CHAPTER 2

Simple Random Sampling: Estimation of Means and Totals

2.1 Introduction

Simple random sampling is the simplest probability sampling procedure. In this method of selecting a sample of the population units, every sample of a fixed size is given an equal chance to be selected. Every population unit is given an equal chance of appearing in the sample. Similarly, every pair of units has an equal chance of appearing in the sample. In general, every collection of units of a fixed size has an equal chance of being selected.

In the following sections, procedures for selecting a simple random sample and for using it to estimate the population means, totals, and variances of the characteristics of interest are presented. Properties of this sampling procedure are studied in detail. Methods for determining the sample size required for a survey are also examined.

2.2 Population total, mean, and variance

For the sake of illustration, the verbal and math scores in the Scholastic Aptitude Test (SAT) for a small population of $N = 6$ students are presented in Table 2.1. One can denote the **population units**, students in this illustration, by $U_1, U_2, \ldots, U_N$ or briefly as $U_i, i = 1, 2, \ldots, N$. The verbal scores of these units can be denoted by $x_i, i = 1, 2, \ldots, N$ and the math scores by $y_i, i = 1, 2, \ldots, N$.

Total and mean

For the y-characteristic, the population total and mean are given by

$$Y = \sum_{1}^{N} y_i = (y_1 + y_2 + \cdots + y_N) \tag{2.1}$$

Table 2.1. SAT verbal and math scores.

Student i	Verbal x_i	Math y_i
1	520	670
2	690	720
3	500	650
4	580	720
5	530	560
6	480	700
Total	3300	4020
Mean	550	670
Variance		
σ^2	4866.67	3066.67
S^2	5840	3680
S	76.42	60.66
C.V. (%)	13.89	9.05

C.V. = coefficient of variation.

and

$$\bar{Y} = Y/N. \tag{2.2}$$

The mean is a measure of the center or **location** of the N population units. From Table 2.1, for the math scores, $Y = (670 + 720 + \cdots + 720) = 4020$ and $\bar{Y} = 4020/6 = 670$.

Variance and standard deviation (S.D.)

The variance of the y-characteristic is defined as

$$\sigma^2 = \frac{\sum_1^N (y_i - \bar{Y})^2}{N} = \frac{(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \cdots + (y_N - \bar{Y})^2}{N} = \frac{\sum_1^N y_i^2 - N\bar{Y}^2}{N}. \tag{2.3}$$

or as

$$S^2 = \frac{\sum_1^N (y_i - \bar{Y})^2}{N-1}. \tag{2.4}$$

For convenience, in sample surveys, the expression in (2.4) is frequently used for the variance. Either of the above expressions can be obtained from the other, since $S^2 = (N-1)\sigma^2/N$.

The standard deviation, σ or S, is obtained from the positive square roots of (2.3) and (2.4). The variance and standard deviation describe the dispersion or spread among the observations of the population units. For the math scores, $S^2 = [(670 - 670)^2 + (720 - 670)^2 + \cdots (700 - 670)^2]/5 = 3680$ and $S = 60.7$. An alternative expression for the variance is presented in Exercise 2.12, which explicitly expresses the variance as the average of the squared differences of all the pairs of observations.

Coefficient of variation (C.V.)

The C.V. of a characteristic is the ratio of the standard deviation to its mean. This index, unlike the standard deviation and the variance, is not affected by the unit of measurement. For example, it will be the same if the incomes of a group of people are measured in thousands of dollars or tens of thousands of dollars. Similarly, the C.V. will be the same if the incomes are expressed in dollars, British pounds, or any other currency. For the math scores, it is $(S/\bar{Y}) = (60.66/670) = 0.0934$, or 9.34%.

For the x-characteristic, the mean, variance, standard deviation, and C.V. are defined similarly. They are presented in Table 2.1 for both the math and verbal scores. Although the mean is larger for the math scores, the standard deviation and C.V. are larger for the verbal scores.

2.3 Sampling without replacement

Consider estimating the mean and variance of the math scores by selecting a sample of size $n = 2$ randomly from the $N = 6$ students. One can select the sample by writing the names, numbers, or labels of the six students on pieces of paper or cards, thoroughly mixing them, selecting one **randomly**, setting it aside, and then selecting the second

one **randomly** from the remaining five. This is one of the procedures for selecting a sample randomly **without replacement**. Alternatively, one can list all the ${}_6C_2$ (six choose two), that is, 6!/2!(6 – 2)! = 15 possible samples and select one of the samples randomly.

Random samples from a population, especially large, can also be selected from random number tables or computer software packages. A set of random numbers generated through a computer program is presented in Table T1 in the Appendix. In any table of random numbers, each of the digits from zero to 9 has a 1 in 10 chance of appearing. To select a sample of size 20 from a population of 200 units, for example, first label the units from 0 to 199. Start with any three columns together anywhere in the table, select some numbers, rejecting a number that exceeds 199 and that has already been selected, move to another set of three columns, select some more numbers in the same manner, and continue the procedure until 20 numbers are selected. The numbers can also be selected, three at a time, from the rows or both from the rows and columns of the random numbers. For example, a sample of two from the population of six units can be selected by starting with a single row or column in the random number table. Throughout the chapters, a sample selected randomly without replacement is referred to as a **simple random sample**.

As will be seen in the following sections, random sampling and statistical procedures enable one to estimate population quantities such as the total and mean, assess the error in the estimates, and also find intervals for these quantities for specified probabilities.

2.4 Sample mean and variance

The observations of a sample of size n can be denoted by y_i, $i = 1, 2, \ldots, n$. The **sample mean** and **variance** of the y-characteristic are

$$\bar{y} = \sum_1^n y_i/n = (y_1 + y_2 + \cdots + y_n)/n \tag{2.5}$$

and

$$s^2 = \frac{\sum_1^n (y_i - \bar{y})^2}{n - 1}. \tag{2.6}$$

By either of the procedures described in Section 2.3, if units (U_1, U_5) are selected, for the math scores $\bar{y} = (670 + 560)/2 = 615$ and $s^2 = [(670 - 615)^2 + (560 - 615)^2]/(2 - 1) = 6050$. These sample **estimates** 615 and 6050 differ from 670 and 3680, the actual population mean and variance. These differences should be anticipated, since the sample is only a fraction of the population. For increased sample sizes, the differences can be small.

2.5 Properties of simple random sampling

For this type of sampling, the probability of

1. U_i, $i = 1, 2,\ldots,N$ appearing at any draw of the sample is $P_i = 1/N$,
2. U_i appearing at a specified draw conditional on U_j $(i \neq j)$ appearing at another draw is $1/(N - 1)$, and
3. U_i and U_j appearing at two specified draws is $P_{ij} = 1/N(N - 1)$.

The last result follows from (1) and (2).

To illustrate further properties of this type of sampling, 15 samples along with their means, variances, and summary figures for the math scores are presented in Table 2.2. Note, however, that in practice only one of these samples is actually selected. Two additional properties of simple random sampling are as follows.

1. There are ${}_NC_n$ possible samples and each population unit U_i, $i = 1, 2,\ldots,N$, appears in ${}_{(N-1)}C_{(n-1)}$ samples. Hence, the probability of a population unit appearing in the samples is ${}_{(N-1)}C_{(n-1)}/{}_NC_n = n/N$. In this illustration, each of the six units appears in 5 of the 15 samples and hence it has a chance of 1/3 of appearing in the samples.
2. Each pair of the N population units appears in ${}_{(N-2)}C_{(n-2)}$ of the samples and has a chance of ${}_{(N-2)}C_{(n-2)}/{}_NC_n = n(n - 1)/N(N - 1)$ of appearing in the samples. In the illustration, each pair has a chance of 1/15 of appearing in the samples.

Similar results hold for the appearance of different combinations of the population units in the sample of any size selected through simple random sampling.

Table 2.2. All possible samples of size two; means and variances of the math scores.

Sample	Units	Observations	Mean	Variances	S.D.
1	1, 2	670, 720	695	1,250	35.36
2	1, 3	670, 650	660	200	14.14
3	1, 4	670, 720	695	1,250	35.36
4	1, 5	670, 560	615	6,050	77.78
5	1, 6	670, 700	685	450	21.21
6	2, 3	720, 650	685	2,450	49.50
7	2, 4	720, 720	720	0	0
8	2, 5	720, 560	640	12,800	113.14
9	2, 6	720, 700	710	200	14.14
10	3, 4	650, 720	685	2,450	49.50
11	3, 5	650, 560	605	4,050	63.64
12	3, 6	650, 700	675	1,250	35.36
13	4, 5	720, 560	640	12,800	113.14
14	4, 6	720, 700	710	200	14.14
15	5, 6	560, 700	630	9,800	99.00
Expected value of the sample mean			670		
Expected value of the sample variance				3,680	
Expected value of the sample standard deviation					49.03

2.6 Unbiasedness of the sample mean and variance

Sample mean

The mean of each of the possible simple random samples can be denoted by $\bar{y}_k$, $k = 1,\ldots,{}_N C_n$, with the associated probability $P_k = 1/{}_N C_n$ for each. Following the definition in Appendix A2.1, the expected value of the sample mean is

$$E(\bar{y}) = \sum_k P_k \bar{y}_k = \frac{1}{\binom{N}{n}} \frac{1}{n} \left[\left(\sum_1^n y_i \right)_1 + \left(\sum_1^n y_i \right)_2 + \cdots + \left(\sum_1^n y_i \right)_t \right], \quad (2.7)$$

where $t = {}_N C_n$.

Since every population unit appears in ${}_{(N-1)}C_{(n-1)}$ of the possible ${}_N C_n$ samples, the summation in the square brackets is equal to

$_{(N-1)}C_{(n-1)}(\Sigma_1^N y_i)$. With this expression,

$$E(\bar{y}) = \bar{Y}. \tag{2.8}$$

Thus, the sample mean $\bar{y}$ is **unbiased** for the population mean $\bar{Y}$. As a verification, the average of the 15 sample means of the math scores in Table 2.2 is identically equal to the population mean.

Sample variance

The sample variance in (2.6) can be expressed as

$$s^2 = \frac{1}{n-1}\left[\sum_1^n y_i^2 - n\bar{y}^2\right]. \tag{2.9}$$

Appendix A2.3 shows that

$$E(s^2) = \frac{1}{N-1}\left[\sum_1^N y_i^2 - N\bar{Y}^2\right] = S^2. \tag{2.10}$$

Thus, s^2 is unbiased for S^2. For the math scores, as can be seen from Table 2.2, the average of the 15 sample variances is equal to the population variance.

Note that if all N population values are known, there is no need for sampling to estimate the total, mean, variance, or any other population quantity. Second, in practice only one sample of size n is selected from the population. All six population values are presented in Table 2.1 and all the possible samples of size two are presented in Table 2.2 just to illustrate and demonstrate the important properties of simple random sampling. The unbiasedness of the sample mean and variance as seen through (2.8) and (2.10) are theoretical results, and they are verified through Table 2.2. The samples in this table are used to illustrate other properties of this sampling procedure.

2.7 Standard error of the sample mean

A large portion of the 15 sample means in Table 2.2 differ from each other, although their expected value is the same as the population mean.

The variance among all the ${}_NC_n$ possible sample means is

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2 = E(\bar{y}^2) - 2\bar{Y}E(\bar{y}) + \bar{Y}^2 = E(\bar{y}^2) - \bar{Y}^2. \quad (2.11)$$

As shown in Appendix A2.4, this variance can be expressed as

$$V(\bar{y}) = \frac{N-n}{Nn}S^2 = (1-f)\frac{S^2}{n}, \quad (2.12)$$

where $f = n/N$ is the **sampling fraction** and $(1 - f)$ is the **finite population correction** (fpc).

The standard deviation of $\bar{y}$, known as its **standard error** (S.E.), is obtained from

$$\text{S.E.}(\bar{y}) = \sqrt{V(\bar{y})} = \sqrt{\frac{(1-f)}{n}}\,S. \quad (2.13)$$

For the math scores, since $S^2 = 3680$, the variance of $\bar{y}$ is equal to (6 – 2) (3680)/12 = 1226.67, and hence it has an S.E. of 35.02.

As can be seen from the definition in (2.11) and (2.13), $V(\bar{y})$ and S.E.$(\bar{y})$ measure the average departure of the sample mean from the population mean. Further, as the S.E.$(\bar{y})$ becomes small, $\bar{y}$ becomes closer to $\bar{Y}$. This observation can also be made from replacing the random variable X by $\bar{y}$ in the Tschebycheff inequality in Appendix A2.2. Notice from (2.13) that the S.E.$(\bar{y})$ will be small if the sample size is large. If the population size is large, the variance in (2.12) and the S.E. in (2.13) become S^2/n and $S/\sqrt{n}$, which will not differ much from σ^2/n and $\sigma^2/\sqrt{n}$.

Since s^2 is unbiased for S^2, an unbiased estimator for $V(\bar{y})$ is obtained from

$$v(\bar{y}) = \frac{(1-f)}{n}s^2. \quad (2.14)$$

From the sample observations, the S.E. in (2.13) can be estimated from $(1 - f)s/\sqrt{n}$.

With the sample units (U_1, U_5), for the math scores, $s^2 = 6050$. From (2.14), $v(\bar{y}) = [(6 - 2)/6 \times 2](6050) = 2016.67$, and hence the S.E. of $\bar{y}$ from the sample is equal to 44.91.

Notice from Table 2.2 that the variances of the 15 possible samples vary widely, although their expected value is the same as the population variance. Thus, the sampling error of s^2 for estimating S^2 can be large, unless the sample size is large.

2.8 Distribution of the sample mean

The distribution of the 15 sample means of the math scores in Table 2.2 is presented in Table 2.3. Grouping of the means as in Table 2.4 into six classes of equal width of 20 results in an approximation to this distribution. The histogram of the distribution with the mid-values 602.5, 627.5, 652.5, 677.5, 702.5, and 727.5 and the frequencies is presented in Figure 2.1.

For an infinite population, from the **central limit theorem**, the sample mean follows the normal distribution provided the sample size is large. As a result, $Z = (\bar{y} - \bar{Y})/\text{S.E.}(\bar{y})$ approximately follows the standard normal distribution with mean zero and variance unity. For this distribution, which is tabulated and also available through computer software programs, $(1 - \alpha)\%$ of the probability (area) lies between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$. For example, 90% of the area lies between -1.65 and 1.65, 95% between -1.96 and 1.96 or roughly between –2 and 2, and 99% between -2.58 and 2.58.

Table 2.3. Distribution of the Sample mean.

Mean	Frequency	Mean	Frequency
605	1	675	1
615	1	685	3
630	1	695	2
640	2	710	2
660	1	720	1

Table 2.4. Frequency distribution.

Mean	Frequency
590–615	1
615–640	2
640–665	3
665–690	4
690–715	4
715–740	1

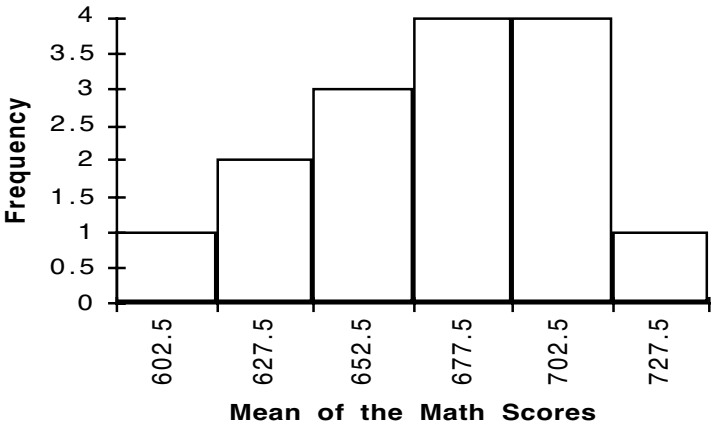

Figure 2.1. Approximate distribution of the sample mean for the math scores.

2.9 Confidence limits for the mean

The Tschebycheff inequality in Appendix A2.2 may be used for this purpose, by replacing X by the sample mean $\bar{y}$. With $\gamma = 2$ for example, 75% of the sample means falls within 2S.E.($\bar{y}$) of the population mean $\bar{Y}$.

With the normal approximation, the probability of $(\bar{y} - \bar{Y})$/S.E.($\bar{y}$) falling between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ is $(1 - \alpha)$, that is,

$$P[\bar{y} - Z_{\alpha/2}\text{S.E.}(\bar{y}) < \bar{Y} < \bar{y} + Z_{\alpha/2}\text{S.E.}(\bar{y})] = (1 - \alpha). \qquad (2.15)$$

From this expression, the **lower** and **upper confidence limits** for $\bar{Y}$ are obtained from $C_L = \bar{y} - Z$ S.E.($\bar{y}$) and $C_U = \bar{y} + Z$ S.E.($\bar{y}$), and the **confidence width** is given by $(C_U - C_L) = 2Z$ S.E.($\bar{y}$); the subscript $\alpha/2$ is suppressed for convenience. Confidence limits with small widths are desirable; otherwise, they provide only vague information regarding the unknown population mean $\bar{Y}$. This width is small if S.E.($\bar{y}$) is small, that is, if the sample size is large.

For the math scores, if (U_1, U_5) are selected into the sample, 95% confidence limits for the population mean are $C_L = 615 - 1.96(35.02) = 546.36$, $C_U = 615 + 1.96(35.02) = 683.64$, and the confidence width is equal to 137.28.

The above limits are obtained with the known S^2. When this variance is estimated from the sample, as seen earlier, S.E.($\bar{y}$) = 44.91. With this estimate, 95% confidence limits for the population mean are $C_L = 615 - 1.96(44.91) = 526.98$ and $C_U = 615 - 1.96(44.91) = 703.02$. The confidence width now is equal to 176.04.

For a large population, assuming that the characteristic y follows a normal distribution with mean $\bar{Y}$ and variance S^2, estimating S.E.($\bar{y}$) from the sample, $t_{n-1} = (\bar{y} - \bar{Y})$/S.E. ($\bar{y}$) follows Student's t-distribution with $(n - 1)$ degrees of freedom (d.f.). The percentiles of this distribution are tabulated and are also available in software programs for statistical analysis. Now, the $(1 - \alpha)$% confidence limits for $\bar{Y}$ are obtained from $\bar{y} - t$ S.E.($\bar{y}$) and $\bar{y} + t$ S.E.($\bar{y}$), with the value of t corresponding to the $(1 - \alpha)$ percentile of the t-distribution with $(n - 1)$ d.f. If the sample size is also large, this percentage point does not differ much from that of the normal distribution, for example, 1.96 when $\alpha = 0.05$.

2.10 Estimating the total

It is frequently of interest to estimate the total Y of characteristics such as the household incomes and expenditures in a region, production of a group of industries, sales of computers in a city or state, and so on.

An estimator of Y is

$$\hat{Y} = N\bar{y} = \frac{N}{n}\sum_{1}^{n} y_i. \tag{2.16}$$

Since $E(\bar{y}) = \bar{Y}$, $E(\hat{Y}) = N\bar{Y} = Y$ and hence $\hat{Y}$ is unbiased for Y. Note that $\hat{Y}$ is obtained by multiplying, or **inflating**, the sample total $\Sigma_1^n y_i$ by (N/n), which is the reciprocal of the probability of selecting a unit into the sample.

The variance of $\hat{Y}$ is

$$V(\hat{Y}) = N^2 V(\bar{y}) = \frac{N(N-n)}{n} S^2. \tag{2.17}$$

An unbiased estimator of this variance is given by

$$v(\hat{Y}) = N^2 v(\bar{y}) = \frac{N(N-n)}{n} s^2. \tag{2.18}$$

The standard error of $\hat{Y}$ is obtained from

$$\text{S.E.}(\hat{Y}) = \sqrt{\frac{N(N-n)}{n}}S, \tag{2.19}$$

and it is estimated by replacing S with the sample standard deviation s.

For the math scores, with the sample units (U_1, U_5), $\hat{Y} = 6(615) = 3690$. With the known value of S^2, $V(\hat{Y}) = 36(1226.67) = 44160.12$ and $\text{S.E.}(\hat{Y}) = 210.14$. When S^2 is estimated by s^2 , $v(\hat{Y}) = 36(2016.67) = 72600.12$ and $\text{S.E.}(\hat{Y}) = 269.44$.

As in the case of the mean, $(1 - \alpha)\%$ confidence limits for Y are obtained from

$$\hat{Y} - Z\,\text{S.E.}(\hat{Y}),\ \hat{Y} + Z\,\text{S.E.}(\hat{Y}). \tag{2.20}$$

These limits can also be obtained by multiplying the limits for the mean by N. If the observations follow the normal distribution, as before, Z is replaced by the percentile of the t-distribution.

For the math scores, when S^2 is known, with the sample units (U_1,U_5), 95% confidence limits for the total are $3690 - 1.96\,(210.14) = 3278.13$ and $3690 + 1.96\,(210.14) = 4101.87$, with the width of 823.74. With the estimated variance, these limits become $3690 - 1.96\,(269.44) = 3161.90$ and $3690 + 1.96\,(269.44) = 4218.10$, and now the confidence width is equal to 1056.20. Except for the round-off errors, the same limits and widths are obtained by multiplying the limits and widths for the population mean obtained in the last section by the population size, six in this illustration.

2.11 Coefficient of variation of the sample mean

The C.V. of the sample mean $\bar{y}$ is

$$\text{C.V.}(\bar{y}) = \text{S.E.}\,(\bar{y})/\bar{Y} = \sqrt{\frac{(1-f)}{n}}(S/\bar{Y}), \tag{2.21}$$

which expresses the S.E. of $\bar{y}$ as a percentage of its expected value, the population mean $\bar{Y}$. To estimate this C.V., $(S/\bar{Y})$ is replaced by $(s/\bar{y})$. Note that for large n, the C.V. of $\bar{y}$ is obtained by dividing the C.V. of y by n. Second, the C.V. of $\hat{Y}$, $\sqrt{V(\hat{Y})}\,/Y$, is the same as that of $\bar{y}$.

As the variance of $\bar{y}$ and its S.E., the C.V. in (2.21) is also a measure of the **sampling error** of the mean. For the math scores, C.V. of $\bar{y}$ is (35.02/670) = 0.0523 or 5.23% when S^2 is known, and it is equal to (44.91/615) = 0.0730 or 7.30% with the estimated variance.

2.12 Sample size determination

The sample size for every survey has to be determined before it goes into operation. To estimate $\overline{Y}$ or Y, the size of the sample is usually needed for a prescribed value of the (1) S.E. of $\bar{y}$, (2) error of estimation, (3) confidence width, (4) C.V. of $\bar{y}$, or (5) the relative error, which is given by $|\bar{y} - \overline{Y}|/\overline{Y}$. These types of prescriptions are frequently employed in industrial research and commercial surveys. Information regarding the standard deviation S should be known for these critera. Past data or a preliminary sample of small size, also known as the pilot survey, can be used for this purpose.

Precision

The precision of an estimator is measured by the reciprocal of its variance. Any requirements for this precision can be translated into a prescription on the variance, and vice versa.

If the variance of $\bar{y}$ should not exceed a given value V, the sample size is obtained from

$$V(\bar{y}) \leq V. \tag{2.22}$$

If N is large, this inequality can be expressed as $S^2/n \leq V$, and with equality the required sample size is obtained from

$$n_l = \frac{S^2}{V}. \tag{2.23}$$

If N is not large, from (2.22), the smallest value of the required sample size is given by

$$n = \frac{n_l}{1 + (n_l / N)} \tag{2.24}$$

As expected, larger sample sizes are needed to estimate the mean with smaller S.E.

Error of estimation

With prior knowledge about $\bar{Y}$, it can be required that $\bar{y}$ should not differ from it by more than a specified amount of absolute error e, a small quantity. This requirement, however, can be satisfied only with a chance of $(1 - \alpha)$. Formally, this requirement can be expressed as the probability statement,

$$P_r\{\,|\bar{y} - \bar{Y}| \leq e\,\} = (1 - \alpha). \tag{2.25}$$

From the correspondence of this probability with the Tschebycheff inequality in Appendix A2.2, $\gamma^2 = (1/\alpha)$ and $\gamma^2 V(\bar{y}) = e^2$. Solving these equations, n is obtained from (2.24) with $n_l = (\gamma S/e)^2$.

Alternatively, note that the above probability statement is the same as

$$P_r[|\bar{y} - \bar{Y}|/\text{S.E.}(\bar{y}) \leq e/\text{S.E.}(\bar{y})] = (1 - \alpha). \tag{2.26}$$

With the assumption of normality, as noted before, the term on the left side in the brackets follows the standard normal distribution. As a result, $Z = e/\text{S.E.}(\bar{y})$ or $Z^2 V(\bar{y}) = e^2$, where Z is the $(1 - \alpha)$ percentile of this distribution. Now, n is obtained from (2.24) with

$$n_l = (ZS/e)^2. \tag{2.27}$$

As can be expected, the sample size will be large if the specified error e is small.

Confidence width

If it is required that the confidence width for $\bar{Y}$, with a confidence probability of $(1 - \alpha)$, should not exceed a prescribed amount w, the sample size is obtained from

$$2Z\ \text{S.E.}(\bar{y}) \leq w. \tag{2.28}$$

Solution of this equation results in (2.24) for the sample size with

$$n_l = 4(ZS/w)^2. \tag{2.29}$$

Coefficient of variation

If it is desired that the C.V. of $\bar{y}$ should not exceed a given value C_0, n should be obtained from

$$[(1 - f)/n]C^2 \leq C_0^2, \tag{2.30}$$

where $C = S/\overline{Y}$ is the population C.V.

The required sample size is again obtained from (2.24) with

$$n_l = (C/C_0)^2. \tag{2.31}$$

Note that this criterion can be satisfied only with some knowledge of the population C.V. Larger sample sizes are needed to estimate the population mean with a smaller C.V.

Relative error

The error of estimation relative to the population mean is $|\bar{y} - \overline{Y}|/\overline{Y}$. The requirement that this error should not exceed a prescribed value **a**, except for a chance of $(1 - \alpha)$, can be expressed as

$$\Pr\{|\bar{y} - \overline{Y}| / \text{S.E.}(\bar{y}) \leq \boldsymbol{a}\overline{Y}/\text{S.E.}(\bar{y})\} = (1 - \alpha) \tag{2.32}$$

with the normality assumption, $Z^2 = a^2\overline{Y}^2/V(\bar{y})$. Solving this equation, n is obtained from (2.24) with

$$n_l = (ZC/\boldsymbol{a})^2. \tag{2.33}$$

The population C.V., $C = S/\overline{Y}$, should be known for this criteria also.

Example 2.1. Sample size: Consider a camera club with 1800 members, where it is required to estimate the average number of rolls of film used during a year. Consider also the information from the past that the average and standard deviation of the number of rolls of film have been around 6 and 4, respectively.

a. To estimate $\bar{Y}$ with an S.E. not exceeding 0.5, $n_l = 16/(0.5)^2 = 64$ from (2.23) and $n = 62$ from (2.24).
b. To estimate $\bar{Y}$ with an error not exceeding 1, except for a 5% chance, with the normal approximation, $n_l = (1.96)^2 16/1 = 62$ from (2.27) and $n = 60$ from (2.24).
c. With $w = 2$ and $(1 - \alpha) = 0.95$, $n_l = 62$ from (2.29) and $n = 61$ from (2.24).
d. With the information on the mean and standard deviation, $C = 4/6$ or 67% approximately. If the C.V. of $\bar{y}$ should not exceed 8%, $n_l = (0.67/0.08)^2 = 70$ from (2.31) and $n = 68$ from (2.24).
e. If the relative error should not exceed (1/10) except for a 5% chance, $n_l = (1.96)^2(0.67)^2/0.01 = 173$ from (2.33) and $n = 158$ from (2.24).

2.13 Sample sizes for individual groups

Frequently, it becomes important to estimate the means and totals for two or more groups or subpopulations of a population. The sample sizes for the groups may be found through specifications of the type described in Section 2.12. Alternatively, the S.E., errors of estimation, and confidence widths for each of the groups for suitable division of the overall sample size required for the entire population may be examined. These two procedures are illustrated in the following example. Chapters 5 and 6 examine in detail sample size determination for two or more groups and also their differences.

Example 2.2. Sample sizes for groups: Consider the 1800 members in Example 2.1 to consist of 1200 old and 600 new members with the standard deviations of 2 and 4 for the number of cartons of film. As in Example 2.1(b), sample sizes required to estimate the means of these groups with the error of estimation not exceeding 1 for each group can be found. Now, for the old group, $n_l = (1.96)^2 \times 4/1 = 15.4$ and $n = 15$ approximately. Similarly, for the new group, $n_l = (1.96)^2 \times 16/1 = 61.5$ and $n = 56$ approximately. The total sample size of 71 required now is larger than the size of 62 for the entire population, found in Example 2.1(b).

Exercises

2.1. Among the 100 computer corporations in a region, average of the employee sizes for the largest 10 and smallest 10 corporations were known to be 300 and 100, respectively. For a sample of 20 from the remaining 80 corporations, the mean and standard deviation were 250 and 110, respectively.

For the total employee size of the 80 corporations, find the (a) estimate, (b) the S.E. of the estimate, and (c) the 95% confidence limits.

2.2. Continuing with Exercise 2.1, for the average and total of the 100 corporations, find the (a) estimate, (b) the S.E. of the estimate, and (c) the 95% confidence limits.

2.3. During the peak season, the mean and standard deviation of the late arrivals for a sample of 30 flights of one airline were 35 and 15 minutes, respectively. For the average delay of all the flights of this airline, find (a) the 95% confidence limits with the t-distribution and (b) the C.V. of the sample mean.

2.4. As in Exercise 2.3, the mean and standard deviation of the late arrivals for a sample of 30 flights of another airline were 30 and 20, respectively. Which of these airlines would one prefer if (a) the upper 95% confidence limit for the average with the t-distribution should not exceed 40 minutes and (b) the coefficient of variation for the sample mean should not exceed 10%?

2.5. The total profit for $N_1 = 5$ of the largest computer companies was known to be $Y_1 = 500$ (million dollars). For a sample of $n_2 = 9$ from the remaining $N_2 = 45$ computer companies, the mean and standard deviation were $\bar{y}_2 = 30$ and $s_2 = 15$. An estimate suggested for the total profit of the $N = 50$ companies was $t_1 = N(Y_1 + n_2\bar{y}_2)/(N_1 + n_2)$. From the above results, $t_1 = (50/14)(500 + 9 \times 30) = 1882.14$. (a) Find an expression for the variance of this estimator, and find its S.E. from the sample observations. (b) Is this an unbiased estimator for the total profit?

2.6. Another estimator for the total profit of the 50 companies in Exercise 2.5 is $t_2 = Y_1 + N_2\bar{y}_2$, and from the sample observations it is equal to $500 + 45 \times 30 = 1850$. (a) Find an expression for the variance of this estimator and find its S.E. from the sample. (b) Is this an unbiased estimator for the total profit?

2.7. Find the sample size required for the problem in Example 2.1 if (a) the S.E. should not exceed 0.25, (b) the sample mean should not differ from the actual mean by more than one roll of film, except for 10% chance, and (c) the relative error should not exceed 7% except for a 5% chance.

2.8. The metropolian area and the suburbs together in a region consist of 5, 10, and 10 thousand families with one, two,

and three or more children. For these three types of families, preliminary estimates of the averages and standard deviations of the number of hours of television watching in a week are (10, 15, 20) and (6, 10, 15), respectively. Find the sample sizes required to estimate the above average for each group if the error of estimation should not exceed 2 hours in each case, except for a 5% chance.

2.9. Continuing with Exercise 2.8, find the sample sizes if the error of estimation should not exceed 10% of the actual average in each case, except for a 5% chance.

2.10. *Project*. Select all the 20 samples of size three from the population of six students in Table 2.1 without replacement and verify the five properties described in Section 2.5 for the probabilities of selecting the units and their appearance in the sample. From each sample, find the 95% confidence limits for the population mean of the math scores with the known population variance and its estimates; use the normal deviate $Z = 1.96$ in both cases. (a) For both the procedures, find the proportion of the confidence intervals enclosing the actual population mean, that is, the **coverage probability**. (b) Compare the average of the confidence widths obtained with the estimates of variance with the exact width for the case of known variance.

2.11. (a) Show that $\Sigma_{i<j}^{N}\Sigma^{N}(y_i - y_j)^2 = N\Sigma_1^N(y_i - \bar{Y})^2$, and hence S^2 in (2.4) is the same as $\Sigma_{i<j}^{N}\Sigma^{N}(y_i - y_j)^2/N(N-1)$.
(b) Similarly, show that s^2 in (2.6) can be expressed as $\Sigma_{i<j}^{N}\Sigma^{N}(y_i - y_j)^2/n(n-1)$.
(c) From these expressions, show that for simple random sampling without replacement, s^2 is unbiased for S^2.

2.12. For simple random sampling without replacement, starting with the expectation of $\Sigma_1^n(y_i - \bar{Y})^2$, show that $V(\bar{y}) = (1 - f)S^2/n$.

Appendix A2

Expected value, variance, and standard deviation

The expected value of a random variable X taking values $(x_1, x_2, \ldots, x_k)$ with probabilities $(p_1, p_2, \ldots, p_k)$, $p_1 + p_2 + \cdots + p_k = 1$, is

$$\mu = E(X) = x_1p_1 + x_2p_2 + \cdots + x_kp_k.$$

For two constants c and d, the expected value of $Y = (cX + d)$ is

$$E(cX + d) = cE(X) + d = c\mu + d.$$

The variance of X is

$$\sigma^2 = V(X) = E(X - \mu)^2 = (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \cdots + (x_2 - \mu)^2 p_k.$$

The variance can also be expressed as $E(X^2) - \mu^2$.

The standard deviation σ is obtained from the positive square root of the variance. The variance of $Y = cX + d$ is

$$\begin{aligned} V(Y) &= E[y - E(Y)]^2 = E[(cX + d) - (c\mu + d)]^2 \\ &= c^2 E(X - \mu)^2 = c^2 V(x). \end{aligned}$$

Adding a constant to a random variable does not change its variance. When the random variable is multiplied by a constant c, the variance becomes c^2 times the original variance.

Tschebycheff inequality

For a random variable X, this inequality is given by

$$\Pr[|X - E(X)| \leq \gamma \sqrt{V(X)}] \geq 1 - \frac{1}{\gamma^2},$$

where γ is a chosen positive quantity. For instance, if $\gamma = 2$, 75% of the observations fall within two standard deviations of $E(X)$, the population mean.

Unbiasedness of the sample variance

The sample variance in (2.9) can be expressed as

$$s^2 = \frac{1}{n-1}\left[\sum_1^n y_i^2 - n\bar{y}^2\right] = \frac{1}{n}\sum_1^n y_i^2 - \frac{1}{n(n-1)} \sum_{i \neq j}^n \sum^n y_i y_j.$$

Noting that each of the samples is selected with a probability of $1/{}_N C_n$, following the definition for the expectation,

$$E(s^2) = \frac{1}{\binom{N}{n}} \frac{1}{n} \left[\left(\sum_i y_i^2 \right)_1 + \left(\sum_i y_i^2 \right)_2 + \cdots + \left(\sum_i y_i^2 \right)_t \right]$$

$$- \frac{1}{\binom{N}{n}} \frac{1}{n(n-1)} \left[\left(\sum_{i \neq j} \sum y_i y_j \right)_1 + \left(\sum_{i \neq j} \sum y_i y_j \right)_2 + \cdots + \left(\sum_{i \neq j} \sum y_i y_j \right)_t \right].$$

As in the case of the mean, the first summation is equal to ${}_{(N-1)}C_{(n-1)}$ $\Sigma_1^N y_i^2$. Similarly, the second summation is equal to ${}_{(N-2)}C_{(n-2)}$ $\Sigma_{i \neq j}^N \Sigma^N y_i y_j$. Substituting these expressions,

$$\begin{aligned} E(s^2) &= \frac{1}{N} \sum^N y_i^2 - \frac{1}{N(N-1)} \sum_{i \neq j}^N \sum^N y_i y_j \\ &= \frac{1}{N-1} \left[\sum_1^N y_i^2 - N \bar{Y}^2 \right] = S^2. \end{aligned}$$

Variance of the sample mean

Since

$$\bar{y}^2 = \frac{1}{n^2} \left[\sum_i^n y_i^2 + \sum_{i \neq j}^n \sum^n y_i y_j \right],$$

$$\begin{aligned} E(\bar{y}^2) &= \frac{1}{n^2} \left[\frac{n}{N} \sum_i^N y_i^2 + \frac{n(n-1)}{N(N-1)} \sum_{i \neq j}^N \sum^N y_i y_j \right] \\ &= \frac{(N-n)}{nN(N-1)} \sum^N y_i^2 + \frac{(n-1)}{n(N-1)} \bar{Y}^2. \end{aligned}$$

Now, with simplification,

$$V(\bar{y}) = E(\bar{y}^2) - \bar{Y}^2 = (1-f) S^2/n.$$