CHAPTER 2

Probability and Sampling Distributions

CONTENTS

2.1 Introduction 68
2.2 Probability 71
2.3 Discrete Probability Distributions 79
2.4 Continuous Probability Distributions 86
2.5 Sampling Distributions 97
2.6 Other Sampling Distributions 108
2.7 Chapter Summary 116
2.8 Chapter Exercises 116

■ Example 2.1

A quality control specialist for a manufacturing company that makes complex aircraft parts is concerned about the costs generated by defective screws at two points in the production line. These defective screws must be removed and replaced before the part can be shipped. The two points in the production operate independent of each other, but a single part may have defective screws at one or both of the points. The cost of replacing defective screws at each point, as well as the long-term observed proportion of times defective screws are found at each point, is given in Table 2.1.

On a typical day, 1000 parts are manufactured by this production line. The specialist wants to estimate the total cost involved in replacing the screws. This example illustrates the use of a concept called probability in problem solving. While the main emphasis of this chapter is to develop the use of probability for statistical inference, there are other uses such as that illustrated in this example. The solution is given in Section 2.3 where we discuss discrete probability distributions.

Statistical Methods, DOI: 10.1016/B978-0-12-374970-3.00002-0

Table 2.1 Summary of Defective Screws

Point in the Production Line	Proportion of Parts Having Defective Screws	Cost of Replacing Defective Screws
A	0.008	$0.23
B	0.004	$0.69

2.1 INTRODUCTION

Up to now, we have used numerical and graphical techniques to describe and summarize sets of data without differentiating between a sample and a population. In Section1.8 we introduced the idea of using data from a sample to make inferences to the underlying population, which we called statistical inference, and is the subject of most of the rest of this text. Because inferential statistics involves using information obtained from a sample (usually a small portion of the population) to draw conclusions about the population, we can never be 100% sure that our conclusions are correct. That is, we are constantly drawing conclusions under conditions of uncertainty. Before we can understand the methods and limitations of inferential statistics we need to become familiar with uncertainty. The science of uncertainty is known as **probability** or probability theory. This chapter provides some of the tools used in probability theory as measures of uncertainty, and particularly those tools that allow us to make inferences and evaluate the reliability of such inferences.

Subsequent chapters deal with the specific inferential procedures used for solving various types of problems.

In statistical terms, a population is described by a distribution of one or more variables. These distributions have some unique characteristics that describe their location or shape.

Definition 2.1 *A **parameter** is a quantity that describes a particular characteristic of the distribution of a variable. For example, the mean of a variable (denoted by μ) is the arithmetic mean of all the observations in the population.*

Definition 2.2 *A **statistic** is a quantity calculated from data that describes a particular characteristic of the sample. For example, the sample mean (denoted by $\overline{y}$) is the arithmetic mean of the values of the observations of a sample.*

In general, statistical inference is the process of using sample statistics to make deductions about a population probability distribution. If such deductions are made on population parameters, this process is called *parametric* statistical inference. If the deductions are made on the entire probability distribution, without reference to particular parameters, the process is called *nonparametric* statistical inference. The majority of this text concerns itself with parametric statistical inference (with the exception of Chapter 14). Therefore, we will use the following definition:

Definition 2.3 *Statistical inference is the process of using sample statistics to make decisions about population parameters.*

An example of one form of statistical inference is to estimate the value of the population mean by using the value of the sample mean. Another form of statistical inference is to postulate or hypothesize that the population mean has a certain value, and then use the sample mean to confirm or deny that hypothesis. For example, we take a small sample from a large population with unknown mean, μ, and calculate the sample mean, $\overline{y}$, as 5.87. We use the value 5.87 to estimate the unknown value of the population mean. In all likelihood the population mean is not exactly 5.87 since another sample of the same size from the same population would yield a different value for $\overline{y}$. On the other hand, if we were able to say that the true mean, μ, is between two values, say 5.70 and 6.04, there is a larger likelihood that we are correct. What we need is a way to quantify this likelihood. Alternatively, we may hypothesize that μ actually had the value 6 and use the sample mean to test this hypothesis. That is, we ask how likely it is that the sample mean was only 5.87 if the true mean has a value of 6? In order to answer this question, we need to explore a way to actually calculate the probability that $\overline{y}$ is as small as 5.87 if $\mu = 6$. We start the discussion of how to evaluate statistical inferences on the population mean in Section 2.5.

Applications of statistical inferences are numerous, and the results of statistical inferences affect almost all phases of today's world. A few examples follow:

1. The results of a public opinion poll taken from a sample of registered voters. The statistic is the sample proportion of voters favoring a candidate or issue. The parameter to be estimated is the proportion of all registered voters favoring that candidate or issue.
2. Testing light bulbs for longevity. Since such testing destroys the product, only a small sample of a manufacturer's total output of light bulbs can be tested for longevity. The statistic is the mean lifetime as computed from the sample. The parameter is the actual mean lifetime of all light bulbs produced.
3. The yield of corn per acre in response to fertilizer application at a test site. The statistic is the mean yield at the test site. The parameter is the mean yield of corn per acre in response to given amounts of the fertilizer when used by farmers under similar conditions.

It is obvious that a sample can be taken in a variety of ways with a corresponding variety in the reliability of the statistical inference. For example, one way of taking a sample to obtain an estimate of the proportion of voters favoring a certain candidate for public office might be to go to that candidate's campaign office and ask workers there if they will vote for that candidate. Obviously, this sampling procedure will yield less than unbiased results. Another way would be to take a well-chosen sample of registered voters in the state and conduct a carefully controlled telephone poll. (We discussed one method of taking such a sample in Section 1.9, and called it a random sample.) The difference in the credibility of the two estimates