1

Review of Basic Statistical Concepts and Matrices

The material in this chapter is generally taught in a first course in statistical methods, which is prerequisite for studying this book. So, depending on the reader's background, this chapter may be treated as a quick review or may be studied in depth.

1.1 Overview

What Is Statistics?

Statistics, in a narrow sense, is a branch of science that deals with making inferences about populations based on samples. In a broader sense, statistics encompasses collection, organization, and summarization of data; presentation of data in tabular and graphical form; developing models for the purpose of understanding random and nonrandom phenomena; use of models for prediction; mathematical approaches to decision making and evaluation of risks; and so forth. Regression (and correlation) analysis is an area of statistics that deals with methods for investigating the existence of associations and, if present, the nature of the associations, among various observable quantities. For instance, one might be interested in knowing whether there is any association between age and blood pressure and, if so, what the nature of this association is. One might also be interested in expressing associations in the form of mathematical equations. Regression analysis is one method of investigating the presence of associations if appropriate data are available. The discovery of associations and the ability to express such associations in a precise mathematical form *may* enable one to predict the unobservable value of a variable based on the observed value of one or more associated or related variables. They may also help determine how one might control the values of one variable by manipulating the values of a related variable. For instance, one might be able to manipulate or control one's blood pressure by controlling one's diet if the nature of the association between blood pressure and the dietary components is understood. We do not make any statements regarding that elusive and controversial concept of cause and effect. Rather, we simply say that it *may* be possible to control the values of one variable by manipulating the

values of a related variable. This can be confirmed only by controlled experimental investigations.

In this book we present the fundamental ideas that form the core of regression (and correlation) analysis, and we give numerous practical applications of the methods developed.

Prerequisites

We assume that you are familiar with the material that is usually taught in an introductory course in statistical methods, which includes concepts of populations, samples, point estimation, confidence interval estimation, tests, and Gaussian distribution (also called normal distribution) models. These are reviewed briefly in Sections 1.2 through 1.6. We further assume that you are acquainted with functions and functional notation and elementary topics in matrix arithmetic. These topics are reviewed in Sections 1.7 and 1.8, respectively.

A brief introduction to the multivariate Gaussian distribution model (also referred to as the multivariate normal distribution) appears in Section 1.9. This topic is not usually taught in an introductory statistics course, but we feel that it is useful background material for studying regression analysis.

Remarks on Computing

Regression analysis requires a great deal of computing, and it will be helpful for you to learn to use computers and statistical computing packages. However, since some of you will not have access to a computer or a suitable computing package, we present most of the basic computations and require you to do only computations that can be done easily on a hand-held calculator. Thus you are not *required* to have knowledge of a computing language or a statistical computing package to read and understand the material in this book. Nevertheless, *we strongly urge that you use a computing package if one is available.*

Practically every statistical computing package contains programs for regression analyses. Some of the popular statistical packages are MINITAB, SAS, SPSS, BMDP, S-PLUS, etc. This textbook is accompanied by a laboratory manual that explains how to use one of the statistical packages—MINITAB or SAS—to perform the calculations needed for regression analysis. The reason that we wrote a laboratory manual using MINITAB is that this system is very easy to learn and does not detract from the main topic, regression. Also, many colleges and universities use this package as a tool for teaching statistics. The reason that we wrote a laboratory manual using SAS is that SAS is widely used by data analysts and researchers in colleges, universities, government, and industry. Also, many graduate students who study regression are already familiar with SAS and would like to analyze their own research using SAS. If computations are needed for the procedures discussed in any section of this book, the corresponding sections of the accompanying laboratory manual explain how to do the calculations in MINITAB or SAS. For example, in Section 3.6 we discuss confidence intervals for simple linear regression, and in Section 3.6 of the laboratory manual we show how to use MINITAB or SAS to perform

the calculations needed for confidence intervals. If MINITAB or SAS is not available, you can either ignore the laboratory manual or, if you have another statistical package available, you can work through the assignments in the laboratory manual using that package. Nothing that follows in this book requires knowledge of or the use of computers and statistical computing packages.

1.2 Basic Ingredients for Statistical Inference

One of the aims of science is to relate, describe, and predict events in the world in which we live. These activities are also important in business and our everyday affairs. In almost every aspect of human endeavor, it is useful to be able to predict future events based on present and past information.

To describe situations of interest, we must define them precisely and decide what we want to determine or predict. To illustrate, we first use a very simple example. You are handed a box that contains 10,000 marbles that are indistinguishable except for color—some are white and some are green—and you are asked to determine the proportion of green marbles in the box. You are not allowed to look in the box, but you are allowed to select 20 marbles at random from the box and note their colors. On the basis of this information (the colors of these 20 marbles), you must estimate what proportion of the 10,000 marbles are green. Of course you do not expect that by examining only 20 marbles you will be able to determine exactly the proportion of the entire 10,000 marbles that are green, but you would like to use scientific reasoning to come as close to the true answer as possible. The science of statistics and probability can be useful for making a decision in this situation and for attaching a measure of uncertainty (or certainty) to your result.

The preceding example may seem trivial, but it contains the basic ingredients of the most complicated of problems. Mother Nature has her secrets "locked up in a box" and, in order to make intelligent decisions about everyday activities, we may want to describe or predict the contents of that box. It may be possible to observe a part of Mother Nature's box (observe the colors of the 20 marbles) and make an intelligent decision as to its contents.

The Four Basic Ingredients

From this example we abstract four fundamental concepts that will be useful in examining more complicated situations: **populations, models, parameters, and samples and inferences.** These are summarized in Table 1.2.1.

We now consider a more realistic illustration.

EXAMPLE 1.2.1

Suppose a company that owns a chain of department stores is considering opening a store in city A. There are many things the company wants to know about city A so that it can decide whether it will be profitable to build a store there. One item of interest to the company is last year's average annual income per household in

TABLE 1.2.1
Four Fundamental Concepts

Concept	Description
Population	A **population** of items (sometimes also called a population of units) is a specified set of items or units in which an investigator is interested. The population may be **real** or it may be **conceptual**. It may be a population that existed in the past, or exists now, or will exist in the future, or exists only in one's imagination. A **real** population exists now and the entire population or any part of the population is available now and may be examined. A **conceptual** population is a population that is not available now and cannot be examined. Typically, a conceptual population is one that will exist in the future or exists only in one's imagination. A population may remain constant, or it may change slowly with time, or it may change rapidly with time. In the preceding illustration, the set of 10,000 marbles is the population of items. It is a real population since it is available for study now. Also it is a population that will remain constant, i.e., the colors of the marbles (white or green) will not change with time.
Model	A population **model** is a description of the quantities (numbers or attributes) of interest associated with each item in the population. In the illustration above, the attribute of interest is the color of the marbles and the description is that the marbles are indistinguishable except for color, with some being white and others green.
Parameter	**Parameters** are summary numbers that characterize various aspects of the population and are usually the quantities of interest to the investigator. In the preceding illustration, the proportion p of marbles that are green, characterizes one aspect of the population and is the quantity of interest in the investigation.
Sample	A **sample** is a set of items selected from the population, and observations are made on this set of items. On the basis of this sample a decision is made about the values of the parameters of interest, and this decision is usually accompanied by a measure of the uncertainty in the answer. In the preceding illustration, the set of 20 marbles is a sample from the population of 10,000 marbles. The proportion of green marbles in the sample may be used as an *estimate* of the proportion of all 10,000 marbles in the box that are green. Thus one uses the information contained in the sample to make **inferences** about population characteristics of interest.

the city. The four fundamental concepts outlined in Table 1.2.1 as they apply to this investigation are as follows:

1. The **population** of interest is the set of households in city A last year. A household must be precisely defined (which may not be an easy task!). For all practical purposes, we can regard this population as being constant although families may have moved in or out of the city during the course of the year. Consequently this is a real population that can be studied now.
2. The population **model** may be that the set of annual incomes of the households last year is a Gaussian population with mean μ and standard deviation σ, both of which are unknown.

3 The parameter of interest is μ, the average income of the households in the population.

4 A random sample of some specified number of households can be selected and last year's annual income of each of the selected households recorded. Based on these data, a decision can be made about the value of μ, the parameter of interest. ■

In the next four sections we discuss in detail each of the four basic ingredients:

- Population
- Model
- Parameters
- Samples and inferences

1.3 Population

A population of items is defined to be any set of items that one wants to study (refer to Table 1.2.1.). In Example 1.3.1 we describe several populations of items.

EXAMPLE 1.3.1

The following are some examples of populations of items.

a The set of all U.S. citizens who paid federal income tax last year.

b The set of all automobiles that will be made by manufacturer A next year.

c The set of all U.S. citizens who were diagnosed as having lung cancer three years ago and are still alive today.

d The set of all trees on a specified tree farm on July 1 two years hence.

e The set of all farms in the states of Iowa and Nebraska last year.

f The set of all grocery stores in the U.S. next year.

g The set of all human beings who will be born in Los Angeles, California, in the year 2000.

h The set of all ball bearings that will be made in a specified production plant next month.

i The set of all plastic containers that a specified manufacturer may make next year, using every possible process temperature between 300° F and 400° F.

j The set of all automobile tires that will be manufactured by a certain company next year using a newly developed tread design.

k The set of all daily record sheets that will be maintained by a particular power plant, containing information including the amount of sulfur-dioxide emitted and the amount of electrical power generated for each day that the power plant will be in operation next year.

l The set of all daily data records that will be kept by a particular monitoring station, containing information about the total daily flow values at a given monitoring point, and the total daily precipitation recorded at a certain gauge, for a specified river, for next year.

m The set of all *measured* values of the length of a single bolt.

n The set of all *measured* values of the breaking strength of a given metal rod. ■

The set of items to be studied, called the *population of items*, must be precisely defined. The number of items in a population can be finite or infinite; if this number is finite we denote it by N. This number N is generally unknown, but in most practical problems it is large.

Notice that the dates are specified for each of the populations of items described in Example 1.3.1; otherwise the set of items would not be precisely defined. For instance, in Example 1.3.1(e) what was a farm last year may not be a farm this year; in (f) a store that is a grocery store next year may not be a grocery store two years hence, etc.

Note In Example 1.3.1 the populations in (a)–(h) and (j)–(l) are all finite, whereas in (i), (m), and (n) they may be infinite. The populations described in (a), (c), and (e) are *real populations*, and we can study these populations now. The populations described in (b), (d), (f), (g), (h), (j), (k), and (l) are *conceptual populations* since they are future populations. They do not exist at the present time and cannot be observed *now*. The populations described in (i), (m), and (n) are also conceptual populations. In fact, they are *imagined* populations; the entire population will *never* become available. In the population described in (i), it is impossible for the manufacturer to actually manufacture plastic containers using every possible process temperature between 300° F and 400° F. But we can certainly *imagine* this population and ask questions about it. For instance, we may want to know what process temperature would lead to production of plastic containers having the required strength. For the population described in (m), it is possible to examine part of the population, viz., the next several measured values, but the entire population can only be *imagined* and will never become available. The population described in (n) is also an *imagined population*. The very act of measurement will destroy the given metal rod and so only *one* value can be observed from this imagined population. Nevertheless, we may want to make inferences about these conceptual populations so that decisions can be made or actions can be taken *now*.

Each item in a population of items possesses one or more characteristics that may be of interest in an investigation, as you can see in Example 1.3.2.

EXAMPLE 1.3.2

In Example 1.3.1(a) we may be interested in the age of each person, the I.Q. of each person, the weight of each person, the sex of each person, the political affiliation of each person, the marital status of each person, etc. In Example 1.3.1(b) we may be interested in studying the miles per gallon each automobile will get, the first-year

maintenance cost for each automobile, the number of miles each car will be driven the first year after its purchase, etc. In Example 1.3.1(d) we may be interested in the age of each tree, the diameter of each tree measured 4.5 feet from the ground, the height of each tree, etc. In Example 1.3.1(g) we may be interested in the number of years each person will live, the number of dollars each person will spend on education, the number of siblings each person will have, etc. ■

Associated with each item in a population are one or more numbers (age, height, income, etc.) or attributes (sex, marital status, political affiliation, etc.) of interest. In this book we are mainly concerned with numerical quantities associated with each population item. This *set of numbers* is called a *population of numbers* and will be referred to as the **population.** If a single number is associated with each item in the population, then we say that this population of numbers is a **univariate population.** If more than one number is associated with each item in the population, then we say that this population of numbers is a **multivariate population.** A **bivariate population** is a special case of a multivariate population, where each population item has two numbers of interest associated with it. A **trivariate population** is a special case of a multivariate population, where each population item has three numbers of interest associated with it. More generally, if each item has k numbers of interest associated with it, we refer to the population of numbers as a ***k*-variate** or a ***k*-variable population.**

In Example 1.3.3 we list specific **sets of numbers (populations)** that can be associated with each population of items listed in Example 1.3.1.

EXAMPLE 1.3.3

The following are populations (of numbers):

a The amount of money in dollars that each U.S. citizen earned as interest income last year.

b The maintenance cost in dollars and the number of miles each car will be driven during its first year.

c The age of each person and the average number of cigarettes each one smoked per day for the five years prior to diagnosis of lung cancer.

d The height, the diameter measured at 4.5 feet above ground level, and the dollar value of each tree, on July 1 two years hence.

e The size of each farm in acres and the profit each farm made.

f The profit, in dollars, that each grocery store will make.

g The number of dollars each person will spend on health care and the number of years each person will live.

h The diameter of each ball bearing.

i The strength of each plastic container and the temperature at which it is made.

j The set of all tread-depths remaining at the end of 50,000 miles of driving of all the automobile tires that will be manufactured by the company next year.

k The set of all total daily sulfur dioxide emission values and the amount of electrical power generated for each day of next year.

l The set of all total daily flow values and corresponding daily precipitation values that will be recorded by the monitoring station next year.

m The set of all measured values of the length of the bolt.

n The set of all measured values of the breaking strength of the rod. ■

Note that in Example 1.3.3 (a), (f), (h), (j), (m), and (n) are examples of univariate populations, (b), (c), (e), (g), (i), (k), and (l) are examples of two-variable (bivariate) populations, and (d) is an example of a 3-variable (trivariate) population.

Notation

At the beginning of an investigation, a target population of items is defined and one or more numbers associated with each item are identified as being of interest. *This resulting population (of numbers) is what an investigator is interested in.*

If a population is a univariate population consisting of N numbers, this population of numbers is written symbolically as $\{Y_1, Y_2, \ldots, Y_N\}$, where Y_I represents the Ith number (the number associated with the Ith item) in the population. Alternatively, we may write $\{Y\}$ without specifying the number of items in the population. If the population is a bivariate population consisting of N pairs of numbers, this population is written symbolically as $\{(Y_1, X_1), (Y_2, X_2), \ldots, (Y_N, X_N)\}$ (or simply as $\{(Y, X)\}$) where Y_I refers to the first quantity and X_I refers to the second quantity associated with the Ith item.

Sometimes a double subscript notation is used, as in $\{(Y_{11}, Y_{12}), (Y_{21}, Y_{22}), \ldots, (Y_{N1}, Y_{N2})\}$. Here Y_{IJ} stands for the Jth quantity of interest associated with the Ith population item. When working with multivariate populations, the double subscript notation is almost always used. For instance, a k-variable population may be denoted by $\{(Y_{11}, Y_{12}, \ldots, Y_{1k}), \ldots, (Y_{N1}, Y_{N2}, \ldots, Y_{Nk})\}$ or simply by $\{(Y_1, Y_2, \ldots, Y_k)\}$. Other symbols such as X or Z may be used instead of Y when convenient.

The population of numbers that is of interest to the investigator is called the **target population.** In some situations, the target population is a real population and is available for study. This is the case for the populations (a), (c), and (e) in Example 1.3.3. But, in other situations, the target population of interest is a conceptual population and is not available for study, as in the case of populations (b), (d), (f), (g), (h), (i), (j), (k), (l), (m), and (n) of Example 1.3.3. Even when the target population is a real population, it may sometimes be impractical or too expensive to study this population, and thus it is unavailable for study. When the target population is unavailable for study, one is sometimes able to find another population that resembles the target population and is a real population that is available for study now. Such a population is called a study population. We discuss these in some detail next.

Target Population and Study Population

Target Population

The target population is the population that is the *target* of a study. Conclusions from an investigation are to be applied to this population. As stated earlier, the target population is often unavailable for study. In many investigations, the numbers in the target population are **future values**, but we want to make decisions about them now. Example 1.3.3(d) is a good illustration. Suppose we are interested in what the dollar value of each tree will be on July 1 two years hence, but we must make this determination now. The target population in this case consists of the dollar values of each of the trees on July 1 two years hence, so this population is unavailable now. Therefore it is not possible to study the target population now. This is also the situation in Example 1.3.3(b), where we want to predict the average first-year maintenance cost of cars that will be produced by manufacturer A *next* year. We must make the prediction at the beginning of the year so we can plan a maintenance budget. But the target population is a future population and is not available for study now.

In fact it is quite common for a target population, which is the population of interest, to be unavailable for study, either because it consists of *future values* or because it is *impractical*, *inconvenient*, or *too expensive* to study this population. Thus we are led to consider another population, the study population.

Study Population

When the target population is unavailable for study, we are sometimes able to find another population that *resembles* the target population and is available for study now. Such a population is called a **study population**. Thus the study population is the population that is actually studied during an investigation. In those situations where the target population itself is available for study, the target population and the study population are one and the same. When the target population is unavailable for study, the study population should be chosen to *resemble* the target population as closely as possible.

To illustrate, consider Example 1.3.3(d). The target population is in the future, so it is not available for study now. The study population can be defined as the set of dollar values of all trees as of the present date. This is a real population that is available for study now. In Example 1.3.3(b) we can use as the study population the first-year maintenance costs for the same make of automobiles last year. In Example 1.3.3(i), no suitable study population may be available. In this case the investigator can produce plastic containers in a pilot plant under several different process temperatures between 300° F and 400° F and measure the strengths of these containers. The data thus generated are all that is available and this collection of pairs of numbers, i.e., the strength of a plastic container and the temperature at which it was made, can be considered the study population.

Methods are available for making *valid statistical inferences* about the study population. When the study population is different from the target population, generalizations from the study population to the target population are subject-matter considerations, and we have to rely on the investigator's judgments. All statistical inference procedures discussed in this book refer to the study population.

1.4 Model

A univariate population to be studied consists of N numbers, where N is generally quite large. In Example 1.3.3(a), for instance, the N numbers are the amounts, in dollars, earned as interest income by each of the N citizens; in Example 1.3.3(f), the N numbers are the profits, in dollars, each grocery store will make next year.

To study a population of N numbers one must have some way of organizing these N numbers. One useful approach is to organize them into a **probability histogram** and find a mathematical function that approximates this histogram. Such a mathematical function is usually called a **probability density function.** Theoretical statisticians use a variety of probability density functions to study different types of populations, the most important among them being **Gaussian** probability density functions. They are the subject of our discussion in this section. The Gaussian population model for multivariate populations is discussed in Section 1.9.

Gaussian Populations

A **theoretical Gaussian population** $\{Y\}$ is completely specified by its mean, denoted by μ_Y, and its standard deviation, denoted by σ_Y. The distribution of this population is described by the probability density function

$$f(Y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left[-\frac{(Y-\mu_Y)^2}{2\sigma_Y^2}\right] \quad \text{for} \quad -\infty < Y < \infty \qquad \textbf{(1.4.1)}$$

The area under the curve, defined by the function in (1.4.1) between the values a and b ($a < b$), gives the proportion of population values that are greater than a but less than or equal to b (see Figure 1.4.1).

The probability density functions corresponding to two different theoretical Gaussian populations are shown in Figure 1.4.2. These two populations have different values for μ_Y, but they have the same standard deviation σ_Y. It is clear from this figure that changing the value of μ_Y changes only the location of the population distribution and not its shape (curves (a) and (b) in Figure 1.4.2 have the same shape but different locations). On the other hand, changing the value of σ_Y changes the shape of the population distribution but not the location, as exemplified by curves (c) and (d) in Figure 1.4.3. These two curves have the same location (same value of μ_Y) but different shapes (different values of σ_Y).

FIGURE 1.4.1

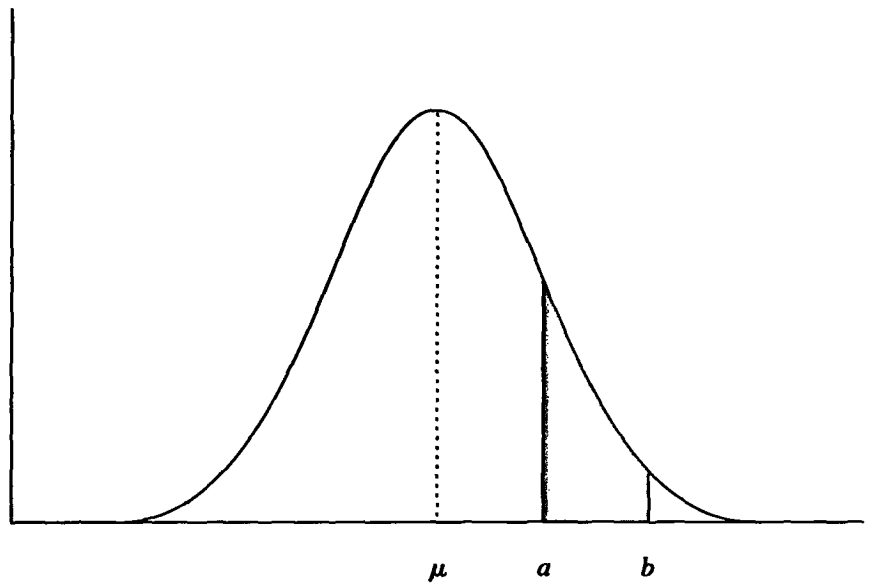


FIGURE 1.4.2

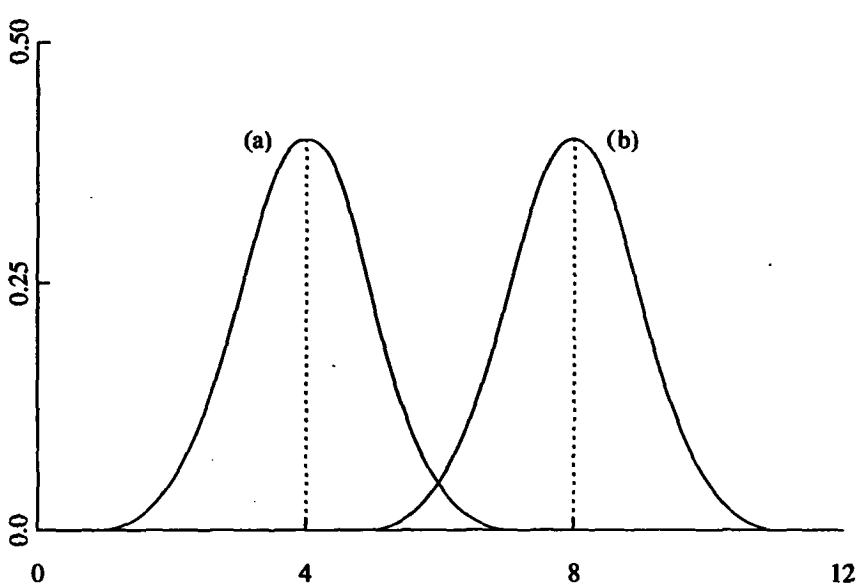
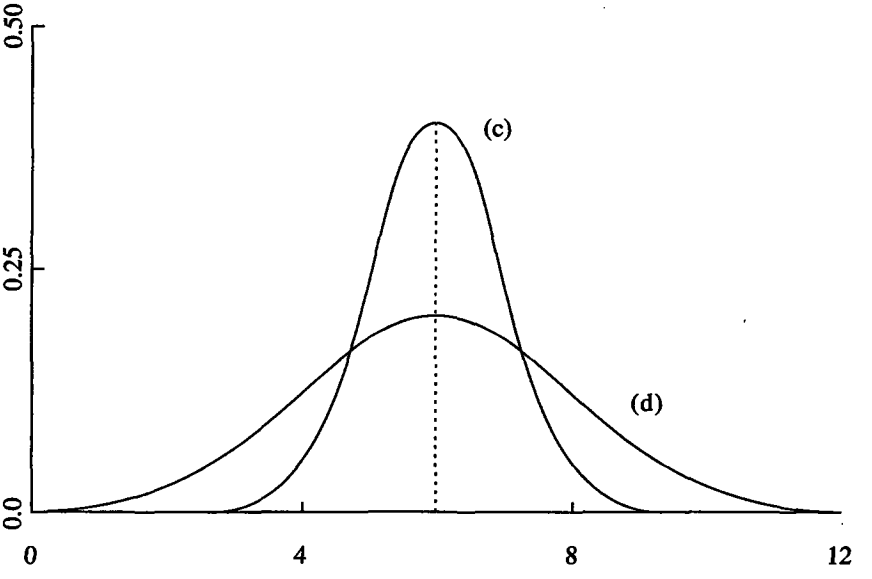


FIGURE 1.4.3



For the sake of completeness we now give the definitions of the **mean**, the **standard deviation**, and the **variance** of a population $\{Y\}$ consisting of the N numbers $Y_1, Y_2, \ldots, Y_N$.

DEFINITION **Mean**

The mean (also called the *average*) of a population $\{Y\}$ is denoted by μ_Y and is defined by

$$\mu_Y = \frac{1}{N}\sum_{I=1}^{N} Y_I \quad \blacksquare \tag{1.4.2}$$

DEFINITION **Standard Deviation**

The standard deviation of a population $\{Y\}$ is denoted by σ_Y and is defined by

$$\sigma_Y = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(Y_I - \mu_Y)^2} \quad \blacksquare \tag{1.4.3}$$

D E F I N I T I O N **Variance**

The variance of a population $\{Y\}$ is the square of its standard deviation. It is denoted by σ_Y^2 and is defined by

$$\sigma_Y^2 = \frac{1}{N}\sum_{I=1}^{N}(Y_I - \mu_Y)^2 \quad \blacksquare \tag{1.4.4}$$

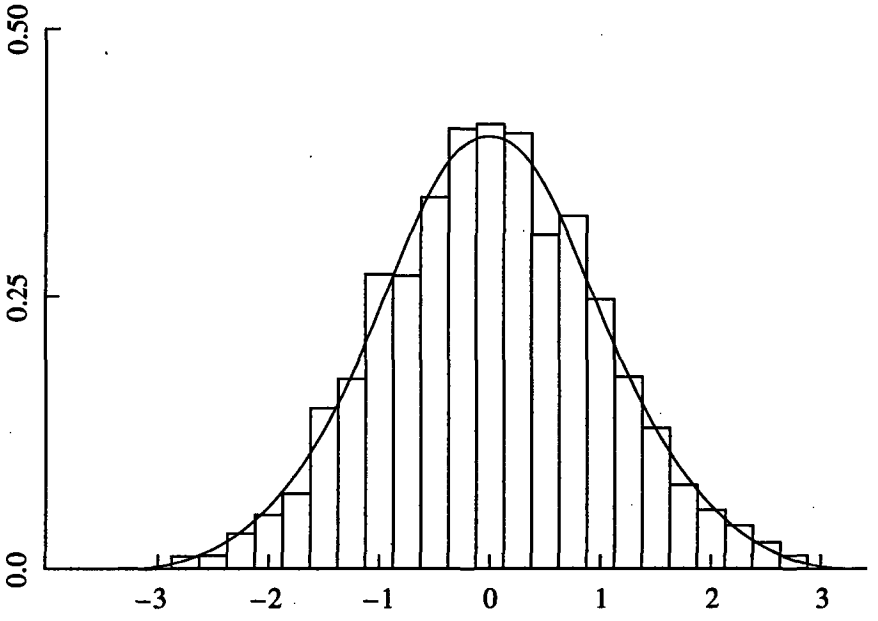
The definitions of mean and standard deviation for infinite populations require concepts from calculus and some knowledge about convergence of infinite series. If you are interested in them you should refer to more advanced books [15], [19], [25], [31].

As stated earlier, in practical problems the N unknown numbers that make up a population could conceptually be used to form a **probability histogram** (a histogram for which the total area equals 1). If this histogram is well approximated by the mathematical function defined in (1.4.1) when the value of the mean and the standard deviation of these N population numbers are substituted for μ_Y and σ_Y, then *we may proceed as if the population under study is Gaussian.* For example, Figure 1.4.4 shows a probability histogram of a population of 4,000 numbers (generated using a computer), with $\mu_Y = 0$ and $\sigma_Y = 1$. The probability density curve of the theoretical Gaussian population with $\mu_Y = 0$, $\sigma_Y = 1$ in (1.4.1) is also displayed there. From these we can see that the theoretical Gaussian population described by equation (1.4.1) appears to be a good approximation to this finite population of size 4,000. Thus, for most practical purposes, we can regard this finite population consisting of 4,000 numbers as a Gaussian population.

The theoretical Gaussian population is a mathematical abstraction and no such population can exist in real investigations, but it is useful as an approximation to finite populations in many applied problems. It is also used in theoretical statistics to derive point estimates, confidence intervals, and tests for μ_Y and σ_Y. Similar approximations are common in other situations. For example, a circle, which can be defined mathematically, does not exist in the real world, but it is useful as an approximation to the wheel; a rectangle, which can be defined mathematically, does not exist in the real world, but it is useful as an approximation to obtain the area of a tabletop, a farmer's field, the side of a building, etc.

Because we never know the population values in a real problem, it is impossible to be certain that the population under study is actually Gaussian. However, even when the population is not exactly Gaussian, the statistical inference procedures are often accurate enough for making decisions as long as the population is *approximately* Gaussian. In some instances statistical procedures are available for detecting serious violations of the Gaussian assumption.

FIGURE 1.4.4



1.5 Parameters (Summary Numbers)

In Table 1.2.1, we described parameters as numbers that summarize certain important characteristics of a population. Even if we had the entire population available (which is rarely the case in a real problem), there would be too many numbers for an investigator to use for making decisions without summarizing them into a smaller set. Thus a judicious summarization of population characteristics is extremely important. Here we discuss this in some detail, first for univariate populations and then for multivariate populations.

Parameters for Univariate Populations

Consider the univariate population $\{Y\}$ and assume for the moment that all the numbers in the population are available. To understand the important characteristics of the population of numbers, it is convenient to summarize them by a smaller set of numbers or perhaps by using suitable graphical techniques. Probability histograms and cumulative frequency curves are two of the commonly used graphical descriptions of populations.

Generally we would like to use one or more (say m) summary numbers, $\theta_1, \ldots, \theta_m$, to describe various characteristics of the entire population. These m

numbers, called (population) **parameters**, would be computed from the entire population if it were known, but in almost all real problems they have to be estimated based on sample data. In this book population parameters are almost always denoted by Greek letters.

Although the m parameters $\theta_1, \ldots, \theta_m$ cannot, in general, tell us everything about the whole population, they often adequately summarize certain important characteristics of the population that are relevant to the study at hand. The **mean** and the **standard deviation** are two particularly useful and important population parameters, especially in the case of Gaussian populations, and we discuss these next.

Mean

Often we would like to use a *single parameter* to *represent* the entire population of numbers. In this book we use the **mean** of the population (also called the **average**), defined in (1.4.2), as the single number that best represents the entire population of numbers. The symbol μ_Y (μ is the Greek letter mu) represents a population mean, and the subscript Y indicates which population is under study. The mean μ_Y is often also used to predict the Y value of any randomly chosen population item. In most real problems the mean of a population is not known because not all of the population elements $\{Y\}$ are known. Nevertheless, it is a number we would like to have available to use to make decisions about a population. So we select a sample from the population and estimate μ_Y. This is discussed later.

In many situations, single numbers such as the mean (or the **median**) are often used to represent the value of each item in a population. For instance, if you plan to retire in the United States, it would be useful to know something about the cost of housing in the various cities where you might choose to live. The mean (average) price of new homes in each city would provide useful information. Also, the average annual cost of living in each city would be helpful, as would the average age of people who live in the area. In many situations the mean is a useful summary of populations: (1) average income of men and women, (2) average number of children per household, (3) average gas mileage for a certain make of cars, (4) average yield per acre of corn in a certain state, (5) the batting average of a certain baseball player, etc.

Standard Deviation

Although the mean μ_Y is often used as the best single number to represent the entire population of numbers $\{Y\}$, no single number can adequately describe or represent an entire population. Therefore an additional summary number is used to tell us how well μ_Y represents the entire population. This number is σ_Y, the standard deviation of the population, which was defined in (1.4.3). Intuitively, the smaller σ_Y is, the more useful μ_Y is as a representative value for the entire population. Likewise, the larger σ_Y is, the less useful μ_Y is as a representative of the whole population. Note that if $\sigma_Y = 0$, then all the numbers in the population $\{Y\}$ are the same and equal to μ_Y, and the mean is a perfect representation of the entire population. But if a substantial proportion of the values of Y are much smaller than the mean and others

are much larger, then σ_Y will be large and μ_Y may not adequately represent each population value.

Chebyshev's Theorem

Recall that about 95% of the numbers in a Gaussian population lie between $\mu_Y - 2\sigma_Y$ and $\mu_Y + 2\sigma_Y$, and about 99% of the numbers lie between $\mu_Y - 3\sigma_Y$ and $\mu_Y + 3\sigma_Y$, so if σ_Y is "small," μ_Y does indeed represent the population quite well. In fact, for any population $\{Y\}$, the following fundamental result, due to the mathematician Chebyshev, is valid.

Chebyshev's Theorem

Let $\{Y\}$ be *any* one-variable population with mean μ_Y and standard deviation σ_Y. Then, for any positive number c, the proportion of population values that are greater than $\mu_Y - c\sigma_Y$ but less than $\mu_Y + c\sigma_Y$ (i.e., that are less than c standard deviations away from the mean) is greater than or equal to $1 - 1/c^2$.

For example, if $c = 3$, Chebyshev's theorem says that for any univariate population $\{Y\}$, at least $1 - 1/3^2 = 8/9 = 88.9\%$ of the population values are within 3 standard deviations of the mean. In particular, for a population $\{Y\}$ whose mean is μ_Y and whose standard deviation σ_Y is 2, we can say that at least 88.9% of the population values are less than 3 standard deviations, i.e., 6 units, away from the mean. Similarly, at least $1 - 1/4^2 = 15/16 = 93.75\%$ of the population values are less than 4 standard deviations (8 units) away from the mean. The exact value of such proportions cannot be found without knowing more about the population, but Chebyshev's theorem does give us a bound for such proportions. From the preceding discussion it is clear that μ_Y and σ_Y are two summary numbers that tell us a great deal about a population and that σ_Y can be used to determine how well μ_Y represents the population values (i.e., how close the population values are to μ_Y). We now consider parameters for multivariate populations.

Parameters for Multivariate Populations

Suppose a population of size N is a k-variate population ($k > 1$). Then there are k quantities associated with each population item, resulting in Nk numbers in all. Each of the k quantities associated with the population items gives rise to a univariate population. Let the k quantities for item I be denoted by $X_{I1}, \ldots, X_{Ik}$. Then the numbers $X_{11}, \ldots, X_{N1}$ form a univariate population with mean μ_1 and standard deviation σ_1, the numbers $X_{12}, \ldots, X_{N2}$ form another univariate population with mean μ_2 and standard deviation σ_2, etc. Thus $\mu_1, \ldots, \mu_k$, and $\sigma_1, \ldots, \sigma_k$ are parameters associated with the k-variate population. The Nk numbers may be schematically represented as in Table 1.5.1.

TABLE 1.5.1
Schematic Representation of a k-Variate Population of Size N

	k Measurements on Each Item			
Items	1	2	⋯	k
1	X_{11}	X_{12}	⋯	X_{1k}
2	X_{21}	X_{22}	⋯	X_{2k}
⋮	⋮	⋮	⋮	⋮
I	X_{I1}	X_{I2}	⋯	X_{Ik}
⋮	⋮	⋮	⋮	⋮
N	X_{N1}	X_{N2}	⋯	X_{Nk}
Mean	μ_1	μ_2	⋯	μ_k
Standard deviation	σ_1	σ_2	⋯	σ_k

For an illustration, consider Example 1.3.3(d). Let X_{I1}, X_{I2}, and X_{I3} denote the *height, the diameter at 4.5 feet above ground level, and the dollar value of the Ith* tree in the population. We thus have a three-variable population, and μ_1, μ_2, and μ_3 represent the mean height, the mean diameter, and the mean dollar value of the trees in the population. Likewise, σ_1, σ_2, and σ_3 represent the standard deviations of the heights, the diameters, and the dollar values, respectively, of the trees in the population.

Coefficient of Correlation

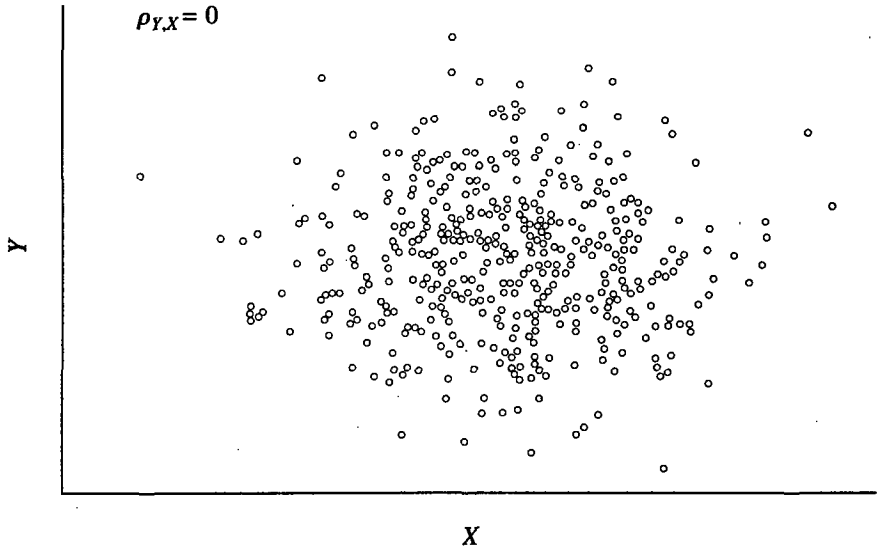
When dealing with multivariate populations there is often a need to summarize associations among various quantities measured on the same item. For instance, how can we summarize the association that may exist between the height of a tree and the diameter of that tree at 4.5 feet above ground level? How can we summarize the association or relationship between the number of miles a car is driven per year and the yearly maintenance cost for the car? One summary measure that is sometimes used for this purpose is the **coefficient of correlation** between two variables (also called the **Pearson correlation** or the **product moment correlation**). This measure of association is denoted by $\rho_{Y,X}$ (ρ is the Greek letter rho) when summarizing the relationship between the variables Y and X. It is defined by

$$\rho_{Y,X} = \frac{\sum_{I=1}^{N}(Y_I - \mu_Y)(X_I - \mu_X)}{\sqrt{\left[\sum_{I=1}^{N}(Y_I - \mu_Y)^2\right]\left[\sum_{I=1}^{N}(X_I - \mu_X)^2\right]}} \tag{1.5.1}$$

for a bivariate population $\{(Y, X)\}$ of N items. Note that (1.5.1) implies $\rho_{Y,X} = \rho_{X,Y}$; i.e., the coefficient of correlation between Y and X is the same as the coefficient of correlation between X and Y. It can be shown that $\rho_{Y,X}$ is a number between -1 and

+1. It is equal to +1 when the population values (Y_I, X_I) all lie on a straight line that has a positive slope, and it is equal to −1 when all the population values lie on a straight line with a negative slope. Figures 1.5.1–1.5.3 are *scatterplots* of bivariate populations $\{(Y, X)\}$ consisting of 500 items, each with a different value of $\rho_{Y,X}$.

FIGURE 1.5.1



Observe that a positive value of the coefficient of correlation between Y and X indicates that, generally speaking, larger values of Y are associated with larger values of X, and smaller values of Y are associated with smaller values of X. Likewise, when the correlation coefficient is negative, we find that larger values of Y are associated with smaller values of X, and smaller values of Y are associated with larger values of X. Notice that a general lack of *linear* association is indicated when the magnitude of the correlation coefficient is close to zero. *Whereas correlation coefficients may have useful interpretations for some problems (particularly when Y and X are approximately linearly related), they may provide no useful interpretation and, in fact, be a misleading summary quantity in other problems.*

In summary, when dealing with a k-variate population, say $\{(X_1, \ldots, X_k)\}$, the basic summary quantities or parameters that are often used are the means, $\mu_1, \ldots, \mu_k$, and standard deviations, $\sigma_1, \ldots, \sigma_k$, of the k univariate populations, along with the $\binom{k}{2} = k(k-1)/2$ coefficients of correlation, ρ_{X_1,X_2}, $\rho_{X_1,X_3}, \ldots,$ ρ_{X_{k-1},X_k}, between the $k(k-1)/2$ pairs of variables (X_1, X_2), $(X_1, X_3), \ldots, (X_{k-1}, X_k)$.

FIGURE 1.5.2

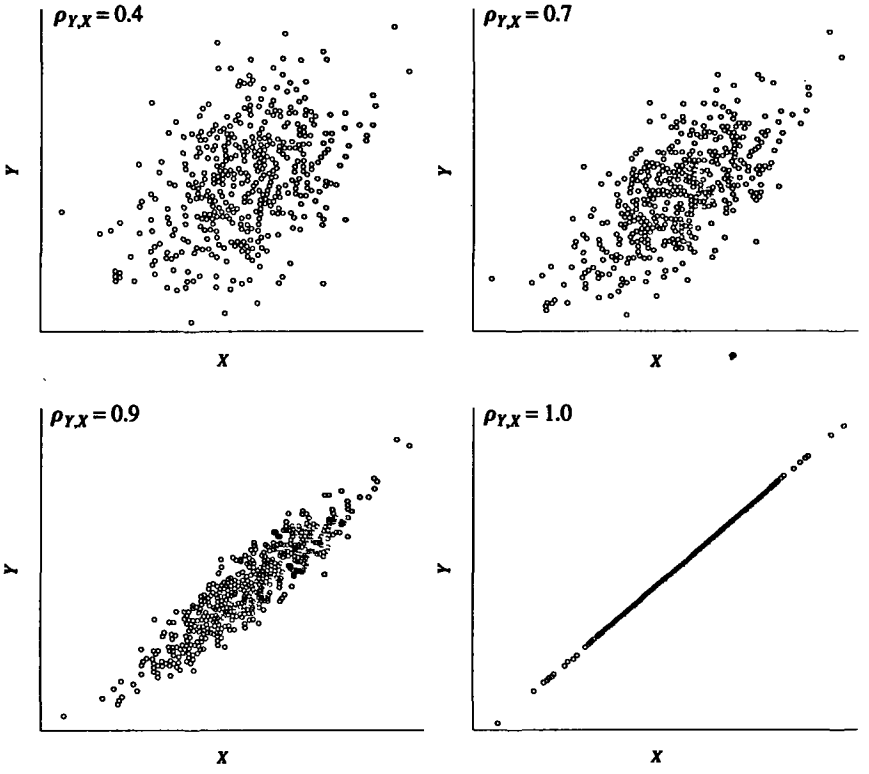
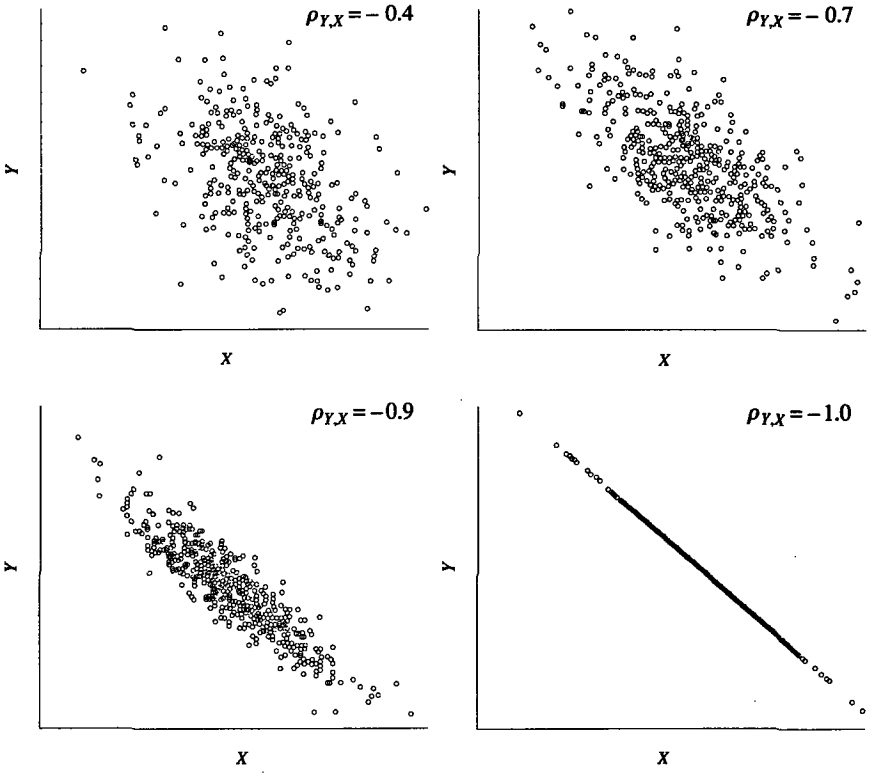


FIGURE 1.5.3



1.6 Samples and Inferences

As stated earlier, an investigator must first define the study population. Whenever possible the target population itself should be the study population. If this is not possible, then the study population should resemble the target population as closely as possible. After the study population is defined, it is described with a model, and parameters that are needed in order to make decisions are identified. The next step is to determine the values of these parameters. Since the population is never completely known in a real situation, an investigator can never know the parameter values exactly. A commonly used procedure is to select a subset of the items, referred to as a **sample**, from the study population and to use the measurements associated with these sample items to *infer* the values of population parameters of interest. If the sample is selected from the population using one of several random sampling procedures, then it is possible to assign a *measure of uncertainty* to the conclusions derived using such a sample. The process of making inferences about the values of population parameters based on random samples is called *statistical inference.*

We use the symbols $y_1, y_2, \ldots, y_n$ to denote the n randomly sampled values from the population $\{Y\}$. Note that lowercase letters (i.e., y) are used to denote sample values, and lowercase n represents the sample size. The values $y_1, y_2, \ldots, y_n$ and n are known numbers.

Consider the population in Example 1.3.3(a). An investigator may be interested in μ_Y, the average interest income earned by the individuals in this population last year. She can select a random sample of n people, say $n = 1{,}000$, from this population and obtain the amount of interest income earned by each person last year. From these data, she can make inferences on μ_Y. In Example 1.3.3(c) the investigator may be interested in μ_Y, the average age of people who were diagnosed as having lung cancer three years ago and are still alive today, or he may be interested in p, the proportion of persons diagnosed as having lung cancer three years ago who are less than 25 years old now. He can select a random sample of 75 persons (say) from this population of persons and obtain the age of each. From these data, he can make inferences about μ_Y and p. In Example 1.3.3(e) the investigator may be interested in μ_Y, the average profit for all farms in Iowa and Nebraska, so a random sample of n farms, say $n = 80$, can be selected and their profits recorded. From these data, the investigator can make inferences about μ_Y.

In each of these cases, the target population and the study population are one and the same. We now discuss situations where this is not so.

In Example 1.3.3(b), an investigator may be interested in the first-year maintenance costs of the cars in the population. Recall that this population is a conceptual population since it is a future population. It may be reasonable to use the collection of first-year maintenance costs of all cars manufactured by company A *last year* as the study population. A random sample of n cars, say $n = 25$, can be obtained from this study population, and their first-year maintenance costs can be determined by contacting the owners of the cars in the sample. From these data, the investigator can make inferences about the study population. If an investigator feels that the study population resembles the target population sufficiently closely, then he/she can use the conclusions to make decisions about the target population.

In Example 1.3.3(j) an investigator may be interested in the remaining tread-depth (in millimeters) of tires after 50,000 miles of driving. Again the target population is a future population. However, no suitable study population is available because there is no existing information regarding tread-wear for tires manufactured using the new tread design. So the investigator may decide to select the first 100 tires manufactured using the new tread design, mount them on different cars that will then be driven 50,000 miles under conditions similar to those typical customers might experience, and measure the tread-depth remaining. These are the only data available to the investigator. *If it is reasonable to regard these data as a random sample from the target population, then we can use these data to make valid statistical inferences about the target population.*

Simple Random Sample

The size n of the sample to be selected is determined by the investigator based on a careful consideration of costs and the objectives of the study. Samples can be

selected from the study population using any one of several random sampling procedures that are discussed in textbooks on sampling methods [32]. A thorough understanding of the advantages and disadvantages of the various sampling methods is required before one of the procedures is selected, and investigators would be wise to consult a professional statistician for advice. The simplest of all sampling procedures is one called *simple random sampling*, and a sample obtained using this procedure is called a **simple random sample**. We assume throughout, unless specifically stated otherwise, that samples are drawn using the simple random sampling procedure. The definition of a simple random sample follows:

D E F I N I T I O N **Simple Random Sample**

If a population has N elements, there are

$$H = \binom{N}{n} = \frac{N \times (N-1) \times \cdots \times (N-n+1)}{1 \times 2 \times \cdots \times n}$$

distinct samples of size n that can be obtained. If each of these H samples has an equal chance of being selected, then the sample actually obtained is called a *simple random sample of size n.* ■

Three Types of Inference Procedures

There are three general types of statistical inference procedures that are commonly used. They are

- Point estimation
- Confidence intervals
- Statistical tests of hypotheses

We discuss these next.

Point Estimation

Suppose θ is an unknown population parameter (θ could be μ_Y, σ_Y, σ_Y^2, $\rho_{Y,X}$, etc.). A **point estimate** of θ is a *number* computed from sample data, to be used by an investigator as the value of θ (because θ is unknown and hence unavailable) in making decisions. Estimation procedures for calculating point estimates are useful when their values are close to the actual values of the unknown population parameters.

Estimates of the Mean and the Standard Deviation of a Univariate Population If $y_1, y_2, \ldots, y_n$ is a simple random sample from a population $\{Y\}$ whose mean is μ_Y and whose standard deviation is σ_Y, the commonly used estimates of μ_Y and σ_Y, denoted by $\hat{\mu}_Y$ and $\hat{\sigma}_Y$, respectively, are

$$\hat{\mu}_Y = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (1.6.1)$$

and

$$\hat{\sigma}_Y = \sqrt{\frac{SSY}{(n-1)}} \tag{1.6.2}$$

where SSY is called the corrected sum of squares for Y, and is defined by

$$SSY = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

The estimate of σ_Y^2 is $\hat{\sigma}_Y^2$, the square of the estimated standard deviation $\hat{\sigma}_Y$. Note that

$$SSY = (n-1)\hat{\sigma}_Y^2$$

Estimate of the Coefficient of Correlation $\rho_{Y,X}$ in a Bivariate Population If $(y_1, x_1), \ldots, (y_n, x_n)$ is a simple random sample from a bivariate population $\{(Y, X)\}$, the estimate of the coefficient of correlation between Y and X that is widely used is given by

$$\hat{\rho}_{Y,X} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{1.6.3}$$

Note *Throughout this book we use a hat symbol ˆ over an unknown parameter to indicate a point estimate of that unknown parameter. For example, $\hat{\sigma}_Y$ represents a point estimate of the unknown population parameter σ_Y.*

Unbiased Estimates To assess whether a *procedure* for estimating a population parameter is a good procedure, *we must investigate the estimates given by the procedure for every sample that could have been obtained.* Consider the procedure just given of using the sample mean for estimating the population mean. If a population $\{Y\}$ has N items and if the sample size is n, then there are $H = \binom{N}{n}$ possible samples of size n, any one of which could have been selected and, under simple random sampling, each of these possible samples has the same probability of being the actual sample chosen. Conceptually, for each of the H possible samples of size n, we could compute the sample mean. This would result in H sample means $\bar{y}_1, \ldots, \bar{y}_H$, some of which would be close to the population mean and others would not be; some would be larger than the population mean and others would be smaller, but any one of the H possible sample means could end up as the mean of the actual sample selected by the investigator. It can be shown mathematically that the average of all of the H possible sample means is equal to the population mean. We express this fact by saying that the *sample mean is an unbiased estimate of the population mean.* Many investigators and statisticians consider unbiased estimation procedures desirable.

More generally, suppose we want to estimate a parameter θ of a population. We select a random sample of size n from the population and compute an estimate $\hat{\theta}$ of θ using the sample values according to some procedure or formula. In principle, for every possible sample of size n (recall there are H of these), an estimate can be obtained using the same procedure. This would result in H estimates $\hat{\theta}_1, \ldots, \hat{\theta}_H$. If the mean of these H estimates, $\hat{\theta}_1, \ldots, \hat{\theta}_H$, of θ is *equal* to θ, then the procedure

used to compute these estimates is said to be an *unbiased estimation procedure*. Of course, in any given problem we compute only one of these H values of $\hat{\theta}$ because we have only one sample of size n available. If we compute this one value $\hat{\theta}$ using an unbiased estimation procedure, then we say that $\hat{\theta}$ is an **unbiased estimate** of θ.

We stated earlier that the sample mean is an unbiased estimate of the population mean. If the sample size n is very large, then $\hat{\sigma}_Y$ as defined in (1.6.2) is approximately an unbiased estimate of σ_Y. For further discussion of the concept of unbiasedness and other desirable properties of estimators, you should consult books on mathematical statistics [15], [19], [25], [31].

To summarize, a point estimate of a parameter θ is a number $\hat{\theta}$, obtained from sample data, that is often used as the value of θ for making various decisions. In addition to a point estimate of θ, we would like to have some way to determine how close $\hat{\theta}$ is to the unknown parameter θ. For most applications in this book, *confidence interval procedures* provide one way of obtaining such information; this is our next topic of discussion.

Confidence Intervals

Suppose an investigator is interested in determining the value of some parameter θ associated with a population under study. As discussed earlier, a point estimate of θ gives us a single value that can be used as the value of θ for making decisions. An alternative approach to estimating the value of θ is to provide, not a single value, but a range of values, say an interval of values, and specify how confident we are that θ is contained in this interval. Such an interval is called a **confidence interval** for θ. This is a reasonable approach because it is often the case that only a range of *possible* or *plausible* values of θ is needed for making decisions and not its *exact* value.

(1.6.4) A **two-sided confidence interval** for a population parameter, say θ (where θ could be μ_Y, σ_Y, σ_Y^2, etc.), is an interval whose lower endpoint, say L, and upper endpoint, say U, are computed from sample data. The procedure for computing the confidence interval has the property that the probability of the resulting interval actually containing θ is a prescribed value $1 - \alpha$, which is referred to as the **confidence coefficient** or the **confidence level** associated with the computed confidence interval. We say we are $100(1 - \alpha)\%$ confident that $L \leq \theta \leq U$, and we write $C[L \leq \theta \leq U] = 1 - \alpha$.

As an illustration, suppose the computations lead to the statement $C[3.86 \leq \theta \leq 6.12] = 0.80$. Then an investigator has 80% confidence that the interval 3.86 to 6.12 contains θ; i.e., we have 80% confidence that θ, the unknown population parameter, is between 3.86 and 6.12.

The confidence level $1 - \alpha$ is selected by the investigator. Values of α that are typically used are 0.01, 0.05, and 0.10, with corresponding confidence levels $1 - \alpha$

of 0.99, 0.95, and 0.90. The correct value of α to select will depend on the particular problem under consideration, and it need not be restricted to one of the values 0.99, 0.95, 0.90, etc. It may be the case that $1-\alpha = 0.80$ or 0.85 is appropriate for a particular problem. The value of α (and of $1-\alpha$) must, of course, lie in the interval from 0 to 1.

The Meaning of a $1-\alpha$ Confidence Interval

The meaning of a $1-\alpha$ confidence interval for an unknown population parameter θ, computed using a simple random sample of size n from a population of N items, is as follows:

There are $H = \binom{N}{n}$ possible samples of size n that could be selected from the population. Conceptually, every possible sample of size n could be selected from the population, and a confidence interval for θ could be computed from each sample; there would be H confidence intervals. *The proportion of these H intervals that would include the unknown parameter θ is equal to $1-\alpha$.* Of course in a real problem an investigator selects only one sample of size n and computes only one confidence interval for θ. Our level of confidence can be quantified by the number $1-\alpha$ because it is known that a $1-\alpha$ proportion of all H possible intervals would include the true unknown value of the parameter θ. (1.6.5)

Equal-Tailed Confidence Intervals and One-sided Confidence Bounds

In (1.6.5) the proportion of the H confidence intervals that will fail to contain θ is α. Suppose that in half of these cases the actual value of θ is greater than the corresponding computed upper endpoint U and that in the other half of these cases θ is smaller than the corresponding computed lower endpoint L. Then the confidence interval procedure is said to be **equal-tailed.** Nearly all of the confidence interval procedures discussed in this book are equal-tailed. In this case a two-sided confidence statement

$$C[L \le \theta \le U] = 1-\alpha$$

has the following additional interpretations.

- If we know before looking at the data that only an upper bound for θ is needed for making decisions, then we can be $100(1-\alpha/2)\%$ confident that θ is less than or equal to U. We will write this

as $C[\theta \leq U] = 1 - \alpha/2$ and say that U is an **upper confidence bound** for θ with confidence coefficient equal to $1 - \alpha/2$.

- If we know before looking at the data that only a lower bound for θ is needed for making decisions, then we can be $100(1 - \alpha/2)\%$ confident that θ is greater than or equal to L. We write this as $C[L \leq \theta] = 1 - \alpha/2$ and say that L is a **lower confidence bound** for θ with confidence coefficient equal to $1 - \alpha/2$. **(1.6.6)**

Note If, before looking at the data, we do not know whether the upper bound or the lower bound for θ is needed for making decisions, but after computing the two-sided $1 - \alpha$ (equal-tailed) confidence interval and looking at L and U we decide to use only L or only U to make a decision, then the appropriate confidence coefficient associated with the decision is $1 - \alpha$ and not $1 - \alpha/2$.

Symmetric Confidence Intervals

In most applications discussed in this book (important exceptions are confidence intervals for standard deviations, variances, and correlation coefficients), a two-sided confidence interval for the unknown parameter θ will be **symmetric** about the best point estimate $\hat{\theta}$ of θ; i.e., the values L and U will be of the form $L = \hat{\theta} - D$ and $U = \hat{\theta} + D$, where $\hat{\theta}$ and D (and hence L and U) are computed from the data, and $D = (U - L)/2$. In these cases the following is a useful way to view a two-sided confidence interval associated with the point estimate $\hat{\theta}$:

If $\hat{\theta}$ is a point estimate of θ and $L \leq \theta \leq U$ is a $1 - \alpha$ **two-sided confidence interval** for θ that is symmetric about $\hat{\theta}$, then we have $100(1 - \alpha)\%$ confidence that $\hat{\theta}$ is within $D = (U - L)/2$ units of the true unknown value θ. **(1.6.7)**

For example, a useful alternative way to describe the confidence statement $C[3.86 \leq \theta \leq 6.12] = 0.80$, symmetric about a point estimate $\hat{\theta}$ of θ, is to say *we have 80% confidence that $\hat{\theta}$ is within* $(6.12 - 3.86)/2 = 1.13$ *units of* θ.

Note Some authors use the symbol P instead of C; for example, some authors write $P[\theta \leq 9.67] = 0.80$ and say that the probability is 80% that θ is less than or equal to 9.67. This is an incorrect use of the word *probability* because the statement $\theta \leq 9.67$ is either true or false and hence $P[\theta \leq 9.67]$ is either 0 or 1 (we do not know which), but it is certainly not equal to 0.80. To avoid this incorrect usage of the term *probability*, we use the symbol C instead of P and say confidence instead of probability where the word confidence has the meaning given in (1.6.5).

For many, and in fact most, applications in this book, a $1 - \alpha$ confidence interval is of the form

$$\hat{\theta} - \text{table-value} \times SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + \text{table-value} \times SE(\hat{\theta}) \qquad (1.6.8)$$

where $SE(\hat{\theta})$ denotes the standard error of $\hat{\theta}$, which is a measure of how precisely $\hat{\theta}$ estimates θ. The detailed meaning of $SE(\hat{\theta})$ is as follows:

Suppose the population under study has N items, and a simple random sample of size n is selected. The total number of different possible samples of size n is denoted by H, which equals $\binom{N}{n}$. Conceptually, each possible sample of size n could be selected from the population, and both an estimate $\hat{\theta}$ of θ and the error of estimation $\hat{\theta} - \theta$ could be computed from each sample. The standard deviation of the collection $\{\hat{\theta} - \theta\}$ of estimation errors, which we may write as $SD(\hat{\theta} - \theta)$, tells us how good the estimation procedure is for estimating θ. Calculation of $SD(\hat{\theta} - \theta)$ requires that the entire population of numbers be available. However, a valid estimate of $SD(\hat{\theta} - \theta)$ can often be computed from sample data; it is denoted by $SE(\hat{\theta} - \theta)$, or simply by $SE(\hat{\theta})$ for ease of notation, and it is called the **standard error of $\hat{\theta}$.** (1.6.9)

Authors' Recommendation

When the confidence interval for an unknown parameter θ is of the form given in (1.6.8), the investigator should be given the following:

1 The point estimate $\hat{\theta}$ of θ.

2 The standard error of $\hat{\theta}$, viz., $SE(\hat{\theta})$.

3 The degrees of freedom (df) associated with $SE(\hat{\theta})$.

With this information the investigator can easily compute confidence intervals for θ with any confidence coefficient $1 - \alpha$ as needed.

We now illustrate the procedures for computing point estimates and confidence intervals.

EXAMPLE 1.6.1

Company A manufactures compact cars whose gas mileages Y (in miles per gallon) form a Gaussian population with mean μ_Y and standard deviation σ_Y, both of which are unknown. The manager of a car rental company is interested in purchasing a large fleet of these cars from company A next year. In order to adhere to company policy, she will do so only if she is reasonably confident that the average gas mileage

of all cars of this type, to be manufactured by company A next year, will be at least 25 miles per gallon when driven over a test route. Thus the target population of numbers is the collection of gas mileages of all cars of this type manufactured by company A next year. This population is unavailable for study now, so the manager decides to study the population of gas mileages of all cars of the type required, manufactured by company A *last year*, because she feels that the study population resembles the target population for all practical purposes. So she obtains a simple random sample of 10 cars manufactured by company A last year and determines their gas mileages. The data are shown in Table 1.6.1 and are also given in the file **table161.dat** on the data disk.

Only one-sided lower confidence bounds are of interest in this problem. The manager wants a 90% lower confidence bound for μ_Y, and she will buy the fleet of cars from company A only if the lower bound exceeds 25. The required lower bound can be obtained from a two-sided 80% confidence interval as explained in (1.6.6). We first compute the following quantities:

$$\hat{\mu}_Y = \bar{y} = 25.302$$
$$SSY = \sum_{i=1}^{10}(y_i - \bar{y})^2 = 3.67796$$
$$\hat{\sigma}_Y = 0.63927$$
$$SE(\hat{\mu}_Y) = \hat{\sigma}_Y/\sqrt{n} = 0.63927/\sqrt{10} = 0.20215$$

A two-sided 80% confidence statement for μ_Y (using $t_{1-\alpha/2:n-1} = t_{0.90:9} = 1.383$ obtained from Table T-2 in the appendix) is

$$C[\hat{\mu}_Y - t_{0.90:9}SE(\hat{\mu}_Y) \le \mu_Y \le \hat{\mu}_Y + t_{0.90:9}SE(\hat{\mu}_Y)]$$
$$= C[25.02 \le \mu_Y \le 25.58] = 0.80$$

Thus the required one-sided 90% lower confidence bound for μ_Y is 25.02. Observe that the lower endpoint of 25.02 in the two-sided 80% confidence statement can in fact be used as a one-sided 90% lower confidence bound since we knew without looking at the data that it is the lower bound that is needed to make a decision. Thus the manager can be 90% confident that the average gas mileage for cars manufactured last year by company A is greater than or equal to 25.02 mpg. To extrapolate this result to cars that will be manufactured next year by company A is a judgment decision.

Note In this book, when we demonstrate computations, we may carry several significant digits for all intermediate calculations and report the final results to more

TABLE 1.6.1

Car number	1	2	3	4	5	6	7	8	9	10
Gas mileage (mpg)	25.72	25.24	25.19	25.88	26.42	24.48	25.11	24.29	25.06	25.63

significant digits than is perhaps necessary. In actual applications we advise that the *final results* be rounded to an appropriate number of significant digits. ■

Prediction Intervals

In many problems the interest is not in estimating the mean μ_Y of a population $\{Y\}$ but in predicting the Y value of an item that is yet to be chosen from the population. We denote this value by Y_0 and call it a future value to be randomly chosen from the population $\{Y\}$. For instance, consider Example 1.6.1 and suppose you plan to purchase a new car to be made by manufacturer A. You would like to know what you can expect your maintenance cost to be the first year after you purchase it. Thus a future observation Y_0 is to be randomly chosen from a population $\{Y\}$, and you want to predict its value now.

Suppose $y_1, y_2, \ldots, y_n$ is a simple random sample of size n from this population. The predicted value of Y_0 using these sample data is $\bar{y}$, the sample mean. Suppose you also want to obtain an interval, say $L \le Y_0 \le U$, such that you are 95% confident this interval will contain the value Y_0 that will be selected from the population. It may be tempting to use a 95% confidence interval for μ_Y, the mean first-year maintenance cost of *all* cars in the population, as the required interval for Y_0. *This is incorrect.* Whereas we can have 95% confidence that the interval for μ_Y will in fact contain μ_Y, our confidence will be lower than 95% that the interval for μ_Y will contain the future value Y_0 because individual values in the population can be, and generally are, different from the population mean. To account for this fact, a 95% confidence interval for a future value Y_0 has to be *wider* than a 95% confidence interval for μ_Y. This interval for Y_0 is called a 95% **prediction interval** and is given by (1.6.8) with θ replaced by Y_0 and $\hat{\theta}$ replaced by $\hat{Y}_0 = \bar{y}$. Also $SE(\hat{\theta})$ is replaced by $SE(\hat{Y}_0) = \hat{\sigma}_Y\sqrt{1 + (1/n)}$. Thus a $1 - \alpha$ confidence statement for Y_0 is

$$C[\hat{Y}_0 - t_{1-\alpha/2:n-1}SE(\hat{Y}_0) \le Y_0 \le \hat{Y}_0 + t_{1-\alpha/2:n-1}SE(\hat{Y}_0)] = 1 - \alpha \qquad (1.6.10)$$

We generally distinguish between a prediction interval and a confidence interval because a confidence interval is for a *fixed unknown parameter* in a population, and a prediction interval is for a *future observation* to be randomly selected from a population.

The Meaning of a $1 - \alpha$ Prediction Interval

The meaning of a $1 - \alpha$ prediction interval for Y_0, a future random observation to be selected from a population $\{Y\}$, computed using a simple random sample of size n from a population of N items, is as follows:

Choose a simple random sample of size n and use this to compute a $1 - \alpha$ two-sided prediction interval for a single future observation. This prediction interval may or may not contain the future observation. Repeat this procedure over and over where each time a new

simple random sample of size n is selected and a prediction interval is computed for a new future observation. Then in the long run, $100(1-\alpha)\%$ of all the prediction intervals will include the corresponding future observation. Our level of confidence that this interval will include the single future observation can be quantified by the number $1-\alpha$. **(1.6.11)**

EXAMPLE 1.6.2

Consider Example 1.6.1 and suppose that you are planning to purchase a new car to be made by company A next year. You want to predict what the miles per gallon (mpg) will be for the car you will buy. You also want an interval for the miles per gallon this car will get so you can have 80% confidence that your interval is correct. The gas mileage for the car you will buy is a future random observation obtained from the population of gas mileages of all cars that will be manufactured by company A next year. If we believe that the study population (last year's cars) closely resembles the target population (next year's cars), we can use last year's data for inference.

From Example 1.6.1 we get $\hat{\mu}_Y = \bar{y} = 25.302$ and $\hat{\sigma}_Y = 0.63927$. Thus $\hat{Y}_0 = 25.302$ and $SE(\hat{Y}_0) = 0.63927\sqrt{1+(1/10)} = 0.670$. Hence we get

$$C[24.37 \leq Y_0 \leq 26.23] = 0.80$$

and you have 80% confidence that the new car you will purchase will get between 24.37 and 26.23 miles per gallon of gasoline. ▪

For convenience, in Table 1.6.2 we have summarized the procedures for computing confidence intervals for the mean μ_Y and the standard deviation σ_Y of a Gaussian population. In the same table we also describe the procedure for obtaining a confidence interval (prediction interval) for a randomly chosen value Y_0 from a Gaussian population.

Statistical Tests of Hypotheses

Often an investigator conjectures that a parameter θ of a population is equal to, less than, or greater than a specified value q. Statistical tests, performed using sample data, are often used to help decide whether or not the data provide evidence *against* the conjecture. The investigator formulates an appropriate pair of hypotheses, one of which is designated as the null hypothesis (NH) and the other as the alternative hypothesis (AH), and a statistical test is used to determine what evidence the sample data can provide *against NH in favor of AH*.

A statistical test typically consists of the four steps below.

1 For a population parameter of interest, say θ, and for a specified value q, we suppose that an investigator is interested in one of the three pairs of hypotheses

TABLE 1.6.2
Point Estimates and Confidence Intervals for μ_Y, σ_Y, and Y_0 in a One-Variable Gaussian Population

Notation: $\bar{y} = \frac{1}{n}\sum y_i$; $SSY = \sum (y_i - \bar{y})^2$	
Inference	**Formulas and Procedures**
Point estimate of μ_Y, σ_Y	$\hat{\mu}_Y = \bar{y}$ $\hat{\sigma}_Y = \sqrt{SSY/(n-1)}$
Two-sided $1-\alpha$ confidence intervals for μ_Y	$\hat{\mu}_Y - t_{1-\alpha/2:n-1} SE(\hat{\mu}_Y) \le \mu_Y \le \hat{\mu}_Y + t_{1-\alpha/2:n-1} SE(\hat{\mu}_Y)$ where $SE(\hat{\mu}_Y) = \frac{\hat{\sigma}_Y}{\sqrt{n}}$
Two-sided $1-\alpha$ confidence intervals for σ_Y	$\sqrt{\frac{SSY}{\chi^2_{1-\alpha/2:n-1}}} \le \sigma_Y \le \sqrt{\frac{SSY}{\chi^2_{\alpha/2:n-1}}}$
Two-sided $1-\alpha$ confidence intervals for Y_0	$\hat{\mu}_Y - t_{1-\alpha/2:n-1}\hat{\sigma}_Y\sqrt{1+\frac{1}{n}} \le Y_0 \le \hat{\mu}_Y + t_{1-\alpha/2:n-1}\hat{\sigma}_Y\sqrt{1+\frac{1}{n}}$

(a)–(c) in Table 1.6.3 and wants to determine what evidence the data provide against NH in favor of AH.

2 A random sample of size n is selected from the study population and an appropriate number, say Q_C, called the **test statistic**, is computed using the sample data (the subscript C in Q_C stands for computed value).

3 This number Q_C is referred to an appropriate table of *percentiles* of a theoretical distribution, known as the *reference distribution*, and a number P is determined such that *if the NH is indeed true, then the probability of obtaining a value of Q_C, as unfavorable as or more unfavorable than the value actually obtained,*

TABLE 1.6.3

	NH	AH
(a)	$\theta = q$	$\theta \neq q$
(b)	$\theta \leq q$	$\theta > q$
(c)	$\theta \geq q$	$\theta < q$

is equal to P. This number P is called the **significance probability** or the **P-value** associated with the test of NH versus AH. Since P is a probability, it is a number between 0 and 1. The value of P is a measure of the evidence that the data provide against NH in favor of AH. Small values of P indicate that the data provide evidence against the null hypothesis in favor of the alternative hypothesis, whereas large values of P imply that the data *do not* provide evidence against the null hypothesis (note that large values of P do not necessarily imply that the data provide evidence in support of the null hypothesis).

Note Computation of exact P-values usually requires detailed statistical tables for the t-distribution, the χ^2-distribution, the F-distribution, etc. Such detailed tables are not easily available, but statistical packages such as MINITAB and SAS may be used to obtain exact P-values in most situations. Alternatively, approximate P-values may be computed by suitably interpolating table values. Appendix T has tables that can be used to obtain bounds for P-values, and these are generally adequate.

4 The investigator decides, based on a detailed knowledge of the problem, what values of P are to be considered small. It seems to be common practice among many investigators to *arbitrarily* select a number α (called the *size* of the test) equal to 0.05 or 0.01 and to consider the value of P to be small if it is less than α. The investigator would then **reject** NH if $P \leq \alpha$ and would *not reject* NH if $P > \alpha$. If this procedure is followed by the investigator, then the probability of rejecting NH when NH is in fact true is guaranteed to be no greater than α. However, the probability of rejecting NH when it is actually false can generally not be determined without further computations.

EXAMPLE 1.6.3

For an illustration consider Example 1.2.1. Suppose the company is interested in building a store in city A only if the average annual income per family is greater than \$20,000. Here $\theta = \mu_Y$, the average annual income per family in city A, and $q = \$20{,}000$. The null hypothesis and the alternative hypothesis that are appropriate for this problem are

$$\text{NH}: \quad \mu_Y \leq 20{,}000$$
$$\text{AH}: \quad \mu_Y > 20{,}000$$

If the investigator chooses $\alpha = .05$, then the probability of rejecting NH (and deciding $\mu_Y > 20{,}000$) when μ_Y is indeed less than or equal to \$20,000 is at most 5%. By choosing the value of α appropriately and using a statistical test, the investigator can control the probability of incorrectly deciding to build a store if the average annual income is too low (i.e., the probability of rejecting NH if NH is indeed true).

However, it should be pointed out that the probability of the company deciding to build a store when in fact the average annual income is greater than \$20,000 may be unacceptably small unless the sample is sufficiently large. This probability is determined by computing the *power* of the test. ■

The test statistic to compute and the appropriate table to consult for various statistical testing situations have been determined by mathematical theory, and they are described as appropriate. In particular, Boxes 1.6.1 and 1.6.2 summarize statistical tests for the mean μ_Y and the standard deviation σ_Y, respectively, of a univariate Gaussian population.

BOX 1.6.1 Hypothesis Tests for μ_Y

Let q be a number specified by the investigator. Compute the statistic

$$t_C = \frac{\hat{\mu}_Y - q}{(\hat{\sigma}_Y/\sqrt{n})}$$

a For testing NH: $\mu_Y = q$ versus AH: $\mu_Y \neq q$, the P-value is the value of α such that $|t_C| = t_{1-\alpha/2:n-1}$.

b For testing NH: $\mu_Y \leq q$ versus AH: $\mu_Y > q$, the P-value is the value of α such that $t_C = t_{1-\alpha:n-1}$.

c For testing NH: $\mu_Y \geq q$ versus AH: $\mu_Y < q$, the P-value is the value of α such that $-t_C = t_{1-\alpha:n-1}$.

BOX 1.6.2 Hypothesis tests for σ_Y

Let q be a positive number specified by the investigator. Compute the statistic

$$\chi^2_C = \frac{(n-1)\hat{\sigma}^2_Y}{q^2}$$

a For testing NH: $\sigma_Y = q$ versus AH: $\sigma_Y \neq q$, the P-value is equal to α, where α is a number between 0 and 1 and satisfies one of the following two conditions (it is not possible for both conditions to be satisfied unless $\alpha = 1$):

$$\chi_C^2 = \chi_{1-\alpha/2:n-1}^2$$

or

$$\chi_C^2 = \chi_{\alpha/2:n-1}^2$$

b For testing NH: $\sigma_Y \leq q$ versus AH: $\sigma_Y > q$, the P-value is the value of α such that $\chi_C^2 = \chi_{1-\alpha:n-1}^2$.

c For testing NH: $\sigma_Y \geq q$ versus AH: $\sigma_Y < q$, the P-value is the value of α such that $\chi_C^2 = \chi_{\alpha:n-1}^2$.

Relationship Between Tests and Confidence Intervals There is a relationship between size α tests about θ and $1 - \alpha$ equal-tailed confidence intervals for θ, and it is as follows.

1. Suppose we want a size α test of NH: $\theta = q$ against AH: $\theta \neq q$. Then NH is rejected if and only if the two-sided $1 - \alpha$ confidence interval for θ does not contain q.
2. Suppose we want a size α test of NH: $\theta \leq q$ against AH: $\theta > q$. Then NH is rejected if and only if q is less than the one-sided $1 - \alpha$ lower confidence bound L. Recall that this lower confidence bound L is actually the lower endpoint of the $1 - 2\alpha$ two-sided (equal-tailed) confidence interval for θ.
3. Suppose we want a size α test of NH: $\theta \geq q$ against AH: $\theta < q$. Then NH is rejected if and only if q is greater than the $1 - \alpha$ one-sided upper confidence bound U. Recall that this upper bound U is actually the upper endpoint of the $1 - 2\alpha$ two-sided (equal-tailed) confidence interval for θ.

Note From a $1 - \alpha$ *confidence interval (if it exists), we can obtain the result of a statistical test of size* α, *but from the result of a statistical test we* ***cannot*** *obtain the corresponding confidence interval.*

For example, suppose we want to test (with $\alpha = 0.05$) NH: $\mu_Y = 6.0$ against AH: $\mu_Y \neq 6.0$ by using a random sample from a Gaussian population with unknown mean μ_Y and unknown standard deviation σ_Y. If a $1 - \alpha = 0.95$ confidence interval for μ_Y is $13.1 \leq \mu_Y \leq 18.9$, then NH is rejected because 6.0 is not contained in this interval. On the other hand, suppose the $1 - \alpha = 0.95$ confidence interval for μ_Y is $3.5 \leq \mu_Y \leq 10.7$; then NH is not rejected because 6.0 is in this interval.

Authors' Recommendation

We recommend that traditional statistical tests of hypotheses for a parameter, say θ (where one rejects or does not reject NH), never be used if a confidence interval for θ is available because confidence intervals are always more informative than tests, and tests alone (without the accompanying confidence intervals) can be misleading. Since tests are taught and widely used by investigators, we discuss them in this book, but as a general rule we advise against their indiscriminate use.

We now illustrate the procedures just discussed for testing statistical hypotheses.

EXAMPLE 1.6.4

Consider Example 1.6.1 where the manager of a car rental company is interested in purchasing a large fleet of cars from company A if it can be determined that μ_Y, the average miles per gallon of all automobiles of this type, is at least 25 when the cars are driven over a specified test route. If the manager uses a statistical test to determine this, the appropriate NH and AH are

$$\text{NH}: \quad \mu_Y \leq 25$$
$$\text{AH}: \quad \mu_Y > 25$$

You may want to refer to Box 1.6.1 for details of the calculations. We obtain

$$t_C = \frac{25.302 - 25}{0.20215} = 1.494$$

From Table T-2 in the appendix we find that the P-value corresponding to this value of t_C is between 0.05 and 0.10. Linear interpolation gives a P-value equal to 0.0877. The P-value obtained from the statistical computing package MINITAB is equal to 0.085 (rounded to 3 decimals).

Suppose the manager chooses $\alpha = 0.10$; i.e., the manager decides to reject NH if $P < 0.10$ (which would correspond to a confidence level of $1 - \alpha = 0.90$ for the lower confidence bound). Then she would reject NH, and decide that the data provide enough evidence to conclude that the gas mileage will exceed 25 mpg. Observe that the significance probability gives us more information than the result of a test using a fixed prechosen value of α. It tells us that NH will be rejected for any value of α greater than 0.0847, but NH will not be rejected for any value of α less than or equal to 0.0847. Notice, however, that much more information is obtained from the two-sided 80% confidence interval for μ_Y. It actually tells us what the value of μ_Y is likely to be. In particular, we recall from Example 1.6.1 that the two-sided 80% confidence interval for μ_Y is given by the confidence statement

$$C[25.02 \leq \mu_Y \leq 25.58] = 0.80$$

On the basis of this confidence interval, the manager may conclude that μ_Y is close enough to 25 to be considered to be equal to 25 mpg for all practical purposes. ■

Simultaneous Tests (Tests about Several Parameters) and Simultaneous Confidence Intervals (Confidence Intervals for Several Parameters)

We have discussed how to test a hypothesis about a single parameter θ, how to obtain a confidence interval for a single parameter θ, and we have discussed the relationship between tests and confidence intervals. Researchers often conduct investigations where several parameters $\theta_1, \ldots, \theta_m$ are involved, and they want to examine the relationships among them. We illustrate with two examples.

EXAMPLE 1.6.5

A company has developed four new chemicals, any one of which can be added to cement in an attempt to increase the strength of cement building blocks. An experiment is conducted to examine the differences among the average strengths, say μ_1, μ_2, μ_3, μ_4, of the cement blocks made with each of the four chemicals; i.e., an experiment is conducted to examine the differences

$$\mu_1 - \mu_2,\ \mu_1 - \mu_3,\ \mu_1 - \mu_4,\ \mu_2 - \mu_3,\ \mu_2 - \mu_4,\ \mu_3 - \mu_4$$

To determine if the average strengths of cement blocks made with the four chemicals are different, many practitioners would conduct a statistical test of

$$\text{NH}: \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{against} \quad \text{AH}: \text{at least one equality does not hold} \tag{1.6.12}$$

■

EXAMPLE 1.6.6

An experiment is conducted to determine how different water temperatures, at which a new fabric is laundered, affect the strength of the fabric after 100 launderings. Three different water temperatures, 100° C, 140° C, and 180° C, were used, and five different pieces of the fabric were laundered at each of these three temperatures (a total of $5 \times 3 = 15$ different pieces of the fabric were used altogether). After the launderings, the strength of each piece of the fabric was measured. The investigator wants to examine the differences among the average strengths of the various pieces of the fabric laundered at different temperatures. To do this, it is common practice to conduct a statistical test of

$$\text{NH}: \mu_1 = \mu_2 = \mu_3 \quad \text{against} \quad \text{AH}: \text{at least one equality fails} \tag{1.6.13}$$

where μ_1, μ_2, and μ_3 are the average strengths of the fabric after being laundered 100 times in water at 100° C, 140° C, and 180° C, respectively. ■

Simultaneous Tests

Suppose $\theta_1, \ldots, \theta_m$ are m parameters of interest and that a researcher conjectures that the equalities $\theta_1 = q_1, \ldots, \theta_m = q_m$ are all true ($q_1, \ldots, q_m$ are specified numbers). This is often expressed by a null hypothesis of the form

$$\text{NH}: \left\{ \begin{array}{l} \theta_1 = q_1 \\ \theta_2 = q_2 \\ \quad \vdots \\ \theta_m = q_m \end{array} \right\}$$

The alternative hypothesis AH states that at least one of the equalities in NH is false. A statistical testing procedure will help decide whether or not the NH should be rejected in favor of AH at a specified level α. Such a testing procedure is called a **simultaneous test** if $m \geq 2$.

In Example 1.6.5 let $\theta_1 = \mu_1 - \mu_2$, $\theta_2 = \mu_1 - \mu_3$, and $\theta_3 = \mu_1 - \mu_4$. The null hypothesis in (1.6.12) can be expressed as

$$\text{NH}: \left\{ \begin{array}{l} \theta_1 = 0 \\ \theta_2 = 0 \\ \theta_3 = 0 \end{array} \right\}$$

The alternative hypothesis states that at least one of θ_1, θ_2, θ_3 is nonzero. Hence the corresponding statistical test is a simultaneous test (i.e., several hypotheses are tested simultaneously).

In Example 1.6.6 let $\theta_1 = \mu_1 - \mu_2$ and $\theta_2 = \mu_1 - \mu_3$. The NH in (1.6.13) can be expressed as

$$\text{NH}: \left\{ \begin{array}{l} \theta_1 = 0 \\ \theta_2 = 0 \end{array} \right\}$$

The alternative hypothesis states that at least one of θ_1 and θ_2 is nonzero. Again this leads to a simultaneous statistical test.

We have recommended that tests about a single parameter θ not be used without the accompanying confidence interval for θ (if it is available). The same recommendation applies for simultaneous tests about several parameters $\theta_1, \ldots, \theta_m$. However, when $m \geq 2$, there are two types of confidence intervals that can be used: (1) one-at-a-time confidence intervals and (2) simultaneous confidence intervals. We discuss these next.

One-at-a-Time Confidence Intervals

For one-at-a-time confidence intervals with confidence coefficient $1 - \alpha$, we compute confidence intervals for each parameter θ_i, and for any *one* of the statements below,

$$L_1 \leq \theta_1 \leq U_1$$
$$L_2 \leq \theta_2 \leq U_2$$

$$\vdots$$

$$L_m \leq \theta_m \leq U_m$$

we have $1 - \alpha$ confidence that it is correct. However, we do not have $1 - \alpha$ confidence (in fact, in most situations we would have less than $1 - \alpha$ confidence) that *all* preceding m statements are *simultaneously correct.*

Simultaneous Confidence Intervals

For simultaneous confidence intervals with confidence coefficient $1 - \alpha$, we compute confidence intervals for each θ_i such that we have $1 - \alpha$ confidence that all of the confidence intervals below are simultaneously correct.

$$L_1^* \leq \theta_1 \leq U_1^*$$
$$L_2^* \leq \theta_2 \leq U_2^*$$
$$\vdots$$
$$L_m^* \leq \theta_m \leq U_m^*$$

The difference between the interpretation of one-at-a-time confidence statements and simultaneous confidence statements is made clear by the following example.

EXAMPLE 1.6.7

Consider a group of 100 students of whom 80 are females and 20 are males. It is known that 70 of the females and 10 of the males are math majors. Suppose a student is to be randomly selected from this group. We are 80% confident that the chosen student will be a female (because 80 of the 100 students are females). We are also 80% confident that the chosen student will be a math major (because 80 of the 100 students are math majors). However, we cannot be 80% confident that the chosen student will be, *simultaneously*, a female and a math major! In fact we can be only 70% confident that the chosen student is a female math major (because 70 of the 100 students are female math majors).

In the same way, we may have $1 - \alpha$ confidence in the statement $L_1 \leq \theta_1 \leq U_1$ and also have $1 - \alpha$ confidence in the statement $L_2 \leq \theta_2 \leq U_2$, but this does not imply that we have $1 - \alpha$ confidence in the simultaneous statement $L_1 \leq \theta_1 \leq U_1$ *and* $L_2 \leq \theta_2 \leq U_2$. Such simultaneous confidence statements can, however, be made if appropriate table-values and computational procedures are used. ■

Authors' Recommendation

For each problem, the investigator must decide which type of confidence interval (one-at-a-time or simultaneous) to use. We recommend that simultaneous confidence intervals be used *only* in situations when an investigator must make a decision that depends on knowing all of the values θ_i simultaneously, with a specified level of confidence. That is, an investigator wants to have $1 - \alpha$ confidence that a decision is correct and, for the decision to be correct, *all* of the confidence intervals, $L_i \le \theta_i \le U_i$, $i = 1, \ldots, m$, must be simultaneously correct. Thus the investigator wants to have $1 - \alpha$ confidence that all m intervals are correct.

We illustrate with an example.

EXAMPLE 1.6.8

To be in compliance with environmental regulations, the manager of a coal-fired power-generating plant makes periodic measurements of emissions from the smokestacks. Using these measurements, she can estimate μ_i, for $i = 1, 2, 3, 4$, the average weekly emissions of four different toxic components. These estimates are used to determine whether the plant is in compliance with the regulations. The manager computes upper confidence bounds

$$\mu_1 \le U_1, \quad \mu_2 \le U_2, \quad \mu_3 \le U_3, \quad \mu_4 \le U_4$$

and she wants to be 99% confident that *all four* of the preceding statements are simultaneously correct. If the value of each U_i is less than the value required by the environmental regulations, she would perhaps conclude that the plant is in compliance. Here, the correctness of the overall decision depends on all of the four confidence statements being simultaneously correct. Thus simultaneous confidence statements are needed. ■

For many of the problems that we discuss, the formula for confidence intervals for $\theta_1, \theta_2, \ldots, \theta_m$ is given by (this formula does not apply to confidence intervals on standard deviations, variances, and correlation coefficients):

$$\hat{\theta}_i - \text{table-value} \times SE(\hat{\theta}_i) \le \theta_i \le \hat{\theta}_i + \text{table-value} \times SE(\hat{\theta}_i) \quad \text{for } i = 1, \ldots, m$$

and the table-value determines whether these m confidence intervals are one-at-a-time or simultaneous. Also, there are several different methods for obtaining simultaneous confidence intervals, each one requiring a different table-value. Which one of these to use depends on the particular problem being studied. We give you the appropriate table-values to use for each situation required in this book.

There is, however, one general method for obtaining simultaneous confidence statements that is easy to apply. This is called the **Bonferroni method**, which is explained in Box 1.6.3.

BOX 1.6.3 Bonferroni Method for Simultaneous Confidence Statements

Suppose m one-at-a-time confidence intervals are computed, each with a confidence coefficient equal to $1 - (\alpha/m)$, leading to the following confidence statements.

$$C[L_1 \le \theta_1 \le U_1] = 1 - \frac{\alpha}{m}$$
$$C[L_2 \le \theta_2 \le U_2] = 1 - \frac{\alpha}{m}$$
$$\vdots$$
$$C[L_m \le \theta_m \le U_m] = 1 - \frac{\alpha}{m}$$

Then the following simultaneous confidence statement is valid.

$$C\begin{bmatrix} L_1 \le \theta_1 \le U_1 \\ L_2 \le \theta_2 \le U_2 \\ \vdots \\ L_m \le \theta_m \le U_m \end{bmatrix} \ge 1 - \alpha$$

Example 1.6.9 illustrates the use of the Bonferroni method.

EXAMPLE 1.6.9

Consider the situation described in Example 1.6.1. Suppose we wish to compute a simultaneous 90% two-sided confidence statement for μ_Y and σ_Y of the form

$$C\begin{bmatrix} L_1 \le \mu_Y \le U_1 \\ \text{and} \\ L_2 \le \sigma_Y \le U_2 \end{bmatrix} \ge 0.90$$

Since we want to make *two* statements with simultaneous confidence coefficient greater than or equal to 0.90, we have $m = 2$ and $\alpha = 0.10$. According to the Bonferroni method, we must compute a confidence interval for μ_Y with confidence coefficient $1 - \alpha/m = 0.95$ and a confidence interval for σ_Y with confidence coefficient $1 - \alpha/m = 0.95$. Then the simultaneous confidence coefficient is *at least* $1 - \alpha = 0.90$. You should verify the following confidence interval calculations.

A two-sided 95% confidence statement for μ_Y can be computed using the formula given in Table 1.6.2. For this we need the table value $t_{0.975:9}$, which can be obtained from Table T-2 in the appendix and is equal to 2.262. The required confidence statement is

$$C[\hat{\mu}_Y - t_{0.975:9}SE(\hat{\mu}_Y) \le \mu_Y \le \hat{\mu}_Y + t_{0.975:9}SE(\hat{\mu}_Y)]$$
$$= C[25.302 - (2.262)(0.20215) \le \mu_Y \le 25.302 + (2.262)(0.20215)]$$

$$= C[24.845 \leq \mu_Y \leq 25.759] = 0.95$$

A two-sided 95% confidence statement for σ_Y, using the table-values $\chi^2_{0.025:9} = 2.699$ and $\chi^2_{0.975:9} = 19.031$ (obtained from Table T-3 in the appendix) in the formula given in Table 1.6.2, is

$$C\left[\sqrt{\frac{SSY}{\chi^2_{0.975:9}}} \leq \sigma_Y \leq \sqrt{\frac{SSY}{\chi^2_{0.025:9}}}\right]$$
$$= C\left[\sqrt{\frac{3.678}{19.023}} \leq \sigma_Y \leq \sqrt{\frac{3.678}{2.700}}\right]$$
$$= C[0.4397 \leq \sigma_Y \leq 1.1674] = 0.95$$

Hence the following simultaneous confidence statement is valid.

$$C\left[\begin{array}{c} 24.845 \leq \mu_Y \leq 25.759 \\ \text{and} \\ 0.4397 \leq \sigma_Y \leq 1.1674 \end{array}\right] \geq 0.90 \quad \blacksquare$$

Just as there is a relationship between tests and confidence intervals in the case of single parameters, *there is also a relationship between simultaneous tests and simultaneous confidence intervals when several parameters are of interest.* This relationship is often somewhat complex, and interested readers should consult more advanced books for details [11].

Note From time to time we present conversations between a professional statistician and an investigator who routinely uses statistical methods to interpret results of experiments. This is intended to emphasize and clarify various topics that are discussed and to show how some of the statistical procedures can be applied to practical problems.

Conversation 1.6

I am a professional statistician with a Ph.D. in statistics. One day I received a telephone call from a person, whom I call the investigator, who wanted to discuss some statistical problems. We set up an appointment and the investigator visited me the following day. Following is an excerpt from our conversation.

Investigator: Good morning, I am an investigator, and I'd like to ask you some questions about statistics.

Statistician: Okay, but first give me some information about your background in statistics, and tell me about the company that employs you.

Investigator: The company I work for makes and sells agricultural products, plastics, soap, over-the-counter medications, canned fruits and vegetables, cosmetics, greeting cards, and many other products. We have a large research and development group and I

work in this group. I have a bachelor's degree in computing science, and I have had three courses in statistics—a course in statistical methods, a course in regression, and a course in statistical computer packages.

Statistician: How can I help you?

Investigator: Our agricultural research division has developed a new commercial fertilizer and has conducted a large experiment to determine whether or not the average yield per acre of corn is greater when using the new fertilizer than when using the old fertilizer. The experiment was conducted in five different locations so that we could examine different climatic conditions and soil types. For each location a statistical test of

$$\text{NH}: \theta = 0 \quad \textit{against} \quad \text{AH}: \theta \neq 0$$

was conducted with $\alpha = .05$, where

$$\theta = \mu_{old} - \mu_{new}$$

is the difference between the average yields of corn in bushels per acre when the two fertilizers, old and new, are used. Here are the results in table form. Will you please examine the entries and see if the computations and conclusions for each location are correct?

Location	*P*-value	Decision (Using $\alpha = 0.05$)
1	0.0474	Reject $\theta = 0$
2	0.0066	Reject $\theta = 0$
3	0.0003	Reject $\theta = 0$
4	0.1280	Do not reject $\theta = 0$
5	0.0772	Do not reject $\theta = 0$

Statistician: Your calculations are correct, but your conclusions would be much more informative if you used 95% confidence intervals instead of 5% tests.

Investigator: What do you mean?

Statistician: A statistical test should never (I repeat, never) be used alone when a confidence interval can be computed, because a confidence interval is always more informative than a statistical test. I can illustrate by using your data and computing a 95% confidence interval for θ for each location. I'll organize the results in a table for each location along with your decision based on a statistical test.

Location	Your Decision Based on a Test	95% Confidence Interval
1	Reject $\theta = 0$	$0.06 \leq \theta \leq 16.67$
2	Reject $\theta = 0$	$10.05 \leq \theta \leq 33.19$
3	Reject $\theta = 0$	$0.101 \leq \theta \leq 0.114$
4	Do not reject $\theta = 0$	$-0.034 \leq \theta \leq 0.021$
5	Do not reject $\theta = 0$	$-0.09 \leq \theta \leq 8.56$

Note The data in this conversation are artificial so that we can make our point in a dramatic fashion. However, results such as these are not uncommon in applied problems.

A test tells you only whether or not you should reject the null hypothesis that $\theta = 0$. In other words, a test tells you what the value of θ is not, but a confidence interval tells you what the plausible values of θ are. And if you know, with a specified confidence (95% in this case), what the plausible values for θ are, then you can make practical decisions based on this knowledge.

For example, for locations 1, 2, and 3, the test indicates that you should reject NH : $\theta = 0$ in each case, but the confidence interval yields the following results. (1) For location 1, farmers who must decide whether or not to use the new fertilizer might conclude that the results are not definitive enough to make a decision, because if $\theta = 0.06$ bu/acre, then the farmer would surely conclude that θ is so small as to be considered negligible in a practical sense. However, if $\theta = 16.67$ bu/acre, then the farmer would surely conclude that θ is not zero and the old fertilizer is better. Since we have 95% confidence that θ is somewhere between 0.06 and 16.67, we need more data to make a definite decision. (2) For location 2, you would surely reject NH that $\theta = 0$ and conclude that $\theta \neq 0$ because we have 95% confidence that $10.05 \leq \theta \leq 33.19$ bu/acre. (3) For location 3, you would undoubtedly decide that θ is so small that it can be considered negligible for all practical purposes, and it wouldn't make any difference which fertilizer was used.

On the other hand, for locations 4 and 5, the test indicates that NH shouldn't be rejected. But for location 4, you would surely accept NH that $\theta = 0$ because we have 95% confidence that θ is so small as to be practically no different from zero for this problem. And finally, for location 5, we see that you don't have enough data to make a definite decision. Thus you can readily see that a confidence interval gives information that a test does not. If, in fact, the only result of this investigation you were allowed to see was either the conclusion of the tests of size $\alpha = .05$ or the 95% confidence intervals, which would you prefer?

Investigator: I think I see your point. I would definitely choose to see the confidence intervals.

Statistician: It is always true that if the result of an α level test of NH : $\theta = q$ is "do not reject," this implies that the corresponding $1 - \alpha$ confidence interval for θ will contain the value q specified by NH, which is zero in your problem.

Investigator: I remember studying that. And I remember that when the result of an α level test of NH : $\theta = q$ is to "reject," then the corresponding $1 - \alpha$ confidence interval won't contain the value q specified by NH. Is this correct?

Statistician: Yes it is. So you can easily see that confidence intervals give all the information that tests do plus a great deal more.

Investigator: I also remember my statistics instructor telling us that if the result of a test is to not reject NH, this doesn't mean that one can accept NH.

Statistician: That's true. But as you can see from the calculations for locations 4 and 5, even if the result of a test is "do not reject NH," when you examine the corresponding confidence interval the practical result could be (a) accept NH for all practical purposes or (b) there are not enough data to make a decision. Thus a confidence interval in these cases gives considerably more information about the population parameter than a test does.

Investigator: I understand—the test says "reject NH," yet when the corresponding confidence interval is examined, the practical result could be (a) reject NH, (b) accept NH, or (c) there aren't enough data to make a definite decision. Similarily, when the test says "do not reject NH," the corresponding confidence interval may indicate (a) accept NH or (b) there aren't enough data to make a definite decision. Is that correct?

Statistician: That's correct.

Investigator: I see what you are saying, but scientists are constantly formulating scientific hypotheses and they want to determine whether they are correct or incorrect. How can I tell them not to test hypotheses?

Statistician: I am not telling you that scientists should not formulate and test scientific hypotheses. But what I am telling you is that, whenever possible, they should use confidence intervals rather than statistical tests to evaluate scientific hypotheses, particularly when a hypothesis concerns a well-defined population parameter.

Investigator: I do have a question about your table. What exactly do you mean by the statement "not enough data" in locations 1 and 5?

Statistician: By "not enough data" I mean that the confidence interval is too wide to make a definitive statement about the value of the parameter under study, which in this case is θ. To make the case more dramatic, suppose that a confidence statement for a location is

$$C[-16.24\ bu/acre \leq \mu_{old} - \mu_{new} \leq 19.67\ bu/acre] = 0.95$$

So we have 95% confidence that $\mu_{old} - \mu_{new}$ is some value in this interval. But if $\mu_{old} - \mu_{new} = -16.24$, the lower endpoint of this interval, this means that μ_{new} is 16.24 bu/acre larger than μ_{old}, so the new fertilizer is certainly better. On the other hand, suppose that $\mu_{old} - \mu_{new} = 19.67$, the upper endpoint of this interval.

This means that μ_{old} is 19.67 bu/acre larger than μ_{new}, and so certainly the old fertilizer is better. But since all that the confidence interval tells us is that $\mu_{old} - \mu_{new}$ is somewhere between -16.24 and 19.67 bu/acre, we can arrive at two different conclusions depending on which value in the interval we use. But if the confidence interval were based on enough data, the width of the interval would be small enough so that, with a specified confidence level (say 95%), a single decision would result using any value of $\mu_{old} - \mu_{new}$ in the confidence interval.

It seems to me that in this problem, what your scientists want to know is how much larger (or smaller) μ_{new} is than μ_{old}, and a two-sided confidence interval will give you this information.

Investigator: I think I understand the importance of what you're saying. But since the scientists in our company have always used statistical tests, it won't be easy to get them to change.

Statistician: If they insist on a statistical test to evaluate their scientific hypotheses, then give them a confidence interval too. I think they'll see that confidence intervals are much more informative and are really what they want to help them make decisions.

Investigator: I'll try to convince them. But before I leave, will you clarify something for me about prediction intervals? It has to do with the interpretation of a prediction interval.

Statistician: Certainly.

Investigator: Suppose I compute a 90% prediction interval for Y_0 and I get $C[15.3 \leq Y_0 \leq 19.4] = 0.90$. Does this mean that I have 90% confidence that all future observations will be between 15.3 and 19.4?

Statistician: *No!* It means that you have 90% confidence that *a single* future observation you obtain will be between 15.3 and 19.4 as explained in (1.6.11).

Investigator: Does it mean that if I select many future observations, 90% of them will be between 15.3 and 19.4?

Statistician: *No!* As I stated above it means that you have 90% confidence that any *one* future observation you obtain will be between 15.3 and 19.4.

Investigator: But I will obtain m future observations, and I want to compute a lower bound, say L, and an upper bound, say U, such that I can be 90% confident that all m future observations are between L and U. Is it possible to compute these bounds L and U?

Statistician: Yes, it is possible. One way to do this is by using the Bonferroni method.

Investigator: You keep referring to future observations. What exactly do you mean by a future observation?

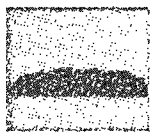
Statistician: I mean any observation that is randomly selected from the population that is being studied and the observation is not known at the time the prediction interval is computed. Generally a future observation is a number to be observed in the future, but it is important to know something about its value before it can be observed for decision-making purposes.

Investigator: One other thing. When you compute a confidence interval for a parameter, say θ, you may *never* know if the interval is correct because θ may never be known. But when you compute a prediction interval for Y_0, a future observation, you can eventually determine whether the interval is correct because, presumably, the future observation will eventually be selected and observed. Is that correct?

Statistician: Yes, you are absolutely right.

Investigator: Thank you. You have certainly helped me understand many fundamental statistical concepts. Can I come to see you again if I have more statistical questions?

Statistician: Certainly. You're welcome anytime.



Problems 1.6

1.6.1 A random sample of size $n = 30$ is obtained from a Gaussian population with unknown mean μ_Y and unknown standard deviation σ_Y. The data are given in Table 1.6.4 and also in the file **table164.dat** on the data disk.

- **a** Compute $\hat{\mu}_Y$ and a two-sided 80% confidence interval for μ_Y.
- **b** Write the confidence statement for the confidence interval in (a).
- **c** Compute a one-sided 95% lower confidence bound for μ_Y.
- **d** Write the confidence statement corresponding to the confidence bound in (c).
- **e** State in words the meaning of the confidence statement in (d).
- **f** Write a short paragraph explaining to an investigator why confidence intervals are more informative than tests.
- **g** Suppose an investigator wants to test NH: $\sigma_Y \leq 5.0$ against AH: $\sigma_Y > 5.0$ with $\alpha = .05$. Perform this test. What is your conclusion?

TABLE 1.6.4

3.48	7.68	12.96	0.65	12.44	0.16	4.34	3.71	4.67	3.47
5.91	6.73	3.64	10.90	6.37	4.62	9.18	11.67	8.29	5.19
4.30	11.67	8.63	7.84	11.92	11.16	6.13	10.21	8.16	3.59

h Compute a 90% two-sided confidence interval for σ_Y. What is your conclusion about the test in (g) using the confidence interval?

i Compute a 99% two-sided confidence interval for μ_Y. Compare the width of this interval with the width of the 80% confidence interval in (a).

1.7 Functional Notation

The concept and notation of functions are used throughout the book. Here we present a short discussion of this topic.

Functions

Let D be a set of numbers. A function $f(\cdot)$ on the set D is a *rule* that describes how numbers in the set D are *changed*, *transformed*, or *mapped* to produce other numbers. If x represents a number in D, then the result of applying the rule, i.e., the function $f(\cdot)$, to x results in a number, say z, and this is symbolically denoted by writing $f(x) = z$ or $z = f(x)$. The set D is called the *domain* of the function $f(\cdot)$.

As an example, let D be the set of positive real numbers and $f(\cdot)$ be a function that is defined by the rule "square each number in the set D." Then the result of applying the function $f(\cdot)$ to the number 2 is 4, and this is denoted by writing $f(2) = 4$. Likewise, $f(0.5) = 0.25$, $f(12) = 144$, etc. In general, the result of applying the function $f(\cdot)$ to a number x in D produces the result x^2. This is described by writing $f(x) = x^2$. If the symbol z is used to denote the result of applying the function $f(\cdot)$ to the number x, then we can also write $z = f(x)$ or $z = x^2$. For another illustration, let $f(\cdot)$ be defined by the equation $f(x) = 6x^2 + x - 5$ for any real number x. Then $f(3) = 6(3)^2 + 3 - 5 = 52$, $f(8) = 387$, etc.

When a function $f(\cdot)$ is specified, its domain D must be specified also. Any letter can be used to represent a function; some examples are $f(\cdot)$, $Y(\cdot)$, and $\mu(\cdot)$. Sometimes a letter with a subscript is used to represent functions; some examples are $f_1(\cdot)$, $f_2(\cdot)$, $\mu_Y(\cdot)$, and $g_t(\cdot)$.

Although, strictly speaking, the symbol $f(\cdot)$ stands for a function and the symbol $f(x)$ stands for the value of the function when applied to the number x in the set D, sometimes people use phrases such as "$f(x)$ is a function" or "let $f(x)$ be a function of x," etc. These phrases are to be interpreted to mean that "$f(\cdot)$ is a function, x is a typical member of the domain D of $f(\cdot)$, and $f(x)$ is the value of the function when applied to the number x." This rarely leads to any confusion because the meaning of the symbol $f(x)$ is usually quite clear from the context.

Independent and Dependent Variables

In the equation $z = f(x)$, x is called the independent variable and z is called the dependent variable. Other letters can be used for dependent and independent variables; for example, we could write the following: $z = f(t), s = v(u), r = g_1(t)$. Some examples of functions appear in (1.7.1), (1.7.2), and (1.7.3).

$$s = f_1(t), \quad 0 \le t \le 3 \quad \text{where} \quad f_1(t) = \frac{2t^2 + 1}{t + 2} \tag{1.7.1}$$

$$z = \mu_Y(x), \quad -5 \le x \le 8 \quad \text{where} \quad \mu_Y(x) = 6x + 4 \tag{1.7.2}$$

$$v = r(u), \quad 0 < u < \infty \quad \text{where} \quad r(u) = \log_e u - 6u + u^2 \tag{1.7.3}$$

The value of s for $t = 3$ in (1.7.1) is $s = f_1(3) = [2(3)^2 + 1]/(3 + 2) = 19/5$. The value of z for $x = 0$ in (1.7.2) is $z = \mu_Y(0) = (6)(0) + 4 = 4$, and for $x = 7$ we get $z = \mu_Y(7) = (6)(7) + 4 = 46$. The value of z in (1.7.2) for $x = 10$ is not defined because $z = \mu_Y(x)$ is defined only for x in the interval $-5 \le x \le 8$.

Functions of Many Variables

The functions just discussed are functions of one (independent) variable, say x. In this book we also need functions of more than one independent variable. A function of three independent variables, say x_1, x_2, x_3, could be denoted by $Y = Y(x_1, x_2, x_3)$, $z = \mu_Y(x_1, x_2, x_3)$, or $z = q(x_1, x_2, x_3)$, etc. Examples are given in (1.7.4) and (1.7.5).

$$z = f(x_1, x_2, x_3), \quad 0 \le x_1 \le 1, \quad 3 \le x_2 \le 8, \quad -4 \le x_3 \le 15 \text{ where} \quad f(x_1, x_2, x_3) = 4x_1^2 + 3x_1x_2 - 6x_3^2 \tag{1.7.4}$$

$$z = \mu_Y(x_1, x_2, x_3), \quad 0 \le x_1 < \infty, \quad 0 \le x_2 < \infty, \quad -\infty < x_3 < \infty \tag{1.7.5}$$
$$\text{where } \mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \text{ and}$$
$$\beta_0, \beta_1, \beta_2, \beta_3 \text{ represent constants}$$

The value of $\mu_Y(x_1, x_2, x_3)$ in (1.7.5) for $x_1 = 1$, $x_2 = 6$, and $x_3 = -4$ is $\mu_Y(1, 6, -4) = \beta_0 + \beta_1(1) + \beta_2(6) + \beta_3(-4) = \beta_0 + \beta_1 + 6\beta_2 - 4\beta_3$, and this is an unknown number unless values are given for $\beta_0, \beta_1, \beta_2$, and β_3. Nevertheless, the symbolic representations $\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and $\mu_Y(1, 6, -4) = \beta_0 + \beta_1 + 6\beta_2 - 4\beta_3$ are useful.

Linear Functions

A function $f(x)$ of a single variable x is said to be **linear in** x if it can be written as

$$f(x) = ax + b \tag{1.7.6}$$

where the values of a and b do not depend on the value of x.

A function $f(x_1, \ldots, x_k)$ of k variables $x_1, \ldots, x_k$ is said to be **simultaneously linear in** $x_1, \ldots, x_k$ if it can be written as

$$f(x_1, \ldots, x_k) = a_0 + a_1 x_1 + \cdots + a_k x_k \tag{1.7.7}$$

where the values of $a_0, a_1, \ldots, a_k$ do not depend on the values of $x_1, \ldots, x_k$.

Consider the function $\mu_Y(x) = 6x + 4$ in (1.7.2). It is of the form $ax + b$ with $a = 6$ and $b = 4$, both of which are free of x. So $\mu_Y(x)$ in (1.7.2) is linear in x. The function $f(x_1, x_2, x_3) = 4x_1^2 + 3x_1x_2 - 6x_3^2$ is not a linear function of the variables x_1, x_2, and x_3 because it cannot be written in the form given in (1.7.7). You should verify that the functions in (1.7.1) and (1.7.3) are not linear functions in the corresponding independent variables.

For further illustration, consider the functions

$$\mu_Y(x_1, x_2, \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 x_1 + e^{\beta_2 x_2} \tag{1.7.8}$$

and

$$\mu_Y(x_1, x_2, \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{1.7.9}$$

It can be verified that the function in (1.7.8) is linear in β_0, linear in β_1, and linear in x_1, but it is not linear in β_2, not linear in x_2, and not simultaneously linear in β_1 and x_1. The function in (1.7.9) is linear in each of the variables $x_1, x_2, \beta_0, \beta_1, \beta_2$; is simultaneously linear in x_1, x_2; is simultaneously linear in $\beta_0, \beta_1, \beta_2$; but is not simultaneously linear in $x_1, x_2, \beta_0, \beta_1, \beta_2$.

Problems 1.7

1.7.1 The function $f(x)$ is defined by $f(x) = 6x^2 + 3x^{1/2} - 9x + 4$ for $4 < x < 49$. Find the following.

- **a** $f(4)$
- **b** $f(16)/f(36)$
- **c** $f(3) + 16$
- **d** $f(34) + f(13)$
- **e** $f(64)$

1.7.2 Problems (a)–(c) refer to the function defined by $\mu_Y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2^3$ for $-\infty < x_1 < +\infty$, $-\infty < x_2 < +\infty$.

- **a** Compute $\mu_Y(6, 1)$.
- **b** Compute $\mu_Y(15, -4)$.
- **c** Is the function $\mu_Y(x_1, x_2)$ linear in x_1? Is it linear in x_2? Is it simultaneously linear in x_1 and x_2?

1.8 Matrices and Vectors

In this section we introduce some of the basic operations of matrix algebra because numerical computations arising in regression analysis can be presented effectively using the language of matrices. You may want to omit this section initially and read appropriate portions of it when matrices are discussed in later chapters.

What Is a Matrix?

A **matrix** is defined as a rectangular array of elements. The elements of a matrix are called *scalars*, and they are either numbers (such as 1.3, −6.0, 0, etc.) or symbols (such as x, $log(x_5)$, xy, etc.) that represent numbers. This array is enclosed in large parentheses or brackets. (We use brackets.) The quantities in (1.8.1)–(1.8.3) are matrices, but (1.8.4) is not because it is not a rectangular array.

$$\begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} \tag{1.8.1}$$

$$\begin{bmatrix} 6 & 2 & 3 \\ 4 & x & y \end{bmatrix} \tag{1.8.2}$$

$$\begin{bmatrix} 5 & x_1 & 4 \\ y_2 & log(x) & 6 \\ 5 & 9 & 6 \end{bmatrix} \tag{1.8.3}$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & \\ 7 & 8 & 9 \end{bmatrix} \tag{1.8.4}$$

The matrix in (1.8.2) has two rows and three columns and is called a "2 by 3 matrix." If a matrix contains r rows and c columns, it is said to be an "r by c matrix" (or a matrix of size "r by c"). The number of rows is always given first when stating the size. The matrix in (1.8.1) is 2 by 2 matrix, and the one in (1.8.3) is a 3 by 3 matrix.

Row Vectors and Column Vectors

If a matrix has only one row, it is usually called a *row vector*. For example, the following 1 by 6 matrix is a row vector.

$$[3 \quad 2 \quad 1 \quad 6 \quad 4 \quad 3] \tag{1.8.5}$$

If a matrix has only one column, it is called a *column vector*. The following 7 by 1 matrix is a column vector.

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{bmatrix} \tag{1.8.6}$$

It is generally clear from the context whether a vector is a row (or a column) vector, and so the word *row* (or *column*) is usually omitted. If a matrix contains only one row and one column, it is a 1 by 1 matrix, say $[x]$. In this case, the brackets are omitted and the matrix $[x]$ is replaced by the number x. For example, [16] is a 1 by 1 matrix, so we write it simply as 16.

We use boldface italic capital letters to represent matrices and boldface italic lowercase letters to represent column vectors. Scalars are not boldface. For example, $\boldsymbol{A}$ is a matrix, and it could denote

$$\boldsymbol{A} = \begin{bmatrix} 6 & 3 & 4 \\ 1 & 9 & 3 \\ 2 & 1 & 7 \\ 4 & 2 & 9 \end{bmatrix}$$

In this case, $\boldsymbol{A}$ is a 4 by 3 matrix. If

$$\boldsymbol{a} = \begin{bmatrix} 6 \\ 9 \\ -1 \\ 0 \\ 2 \end{bmatrix}$$

then $\boldsymbol{a}$ is a column vector of length 5 (i.e., a 5 by 1 matrix).

Matrix Elements

It may be desirable to identify the individual elements of matrices or vectors in some systematic manner, so we sometimes write a matrix $\boldsymbol{G}$ as $[g_{ij}]$ or the matrix $\boldsymbol{K}$ as $[k_{ij}]$, etc., where the quantity in brackets denotes the element in the ith row and jth column of the matrix. For a vector, say $\boldsymbol{a}$, we sometimes write $[a_i]$. A single subscript is used because there is only one row or one column in a vector. Therefore a_i denotes the ith element of a vector $\boldsymbol{a}$. To illustrate, if we let $\boldsymbol{A}$ denote the following 3 by 2 matrix

$$\begin{bmatrix} 3 & 1 \\ 4 & -4 \\ 6 & 0 \end{bmatrix}$$

we could write

$$A = \begin{bmatrix} 3 & 1 \\ 4 & -4 \\ 6 & 0 \end{bmatrix}$$

or we could write

$$[a_{ij}] = \begin{bmatrix} 3 & 1 \\ 4 & -4 \\ 6 & 0 \end{bmatrix}$$

Notice that a_{12} denotes the element in the first row and the second column, and so in this case $a_{12} = 1$. Verify that $a_{31} = 6$.

As a further illustration, we let $\boldsymbol{a}$ or $[a_i]$ denote the vector

$$\boldsymbol{a} = [a_i] = \begin{bmatrix} 6 \\ 4 \\ -2 \\ 0 \end{bmatrix}$$

Here a_3 is -2, $a_1 = 6$, etc.

Equality of Matrices

Two matrices are said to be equal if and only if the following are true:

1 They are the same size.

2 All corresponding elements are equal.

For example, consider the matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ where

$$A = \begin{bmatrix} 6 & 9 \\ 15 & 3 \end{bmatrix} \quad B = \begin{bmatrix} 6 & 9 \\ 18 & 3 \end{bmatrix} \quad C = \begin{bmatrix} 6 & 9 & 0 \\ 15 & 3 & 0 \end{bmatrix} \quad D = \begin{bmatrix} 6 & 9 \\ 18 & 3 \end{bmatrix}$$

Matrix $\boldsymbol{A}$ is not equal to matrix $\boldsymbol{B}$ because element a_{21} is not equal to element b_{21}. Matrix $\boldsymbol{C}$ is not equal to matrix $\boldsymbol{A}$, $\boldsymbol{B}$, or $\boldsymbol{D}$ because it is not the same size. Matrix $\boldsymbol{D}$ is equal to $\boldsymbol{B}$, and we write $\boldsymbol{D} = \boldsymbol{B}$ because all corresponding elements are equal.

In order to effectively use matrices, vectors, and scalars in our discussion, we need to define various operations such as addition, subtraction, multiplication, transposition, etc., for matrices, vectors, and scalars. Because scalars are numbers, the usual rules of arithmetic apply, but matrices and vectors have different definitions.

Transposition of Matrices

For every matrix, say $\boldsymbol{A}$, there is another matrix derived from it called the *transpose* of $\boldsymbol{A}$. The transpose of a matrix $\boldsymbol{A}$ is obtained by replacing each row of $\boldsymbol{A}$ by its

corresponding column. Thus, if A is given by

$$\begin{bmatrix} 3 & -1 \\ 0 & 4 \\ 2 & 6 \end{bmatrix}$$

the transpose of A is

$$\begin{bmatrix} 3 & 0 & 2 \\ -1 & 4 & 6 \end{bmatrix}$$

It is easy to see that if A is a matrix of size r by c, then the transpose of A is a matrix of size c by r. The column vector

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix}$$

has as its transpose the row vector

$$[a_1, a_2, a_3, a_4, a_5]$$

which is obtained by arranging the elements a_i in their original order in a row. Thus the transpose of a column vector is a row vector. If the original vector is of size 5 by 1, its transpose is of size 1 by 5. Verify that the transpose of a row vector is a column vector. The transpose of a vector a or a matrix A is indicated by the notation a^T and A^T (some authors write a' and A'). Notice also that $(A^T)^T = A$; that is, the transpose of the transpose of a matrix A is equal to the matrix A.

An 11 by 2 matrix D and its transpose D^T follow:

$$D = \begin{bmatrix} 52 & 0.62 \\ 43 & 0.74 \\ 36 & 0.65 \\ 32 & 0.71 \\ 27 & 0.68 \\ 26 & 0.59 \\ 22 & 0.49 \\ 37 & 0.67 \\ 24 & 0.64 \\ 19 & 0.56 \\ 13 & 0.51 \end{bmatrix}$$

$$D^T = \begin{bmatrix} 52 & 43 & 36 & 32 & 27 & 26 & 22 & 37 & 24 & 19 & 13 \\ 0.62 & 0.74 & 0.65 & 0.71 & 0.68 & 0.59 & 0.49 & 0.67 & 0.64 & 0.56 & 0.51 \end{bmatrix}$$

Addition and Subtraction of Matrices

Addition is defined for two matrices if and only if they are of the same size. For example, consider the matrices A, B, C, and D where

$$A = \begin{bmatrix} 3 & -1 \\ 0 & 4 \\ 2 & 6 \end{bmatrix} \quad B = \begin{bmatrix} 4 & 0 & 2 & 3 \\ 0 & 2 & 6 & 10 \end{bmatrix} \quad C = \begin{bmatrix} 5 & 9 \\ -16 & 4 \\ 0 & 6 \end{bmatrix} \quad D = \begin{bmatrix} 3 & 10 & 4 \\ -1 & 0 & 6 \end{bmatrix} \tag{1.8.7}$$

Matrices A and C are both 3 by 2 so we can add A and C to form a new matrix, say G. Thus, we write $A + C = G$ or

$$[a_{ij}] + [c_{ij}] = [g_{ij}]$$

and the resultant matrix G is the same size as A (and, of course, the same size as C). We cannot add A and B because they are not the same size, and we cannot add B and C for the same reason. The matrix G that results from adding two matrices A and C is obtained by adding the corresponding elements of A and C; that is, $g_{ij} = a_{ij} + c_{ij}$ for all i and j. We obtain

$$G = A + C = \begin{bmatrix} 3 & -1 \\ 0 & 4 \\ 2 & 6 \end{bmatrix} + \begin{bmatrix} 5 & 9 \\ -16 & 4 \\ 0 & 6 \end{bmatrix} = \begin{bmatrix} 3+5 & -1+9 \\ 0-16 & 4+4 \\ 2+0 & 6+6 \end{bmatrix} = \begin{bmatrix} 8 & 8 \\ -16 & 8 \\ 2 & 12 \end{bmatrix}$$

Notice that $A + C = C + A$.

The matrix S that is the result of subtracting a matrix C from a matrix A, which is written as $S = A - C$, or as $[s_{ij}] = [a_{ij}] - [c_{ij}]$, is obtained by subtracting elements of C from the corresponding elements of A. Hence, subtraction is defined if and only if the two matrices A and C that are involved are the same size; the resulting matrix S is also the same size as A and C. For an illustration, consider the matrices defined in (1.8.7):

$$S = A - C = \begin{bmatrix} 3 & -1 \\ 0 & 4 \\ 2 & 6 \end{bmatrix} - \begin{bmatrix} 5 & 9 \\ -16 & 4 \\ 0 & 6 \end{bmatrix} = \begin{bmatrix} 3-5 & -1-9 \\ 0-(-16) & 4-4 \\ 2-0 & 6-6 \end{bmatrix} = \begin{bmatrix} -2 & -10 \\ 16 & 0 \\ 2 & 0 \end{bmatrix}$$

The transpose of the sum (or difference) of any two matrices is equal to the sum (or difference) of the transpose of each matrix; that is,

$$(A + C)^T = A^T + C^T$$

and

$$(A - C)^T = A^T - C^T$$

You should verify this for the matrices in (1.8.7).

Multiplication of Matrices

The process of adding and subtracting matrices is very similar to that of adding and subtracting numbers. This is not the case for multiplication. For example, with numbers, 6 times 3 and 3 times 6 are both equal to 18, but in multiplying two matrices

it is generally not true that A times B and B times A are equal. The quantity A times B, usually written as AB, is defined as multiplying B on the left by A (or A on the right by B), and BA is defined as multiplying A on the left by B (or B on the right by A). The product AB is defined if and only if the number of columns of A (the matrix on the left) is equal to the number of rows of B (the matrix on the right). The matrix, say C, that results from this multiplication has size r by c, where r is the number of rows of A and c is the number of columns of B.

As an illustration, consider the matrices in (1.8.7). Suppose we want to see which of the multiplications AB, AC, and DC are defined. A convenient procedure is to write the size of the two matrices in their corresponding places; that is, we write

$$\begin{matrix} A & B \\ 3 \text{ by } 2 & 2 \text{ by } 4 \end{matrix} \tag{1.8.8}$$

where the quantity "3 by 2" is the size of A and is placed to the left of the quantity "2 by 4," which is the size of B, because A is to the left of B. Now if the two inner numbers (in this case 2 and 2) are equal, the multiplication is defined, and the matrix that results from this multiplication has size 3 by 4, which is obtained from the two outside numbers in (1.8.8). In general, if the matrix A has size r by s (r rows and s columns) and if B has size t by c (t rows and c columns), then to test whether the multiplication AB is defined, the sizes of A and B are written in the order in which they are to be multiplied; that is, (r by s)(t by c). The multiplication AB is defined if and only if the two inside numbers, s and t, are equal. If AB is defined, the resulting matrix has size r by c, the two outside numbers. Is the multiplication AC in (1.8.7) defined? We get

$$\begin{matrix} A & C \\ 3 \text{ by } 2 & 3 \text{ by } 2 \end{matrix}$$

The two inside numbers 2 and 3 are not equal, and so this multiplication is not defined. The multiplication DC is defined, and the resulting matrix is of size 2 by 2. The same is true for DA. Notice that BA, DB, and BC are not defined; neither are BD and CA.

Before we give the rule for multiplying two matrices, we give the rule for multiplying a row vector a^T on the right by the column vector b, where a and b are both column vectors of length k (so a^T is a row vector of length k). Of course a and b are matrices, because a vector is a special kind of matrix. First we check to see if the multiplication $a^T b$ is defined according to the preceding discussion. We write

$$\begin{matrix} a^T & b \\ 1 \text{ by } k & k \text{ by } 1 \end{matrix}$$

and observe that the middle two numbers are both k so the multiplication is defined; the result is a 1 by 1 matrix that is a scalar; that is, a number. We define $a^T b$ as follows:

$$a^T b = \sum_{i=1}^{k} a_i b_i \tag{1.8.9}$$

For example, if

$$a^T = [6 \quad -2 \quad 0 \quad 3] \quad \text{and} \quad b = \begin{bmatrix} 4 \\ 9 \\ -8 \\ 7 \end{bmatrix}$$

then $a^T b$ is defined and is equal to

$$a^T b = [6 \quad -2 \quad 0 \quad 3] \begin{bmatrix} 4 \\ 9 \\ -8 \\ 7 \end{bmatrix} = (6)(4) + (-2)(9) + (0)(-8) + (3)(7) = 27$$

This method of multiplying a row vector on the left by a column vector on the right is very important because this is the basic calculation needed for multiplying matrices.

If A is a matrix of size r by t and B is a matrix of size t by c, then AB is defined. Let the resulting matrix of size r by c be denoted by P. We write

$$P = AB$$

The (i, j) element of P is obtained by multiplying the ith row of A (the matrix on the left in the multiplication AB) by the jth column of B (the matrix on the right in the multiplication AB) by the rule in (1.8.9). If this is done for every row in A with every column in B, all the elements in the matrix P are obtained. Another way to state this is

$$p_{ij} = \sum_{m=1}^{t} a_{im} b_{mj}$$

where $[p_{ij}] = P$; $[a_{im}] = A$; and $[b_{mj}] = B$.

As an illustration, we compute P, where

$$P = AB$$

and A and B are as in (1.8.7). We obtain

$$P = AB = \begin{bmatrix} 3 & -1 \\ 0 & 4 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} 4 & 0 & 2 & 3 \\ 0 & 2 & 6 & 10 \end{bmatrix} = \begin{bmatrix} 12 & -2 & 0 & -1 \\ 0 & 8 & 24 & 40 \\ 8 & 12 & 40 & 66 \end{bmatrix}$$

For example, the element in the second row and third column of P—that is p_{23}, which is 24—is obtained by multiplying the second row of A by the third column of B according to the rule in (1.8.9). We obtain

$$p_{23} = [0 \quad 4] \begin{bmatrix} 2 \\ 6 \end{bmatrix} = (0)(2) + (4)(6) = 24$$

Similarly, p_{34}, which is equal to 66, is the product of the third row of A and the fourth column of B. We get

$$p_{34} = [2 \quad 6] \begin{bmatrix} 3 \\ 10 \end{bmatrix} = (2)(3) + (6)(10) = 66$$

You should verify the remaining elements in P.

For another illustration, let us evaluate S, where $S = RQ$ and R and Q are defined as follows:

$$R = \begin{bmatrix} 3 & 5 & -1 \\ 2 & 6 & 0 \end{bmatrix} \qquad Q = \begin{bmatrix} 5 & 4 & 9 & 1 \\ -3 & 0 & 5 & 4 \\ 6 & 1 & 2 & 6 \end{bmatrix}$$

We get

$$S = RQ = \begin{bmatrix} -6 & 11 & 50 & 17 \\ -8 & 8 & 48 & 26 \end{bmatrix}$$

As a final example, let us evaluate Ax, where A is a 2 by 2 matrix (that is, a matrix of size 2 by 2) and x is a 2 by 1 matrix (x is therefore a column vector) defined as follows:

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

We get

$$Ax = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 + x_2 \\ x_1 + 3x_2 \end{bmatrix} \tag{1.8.10}$$

Clearly, the result of Ax in (1.8.10) is a 2 by 1 matrix, which, of course, is a column vector with two elements; the first element is $2x_1 + x_2$, and the second element is $x_1 + 3x_2$. If we set the product Ax equal to a 2 by 1 vector b where

$$b = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

we get

$$Ax = b$$

or

$$\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

This becomes (by performing the multiplication Ax)

$$\begin{bmatrix} 2x_1 + x_2 \\ x_1 + 3x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

Because two matrices are equal if and only if the corresponding elements are equal, we get the two equations

$$\begin{aligned} 2x_1 + x_2 &= 4 \\ x_1 + 3x_2 &= 7 \end{aligned}$$

as a result of $Ax = b$. In other words, matrix equations can be used to represent systems of linear equations.

To multiply more than two matrices, we simply multiply any adjoining pair, then multiply this result by any adjoining matrix, and continue until all multiplications are completed. For example, suppose we want to evaluate P, where $P = ABCD$ and

where we assume that the sizes are such that all multiplications are defined. One way of calculating $\boldsymbol{P}$ is to first compute $\boldsymbol{AB}$, then multiply this result on the right by $\boldsymbol{C}$, and then multiply this result on the right by $\boldsymbol{D}$.

The Transpose of Products of Matrices

The transpose of the product of two or more matrices is equal to the product in *reverse order* of the transposes; that is, $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$, $(\boldsymbol{CDE})^T = \boldsymbol{E}^T\boldsymbol{D}^T\boldsymbol{C}^T$, etc.

The Multiplication of a Matrix and a Scalar

For further development of the arithmetic of matrices, we define the multiplication of a scalar and a matrix. This multiplication is defined so that any matrix can be multiplied by any scalar. Also, the result is the same whether the scalar is multiplied on the left or the right of the matrix. The product of a scalar and a matrix is obtained by multiplying every element in the matrix by the scalar. For example, let a scalar c and matrix $\boldsymbol{A}$ be defined as follows:

$$c = 20, \quad \boldsymbol{A} = \begin{bmatrix} 4 & -1 & 0 \\ 3 & 2 & 1 \end{bmatrix}$$

Then

$$c\boldsymbol{A} = \boldsymbol{A}c = 20\begin{bmatrix} 4 & -1 & 0 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & -1 & 0 \\ 3 & 2 & 1 \end{bmatrix}20$$

$$= \begin{bmatrix} 20 \times 4 & 20 \times -1 & 20 \times 0 \\ 20 \times 3 & 20 \times 2 & 20 \times 1 \end{bmatrix} = \begin{bmatrix} 80 & -20 & 0 \\ 60 & 40 & 20 \end{bmatrix}$$

Note that the operation of dividing vectors or matrices by nonzero scalars is equivalent to multiplying the vector or matrix by the reciprocal of the divisor. Thus, if division by 20 had been intended in the previous example, the matrix $\boldsymbol{A}$ would be multiplied by 1/20, or 0.05. This would lead to the result

$$\frac{\boldsymbol{A}}{20} = \frac{1}{20}\boldsymbol{A} = \boldsymbol{A}\frac{1}{20} = 0.05\begin{bmatrix} 4 & -1 & 0 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & -1 & 0 \\ 3 & 2 & 1 \end{bmatrix}0.05$$

$$= \begin{bmatrix} 0.05 \times 4 & 0.05 \times -1 & 0.05 \times 0 \\ 0.05 \times 3 & 0.05 \times 2 & 0.05 \times 1 \end{bmatrix} = \begin{bmatrix} 0.20 & -0.05 & 0 \\ 0.15 & 0.10 & 0.05 \end{bmatrix}$$

Special Matrices

Certain matrices arise quite often in applications, and they have been given special names. Some of these special matrices are defined next.

Square Matrix

A matrix having the same number of rows as columns is called a *square matrix*. The matrix

$$A = \begin{bmatrix} 3 & -1 \\ 0 & 4 \\ 2 & 6 \end{bmatrix}$$

in (1.8.7) is not a square matrix but

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

in (1.8.10) is a square matrix. In a square matrix A, the elements a_{ii} are called the *diagonal elements*. In the square matrix A in (1.8.10), the numbers 2 and 3 are the diagonal elements of A. The elements not on the diagonal (nondiagonal elements) are called *off-diagonal elements*.

Identity Matrix

A square matrix whose diagonal elements are each equal to 1 and whose off-diagonal elements are equal to 0 is called an *identity matrix* and is denoted by I. The size of I is usually clear from the context. For example,

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

are both identity matrices. An important feature of an identity matrix is that when it is multiplied by any matrix K for which the multiplication is defined, the result is K. This is written as $KI=IK=K$. Readers can verify this fact for the matrices in (1.8.7) and for identity matrices of appropriate sizes. In many cases, the identity matrix plays a role for matrices similar to that played by the number 1 for numbers. It is also the case that $I^T = I$.

Zero Matrix

A matrix whose elements are all 0 is called a *zero matrix* and is denoted by 0. The size of a zero matrix is usually clear from the context of the discussion. In many cases, the zero matrix plays a role for matrices similar to that played by the number 0 for numbers.

Diagonal Matrix

A matrix is a diagonal matrix if and only if (1) it is a square matrix and (2) all the off-diagonal elements are equal to zero. For example, I is a diagonal matrix, and any square zero matrix is also a diagonal matrix. For another illustration, the following

matrix is also a diagonal matrix.

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Symmetric Matrix

A matrix is called a *symmetric matrix* if it is equal to its transpose. Thus, if $\boldsymbol{A}$ is symmetric, $\boldsymbol{A}^T = \boldsymbol{A}$. This requires a symmetric matrix to be a square matrix. Notice that $\boldsymbol{I}$ is a symmetric matrix, and every diagonal matrix is also a symmetric matrix. The following matrix is symmetric.

$$\begin{bmatrix} 6 & -2 & 3 \\ -2 & 5 & 0 \\ 3 & 0 & 4 \end{bmatrix}$$

Matrix Inversion

Although matrices can be added, subtracted, and multiplied (when their sizes allow these operations), *division of matrices* is not defined. However, an operation that is similar to division, called the *inversion* of a matrix, takes the place of division in certain situations. It is a basic operation in matrix algebra, and it is much used in regression calculations.

Let $\boldsymbol{A}$ denote a square matrix. If there exists another matrix $\boldsymbol{B}$ such that $\boldsymbol{AB} = \boldsymbol{I}$, then $\boldsymbol{B}$ is called the inverse of $\boldsymbol{A}$ and is generally denoted by $\boldsymbol{A}^{-1}$. Thus we have $\boldsymbol{AA}^{-1} = \boldsymbol{I}$. If matrix $\boldsymbol{A}$ has an inverse $\boldsymbol{A}^{-1}$, then $\boldsymbol{AA}^{-1} = \boldsymbol{I}$ and $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$. For example, let $\boldsymbol{A}$ be defined by

$$\boldsymbol{A} = \begin{bmatrix} 7 & 2 \\ 10 & 3 \end{bmatrix}$$

This matrix has an inverse, which is given by

$$\boldsymbol{A}^{-1} = \begin{bmatrix} 3 & -2 \\ -10 & 7 \end{bmatrix}$$

You should verify that for this matrix $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{AA}^{-1} = \boldsymbol{I}$.

Not all matrices have inverses; for example, the zero matrix does not have an inverse. One difficult problem in using matrices is finding the inverse of a matrix. A computer is generally used for finding inverses except when we are dealing with small matrices (2 by 2 or perhaps 3 by 3 matrices).

Matrices are very useful for solving systems of linear equations, a frequent step in regression analysis. To illustrate this let us consider the equations

$$4\hat{\beta}_0 + 8\hat{\beta}_1 = 16 \qquad \textbf{(1.8.11)}$$

$$8\hat{\beta}_0 + 20\hat{\beta}_1 = 36 \qquad \textbf{(1.8.12)}$$

Using elementary algebra, we can see that the solution to the preceding system of two linear equations is given by $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = 1$. If we use matrices and vectors,

we can write these equations as

$$S\hat{\beta} = g \tag{1.8.13}$$

where

$$S = \begin{bmatrix} 4 & 8 \\ 8 & 20 \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}, \quad g = \begin{bmatrix} 16 \\ 36 \end{bmatrix}$$

The objective is to find the numbers $\hat{\beta}_0$ and $\hat{\beta}_1$ (in other words, to find the vector $\hat{\beta}$) that satisfy the equations in (1.8.11) and (1.8.12). To find the solution, we use (1.8.13) and multiply both sides of the equation by S^{-1} if it exists. This gives us

$$S^{-1}S\hat{\beta} = S^{-1}g$$

i.e.,

$$I\hat{\beta} = S^{-1}g$$

i.e.,

$$\hat{\beta} = S^{-1}g$$

Therefore, if we find S^{-1} and multiply it on the right by g, we get the desired result. We have not yet explained how to obtain the inverse of a matrix, but you can verify that the following is the inverse of S by showing that $SS^{-1} = I$.

$$S^{-1} = \begin{bmatrix} 1.25 & -0.50 \\ -0.50 & 0.25 \end{bmatrix}$$

Then

$$\hat{\beta} = S^{-1}g = \begin{bmatrix} 1.25 & -0.50 \\ -0.50 & 0.25 \end{bmatrix} \begin{bmatrix} 16 \\ 36 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

or

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

and so $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = 1$ as before. Notice that in this illustration, S is a symmetric matrix and so is S^{-1}. It is always the case that the inverse of a symmetric matrix is a symmetric matrix (if the inverse exists).

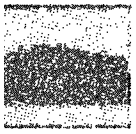
It is easy to write the inverse of a 2 by 2 matrix when it exists.
Let

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

and $d = a_{11}a_{22} - a_{12}a_{21}$. If $d = 0$ then the matrix A does not have an inverse. If $d \neq 0$, then the inverse of A exists and is given by

$$A^{-1} = \begin{bmatrix} a_{22}/d & -a_{12}/d \\ -a_{21}/d & a_{11}/d \end{bmatrix} \tag{1.8.14}$$

By straightforward multiplication we can see that $AA^{-1} = A^{-1}A = I$.



Problems 1.8

1.8.1 Problems (a) and (b) refer to the matrices A, B, and C defined as follows:

$$A = \begin{bmatrix} 9 & 4 & 3 \\ 4 & 16 & 8 \\ 3 & 8 & 12 \end{bmatrix} \qquad B = \begin{bmatrix} 12 & 14 & 3 \\ 4 & 31 & 5 \\ 5 & 13 & 21 \\ 6 & 2 & 31 \end{bmatrix} \qquad C = \begin{bmatrix} 12 & 23 & 17 & 22 \\ 24 & 28 & 19 & 20 \\ 31 & 30 & 41 & 27 \end{bmatrix}$$

a Compute A^T.

b Which of the following operations are defined?

$$C + B^T, \quad B + C, \quad B + C^T, \quad AC, \quad CA, \quad B - C^T$$

c In (b), evaluate the expressions that are defined.

1.8.2 Let

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \qquad y = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$$

Find the 2 by 1 vector $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$ such that $A\hat{\beta} = y$.

1.8.3 Problems (a) through (d) refer to the matrices X and y defined as follows:

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 1 \\ 1 & 4 \end{bmatrix} \qquad y = \begin{bmatrix} 1 \\ 6 \\ 3 \\ 2 \end{bmatrix}$$

a Find $X^T X$.

b Find $X^T y$.

c Find $(X^T X)^{-1}$.

d Find the 2 by 1 vector $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$ such that $X^T X\hat{\beta} = X^T y$.

1.9 Multivariate Gaussian Populations

In Section 1.4 we discussed univariate Gaussian populations. In this section we discuss multivariate Gaussian populations. This section can be studied now or later when the k-variate Gaussian population is discussed in Chapters 3 and 4.

We consider a k-variable population

$$\{(X_{11}, X_{12}, \ldots, X_{1k}), \ldots, (X_{N1}, X_{N2}, \ldots, X_{Nk})\}$$

written $\{(X_1, X_2, \ldots, X_k)\}$ for short.

We say that this k-variable population is a ***k-variate Gaussian population*** *(also called a k-variate normal population) if the collection of numbers*

$$Z_I = a_1X_{I1} + a_2X_{I2} + \cdots + a_kX_{Ik},$$

for $I = 1, 2, \ldots, N$ *is a one-variable Gaussian population for every possible choice of the values of the constants* $a_1, a_2, \ldots, a_k$.

In reality a population can be Gaussian only if N is infinite as discussed in Section 1.4. So we use Gaussian populations only as approximations. It is easier to determine whether a k-variate population is approximately Gaussian when $k = 1$ than it is when $k > 1$. So to determine if a k-variate ($k > 1$) population is approximately Gaussian, we examine $\{Z\}$ to see if it is approximately a univariate Gaussian population for every choice of the constants $a_1, a_2, \ldots, a_k$. If we find that $\{Z\}$ is approximately Gaussian for every choice of the constants a_i, then we can conclude that the k-variate population $\{(X_1, X_2, \ldots, X_k)\}$ is approximately Gaussian.

In particular, this implies that if $\{(X_1, X_2, \ldots, X_k)\}$ is a k-variate Gaussian population, then each of the k univariate populations $\{X_1\}, \{X_2\}, \ldots, \{X_k\}$ must be a Gaussian population. For instance, by choosing $a_j = 0$ for all $j \neq i$ and taking $a_i = 1$, we can conclude that the population $\{X_i\}$ is a Gaussian population. However, even if each of the k univariate populations $\{X_1\}, \{X_2\}, \ldots, \{X_k\}$ is a Gaussian population, we cannot conclude that $\{(X_1, X_2, \ldots, X_k)\}$ is a k-variate Gaussian population unless we also verify that the collection of numbers $a_1X_{I1} + a_2X_{I2} + \cdots + a_kX_{Ik}$ for $I = 1, 2, \ldots, N$ is a one-variable Gaussian population for every possible set of values for the constants $a_1, a_2, \ldots, a_k$. We give two illustrations, one of a bivariate Gaussian population and the other of a bivariate non-Gaussian population.

EXAMPLE 1.9.1

We consider a bivariate population $\{(X_1, X_2)\}$ of size 1,000 stored on the data disk in a file named **bivgauss.dat** (**bivariate Gaussian**). Of course a theoretical two-variable Gaussian population must contain an infinite number of elements, but for practical applications this is a close enough approximation to a theoretical two-variable Gaussian population. A scatter plot of the data is displayed in Figure 1.9.1.

The mean of the population of X_1 values is $\mu_1 = 2.0573$, and the mean of the population of X_2 values is $\mu_2 = 2.9592$. The standard deviations of these two univariate populations are 3.8535 and 9.5729, respectively. Histograms of these two univariate populations are displayed in Figures 1.9.2 and 1.9.3, respectively.

For illustrations, we examine two different linear combinations of X_1 and X_2. First, we calculate the quantity $Z_I = 2X_{I1} + 4X_{I2}$ for each item in the population. The Z_I form a univariate population with mean equal to 15.951 and standard deviation equal to 38.940. A histogram of the population of numbers $\{Z\}$ is shown in Figure 1.9.4.

Likewise, the linear combination $W_I = 3X_{I1} - 2X_{I2}$ is calculated for each item. The mean of the population $\{W\}$ is 0.2535, and its standard deviation is 22.521. A histogram of the population $\{W\}$ is shown in Figure 1.9.5.

In Figures 1.9.2–1.9.5, we have superimposed the theoretical Gaussian distribution curve over the histograms. Note that $\{X_1\}$, $\{X_2\}$, $\{Z\}$, and $\{W\}$ are well

FIGURE 1.9.1

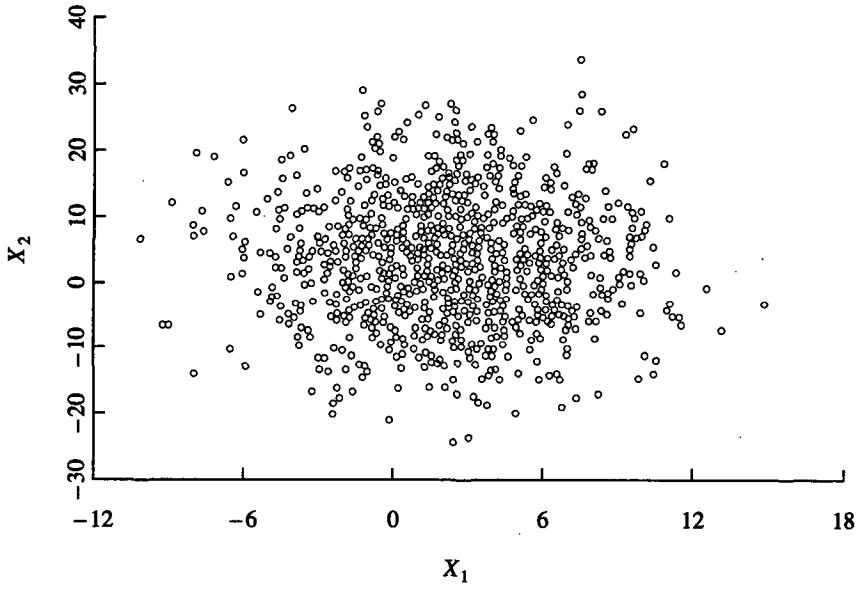


FIGURE 1.9.2

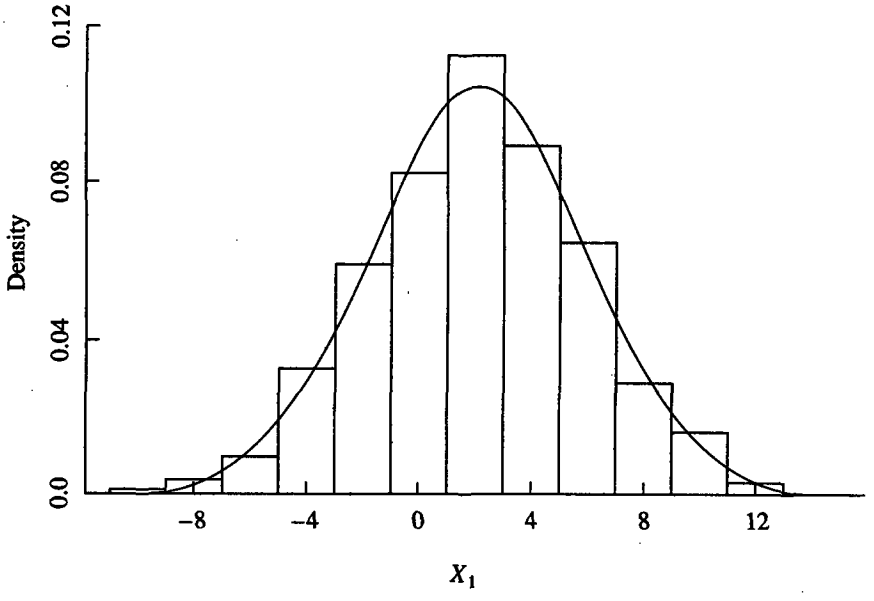


FIGURE 1.9.3

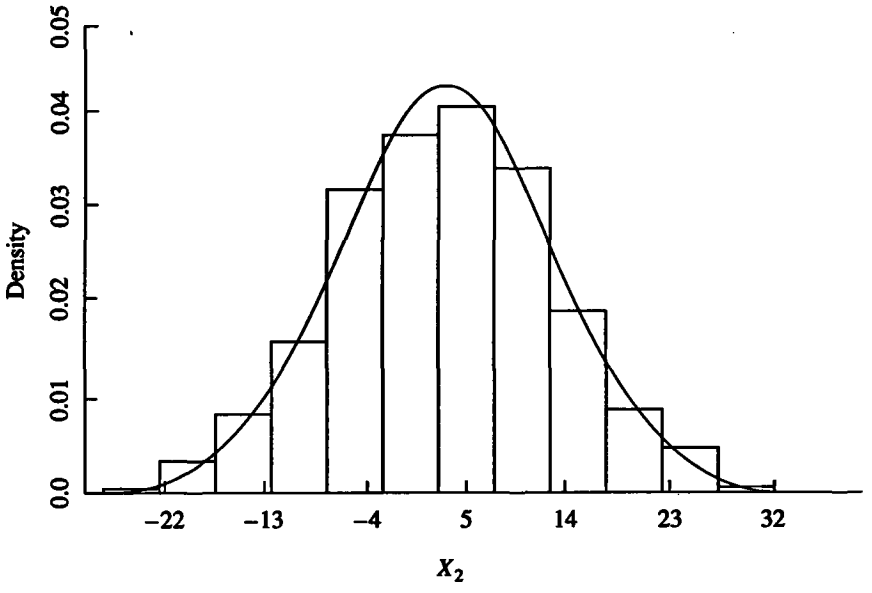


FIGURE 1.9.4

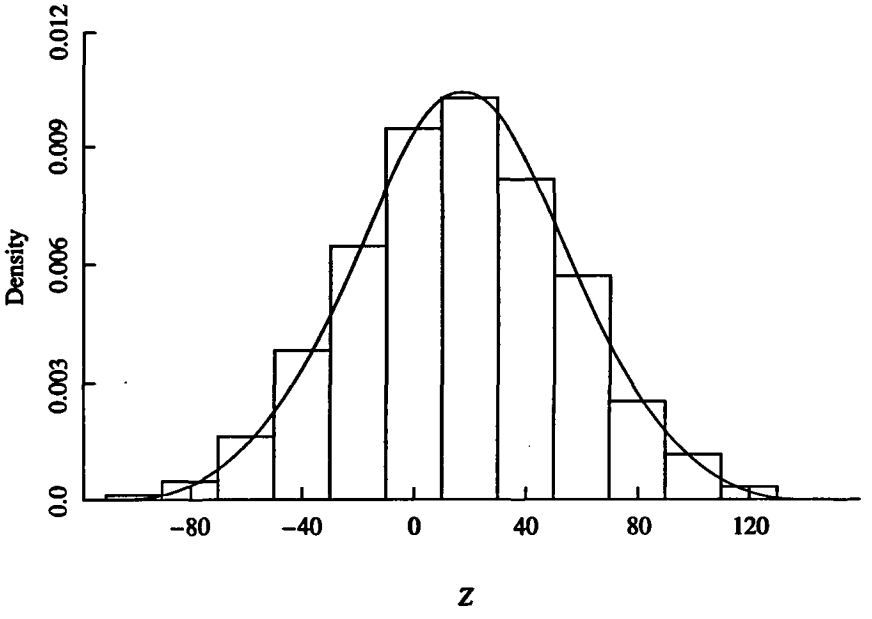
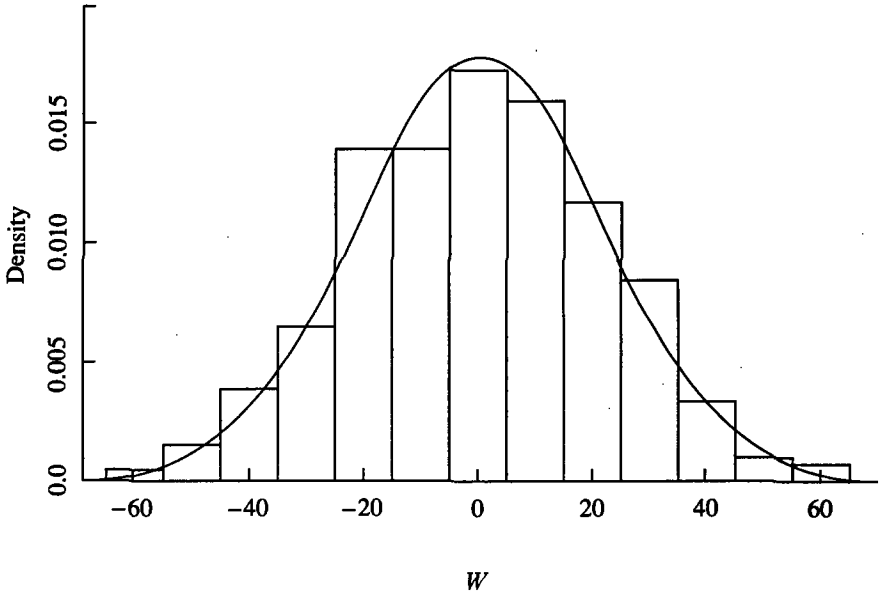


FIGURE 1.9.5



approximated by a Gaussian distribution. It may be verified that any other linear function of X_1 and X_2 is also well approximated by a Gaussian distribution. *Thus the bivariate population* $\{(X_1, X_2)\}$ *is well approximated by a bivariate Gaussian distribution and we say* $\{(X_1, X_2)\}$ *is a bivariate Gaussian population (approximately).* ■

EXAMPLE 1.9.2

Now consider another bivariate population $\{(X_1, X_2)\}$ of size 1,000 stored on the data disk in a file named **bivngaus.dat** (**biv**ariate **non-Gaus**sian). A scatter plot of the data is displayed in Figure 1.9.6.

The mean of the population of X_1 values is $\mu_1 = 2.7958$, and the standard deviation is 3.9842. The mean of the population of X_2 values is $\mu_2 = 1.4921$, and the standard deviation is 11.836. A histogram of the population $\{X_1\}$ is given in Figure 1.9.7 and a histogram of the population $\{X_2\}$ is given in Figure 1.9.8.

Both of these histograms suggest that the two univariate populations $\{X_1\}$ and $\{X_2\}$ are Gaussian populations. But consider the linear function $Z_I = X_{I1} - X_{I2}$. A histogram of the population $\{Z\}$ is given in Figure 1.9.9.

Figure 1.9.10 displays a histogram of the population $\{W\}$, where $W_I = X_{I1} + X_{I2}$, and Figure 1.9.11 shows a histogram of the population $\{V\}$, where $V_I = 8X_{I1} - 3X_{I2}$.

In Figures 1.9.9–1.9.11 we have also displayed the corresponding theoretical Gaussian curve to help us assess the adequacy of the Gaussian distribution as an

FIGURE 1.9.6

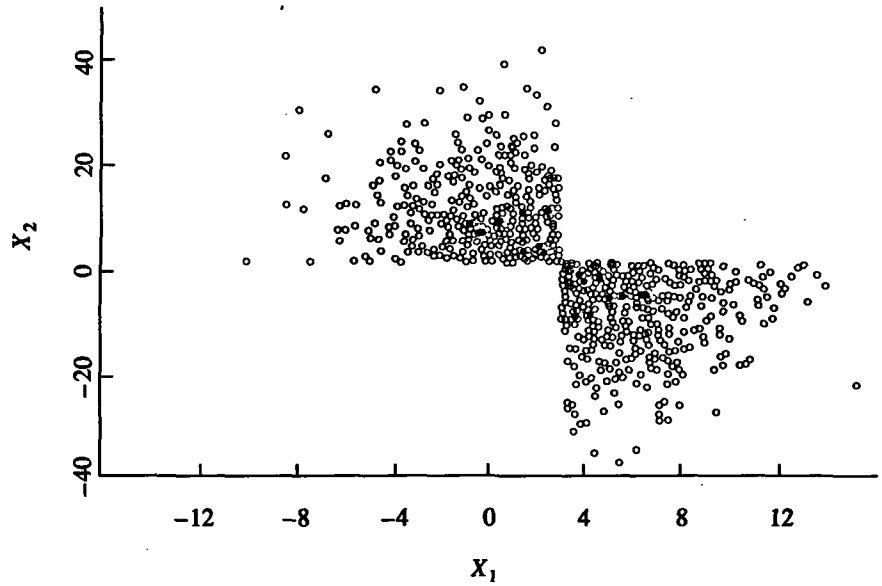


FIGURE 1.9.7

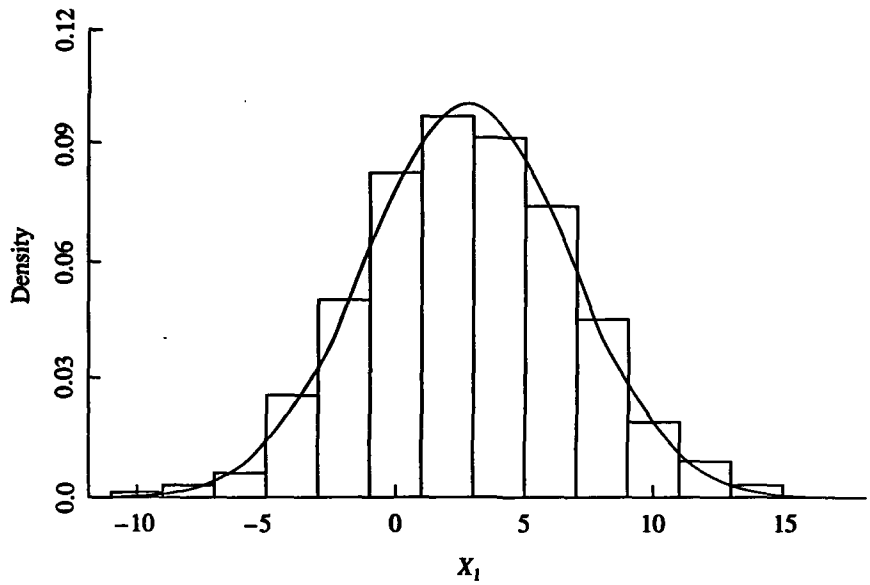


FIGURE 1.9.8

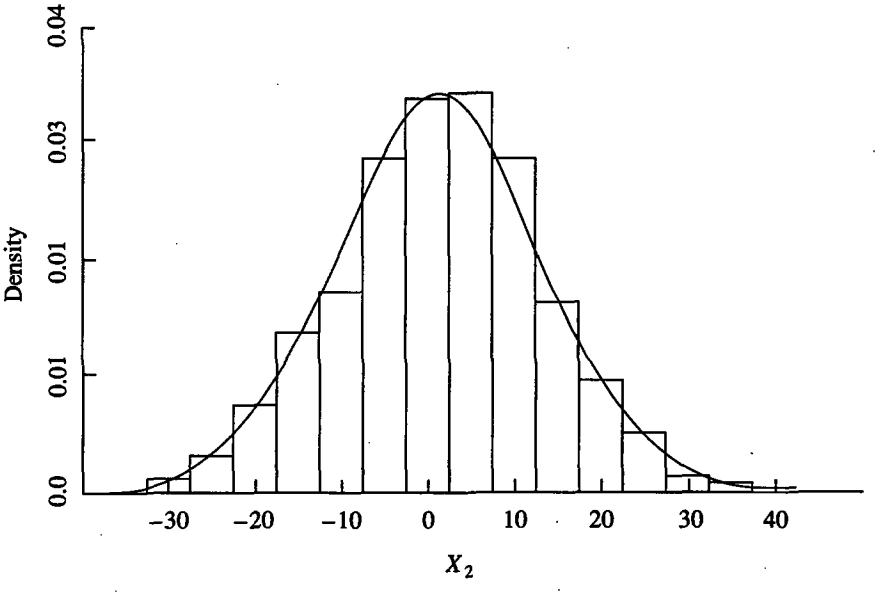


FIGURE 1.9.9

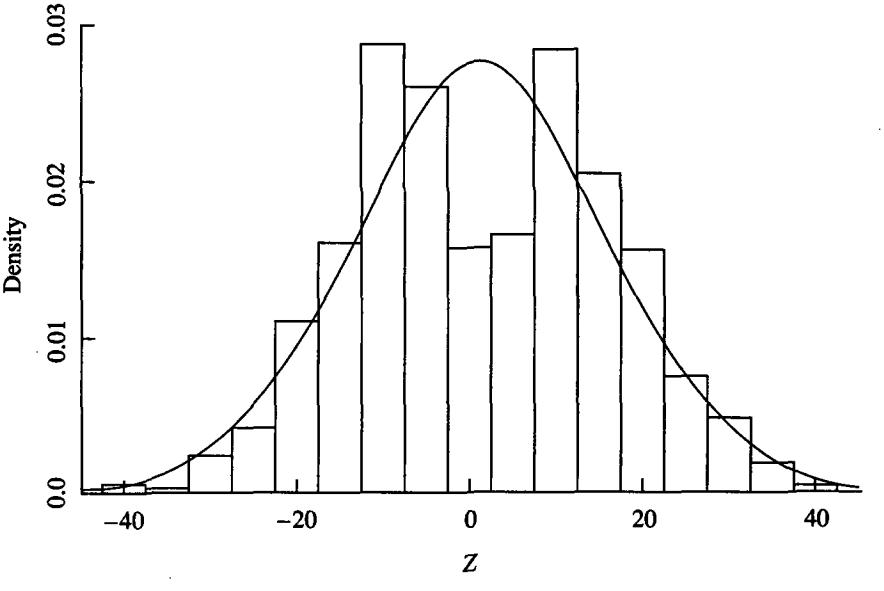


FIGURE 1.9.10

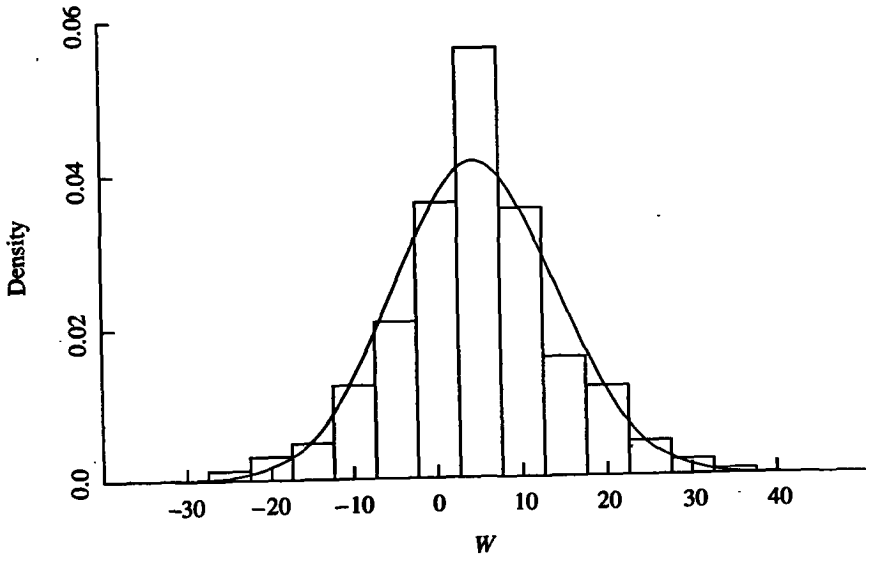
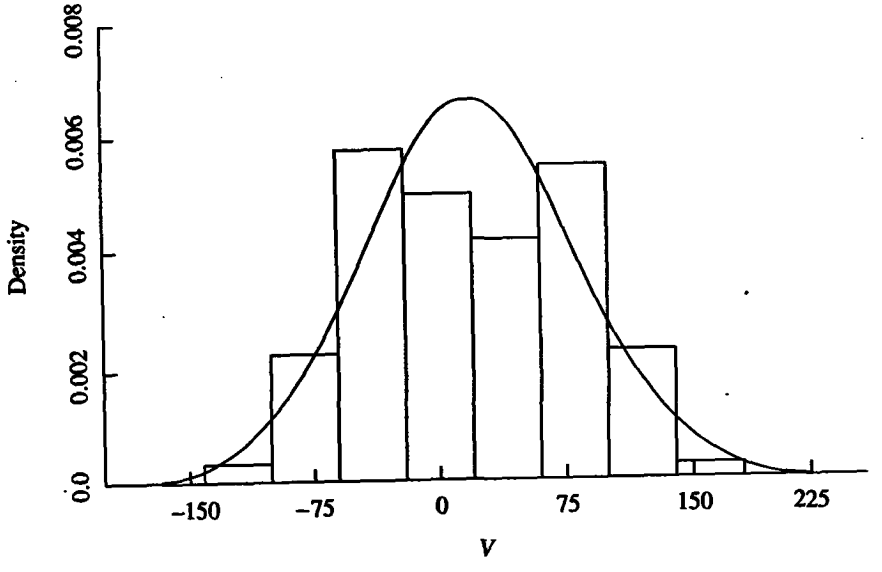


FIGURE 1.9.11



approximation to the populations in question. These figures indicate that none of the three populations $\{Z\}$, $\{W\}$, $\{V\}$ is a Gaussian population. We therefore conclude that the bivariate population $\{(X_1, X_2)\}$ is not a Gaussian population. Thus we have an example of a bivariate population that is not a Gaussian population, and yet the two univariate populations $\{X_1\}$ and $\{X_2\}$ are Gaussian populations (approximately). ▪

You need to refer to more advanced books [11] for further information regarding multivariate Gaussian population models.

1.10 Exercises

1.10.1 A state health agency wants to determine the average number of days of sick leave state employees took last year. Define the target and study populations of items and of numbers. What parameter is to be studied?

1.10.2 In Exercise 1.10.1, the agency wants to predict the average number of days of sick leave state employees will take next year. Define the target and study populations of items and of numbers.

1.10.3 A pharmaceutical company wants to determine what proportion of the bottles of aspirin it will manufacture next month will be damaged before they reach retail stores. Define the target and study populations. What parameter is to be studied?

1.10.4 For an arbitrary population $\{X\}$ of numbers, obtain an upper bound for the proportion of numbers that are more than 3 standard deviations away from the mean (use Chebyshev's theorem). Write down any 20 numbers. Does Chebyshev's theorem apply to this set of numbers (check for many different values of c)?

1.10.5 Consider a population consisting of the following eight numbers (i.e., $N = 8$).

$$3, 7, 5, 2, 9, 8, 5, 6$$

Calculate the mean and the standard deviation for this population. How many distinct samples of size 3 can be obtained from this population? List all possible samples of size 3 that can be obtained from this population. For each sample, compute the sample mean and the sample standard deviation. Is the sample mean an unbiased estimate of the population mean for this population? Is the sample standard deviation an unbiased estimate of the population standard deviation for this population?

1.10.6 Suppose a population is Gaussian with mean 16 and variance 2. Find the proportion of the population that is less than 17.5. Find the proportion of the population that is between 12 and 16. Find the proportion of the population that is greater than 15.

1.10.7 A hospital administrator expects 10,000 patients next year. He wants to determine p, the proportion of patients who will stay less than 7 days. It is assumed that the length of stay Y in days is a Gaussian population (approximately) with mean μ_Y and standard deviation σ_Y. Define the target population of items and of numbers. If $\mu_Y = 8$ and $\sigma_Y = 2$, find p.

1.10.8 A manufacturer of cement blocks runs an experiment to determine the strength Y (in pounds per square inch) of the blocks made by a new method. The target population

consists of the strengths of all blocks that will be made by the new process if it is adopted. Twenty blocks are made by the new process, and it is assumed that they are a simple random sample from the target population, which is Gaussian with unknown mean μ_Y and standard deviation σ_Y. From the sample the following are computed: $\sum y_i = 12{,}000$, $\hat{\sigma}_Y = 14$. A 95% confidence statement is to be used to decide if the blocks are strong enough. Which confidence statement gives the appropriate information—a lower confidence bound, an upper confidence bound, or a two-sided confidence interval?

1.10.9 In Exercise 1.10.8, the cement blocks that were made by the old process produced blocks whose average strength was 540 pounds per square inch. To determine if the new process produces stronger blocks, the investigator wants to carry out a hypothesis test. State the appropriate null and alternative hypotheses.

1.10.10 In Exercise 1.10.9, will the null hypothesis be rejected at the 5% level? What is your conclusion about the new process?

1.10.11 In Exercise 1.10.9, compute the *P*-value for the test. Does this *P*-value give you more or less reason to reject the NH than the result of Exercise 1.10.10 ? Why?

1.10.12 In Exercise 1.10.8, compute a 95% lower confidence bound for μ_Y. State your conclusion based on this confidence bound. Which gives you more information—a confidence statement or a test?

1.10.13 The matrices A, B, C are defined as follows:

$$A = \begin{bmatrix} 6 & -1 & 0 \\ 3 & 1 & 5 \\ 2 & 1 & 9 \end{bmatrix} \qquad B = \begin{bmatrix} 3 & 1 \\ 2 & 6 \\ 1 & 0 \end{bmatrix} \qquad C = \begin{bmatrix} 13 & 2 & 1 \\ 0 & -1 & 4 \end{bmatrix}$$

Which of the following operations are defined? (a) $A + B$ (b) $C - B^T$ (c) AB (d) CA^T

1.10.14 In Exercise 1.10.13, evaluate the expressions that are defined.

1.10.15 For the matrix D defined by

$$D = \begin{bmatrix} 52 & 0.62 \\ 43 & 0.74 \\ 36 & 0.65 \\ 32 & 0.71 \\ 27 & 0.68 \\ 26 & 0.59 \\ 22 & 0.49 \\ 37 & 0.67 \\ 24 & 0.64 \\ 19 & 0.56 \\ 13 & 0.51 \end{bmatrix}$$

find $K = D^T D$.

1.10.16 In Exercise 1.10.15, find K^{-1}.

1.10.17 For the function $Y(x) = 6x^3 - 5x + 8$, $-6 \leq x \leq 5$, find $Y(3)$ and $Y(-2)$.

1.10.18 For the function $\mu_Y(x_1, x_2) = 3x_1^2 + 6x_2^2 - 2x_1x_2 + 4x_1 + 3x_2 - 9$, defined for all real values of x_1, x_2, find $\mu_Y(3, -1)$.

1.10.19 For the function $\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$, defined for $-3 \leq x_i \leq 4$ $(i = 1, 2, 3)$, find $\mu_Y(1, 3, -2)$.

1.10.20 Which of the following functions are simultaneously linear in all β_i?

a $\mu_Y(x) = \beta_0 + \beta_1 \dfrac{x}{(1+x)}$

b $\mu_Y(x_1, x_2) = \beta_0^2 + \beta_1x_1 + \beta_2x_2$

c $\mu_Y(x) = \beta_0 + x\,log|\beta_1| + \beta_2x^2$

d $\mu_Y(x_1, x_2) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$