10

ESTIMATION STUDIES

Inferring the Parameters of a Population from the Statistics of a Sample

10.0 LEARNING OBJECTIVES

In this chapter, we discuss the construction and interpretation of *estimation* studies. Our learning objectives include the following:

- using a sample to estimate the characteristics of a property of a population when the property is qualitative/categorical;
- using a sample to estimate the characteristics of a property of a population when the property is quantitative/scale; and
- a brief discussion of the practical problems of sampling.

10.1 MOTIVATION

In Chapter 2, we described the scenario in which an investigator is interested in characterizing a large set of phenomena with regard to a particular property, but all the phenomena of interest cannot be observed for logistical reasons. For example, a sociologist might be interested in assessing the "feelings of alienation" for the adult population of the United States, but assessing all those individuals is clearly not feasible. In such cases, the investigator might judiciously select a small set of phenomena from the larger set and use the assessments of the values of the property found for this small set of phenomena to characterize the occurrences of the property in the larger set. In this way, the investigator is using *specific* observations to reach a *general* conclusion; thus, the investigator is applying the logic of *inference*. In more formal terms, we have the following:

- The set of phenomena in which the investigator is interested is said to be a *population*, and the judiciously selected phenomena actually observed is said to be a *sample*. As discussed in Chapter 9, the logic by which a sample can be used to represent a population requires the process of selecting the sample set of phenomena to be equivalent to a lottery. This was said to be *random* selection.
- In an objectively assessed property, the occurrences of the different values of that property found in a sample can be summarized as a *statistic*. These *descriptive sample statistics* were described in Chapters 3 to 5. *If they could be observed*, the occurrences of the different values of the property existing in the population could also be summarized in the same manner as the descriptive statistics of a sample. However, these summaries are said to be *parameters* when constructed for a *population*.
- In an *estimation study*, an investigator uses the statistics of a sample to estimate the unknown parameters of the population from which the sample was drawn.

As noted in Chapter 7, the logical basis for making such projections—or statistical inferences—can be found in probability theory. However, probability theory also yields the following *caveat*:

> A sample can be expected to be only an imperfect representation of the population from which it is drawn, and two different samples drawn from the same population might be substantially different with regard to the values of the property of interest held by those selected phenomena. Thus, the inference process is imperfect.

Fortunately, using these same probability models, we can set some limits to the extent of the potential imperfection in using a sample statistic to estimate a population parameter. These "imperfection limits" are often said to be *confidence intervals*. Such confidence intervals take the following form:

> With a "high" probability (often 95%), we can be confident that the population parameter will be between the values (*Statistic* – *a*) and (*Statistic* + *a*), where *Statistic* is the relevant sample statistic and *a* is a measure of variability estimated by the sample.

In a somewhat confusing and arrogant usage, the extent to which a sample statistic might vary from the underlying population parameter is said to be an "error" (the usage is arrogant in that the term *error* presumes a mistake, and no sample can be expected to be a "perfect" representation of a population). This usage was previously encountered in Chapter 9 with the term "standard error of the mean." In this chapter, we discuss the construction of such confidence intervals for different types of estimation studies.

As a second matter of concern, when using a statistic from a sample to estimate a parameter for a population where the property of interest is quantitative, it can be shown that the variance of the sample will underestimate the variance of the population from which the sample was drawn. This underestimation was identified by Friedrich Bessel (1784–1846), and the correction to this underestimation is said to be the *Bessel Correction* (Upton, Graham, and Ian Cook, 2014, *A Dictionary of Statistics*, 3rd ed., Oxford, UK: Oxford University Press). In this chapter, we also discuss this "correction."

10.2 ESTIMATING THE OCCURRENCE OF A QUALITATIVE PROPERTY FOR A POPULATION

Two candidates, *A* and *B*, are competing for election in a state with a *population* of *N* potential voters. To better plan the campaign, the campaign manager of Candidate *A*, Paula, wants to have some sense of the proportion of voters favoring each—or none—of the two candidates. Thus, each voter may be characterized as having one of the following four dispositions:

- favoring *A*, denoted as (*A*);
- favoring *B*, denoted as (*B*);
- rejecting both, denoted as (*C*); and
- undecided, denoted as (*D*).

While Paula would prefer to contact every voter, the logistics of this approach are prohibitive. However, from her understanding of probability theory, Paula is relatively confident that a randomly selected sample of voters might provide a reasonable representation of this population; thus, she conducts a survey of *n* potential voters and finds what is shown in Table 10.1.

Thus, Paula would project these relative frequencies—or proportions—onto the population of *N* voters. However, her justification for doing so requires an appropriate probability model to describe the relationship she can expect between an "unknown" population and samples extracted from that population. In fact, such a model can be constructed using the concept of a *Bernoulli trial* described in Chapter 9. For a detailed explanation of this model, see Box 10.1.

TABLE 10.1 ■ Voting Dispositions of a Randomly Selected Sample of *n* Voters

Disposition	Frequency	Relative Frequency
A	a	$a / n = f(A)$
B	b	$b / n = f(B)$
C	c	$c / n = f(C)$
D	d	$d / n = f(D)$
Total	n	1.00

BOX 10.1

In this population of N voters, we know that some number α favor A, some number β favor B, some number γ reject both (C), and some number δ are undecided (D). Now, even if we do not know the values α, β, γ, and δ, we do know that if we were to randomly choose a potential voter from this population, we have the following potential outcomes and probabilities of occurrence:

- the individual will favor A with a probability of $\alpha / N = \text{p}(A)$;
- the individual will favor B with a probability of $\beta / N = \text{p}(B)$;
- the individual will reject both candidates with a probability of $\gamma / N = \text{p}(C)$; or
- the individual will be undecided with a probability of $\delta / N = \text{p}(D)$.

With this understanding, how might we characterize *our expectations* regarding a sample constructed from a series of n independent selections from this population? While the *Central Limit Theorem* would be directly applicable for a quantitative/scale property, the property in this case is qualitative. However, by applying the model of a *Bernoulli trial* as discussed in Chapter 9, we can make a direct prediction as to what to expect from a sample constructed from this population.

Let us first consider the dispositional value "favors *A*." From this perspective, we can identify two types of individuals in this population:

- those who "favor *A*," which we will denote as (A); and
- those who do not "favor *A*," which we will denote as (~A) and includes those who "favor *B*," those who "reject both candidates," and those who are "undecided."

From this perspective, we can describe the population as shown in Table 10.2.

TABLE 10.2 ■ Population of *N* Voters Described by Their Disposition Toward Candidate *A*

Disposition	Number of Individuals
Does not favor A (~A)	$\beta + \gamma + \delta = N - \alpha$
Favors A (A)	A
Total	N

Now, if we were to randomly select an individual from this population, we could expect outcomes with the probabilities shown in Table 10.3.

TABLE 10.3 ■ Probability Model of a Random Selection of an Individual from a Population of *N* Voters

Disposition	Number of Individuals	Probability of Selection
(~*A*)	*N* – α	(*N* – α) / *N* = 1 – p(*A*) = p(~*A*)
(*A*)	α	α / *N* = p(*A*)
Total	*N*	1

Continuing with our narrative, our interest is in assessing the relative proportion of individuals in the population having the disposition (*A*) using a sample selected from this population. From this perspective, it would then be reasonable to ask *for each selection from the population* the number of individuals who "favor *A*." Here we have two potential outcomes:

- 1 if the individual "favors *A*"; or
- 0 if the individual does not "favor *A*."

Because this outcome will vary for each selection, we can identify the outcome of each selection as a "random variable," which we will denote as $\mathbf{X}_i$. Moreover, for *n* independent selections from this population, the total number of individuals found to "favor *A*" will be equal to the sum of the individual selections, or

$$\mathbf{X}_1 + \mathbf{X}_2 + \ldots + \mathbf{X}_{n-1} + \mathbf{X}_n.$$

We will denote this total as $\mathbf{X}_A$.

Now, for a set of *n* independent selections from this population, how many individuals can we expect to find who "favor *A*"? That is, what is $\mathbf{E(X}_A\mathbf{)}$? To this question, the answer is

$$\mathbf{E(X}_A\mathbf{)} = \mathbf{E(X}_1\mathbf{)} + \mathbf{E(X}_2\mathbf{)} + \ldots + \mathbf{E(X}_{n-1}\mathbf{)} + \mathbf{E(X}_n\mathbf{)}.$$

Moreover, for this population, we can assess the expected value of each selection experiment—denoted as $\mathbf{E(X}_i\mathbf{)}$—as shown in Table 10.4.

TABLE 10.4 ■ Expected Value of $\mathbf{X}_i$, Where X Is the Number of Individuals Who Favor *A* in a Single Random Selection from a Population of *N* Voters

Value	Probability	Contribution
0	p(~*A*) = 1 – *p*(*A*)	0 • (1 – p(*A*)) = 0
1	p(*A*)	1 • p(*A*) = p(*A*)
Total	1.00	p(*A*)

Thus, we would identify *p*(*A*) as the expected value of a single selection from this population of *N* voters, and the expected number of *n* such selections is

$$\mathbf{E(X}_A\mathbf{)} = \mathrm{p}(A) + \mathrm{p}(A) + \ldots + \mathrm{p}(A) + \mathrm{p}(A) = n \bullet \mathrm{p}(A).$$

From this, we can then assess the relative proportion of the sample who "favor *A*" as $\mathbf{X}_A / n$, and the *expected value of the relative proportion of the sample* who "favor *A*" is

$$\mathbf{E(X}_A / n) = \mathbf{E(X}_A\mathbf{)} / n = (n \bullet \mathrm{p}(A)) / n = \mathrm{p}(A).$$

Thus, we can logically justify using the actual relative proportion of those who "favor *A*" in our sample of *n* individuals—denoted earlier as *f*(*A*)—as a reasonable estimate of p(*A*), which is the relative proportion of those who "favor *A*" in the population of *N* voters. In a similar way, we would

- estimate the proportion of individuals in the population "favoring *B*" by using *f*(*B*);
- estimate the proportion of individuals in the population "rejecting both candidates" by using *f*(*C*); and
- estimate the proportion of individuals in the population who are "undecided" by using *f*(*D*).

Now, while Paula is logically justified in using the sample proportions she collected to estimate proportions of supporters and non-supporters of her candidate among the population of potential voters, she also knows that there is a potential for error in those estimations. That is, Paula knows that any two samples taken from a population can vary, and she cannot know whether the sample she has collected has the "right" proportions relative to the population. However, she does have some idea as to the variation she might expect among the potential samples she might collect. That is, using the same logical model of a *Bernoulli trial*, Paula can construct an estimate of the variability that can be expected among the potential samples that may be drawn from a population. For example, with regard to the proportion of the population supporting Candidate *A*—p(*A*)—the expected variability among the potential samples drawn from this population is estimated as

$$\sqrt{}\ ((f(A) - f(A)^2)/n),$$

where $f(A)$ is the observed proportion of voters in the sample supporting Candidate *A* and *n* is the size of the sample. Now, what can Paula infer about the correctness of the sample she has obtained? Paula can be relatively confident that the population proportion of voters supporting Candidate *A*—p(*A*)—is not likely to be less than

$$f(A) - \sqrt{}\ ((f(A) - f(A)^2)/n)$$

and is not likely to be greater than

$$f(A) + \sqrt{}\ ((f(A) - f(A)^2)/n).$$

Similarly, Paula can estimate the correctness of her sample with regard to the other voter dispositions:

- the expected variability among the potential samples drawn from this population with regard to the observed proportion of voters in the sample supporting Candidate *B* is estimated as

  $$\sqrt{}\ ((f(B) - f(B)^2)/n);$$

- the expected variability among the potential samples drawn from this population with regard to the observed proportion of voters in the sample who reject both candidates is estimated as

  $$\sqrt{}\ ((f(C) - f(C)^2)/n);\text{ and}$$

- the expected variability among the potential samples drawn from this population with regard to the observed proportion of voters in the sample who are undecided is estimated as

  $$\sqrt{}\ ((f(D) - f(D)^2)/n).$$

Unfortunately, we cannot use these results to compare the accuracy of the different *relative* proportions because such comparisons violate the model by which these estimates of variability were constructed. For the details of this logical model, see Box 10.2.

BOX 10.2

Although we may be logically justified in using sample proportions to estimate proportions in a population, we need to acknowledge the potential for error in those projections. That is, we know that any two samples taken from a population may vary, and we cannot know whether the sample we have collected has the "right" proportions relative to the population. However, we do have some idea as to the variation we might expect among the potential samples we might collect.

Returning to our preceding discussion, for a sample of n individual selections from our population of N voters, we have the following expectation regarding the relative proportion of those individuals who "favor A":

$$\mathbf{E}(\mathbf{X}_A / n) = (\mathbf{E}(\mathbf{X}_1) + \mathbf{E}(\mathbf{X}_2) + \ldots + \mathbf{E}(\mathbf{X}_{n-1}) + \mathbf{E}(\mathbf{X}_n))/n =$$

$$\mathbf{E}(\mathbf{X}_1/n) + \mathbf{E}(\mathbf{X}_2/n) + \ldots + \mathbf{E}(\mathbf{X}_{n-1}/n) + \mathbf{E}(\mathbf{X}_n/n).$$

Now, we also know that the expected variability—or *variance*—in the sample proportion $\mathbf{X}_A / n$ is

$$\mathbf{Var}(\mathbf{X}_A / n) = \mathbf{Var}(\mathbf{X}_1 / n) + \mathbf{Var}(\mathbf{X}_2 / n) + \ldots + \mathbf{Var}(\mathbf{X}_{n-1} / n) + \mathbf{Var}(\mathbf{X}_n / n) =$$

$$1 / n^2 \bullet \mathbf{Var}(\mathbf{X}_1) + 1 / n^2 \bullet \mathbf{Var}(\mathbf{X}_2) + \ldots + 1 / n^2 \bullet \mathbf{Var}(\mathbf{X}_{n-1}) + 1 / n^2 \bullet \mathbf{Var}(\mathbf{X}_n) =$$

$$1 / n^2 \bullet (\mathbf{Var}(\mathbf{X}_1) + \mathbf{Var}(\mathbf{X}_2) + \ldots + \mathbf{Var}(\mathbf{X}_{n-1}) + \mathbf{Var}(\mathbf{X}_n)).$$

Then, to assess $\mathbf{Var}(\mathbf{X}_i)$, we have Table 10.5. This gives us

$$\mathbf{Var}(\mathbf{X}_A / n) = (1 / n^2) \bullet (n \bullet (p(A) - p(A)^2)) = 1 / n \bullet ((p(A) - p(A)^2).$$

More useful, however, in assessing the potential variability among the sample proportions is the standard deviation σ_n of those proportions $= \surd\ \mathbf{Var}(\mathbf{X}_A / n) = \surd\ ((p(A) - p(A)^2) / n)$. In concept, this assessment of variability is similar to the *standard error of the mean* discussed in Chapter 9.

Now, with this understanding of the potential variability in the samples we might collect from this population, how do we interpret the actual sample proportion $f(A) = a / n$?

1) It is reasonable to expect that the actual sample proportion of $f(A)$ is exactly

TABLE 10.5 ■ Expected Variance in $\mathbf{X}_i$, Where $\mathbf{X}_i$ Is the Number of Individuals Who Favor *A* in a Single Random Selection from a Population of *N* Voters

Value	Difference from E(X_i)	Difference Squared	Probability	Contribution to Variance
0	$0 - p(A)$	$p(A)^2$	$1 - p(A)$	$p(A)^2 \bullet (1 - p(A)) = p(A)^2 - p(A)^3$
1	$1 - p(A)$	$1 - 2p(A) + p(A)^2$	$p(A)$	$p(A) - 2p(A)^2 + p(A)^3$
Total			1.00	$p(A) - p(A)^2$

(Continued)

(Continued)

equal to the unknown population proportion of p(*A*).

2) It is possible, however, that our sample proportion *f*(*A*) might actually be less than the population proportion p(*A*). However, from our understanding of the standard deviation and expectations, we know that our expected sample proportion is not likely to be less than the population proportion minus the potential "error." That is,

$$f(A) \geq p(A) - \surd\ ((p(A) - p(A)^2) / n).$$

This in turn means

$$f(A) + \surd\ ((p(A) - p(A)^2) / n) \geq p(A).$$

That is, the population proportion is likely to be not greater than the expected sample proportion plus the standard error. Now, because we do not know the actual population proportion p(*A*), we can only use our best estimate of p(*A*) = *f*(*A*). Thus, we would reasonably project our population proportion to be no greater than

$$f(A) + \surd\ ((f(A) - f(A)^2) / n).$$

3) Similarly, it is possible that our sample proportion *f*(*A*) might actually be greater than the population proportion p(*A*). However, from our understanding of the standard deviation and expectations, we know that our expected sample proportion is not likely to be greater than the population proportion plus the standard error. That is,

$$f(A) \leq p(A) + \surd\ ((p(A) - p(A)^2) / n).$$

This in turn means

$$f(A) - \surd\ ((p(A) - p(A)^2) / n) \leq p(A).$$

That is, the population proportion is likely to be not less than the expected sample proportion minus the standard error. Now, because we do not know the actual population proportion p(*A*), we can only use our best estimate of p(*A*) = *f*(*A*). Thus, we would reasonably project our population mean to be no less than

$$f(A) - \surd\ ((f(A) - f(A)^2) / n).$$

4) Combining these two sets of expectations, we have

$$f(A) - \surd\ ((f(A) - f(A)^2) / n) \leq$$
$$p(A) \leq f(A) + \surd\ ((f(A) - f(A)^2) / n).$$

In a similar fashion, we would address the projections of the population proportions of those who "favor Candidate *B*," "reject both candidates," and are "undecided":

5) We would project the population proportion of those "favoring Candidate *B*" = p(*B*) to be equal to the sample proportion *f*(*B*), but we would acknowledge the potential for error in this projection by providing the
likely range of the population proportion as

$$f(B) - \surd\ ((f(B) - f(B)^2) / n) \leq$$
$$p(B) \leq f(B) + \surd\ ((f(B) - f(B)^2) / n).$$

6) We would project the population proportion of those "rejecting both candidates" = p(*C*) to be equal to the sample proportion *f*(*C*), but we would acknowledge the potential for error in this projection by providing the likely range of the population proportion as

$$f(C) - \surd\ ((f(C) - f(C)^2) / n) \leq$$
$$p(C) \leq f(C) + \surd\ ((f(C) - f(C)^2) / n).$$

7) We would project the population proportion of those "undecided" = p(*D*) to be equal to the sample proportion *f*(*D*), but we would acknowledge the potential for error in this projection by providing the likely range of the population proportion as

$$f(D) - \surd\ ((f(D) - f(D)^2) / n) \leq$$
$$p(D) \leq f(D) + \surd\ ((f(D) - f(D)^2) / n).$$

Now, having estimated the different proportions of the different voting dispositions projected for the population of potential voters, can Paula proceed with confidence in constructing her candidate's campaign strategy using these assessments? Not quite. It is possible that although Paula has properly assessed the potential variability in her projections, she cannot be confident that the observed differences in these relative proportions are not simply the result of the normal variability of sampling. Using the terminology introduced in Chapter 9, she cannot be certain of the *statistical significance* of these observed differences. Fortunately, Paula can apply another logical model to address this question. However, this logical model requires several analytical steps that are beyond the scope of the current discussion, but these steps will be addressed in the next chapter. For that reason, we defer this matter to Chapter 11.

As a final comment, we will turn our attention to our assessment of the variation that can be expected in our estimate of the population proportion p(A). That variability was assessed to be equal to

$$\sqrt{}\ ((f(A) - f(A)^2)/n).$$

Here it may be noted that the *variability* among the potential samples drawn from a population *decreases* as the size of the sample *increases*, and the size of the sample is under the control of the investigator. This is the basis of the maxim

"The larger the sample, the better the estimate."

By this, we mean

"The larger the sample, the less the probability of obtaining a sample that is a poor representation of the population."

For some investigations, however, the cost in time and the difficulty of sample collection may be sufficiently high where a trade-off is necessary. That is, the investigator is willing to accept a greater uncertainty in the results of the investigation in order to actually conduct the investigation. In those fields in which sample collection is particularly difficult, such as medical research, the relation between sample size and sample uncertainty is often used to determine a minimum sample size necessary to achieve an acceptable level of uncertainty with regard to the correctness of the sample as an estimate of the population parameter. This "acceptable level of uncertainty" is said to be "statistical power."

10.3 ESTIMATING THE OCCURRENCE OF A QUANTITATIVE PROPERTY FOR A POPULATION

Dan Roberts (D.R.) is a newspaper publisher, and he is desirous of knowing the amount of time readers actually spend reading the daily newspaper. Why would he care? Because he can presume that the greater the reading time, the greater the reading depth. If the typical reader spends a relatively short time reading, it is unlikely the reader will "go for depth"; thus, articles must be correspondingly brief to capture the reader's attention. On the other hand, if the typical reader spends a relatively long period of time reading, it is likely the reader is looking for conceptual depth in each article, and a conceptually shallow article will lose the reader's attention. Why would D.R. care about capturing the reader's attention? Because a reader's attention that is not captured is less likely to continue purchasing that newspaper brand.

Because D.R. is concerned with understanding the reading habits of the general population of newspaper readers, his logical model can be described as shown in Table 10.6.

TABLE 10.6 ■ Potential Reading Times (X) of a Population of *N* Readers

Time (hours)	Proportion of Population	Relative Proportion
0	$N_{0.0}$	$N_{0.0}/N = p(0.0)$
0.5	$N_{0.5}$	$N_{0.5}/N = p(0.5)$
1.0	$N_{1.0}$	$N_{1.0}/N = p(1.0)$
1.5	$N_{1.5}$	$N_{1.5}/N = p(1.5)$
2.0	$N_{2.0}$	$N_{2.0}/N = p(2.0)$
. . .	. . .	. . .
23.0	$N_{23.0}$	$N_{23.0}/N = p(23.0)$
23.5	$N_{23.5}$	$N_{23.5}/N = p(23.5)$
24.0	$N_{24.0}$	$N_{24.0}/N = p(24.0)$
Total	N	1.00

In this model, there are 48 potential values of reading time and some unknown proportion of the population (of unknown size) for each one, and by using this model D.R. can address three questions:

(1) For marketing reasons, D.R. might ask, "What is the size of each proportion of readers?" In marketing terms, each of these reading time groups represents a "market segment," and D.R. might develop a strategy to appeal to a particular group of readers.

(2) D.R. might also ask the question, "Can a typical reader be identified with regard to a typical amount of reading time?" In asking this question,

D.R. might attempt a strategy of appealing to the potential readership with a generally acceptable news format. To address this question, D.R. can use the conceptual model we developed in Chapter 4 for identifying a typical phenomenon from a sample of phenomena. In this case, D.R. would conceptually construct a "typical" reader based on the mean reading time (μ) of the unknown population (Table 10.7).

TABLE 10.7 ■ Conceptual Assessment of the Mean Reading Time (μ) of a Population of *N* Readers

Time (hours)	Relative Proportion	Contribution to the Mean
0.0	$p(0.0)$	$0.0 \bullet p(0.0)$
0.5	$p(0.5)$	$0.5 \bullet p(0.5)$
1.0	$p(1.0)$	$1.0 \bullet p(1.0)$
. . .	. . .	. . .
23.5	$p(23.5)$	$23.5 \bullet p(23.5)$
24.0	$p(24.0)$	$24.0 \bullet p(24.0)$
Total	1	μ

However, because the actual proportions and total number of potential readers are unknown, the model remains conceptual, and the unknown mean reading time μ is said to be a *parameter* of the population.

(3) Finally, D.R. might also ask the question, "What is the typical variability in the typical reading time among the potential newspaper readers?" Why would D.R. care about the variability among the readers' habits? The variability among the readers' habits presents a risk to any marketing strategy. That is, if there is a relatively small amount of variability among the readers' habits, a strategy of offering a generally acceptable article length will have a relatively good chance of success. However, if the variability among the readers' habits is large, a strategy of offering a generally acceptable article length based on the mean-time reader will be less likely to succeed because the shorter-time readers and longer-time readers will similarly find articles appealing to the mean-time readers as unacceptable. To address this question, D.R. can use the conceptual model we developed in Chapter 5 for identifying a typical variability found among a set of phenomena. This typical amount of variability was assessed as the *variance* and interpreted as the standard

deviation (Table 10.8). For an unknown population, the variance is denoted as **Var** or σ^2, and the standard deviation, which is the square root of the variance, is denoted as σ.

TABLE 10.8 ■ Calculation of the Variance (σ^2) in the Reading Times (X) of a Population of *N* Readers

Time (hours)	Time − $\mu = \Delta$	Δ^2	Relative Proportion	Contribution to the Variance
0.0	$0.0 - \mu$	$(0.0 - \mu)^2$	$p(0.0)$	$(0.0 - \mu)^2 \bullet p(0.0)$
0.5	$0.5 - \mu$	$(0.5 - \mu)^2$	$p(0.5)$	$(0.5 - \mu)^2 \bullet p(0.5)$
1.0	$1.0 - \mu$	$(1.0 - \mu)^2$	$p(1.0)$	$(1.0 - \mu)^2 \bullet p(1.0)$
. . .	. . .	. . .	. . .	. . .
23.5	$23.5 - \mu$	$(23.5 - \mu)^2$	$p(23.5)$	$(23.5 - \mu)^2 \bullet p(23.5)$
24.0	$24.0 - \mu$	$(24.0 - \mu)^2$	$p(24.0)$	$(24.0 - \mu)^2 \bullet p(24.0)$
Total	. . .	. . .	1.0	σ^2

As with the assessment of the population mean, the actual number of individuals characterized by each "reading time" value is unknown, so the variance and the standard deviation in these reading time values are also unknown. Moreover, like the mean, the variance and the standard deviation are said to be parameters of the population.

Now, because interviewing every potential newspaper reader is not a feasible option, D.R. commissions a study in which a random sample of potential readers are asked to identify the amount of time they spend daily reading the newspaper. The results of this survey are reported in Table 10.9.

TABLE 10.9 ■ Reading Times (X) of a Sample of 1000 Readers

Time (hours)	Frequency	Relative Frequency
0.0	100	0.10
0.5	400	0.40
1.0	200	0.20
1.5	200	0.20
2.0	100	0.10
Total	1000	1.00

With the results of this survey, D.R. can answer the first of his questions as follows:

- The proportion of potential readers who spend zero hours reading the newspaper is not likely to be less than

  $$0.10 - \sqrt{((0.10 - 0.01)/1000)} =$$

  $$0.10 - \sqrt{(0.09/1000)} =$$

  $$0.10 - 0.0095 = 0.0905$$

  and is not likely to be greater than

  $$0.10 + 0.0095 = 0.1095.$$

- The proportion of potential readers who spend 0.5 hour reading the newspaper is not likely to be less than

  $$0.40 - \sqrt{((0.40 - 0.16)/1000)} =$$

  $$0.40 - \sqrt{(0.24/1000)} =$$

  $$0.40 - 0.0155 = 0.3845$$

  and is not likely to be greater than

  $$0.40 + 0.0155 = 0.4155.$$

- The proportion of potential readers who spend 1 hour reading the newspaper is not likely to be less than

  $$0.20 - \sqrt{((0.20 - 0.04)/1000)} =$$

  $$0.20 - \sqrt{(0.16/1000)} =$$

  $$0.20 - 0.0126 = 0.1874$$

  and is not likely to be greater than

  $$0.20 + 0.0126 = 0.2126.$$

- The proportion of potential readers who spend 1.5 hours reading the newspaper is not likely to be less than

  $$0.20 - \sqrt{((0.20 - 0.04)/1000)} =$$

  $$0.20 - \sqrt{(0.16/1000)} =$$

  $$0.20 - 0.0126 = 0.1874$$

and is not likely to be greater than

$$0.20 + 0.0126 = 0.2126.$$

- The proportion of potential readers who spend 2 hours reading the newspaper is not likely to be less than

$$0.10 - \sqrt{((0.10 - 0.01)/1000)} =$$

$$0.10 - \sqrt{(0.09/1000)} =$$

$$0.10 - 0.0095 = 0.0905$$

and is not likely to be greater than

$$0.10 + 0.0095 = 0.1095.$$

With regard to his second question, D.R. understands from the *Central Limit Theorem* of probability theory (Chapter 9) that the mean value of a property found for any random sample collected from a population is likely to be equal to the mean value of that property for the population from which the sample was drawn. This is true regardless of the distribution of the property values existing in that population. Consequently, D.R. reasons that he can use the mean value of the property "reading time" assessed for his sample of 1000 potential newspaper readers (denoted as $\bar{x}$) to estimate the reading time μ of the typical potential newspaper reader in the larger population. To assess the mean reading time $\bar{x}$ for the sample set of individuals, D.R. uses the method described in Chapter 5. This analysis is found in Table 10.10.

Thus, D.R. would estimate the typical reading time for the typical individual of the population as 0.9 hour. However, D.R. is also aware that no sample can be assumed to be a perfect representation of the population from which it was drawn;

TABLE 10.10 ■ Assessing the Mean Reading Time ($\bar{x}$) of a Sample of 1000 Readers

Time (hours)	Frequency	Contribution to the Mean (time • frequency)
0.0	100	0
0.5	400	200
1.0	200	200
1.5	200	300
2.0	100	200
Total	1000	900

$\bar{x}$ = 900 hours / 1000 individuals = 0.9 hour per individual.

thus, D.R. takes the extra step of assessing the potential "error" in this particular estimate of the mean reading time for this population of potential newspaper readers. Now, to make this assessment, D.R. first needs to assess the potential variability in the reading times of the population of potential newspaper readers, and in making this assessment D.R. has no other choice than to use the variability observed in reading times of the 1000 individuals of his survey. In this way, D.R. is also addressing his third question as to the variability in the reading times of the population of potential newspaper readers. To assess the variability of reading times observed for his sample set of individuals, D.R. uses the method of moments described in Chapter 5 (Table 10.11).

As described in Chapter 5, to find the variance, D.R. would find the mean value of Δ^2 by dividing the total value of the observed differences squared by the number of observations. This would give the variance as

$$\mathbf{s}^2 = 340/1000 = 0.340.$$

However, *when using a sample variance to estimate a population variance*, it is standard practice to apply the Bessel Correction, which is to reduce the divisor by 1. This would give the "corrected" variance as

$$\mathbf{s}^2 = 340/999 = 0.3403.$$

Now, based on his understanding of the *Central Limit Theorem*, D.R. knows the following:

- the sample he has collected *might* have a mean value of **X** that is less than the population mean μ; and
- the sample he has collected *might* have a mean value of **X** that is greater than the population mean μ; but

TABLE 10.11 ■ Calculation of the Variance (s^2) in the Reading Times (X) of a Sample of 1000 Readers

Time (hours)	Time – $\bar{x}$ = Δ	Δ^2	Frequency	Contribution to the Variance (Δ^2 • frequency)
0.0	0.0 – 0.9 = –0.9	0.81	100	81
0.5	0.5 – 0.9 = –0.4	0.16	400	64
1.0	1.0 – 0.9 = 0.1	0.01	200	2
1.5	1.5 – 0.9 = 0.6	0.36	200	72
2.0	2.0 – 0.9 = 1.1	1.21	100	121
Total	. . .	. . .	1000	340

- his sample mean is *most likely* to have a mean value of **X** that is *equal* to the population mean μ.

Moreover, even if the sample D.R. has chosen has a mean value of **X** that is not equal to the mean value of **X** for the population, he can be *relatively confident* that the mean value of **X** for his sample is within the following bounds:

- $\bar{\mathbf{x}}$ is not likely to be less than $\mu - (\sigma/\sqrt{n})$; and
- $\bar{\mathbf{x}}$ is not likely to be greater than $\mu + (\sigma/\sqrt{n})$,

where *n* is the size of his sample. Why? Because the value $(\sigma/\sqrt{n})$ is the expected variation in the mean values of **X** that would be found among all of the potential samples that might be drawn from the population of potential newspaper readers. This was described as the *standard error of the mean* in Chapter 9. Thus, D.R. has the following expectations:

- If $\bar{\mathbf{x}}$ is not likely to be less than $\mu - (\sigma/\sqrt{n})$, then $\bar{\mathbf{x}} \geq \mu - (\sigma/\sqrt{n})$. This, in turn, means

$$\bar{\mathbf{x}} + (\sigma/\sqrt{n}) \geq \mu.$$

- If $\bar{\mathbf{x}}$ is not likely to be greater than $\mu + (\sigma/\sqrt{n})$, then $\bar{\mathbf{x}} \leq \mu + (\sigma/\sqrt{n})$. This, in turn, means

$$\bar{\mathbf{x}} - (\sigma/\sqrt{n}) \leq \mu.$$

- Together, he has

$$\bar{\mathbf{x}} - (\sigma/\sqrt{n}) < \mu < \bar{\mathbf{x}} + (\sigma/\sqrt{n}).$$

Now, because σ *is unknown,* D.R. follows the common practice of using the standard deviation of his sample (**s**) to estimate the unknown standard deviation (σ) of the population. In this case,

$$\mathbf{s}^2 = 0.3403 \text{ and}$$

$$\mathbf{s} = \sqrt{0.3403} = 0.5834.$$

Thus, D.R. can estimate the typical reading time μ for the population of potential newspaper readers as

$$\bar{\mathbf{x}} - (\mathbf{s}/\sqrt{n}) \leq \mu \leq \bar{\mathbf{x}} + (\mathbf{s}/\sqrt{n}) \text{ or}$$

$$0.9 - (0.5834/\sqrt{1000}) \leq \mu \leq 0.9 + (0.5834/\sqrt{1000}) \text{ or}$$

$$0.9 - (0.184) \leq \mu \leq 0.9 + (0.184) \text{ or}$$

$$0.716 < \mu < 1.084.$$

Finally, D.R. can address his third question by interpreting the meaning of the standard deviation estimated for the reading habits of the population of potential newspaper readers. Here the mean reading time estimated for the population is 0.9 hour, and the estimated standard deviation in reading time is 0.5834 hour. As discussed in Chapter 5, the standard deviation of a set of observations can be interpreted in two ways:

- the standard deviation can simply reflect the natural variability of the values of the property; or
- the standard deviation can reflect the existence of two distinct groups of values with respective means of $\mu - \sigma$ and $\mu + \sigma$.

To compare these two possible interpretations, D.R. considers the following:

- Was the distribution of reading time values found in the sample multimodal? If so, this fact would support the interpretation of the standard deviation representing two distinct groups. In this case, the answer is "no."
- Was the standard deviation σ smaller than the smallest increment on the scale of measurement? If so, the scale of measurement is insufficiently precise to distinguish between the two groups. In this case, the standard deviation is 0.5834 hour, while the smallest scale increment is 0.5 hour.

With these conflicting assessments, D.R. concludes that there is insufficient evidence to interpret the standard deviation as indicating two distinct groups of individuals with different reading habits.

10.4 SOME NOTES ON SAMPLING

In probability theory, the concept of "random sampling" is intuitively straightforward. However, in empirical practice, approximating the theoretical "random selection process" is fraught with logistical problems, and numerous techniques have been developed to address these problems. While an understanding of such techniques is useful in designing empirical studies, the description of these techniques lies beyond the scope of this text. Nevertheless, two general "problems" in sampling are of particular note *in observing human behavior*.

Selection Bias

Suppose we are interested in peoples' opinions on their favorite *genre* of music, and we have designed a questionnaire—said to be an "instrument"—to assess these opinions. Now, in order to administer this questionnaire, we need to find individuals, capture their attention, and enlist their agreement to volunteer their time to complete the questionnaire. Some individuals will be more accessible than others, and some individuals will be more willing to volunteer their time to complete the questionnaire. Clearly, this reality of sampling does not correspond to the theoretical model of a lottery in which a ball is blindly selected from a bag, and the probability of selecting one individual rather than another is not equal. This is an example of what is said to be *selection bias*, and it is an unavoidable problem in some empirical studies. At best, a practitioner should understand when his or her research project is subject to such bias, and temper his or her conclusions regarding the accuracy of his or her sampling.

Response Bias

Suppose we are interested in people's opinions on several alternative options regarding a controversial policy issue. Having designed a questionnaire "instrument" to assess these opinions, we proceed to administer the instrument among a random sample of individuals. In assessing each individual's opinion, we are presuming that the individual will honestly express his or her true feelings. Unfortunately, this may or may not be the case. For example, suppose there are two policy alternatives, and one has received a great deal of promotional support by a number of prominent social institutions, while the other policy option has been disparaged by those same institutions. Human nature being as it is, it would not be unreasonable to expect some individuals who favor the disparaged policy to be reticent to reflect their preference in their completion of the questionnaire, and this reality of opinion assessment clearly does not correspond to the necessary presumption of "honesty." This is an example of what is said to be *response bias*, and it is an unavoidable problem in some empirical studies. In some cases, a researcher may attempt to elicit "honesty" indirectly through a series of proxy questions, but this introduces two new problems:

- First, the proxy questions may or may not be a proper representation of the individual's direct opinion on the policy.
- Second, the use of such techniques may present an ethical concern regarding the "tricking" of the respondent to reveal his or her true preferences.

While these indirect methods are commonly used, the practitioner should understand the problems in doing so, and a responsible practitioner will accordingly temper his or her conclusions regarding the accuracy of the sample results.

10.5 SPSS TUTORIAL

The techniques for analyzing observations of phenomena described by a qualitative property—presented in Chapter 3—are appropriate for both case studies and estimation studies. For analyzing observations of phenomena described by a quantitative property, most statistical software programs automatically apply the Bessel Correction even if the observations being analyzed are part of a case study rather than an estimation study. This is true of SPSS, so the tutorial presented in Chapter 5 was technically incorrect for the case study applications described but was technically correct for an estimation study.

10.6 Summary

- In an estimation study, an investigator is interested in characterizing a set of phenomena (population) with regard to a property of interest, but observing the full population is infeasible. To estimate the characteristics of the population, the investigator instead collects a random sample of phenomena from the population and uses the characteristics of the sample to estimate the characteristics of the population. The characteristics of the sample are said to be statistics, the characteristics of the population are said to be parameters, and the logical justification by which sample statistics can be used to estimate population parameters is based on probability theory.
- For a population of phenomena described by a qualitative, ordinal, or quantitative property, the population may be characterized by the relative proportions of the values of the property of interest, and the relative proportions of the values of the property found in a random sample of the population are good estimates of the relative proportions of the values of the property in the population. However, these estimates cannot be expected to be perfect and, thus, are subject to some measure of potential error. The potential error in an estimate of a population proportion is equal to

  $$\sqrt{\,((f(A) - f(A)^2) / n)},$$

  where $f(A)$ is the relative proportion estimated from the sample and n is the size of the sample.
- For a population of phenomena described by a quantitative property, the population also may be characterized by the mean value of the property, the variance in the values of the property, and the standard deviation in the values of the property. If a randomly selected set of phenomena is drawn from the population as a sample, the sample mean, the sample variance, and the sample standard deviation all are good estimates of the corresponding population parameters. However, these estimates cannot be expected to be perfect and, thus, are subject to some measure of potential error. The potential error in the estimate of a population mean is equal to

  $$\mathbf{s}/\sqrt{n},$$

  where **s** is the standard deviation of the values of the property found in the sample and n is the size of the sample. This potential error in the estimate of the population is said to be the *standard error of the mean*. As a technical note, when a sample variance is used to estimate a population, the sample variance is subjected to the Bessel Correction, in which the total of the squared differences from the mean is divided by the sample size minus 1 rather than by the sample size.
- In theory, the random selection process is conceived as a lottery. However, approximating this model is fraught with logistical problems, and numerous techniques have been developed to address these problems. While a description of these collection techniques is beyond the scope of this test, we can address two additional problems in sampling where the property of interest involves human behavior. First, because not all of the individuals in a population are equally accessible, or equally willing to be "observed," any sample set of phenomena will be subject to *selection bias*. Second, not all individuals will tell the truth in revealing their preferences or behavioral habits; thus, samples of such properties are subject to *response bias*.

10.7 Exercises

1) Design and conduct a small-scale estimation study in which the property of interest is qualitative.

   a) Identify the population of interest. (If you plan to investigate some aspect of human behavior, do not include individuals under the age of 18 years in your population.)

   b) Identify the property of interest.

   c) Design an assessment instrument.

   d) Design your sampling technique. (If you plan to investigate some aspect of human behavior, do not include individuals under the age of 18 years in your sampling technique.)

   e) Collect the sample.

   f) Estimate the characteristics of the population from the sample, including the potential errors in the estimates.

2) Design and conduct a small-scale estimation study in which the property of interest is quantitative.

   a) Identify the population of interest. (If you plan to investigate some aspect of human behavior, do not include individuals under the age of 18 years in your population.)

   b) Identify the property of interest.

   c) Design an assessment instrument.

   d) Design your sampling technique. (If you plan to investigate some aspect of human behavior, do not include individuals under the age of 18 years in your sampling technique.)

   e) Collect the sample.

   f) Estimate the characteristics of the population from the sample, including potential errors in the estimates.

BOX 10.3

Optional exercises: These exercises demonstrate the relationship between theoretical predictions of probability and statistical outcomes.

3) *A Sampling Experiment*. Obtain 50 3 × 5-inch index cards and cut each one into two halves at the midpoint of the long (5-inch) axis, resulting in 100 cards.

   - On 50 of those cards, write “Candidate *A*.”
   - On 30 of those cards, write “Candidate *B*.”
   - On 10 of those cards, write “Neither, *C*.”
   - On 10 of those cards, write “Undecided, *D*.”

In doing so, we have created a population of 100 potential voters with the following dispositions:

- 50% favor Candidate *A* = p(*A*);
- 30% favor Candidate *B* = p(*B*);
- 10% reject both candidates = p(*C*); and
- 10% are undecided = p(*D*).

Place the 100 cards into a medium paper bag, fold the top, and shake the bag. Then, randomly select a card from the deck, record the disposition written on the card, and return the card to the bag. Repeat this procedure nine times. This will result in a sample set of 10 observations of the property “preference for candidates.” Summarize your sample as in Table 10.12.

   a) From your sample, find $f(A)$ and compare with $p(A) = 0.50$.

   b) Calculate $\surd\,((f(A) - f(A)^2)/10)$.

      - Is $p(A) = 0.50$ greater than $f(A) - \surd\,((f(A) - f(A)^2)/10)$?
      - Is $p(A) = 0.50$ less than $f(A) + \surd\,((f(A) - f(A)^2)/10)$?

TABLE 10.12 ■ Voting Dispositions of a Randomly Selected Sample of 10 Voters

Disposition	Frequency	Relative Frequency
A	*a*	*a* / 10 = *f*(*A*)
B	*b*	*b* / 10 = *f*(*B*)
C	*c*	*c* / 10 = *f*(*C*)
D	*d*	*d* / 10 = *f*(*D*)
Total	10	1.00

c) From your sample, find *f*(*B*) and compare with p(*B*) = 0.30.

d) Calculate $\sqrt{((f(B) - f(B)^2) / 10)}$.

- Is p(*B*) = 0.30 greater than $f(B) - \sqrt{((f(B) - f(B)^2) / 10)}$?
- Is p(*B*) = 0.30 less than $f(B) + \sqrt{((f(B) - f(B)^2) / 10)}$?

e) From your sample, find *f*(*C*) and compare with p(*C*) = 0.10.

f) Calculate $\sqrt{((f(C) - f(C)^2) / 10)}$.

- Is p(*C*) = 0.10 greater than $f(C) - \sqrt{((f(C) - f(C)^2) / 10)}$?
- Is p(*C*) = 0.10 less than $f(C) + \sqrt{((f(C) - f(C)^2) / 10)}$?

g) From your sample, find *f*(*D*) and compare with p(*D*) = 0.10.

h) Calculate $\sqrt{((f(D) - f(D)^2) / 10)}$.

- Is p(*D*) = 0.10 greater than $f(D) - \sqrt{((f(D) - f(D)^2) / 10)}$?
- Is p(*D*) = 0.10 less than $f(D) + \sqrt{((f(D) - f(D)^2) / 10)}$?

4) *Another Sampling Experiment*. In this experiment, we will simulate an economy in which 63% of the working-age adults participate in the workforce and 37% do not. As a note of interest, this participation rate is similar to that estimated for the United States. Thus, among the population of working-age adults, we have the following distribution of workweek hours:

- 37% work zero hours;
- 7% work 10 hours;
- 16% work 20 hours;

TABLE 10.13 ■ Weekly Work Hours for a Population of *N* Workers

Hours	Probability	Contribution to the Mean (μ)
0	0.37	0 • 0.37 = 0.0
10	0.07	10 • 0.07 = 0.7
20	0.16	20 • 0.16 = 3.2
40	0.32	40 • 0.32 = 12.8
60	0.08	60 • 0.08 = 4.8
Total	1.00	21.5

(Continued)

(Continued)

- 32% work 40 hours; and
- 8% work 60 hours.

Moreover, we can describe the population in terms of the distribution of working-age adults by their weekly work hours (Table 10.13).

Thus, for any working-age adult individual randomly selected from this population, the expected number of work hours per week (μ) is 21.5. Moreover, the variability we might expect in this selection—the Variance—is 408.75 hours^2, and the standard deviation is 20.22 hours (see Table 10.14).

For policy purposes, the government needs to know the expected work activities of this population of working-age individuals to project the cost of potential unemployment benefits. However, because it is impractical for the government to canvass every working-age adult, the government relies on random samples collected from the population. In this exercise, you will have an opportunity to check the accuracy of such sampling.

a) Obtain 50 3 × 5-inch index cards, and cut each one into two halves at the midpoint of the long (5-inch) axis, resulting in 100 cards.

- On 37 of those cards, write "0 hours."
- On 7 of those cards, write "10 hours."
- On 16 of those cards, write "20 hours."
- On 32 of those cards, write "40 hours."
- On 8 of those cards, write "60 hours."

TABLE 10.14 ■ Variance in Expected Weekly Work Hours for a Population of *N* Workers

Hours	Hours − $\mu = \Delta$	Δ^2	Probability	Contribution to the Mean = $\Delta^2 \bullet$ Probability
0	−21.5	462.25	0.37	$462.25 \bullet 0.37 = 171.03$
10	−11.5	132.25	0.07	$132.25 \bullet 0.07 = 9.26$
20	−1.5	2.25	0.16	$2.25 \bullet 0.16 = 0.36$
40	18.5	342.25	0.32	$342.25 \bullet 0.32 = 109.52$
60	38.5	1482.25	0.08	$1482.25 \bullet 0.08 = 118.58$
Total			1.00	408.75

TABLE 10.15 ■ Weekly Work Hours for a Sample of 10 Working-Age Individuals

Hours	Relative Frequency
0	$f(0)$
10	$f(10)$
20	$f(20)$
40	$f(40)$
60	$f(60)$
Total	1.00

In doing so, you will have created a population of 100 working-age individuals equivalent to the population of this hypothetical economy.

b) Place the 100 cards into a medium paper bag, fold the top, and shake the bag. Then, randomly select a card from the deck, record the work hours written on the card, and return the card to the bag. Repeat this procedure nine times. This will result in a sample set of 10 observations of the property "weekly work hours." Summarize your sample of 10 observations in a relative frequency distribution (see Table 10.15).

c) Compare the sample proportions with those of the population:

- Compare the sample proportion of those working "0 hours" = *f*(0) with the population proportion of those working "0 hours" = 0.37.
- Compare the sample proportion of those working "10 hours" = *f*(10) with the population proportion of those working "10 hours" = 0.07.
- Compare the sample proportion of those working "20 hours" = *f*(20) with the population proportion of those working "20 hours" = 0.16.
- Compare the sample proportion of those working "40 hours" = *f*(40) with the population proportion of those working "40 hours" = 0.32.
- Compare the sample proportion of those working "60 hours" = *f*(60) with the population proportion of those working "60 hours" = 0.08.

d) Find the mean work hours ($\bar{\mathbf{x}}$) of the sample. Compare this with the population mean = 21.5 hours.

e) Find the standard deviation in work hours (**s**) of the sample. Compare this with the standard deviation in work hours of the population = 20.22.

f) Use the standard deviation of your sample (**s**) to estimate the standard deviation of the population (σ), and find the *standard error of the mean* "for samples of size 10" taken from this population = $\sigma / \sqrt{10}$.

- Is the population mean μ = 21.5 greater than or equal to

  $\bar{\mathbf{x}}$ – standard error of the mean?

- Is the population mean μ = 21.5 less than or equal to

  $\bar{\mathbf{x}}$ + standard error of the mean?