{Y}_2)$, and (S_1^2, S_2^2), respectively. The total of the N population units is $Y = Y_1 + Y_2$. The population mean can be expressed as

$$\begin{aligned}\bar{Y} &= Y/N = (Y_1 + Y_2)/N \\ &= (N_1\bar{Y}_1 + N_2\bar{Y}_2)/N = W_1\bar{Y}_1 + W_2\bar{Y}_2, \end{aligned} \tag{3.22}$$

where $W_1 = N_1/N$ and $W_2 = N_2/N = 1 - W_1$ are the proportions of the units in the two groups.

In a sample of size n selected randomly from the N units, n_1 responses and $n_2 = n - n_1$ nonresponses will be obtained. The probability of (n_1, n_2) is given by

$$P(n_1, n_2) = ({}_{N_1}C_{n_1})({}_{N_2}C_{n_2})/({}_{N}C_{n})\,. \tag{3.23}$$

Denote the sample mean and variance of the n_1 responses by $\bar{y}_1$ and s_1^2. Note that for the n_2 nonrespondents, the sample mean and variance, $\bar{y}_2$ and s_2^2, are not available.

The sample mean $\bar{y}_1$ is unbiased for $\bar{Y}_1$. However, for estimating the population mean, it has a bias of

$$\begin{aligned}B(\bar{y}_1) &= E(\bar{y}_1) - \bar{Y} = \bar{Y}_1 - \bar{Y} \\ &= \bar{Y}_1 - (W_1\bar{Y}_1 + W_2\bar{Y}_2) = W_2(\bar{Y}_1 - \bar{Y}_2). \end{aligned} \tag{3.24}$$

This **bias** is positive if $\bar{Y}_1$ is larger than $\bar{Y}_2$, and negative otherwise.

The absolute value of this bias depends on the difference of these means and on the size N_2 of the nonresponse group, but not on the sample size n.

The variance of $\bar{y}_1$ is $V(\bar{y}_1) = (1 - f_1)\ S_1^2/n_1$, and its MSE is given by

$$\mathrm{MSE}(\bar{y}_1) = V(\bar{y}_1) + B^2(\bar{y}_1)\,. \tag{3.25}$$

The number of respondents n_1 may increase with an increase of the sample size n. As a result, $V(\bar{y}_1)$ and $\mathrm{MSE}(\bar{y}_1)$ may decrease.

3.8 Inference regarding the means

Statistical tests of hypotheses available for large populations can also be used as approximate procedures for sampling from finite populations, unless the population and sample sizes are too small. The following subsections examine these tests for the mean of a population and the difference of the means of two populations.

Population mean

Consider the statement that the $N = 2000$ freshman and sophomore (FS) students in a college use the campus computer facilities on an average at least 6 hours a week. One can examine this statement from a sample of, for example, $n = 25$ students. For these selected students, consider the sample mean $\bar{y} = 8$ and variance $s^2 = 15$. Now, $v(\bar{y}) = [(2000 - 25)/2000](15/25) = 0.5925$, and hence $\mathrm{S.E.}(\bar{y}) = 0.77$, which is almost 10% of the sample mean.

One may examine the null hypothesis, $\bar{Y} \leq 6$ against the **alternative hypothesis**, $\bar{Y} > 6$, assuming that approximately $Z = (\bar{y} - \bar{Y})/\mathrm{S.E.}(\bar{y})$ follows the standard normal distribution. Since $Z = 1.65$ for the significance level $\alpha = 0.05$, the 95% one-sided lower confidence limit for $\bar{Y}$ is given by $\bar{y} - 1.65\ \mathrm{S.E.}(\bar{y}) \leq \bar{Y}$, that is, $\bar{Y} \geq 8 - 1.65(0.77) = 6.73$. From this result, the null hypothesis is rejected at the 5% level of significance, and credence is given to the statement that the FS group utilizes the computer facilities for at least 6 hours a week.

Alternatively, with the null hypothesis value for $\bar{Y}$, one can compute $Z = (\bar{y} - \bar{Y})/\mathrm{S.E.}(\bar{y})$, which is equal to $(8 - 6)/0.77 = 2.6$. Since this value exceeds 1.65, the null hypothesis is rejected at $\alpha = 0.05$ and the same inference as above is reached.

One can also formally examine the above hypotheses from finding the ***p*-value**, which is the probability of observing a sample mean of 8 or more when the null hypothesis is true, that is, $\overline{Y} \leq 6$. This probability is the same as $\Pr(Z \geq 2.6)$, which is equal to 0.0047 from the tables of the standardized normal distribution. Since this probability is small, the null hypothesis is rejected.

Since the population size is not too small, one may examine the above hypotheses through the t-distribution. From the tables of this distribution $t_{24} = 1.7109$ for $\alpha = 0.05$. The lower confidence limit for $\overline{Y}$ is $8 - 1.7109(0.77) = 6.68$. Also, $t = (8 - 6)/(0.77) = 2.6$ exceeds 1.7109. The p-value now is $\Pr(t_{24} > 2.6)$, which is close to 0.0082. Each of these three procedures comes to the same conclusion as above with normal approximation.

Difference of two population means

Continuing with the illustration in the last section, consider the statement that the average number of hours of usage of the computer facilities for the FS group is different from that of the 1800 junior and senior (JS) students. With the subscripts 1 and 2 representing the FS and JS groups, one can denote their means and variances by $(\overline{Y}_1, \overline{Y}_2)$ and (S_1^2, S_2^2). To examine the above statement, the null and alternative hypotheses, respectively, are $\overline{Y}_1 = \overline{Y}_2$ and $\overline{Y}_1 \neq \overline{Y}_2$.

With $D = \overline{Y}_1 - \overline{Y}_2$, one can express these hypotheses as $D = 0$ and $D \neq 0$. Notice that the alternative hypothesis is **two sided**, whereas it was **one sided** in the last section.

For the $N_1 = 2000$ students of the FS group, a sample of size $n_1 = 25$ with $\bar{y}_1 = 8$ and $s_1^2 = 15$ has been considered. For the JS group, consider an independent sample of size $n_2 = 20$ from the $N_2 = 1800$ students with mean $\bar{y}_2 = 10$ and variance $s_2^2 = 12$. The estimate of $V(\bar{y}_2)$ is $v(\bar{y}_2) = [(1800 - 20)/1800](12/20) = 0.5933$.

From the above two samples, an estimate of D is $d = \bar{y}_1 - \bar{y}_2 = -2$. Since $V(d) = V(\bar{y}_1) + V(\bar{y}_2)$, its estimate is $v(d) = v(\bar{y}_1) + v(\bar{y}_2) = 0.5925 + 0.5933 = 1.1858$, and hence $\text{S.E.}(d) = 1.09$. As an approximation, $Z = (d - D)/\text{S.E.}(d)$ follows the standard normal distribution. Now, 95% lower and upper confidence limits for D are $-2 - 1.96(1.09) = -4.14$ and $-2 + 1.96\ (1.09) = 0.14$. Since these limits enclose zero, the null hypothesis value for D, at the 5% level of significance, the null hypothesis that the average number of hours of usage is the same for both the FS and JS groups is not rejected.

Alternatively, $Z = -2/1.09 = -1.83$, which is larger than -1.96. Hence at the 5% level of significance, the null hypothesis is again not rejected. From the tables of Z, the p-value is equal to 0.0672, which may be considered to be large to reject the null hypothesis.

If both the populations are considered to be large with the same variance, the pooled estimate of the variance is obtained from $s^2 = [(n_1 - 1) \times s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$ d.f. with $(n_1 + n_2 - 2)$ d.f. In this case, the variance of d is obtained from $v(d) = s^2(1/n_1 + 1/n_2)$, and the S.E.($d$) is obtained from the square root of this variance. Now, $t = d/\text{S.E.}(d)$ follows the t-distribution with $(n_1 + n_2 - 2)$ d.f, and the above procedures are followed with this approximation.

Correlated samples

Now consider the statement that after improving the computer facilities, the average usage for the FS group increased by at least 2 hours a week.

Denote the number of hours of usage for the $N = 2000$ students initially by y_{1i} and after the improvements by y_{2i}, $i = 1, 2,\ldots, N$. With this notation, the mean and variance initially are $\bar{Y}_1$ and S_1^2. Similarly, the mean and variance after the improvements are $\bar{Y}_2$ and S_2^2. The increase for each student is $d_i = y_{2i} - y_{1i}$, and for the N students the mean of this difference is $D = \bar{Y}_2 - \bar{Y}_1$. The variance of d_i is $S_d^2 = S_2^2 + S_1^2 - 2S_{12}$.

The null and alternative hypotheses now are $D \leq 2$ and $D > 2$. For the $N = 2000$ students of the FS group, a sample of size $n = 25$ with the mean and variance $\bar{y}_1 = 8$ and $s_1^2 = 15$ has been considered. For the same sample, after the improvements, consider the mean $\bar{y}_2 = 11.5$ and variance $s_2^2 = 18$, and the covariance $s_{12} = 12$ for the observations before and after the improvements. From these figures, the estimate of D is given by $d = \bar{y}_2 - \bar{y}_1 = 3.5$, and the estimate of S_d^2 by $s_d^2 = s_2^2 + s_1^2 - 2s_{12} = 11$. The variance of d is $V(d) = (1 - f)S_d^2/n$, where $f = n/N$. From the sample observations, an estimate of this variance is $v(d) = (1 - f)s_d^2/n = (79/80)11/25 = 0.4345$, and hence S.E.($d$) = 0.66.

Now, with $\alpha = 0.05$, the one-sided lower confidence limit for D is given by $D \geq 3.5 - 1.65\,(0.66) = 2.41$. Hence, it may be inferred that for the FS group, the average usage of the computer facilities has increased at least by 2 hours a week. Since $Z = (d - 2)/\text{S.E.}(d) = 2.3$, which exceeds 1.65, the same conclusion is reached from this standardized value. From the tables of Z, the p-value in this case is close to 0.01.

The t-distribution in this case will have 24 d.f. as in Section 3.8, and the p-value now is close to 0.0168.

3.9 Sampling with replacement

If a sample of n units is selected randomly and **with replacement** from the population, any unit U_i, $i = (1, 2, \dots, N)$, may appear more than once in the sample. This procedure is not practical for sampling from a finite population. However, one can compare its properties with sampling randomly without replacement, presented in Section 2.5.

For this procedure, there are N^n possible samples. The probabilities of selection of the units follow the binomial distribution with n trials and the probability of selection of a unit equal to $1/N$ at any trial. The sample mean $\bar{y}$ is unbiased for $\bar{Y}$ as in the case of the without replacement procedure, but $V(\bar{y}) = \sigma^2/n = (N - 1)S^2/Nn$. An unbiased estimator of this variance is $v(\bar{y}) = s^2/n$; see Exercises 3.19 and 3.20.

Notice that the variance of $\bar{y}$ for this method is larger than the variance in (2.12) for the without replacement procedure. The difference between these procedures becomes small if the population size is large. By deleting the duplications, the sample mean obtained from the **distinct** units will be unbiased for $\bar{Y}$, and its variance will be smaller than $(N - 1)S^2/Nn$.

Exercises

3.1. To estimate the mean of a population of 3000 units, consider increasing the sample size from 5 to 10%. For the sample mean, find the relative (a) decrease in the variance, (b) decrease in the standard error, and (c) decrease in the coefficient of variation and (d) increase in the precision. (e) What will be the relative decrease in the confidence width for a specified probability? (f) What will be the changes in (a) through (e) for the estimation of the population total?

3.2. (a) Find the means, variances, covariance, and correlation of the 30 verbal and math scores in Table T2 in the Appendix. (b) Group the data into a bivariate distribution with classes 350 to 470, 470 to 590, and 590 to 710 for the verbal scores and 400 to 520, 520 to 640, and 640 to 760 for the math scores. Find the means, variances, covariance, and correlation from this distribution, and compare them with the results in (a).

3.3. From the information in Exercise 2.1, for the difference in the means of the largest and middle-sized corporations, find (a) the estimate, (b) its S.E., and (c) 95% confidence limits. (d) Explain whether the procedure used for (a) provides an unbiased estimator.

3.4. From the information in Exercise 2.1, for the ratio of the means of the middle-sized to the smallest corporations, find (a) the estimate, (b) its S.E., and (c) the 95% confidence limits. (d) Explain whether the procedure used for (a) provides an unbiased estimator.

3.5. For a random sample of 20 from 2000 candidates, the averages and standard deviations of the SAT verbal scores at the first and second attempts were (520, 150) and (540, 160), respectively. The sample correlation coefficient of the scores on the two occasions was 0.6. For the increase in the average score, find (a) the estimate, (b) its S.E., and (c) the 95% confidence limits.

3.6. Discuss whether the following types of systematic sampling result in higher precision than simple random sampling. (a) Every fifth sentence to estimate the total number of words on a page. (b) Every tenth student entering the bookstore of a university from 12 to 2 P.M. on a Monday to estimate the average amount spent by all the students of that university for books. (c) Every fourth house on each street of a region consisting of one, two, three, or more member families to estimate the total number of children attending schools in the region. (d) Every tenth person leaving a polling booth between 6 and 9 P.M. to predict the winning candidate in a political election.

3.7. Three different types of lists are available for the 1200 employees of an educational institution. In the first one, they appear in alphabetical order. In the second list, they are arranged according to the years of employment. About 600 in the list are employed for 10 or more years, 400 from 5 to 9 years, and the rest for less than 5 years. In the third list, the names appear alphabetically, in succession for the above three groups. It is planned to estimate each of the following items with a 5% simple random or systematic sample: (a) annual savings, (b) number of children in the college, and (c) number of hours spent for sports and recreation during a typical week. For each case, recommend the lists and the type of sampling.

3.8. For the population of the savings in Table 3.3, find the means and S.E. values of $\bar{y}_{sy}$ for (a) one in three and (b) one in eight systematic samples. Compare the precisions of these means with the one in Example 3.1 for $k = 4$ and the means of simple random samples of equivalent sizes.

3.9. In the pilot survey of $n = 25$ selected randomly from the $N = 4000$ students mentioned in Section 1.8, only $n_1 = 15$ responded to the question on the number of hours (y_i) of utilizing the athletic facilities during a week. The mean and standard deviation of the responses were $\bar{y}_1 = 5$ and $s_1 = 3$. The mean $\bar{y}_2$ and standard deviation s_2 for the $n_2 = n - n_1 = 10$ nonrespondents are not known. For the population mean $\bar{Y}$ of the 4000 students, three estimates may be considered: (1) $\bar{y}_1$, (2) substitute $\bar{y}_1$ for each of the n_2 nonrespondents, and (3) substitute zero for each of the nonrespondents. (a) Find expressions to the biases, S.E.values and MSEs of these three estimating procedures. (b) Describe the conditions for which the biases and S.E. values can be large or small.

3.10. Blood pressures of a sample of 20 persons are presented in Table T6. Test the following hypotheses at a 5% level of significance in each case. (a) Before the treatment, the mean systolic pressure is at least 165. (b) Before the treatment, the mean diastolic pressure is at most 90. (c) After the treatment, the mean systolic pressure is 145. (d) After the treatment, the mean diastolic pressure is 75.

3.11. (a) From the sample in Table T6, find the 90% confidence limits for the mean systolic pressure before the treatment. (b) From these limits, is the same conclusion for the hypothesis reached as in Exercise 3.10(a).

3.12. From the sample in Table T6, test the hypotheses at the 5% level of significance that the treatment reduces the means of the systolic and diastolic pressures by 25 and 15, respectively.

3.13. *Project*. Find the covariance s_{xy}, the coefficient of variation $s_y/\bar{y}$, and the correlation coefficient r for each of the 20 samples selected in Exercise 2.10, and verify that s_{xy} is unbiased for S_{xy} but $s_y/\bar{y}$ is biased for the population coefficient of variation $S_y/\bar{Y}$ and r is biased for ρ. For r and s_{xy}, find the (a) bias, (b) variance, and (c) MSE.

3.14. It was shown in Section 3.5 that the sample covariance s_{xy} is unbiased for the population covariance S_{xy}. From the expressions in (3.4) and (3.6), show directly that s_{xy} is unbiased for S_{xy}.

3.15. From the Cauchy–Schwarz inequality in Appendix A3, show that (a) $-1 \le \rho \le 1$ and (b) $-1 \le r \le 1$.

3.16. Show that $1/\bar{y}$ is positively biased for $1/\bar{Y}$ by noting that (a) the arithmetic mean is larger than the harmonic mean, and (b) $\bar{y}$ and $1/\bar{y}$ are negatively correlated, and also (c) from the Cauchy–Schwartz inequality in Appendix A3.

3.17. (a) Show that

$$\sum_{i<j}^{N} \sum^{N} (x_i - x_j)(y_i - y_j) = N \sum_{1}^{N} (x_i - \bar{X})(y_i - \bar{Y}),$$

and hence S_{xy} in (3.4) is the same as

$$\sum_{i<j}^{N} \sum^{N} (x_1 - x_j)(y_i - y_j)/N(N-1).$$

(b) Similarly, show that s_{xy} in (3.6) can be expressed as

$$\sum_{i<j}^{n} \sum^{n} (x_i - x_j)(y_i - y_j)/n(n-1).$$

(c) From these expressions, show that for simple random sampling without replacement, s_{xy} is unbiased for S_{xy}.

3.18. For simple random sampling **without** replacement, show by using the properties in Section 2.5 that (a) $E(y_i) = \bar{Y}$, (b) $V(y_i) = (N-1)S^2/N = \sigma^2$, and (c) $\mathrm{Cov}(y_i, y_j) = -S^2/N$. With these results, derive the formula for $V(\bar{y})$ in (2.12).

3.19. For random sampling **with** replacement, show that (a) $E(y_i) = \bar{Y}$, (b) $V(y_i) = (N-1)S^2/N = \sigma^2$, and (c) $\mathrm{Cov}(y_i, y_j) = 0$. With these results, show that the sample mean $\bar{y}$ is unbiased for $\bar{Y}$ and its variance is given by $V(\bar{y}) = \sigma^2/n$. Note that this variance is larger than the variance for the without replacement procedure.

3.20. *Project*. Select all the $6^2 = 36$ samples of size two randomly with replacement from the six units in Table 2.1. Show that (a) $\bar{y}$ is unbiased for $\bar{Y}$ and its variance is given by

$(N - 1)S^2/Nn$, (b) s^2 is unbiased for σ^2 and (c) the mean of the distinct units in the sample is unbiased for $\bar{Y}$, and its variance is smaller than $(N - 1)S^2/Nn$.

Appendix A3

Joint distribution

Consider two random variables (X, Y) taking values (x_i, y_j) with probabilities $P(x_i, y_j)$, $i = 1, \ldots, k$; $j = 1, \ldots, l$, $\Sigma_1^k \Sigma_1^l P(x_i, y_j) = 1$. The probability for X taking the value x_i is $P(x_i) = \Sigma_j P(x_i, y_j)$, and the probability for Y taking the value y_j is $P(y_j) = \Sigma_i P(x_i, y_j)$. Further, $\Sigma_i P(x_i) = 1$ and $\Sigma_j P(y_j) = 1$.

The data in Table 3.1 provide an illustration. For example, the probability $P(x_1 = 16, y_1 = 1) = 3/51$. The random variable X takes the five values (16, 20, 24, 28, 32), with probabilities (6/51, 18/51, 14/51, 9/51, 4/51). The random variable Y takes the three values (1, 3, 12) with probabilities (24/51, 15/51, 12/51).

Expectation and variance

The **expected value** and **variance** of a discrete random variable X taking values x_i with probabilities $P(x_i)$, $i = 1, \ldots, k$, $\Sigma_1^k P(x_i) = 1$, are

$$E(X) = \sum_1^k x_i P(x_i)$$

and

$$V(X) = E[X - E(X)]^2 = \sum_1^k [x_1 - E(x_i)]^2 P(x_i) .$$

The formula for the variance can also be expressed as

$$V(X) = E(X^2) - [E(X)]^2 = \left[\sum_1^k x_i^2 P(x_i)\right] - \left[\sum_1^k x_i P(x_i)\right]^2 .$$

The above expressions are compact forms of the expectation and variance of a random variable presented in Appendix A2. The expectation and variance of the random variable Y are given by similar expressions.

Covariance and correlation

The **covariance** of X and Y is

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E\{[X - E(X)][Y - E(Y)]\} \\ &= \sum_1^k \sum_1^l [x_i - E(X)][Y_j - E(Y)]P(x_i, y_j).\end{aligned}$$

This formula can also be expressed as

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E(XY) - E(X)E(Y) \\ &= \sum_1^k \sum_1^l x_i y_j P(x_i, y_j) - \left[\sum_1^k x_1 P(x_i)\right]\left[\sum_1^l y_j P(y_j)\right].\end{aligned}$$

The **correlation coefficient** of X and Y is

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}},$$

where $\sqrt{V(X)}$ and $\sqrt{V(Y)}$ are the **standard deviations** of X and Y. The range for ρ falls between -1 and 1.

Conditional and unconditional expectations and variances

In several instances, expectations and variances of the sample quantities for repetitions of the samples, for two or more stages of sampling, will be needed. The following results will be useful on such occasions.

The expectation of Y, $\Sigma_1^l y_j P(y_j)$, can also be expressed as

$$\begin{aligned}E(Y) &= \sum_1^k \sum_1^l y_j P(x_i, y_j) \\ &= \sum_1^k P(x_i) \sum_1^l y_j [P(x_i, y_j)/P(x_i)] \\ &= \sum_1^k P(x_i) \sum_1^l y_j P(y_j \mid x_i),\end{aligned}$$

where $P(y_j \mid x_i)$ is the **conditional probability** of Y taking the value y_j when X takes the value x_i. The second summation on the right-hand side is the conditional mean of Y at $x = x_i$, which can be expressed as $E(Y \mid x_i)$, or as $E(Y \mid X)$. Thus, the above formula can be expressed as

$$E(Y) = E[E(Y \mid X)],$$

where the outer expectation on the right-hand side refers to the expectation with respect to X.

The variance of Y is

$$V(Y) = E(Y^2) - [E(Y)]^2.$$

Further, as in the case of the expectation of Y, $E(Y^2) = E[E(Y^2 \mid X)]$. Hence, $V(Y)$ can be expressed as

$$V(Y) = E[E(Y^2 \mid X)] - \{E[E(Y \mid X)]\}^2.$$

Subtracting and adding $E[E(Y \mid X)]^2$,

$$V(Y) = E\{E(Y^2 \mid X) - [E(Y \mid X)]^2\} + E[E(Y \mid X)]^2 - \{E[E(Y \mid X)]\}^2.$$

The expression in the first braces is the conditional variance of Y, and the third and fourth terms together are equal to the variance of $E(Y \mid X)$. Thus, $V(Y)$ can be expressed as

$$V(Y) = E[V(Y \mid X)] + V[E(Y \mid X)]$$

In this expression, the operators E and V inside and outside the square brackets refer to the expectation and variance conditional and unconditional on X, respectively.

Variance of a linear combination

$$\begin{aligned} S_g^2 = V(ax_i + by_i) &= \sum[(ax_i + by_i + c) - (a\bar{X} + b\bar{Y} + c)]^2/(N-1) \\ &= \sum[a(x_i - \bar{X}) + b(y_i - \bar{Y})]^2/(N-1) \\ &= a^2\sum(x_i - \bar{X})^2/(N-1) + b^2\sum(y_i - \bar{Y})^2/(N-1) \\ &\quad + 2ab\sum(x_i - \bar{X})(y_i - \bar{Y})/(N-1) \\ &= a^2S_x^2 + b^2S_y^2 + 2abS_{xy}. \end{aligned}$$

An expression for S^2

The population variance S^2 can be expressed as

$$\begin{aligned}(N-1)S^2 &= \sum_1^k \sum_1^n (y_{ij} - \overline{Y})^2 = \sum_1^k \sum_1^n [(y_{ij} - \bar{y}_i) + (\bar{y}_i - \overline{Y})]^2 \\ &= \sum_1^k \sum_1^n (y_{ij} - \bar{y}_i)^2 + n \sum_1^k (\bar{y}_i - \overline{Y})^2 \\ &= k(n-1)S_{wsy}^2 + nkV(\bar{y}_{sy}).\end{aligned}$$

Cauchy–Schwartz inequality

For a set of numbers (a_i, b_i), $i = 1, 2, \ldots, k$,

$$\left(\sum_1^k a_i^2 \right) \left(\sum_1^k b_i^2 \right) \geq \left(\sum_1^k a_i b_i \right)^2 ,$$

and the equality occurs if a_i/b_i is a constant for $i = 1, 2, \ldots, k$.