Previous Section Table of Contents Next Section

# 1.2 Overview

**LEARNING OBJECTIVE**

1. To obtain an overview of the material in the text.

The example we have given in the first section seems fairly simple, but there are some significant problems that it illustrates. We have supposed that the 200 cars of the sample had an average value of $8,357 (a number that is precisely known), and concluded that the population has an average of about the same amount, although its precise value is still unknown. What would happen if someone were to take another sample of exactly the same size from exactly the same population? Would he get the same sample average as we did, $8,357? Almost surely not. In fact, if the investigator who took the second sample were to report precisely the same value, we would immediately become suspicious of his result. The sample average is an example of what is called a *random variable*: a number that varies from trial to trial of an experiment (in this case, from sample to sample), and does so in a way that cannot be predicted precisely. Random variables will be a central object of study for us, beginning in Chapter 4 "Discrete Random Variables".

Another issue that arises is that different samples have different levels of reliability. We have supposed that our sample of size 200 had an average of $8,357. If a sample of size 1,000 yielded an average value of $7,832, then we would naturally regard this latter number as likely to be a better estimate of the average value of all cars. How can this be expressed? An important idea that we will develop in Chapter 7 "Estimation" is that of the *confidence interval*: from the data we will construct an interval of values so that the process has a certain chance, say a 95% chance, of generating an interval that contains the actual population average. Thus instead of reporting a single estimate, $8,357, for the population mean, we would say that we are 95% certain that the true average is within $100 of our sample mean, that is, between $8,257 and $8,457, the number $100 having been computed from the sample data just like the sample mean $8,357 was. This will automatically indicate the reliability of the sample, since to obtain the same chance of containing the unknown parameter a large sample will typically produce a shorter interval than a small one will. But unless we perform a census, we can never be completely sure of the true average value of the population; the best that we can do is to make statements of *probability*, an important concept that we will begin to study formally in Chapter 3 "Basic Concepts of Probability".

Sampling may be done not only to estimate a population parameter, but to test a claim that is made about that parameter. Suppose a food package asserts that the amount of sugar in one serving of the product is 14 grams. A consumer group might suspect that it is more. How would they test the competing claims about the amount of sugar, 14 grams versus more than 14 grams? They might take a random sample of perhaps 20 food packages, measure the amount of sugar in one serving of each one, and average those amounts. They are not interested in the true amount of sugar in one serving in itself; their interest is simply whether the claim about the true amount is accurate. Stated another way, they are sampling not in order to estimate the average amount of sugar in one serving, but to see whether that amount, whatever it may be, is larger than 14 grams. Again because one can have certain knowledge only by taking a census, ideas of probability enter into the analysis. We will examine tests of hypotheses beginning in Chapter 8 "Testing Hypotheses".

Several times in this introduction we have used the term “random sample.” Generally the value of our data is only as good as the sample that produced it. For example, suppose we wish to estimate the proportion of all students at a large university who are females, which we denote by $p$. If we select 50 students at random and 27 of them are female, then a natural estimate is $p \approx \hat{p} = 27/50 =$ 0.54 or 54%. How much confidence we can place in this estimate depends not only on the size of the sample, but on its quality, whether or not it is truly random, or at least truly representative of the whole population. If all 50 students in our sample were drawn from a College of Nursing, then the proportion of female students in the sample is likely higher than that of the entire campus. If all 50 students were selected from a College of Engineering Sciences, then the proportion of students in the entire student body who are females could be underestimated. In either case, the estimate would be distorted or biased. In statistical practice an unbiased sampling scheme is important but in most cases not easy to produce. For this introductory course we will assume that all samples are either random or at least representative.

## KEY TAKEAWAY

- Statistics computed from samples vary randomly from sample to sample. Conclusions made about population parameters are statements of probability.