9.6 THE CORRELATION MODEL

In the classic regression model, which has been the underlying model in our discussion up to this point, only Y, which has been called the dependent variable, is required to be random. The variable X is defined as a fixed (nonrandom or mathematical) variable and is referred to as the independent variable. Recall, also, that under this model observations are frequently obtained by preselecting values of X and determining corresponding values of Y.

When both Y and X are random variables, we have what is called the *correlation model*. Typically, under the correlation model, sample observations are obtained by selecting a random sample of the *units of association* (which may be persons, places, animals, points in time, or any other element on which the two measurements are taken) and taking on each a measurement of X and a measurement of Y. In this procedure, values of X are not preselected but occur at random, depending on the unit of association selected in the sample.

Although correlation analysis cannot be carried out meaningfully under the classic regression model, regression analysis can be carried out under the correlation model. Correlation involving two variables implies a co-relationship between variables that puts them on an equal footing and does not distinguish between them by referring to one as the dependent and the other as the independent variable. In fact, in the basic computational procedures, which are the same as for the regression model, we may fit a straight line to the data either by minimizing $\sum (y_i - \hat{y}_i)^2$ or by minimizing $\sum (x_i - \hat{x}_i)^2$. In other words, we may do a regression of X on Y as well as a regression of Y on X. The fitted line in the two cases in general will be different, and a logical question arises as to which line to fit.

If the objective is solely to obtain a measure of the strength of the relationship between the two variables, it does not matter which line is fitted, since the measure usually computed will be the same in either case. If, however, it is desired to use the equation describing the relationship between the two variables for the purposes discussed in the preceding sections, it does matter which line is fitted. The variable for which we wish to estimate means or to make predictions should be treated as the dependent variable; that is, this variable should be regressed on the other variable.

The Bivariate Normal Distribution Under the correlation model, X and Y are assumed to vary together in what is called a *joint distribution*. If this joint distribution is a normal distribution, it is referred to as a *bivariate normal distribution*. Inferences regarding this population may be made based on the results of samples properly drawn from it. If, on the other hand, the form of the joint distribution is known to be nonnormal, or if the form is unknown and there is no justification for assuming normality, inferential procedures are invalid, although descriptive measures may be computed.

Correlation Assumptions The following assumptions must hold for inferences about the population to be valid when sampling is from a bivariate distribution.

1. For each value of X there is a normally distributed subpopulation of Y values.
2. For each value of Y there is a normally distributed subpopulation of X values.
3. The joint distribution of X and Y is a normal distribution called the *bivariate normal distribution*.

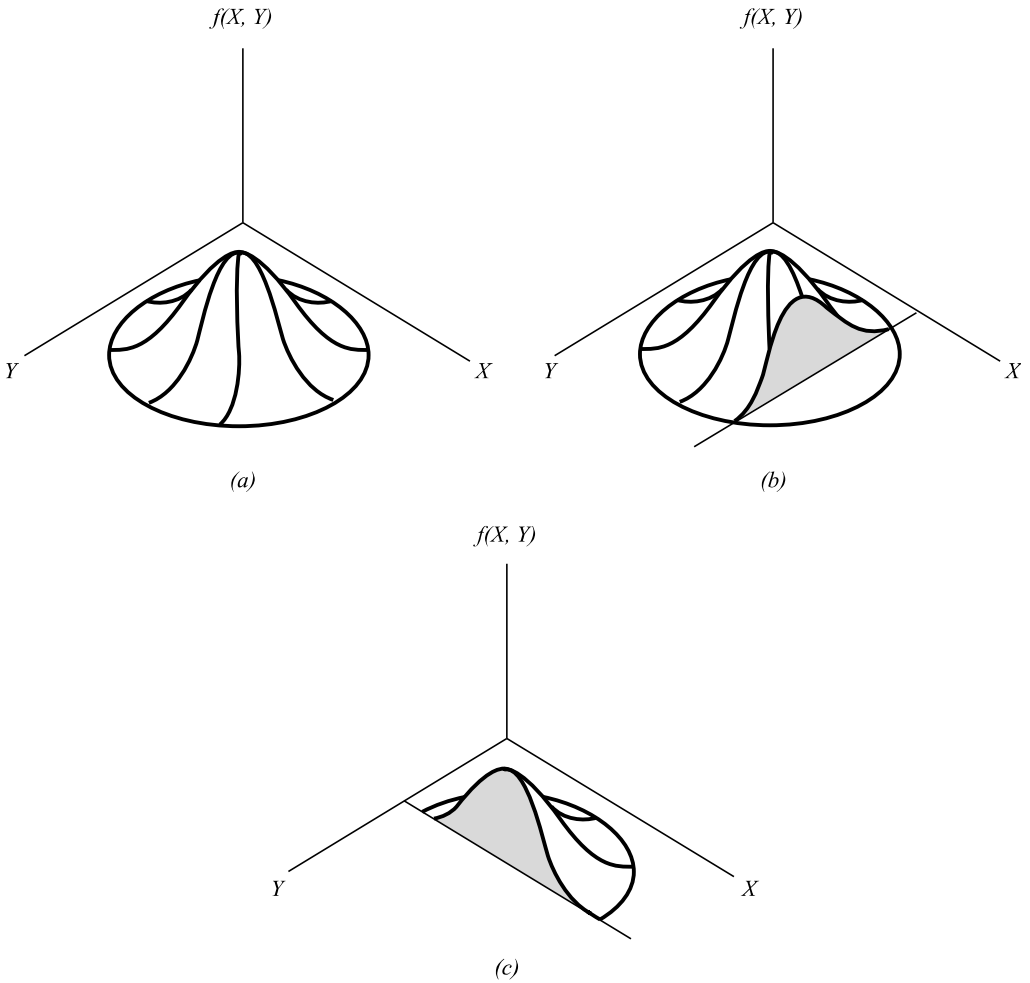

FIGURE 9.6.1 A bivariate normal distribution. (*a*) A bivariate normal distribution. (*b*) A cutaway showing normally distributed subpopulation of *Y* for given *X*. (*c*) A cutaway showing normally distributed subpopulation of *X* for given *Y*.

4. The subpopulations of Y values all have the same variance.
5. The subpopulations of X values all have the same variance.

The bivariate normal distribution is represented graphically in Figure 9.6.1. In this illustration we see that if we slice the mound parallel to Y at some value of X, the cutaway reveals the corresponding normal distribution of Y. Similarly, a slice through the mound parallel to X at some value of Y reveals the corresponding normally distributed subpopulation of X.

9.7 THE CORRELATION COEFFICIENT

The bivariate normal distribution discussed in Section 9.6 has five parameters, σ_x, σ_y, μ_x, μ_y, and ρ. The first four are, respectively, the standard deviations and means associated with the individual distributions. The other parameter, ρ, is called the population

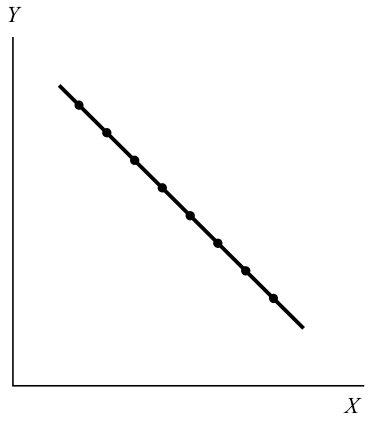

FIGURE 9.7.1 Scatter diagram for $r = -1$.

correlation coefficient and measures the strength of the linear relationship between X and Y.

The population correlation coefficient is the positive or negative square root of ρ^2, the population coefficient of determination previously discussed, and since the coefficient of determination takes on values between 0 and 1 inclusive, ρ may assume any value between -1 and $+1$. If $\rho = 1$ there is a perfect direct linear correlation between the two variables, while $\rho = -1$ indicates perfect inverse linear correlation. If $\rho = 0$ the two variables are not linearly correlated. The sign of ρ will always be the same as the sign of β_1, the slope of the population regression line for X and Y.

The sample correlation coefficient, r, describes the linear relationship between the sample observations on two variables in the same way that ρ describes the relationship in a population. The sample correlation coefficient is the square root of the sample coefficient of determination that was defined earlier.

Figures 9.4.5(d) and 9.4.5(c), respectively, show typical scatter diagrams where $r \to 0 (r^2 \to 0)$ and $r = +1 (r^2 = 1)$. Figure 9.7.1 shows a typical scatter diagram where $r = -1$.

We are usually interested in knowing if we may conclude that $\rho \neq 0$, that is, that X and Y are linearly correlated. Since ρ is usually unknown, we draw a random sample from the population of interest, compute r, the estimate of ρ, and test $H_0 : \rho = 0$ against the alternative $\rho \neq 0$. The procedure will be illustrated in the following example.

EXAMPLE 9.7.1

The purpose of a study by Kwast-Rabben et al. (A-7) was to analyze somatosensory evoked potentials (SEPs) and their interrelations following stimulation of digits I, III, and V in the hand. The researchers wanted to establish reference criteria in a control population. Thus, healthy volunteers were recruited for the study. In the future this information could be quite valuable as SEPs may provide a method to demonstrate functional disturbances in patients with suspected cervical root lesion who have pain and sensory symptoms. In the study, stimulation below-pain-level intensity was applied to the fingers. Recordings of spinal

responses were made with electrodes fixed by adhesive electrode cream to the subject's skin. One of the relationships of interest was the correlation between a subject's height (cm) and the peak spinal latency (Cv) of the SEP. The data for 155 measurements are shown in Table 9.7.1.

TABLE 9.7.1 Height and Spine SEP Measurements (Cv) from Stimulation of Digit I for 155 Subjects Described in Example 9.7.1

Height	Cv	Height	Cv	Height	Cv
149	14.4	168	16.3	181	15.8
149	13.4	168	15.3	181	18.8
155	13.5	168	16.0	181	18.6
155	13.5	168	16.6	182	18.0
156	13.0	168	15.7	182	17.9
156	13.6	168	16.3	182	17.5
157	14.3	168	16.6	182	17.4
157	14.9	168	15.4	182	17.0
158	14.0	170	16.6	182	17.5
158	14.0	170	16.0	182	17.8
160	15.4	170	17.0	184	18.4
160	14.7	170	16.4	184	18.5
161	15.5	171	16.5	184	17.7
161	15.7	171	16.3	184	17.7
161	15.8	171	16.4	184	17.4
161	16.0	171	16.5	184	18.4
161	14.6	172	17.6	185	19.0
161	15.2	172	16.8	185	19.6
162	15.2	172	17.0	187	19.1
162	16.5	172	17.6	187	19.2
162	17.0	173	17.3	187	17.8
162	14.7	173	16.8	187	19.3
163	16.0	174	15.5	188	17.5
163	15.8	174	15.5	188	18.0
163	17.0	175	17.0	189	18.0
163	15.1	175	15.6	189	18.8
163	14.6	175	16.8	190	18.3
163	15.6	175	17.4	190	18.6
163	14.6	175	17.6	190	18.8
164	17.0	175	16.5	190	19.2
164	16.3	175	16.6	191	18.5
164	16.0	175	17.0	191	18.5
164	16.0	176	18.0	191	19.0
165	15.7	176	17.0	191	18.5
165	16.3	176	17.4	194	19.8

(*Continued*)

Height	Cv	Height	Cv	Height	Cv
165	17.4	176	18.2	194	18.8
165	17.0	176	17.3	194	18.4
165	16.3	177	17.2	194	19.0
166	14.1	177	18.3	195	18.0
166	14.2	179	16.4	195	18.2
166	14.7	179	16.1	196	17.6
166	13.9	179	17.6	196	18.3
166	17.2	179	17.8	197	18.9
167	16.7	179	16.1	197	19.2
167	16.5	179	16.0	200	21.0
167	14.7	179	16.0	200	19.2
167	14.3	179	17.5	202	18.6
167	14.8	179	17.5	202	18.6
167	15.0	180	18.0	182	20.0
167	15.5	180	17.9	190	20.0
167	15.4	181	18.4	190	19.5
168	17.3	181	16.4		

Source: Data provided courtesy of Olga Kwast-Rabben, Ph.D.

Solution: The scatter diagram and least-squares regression line are shown in Figure 9.7.2.

Let us assume that the investigator wishes to obtain a regression equation to use for estimating and predicting purposes. In that case the sample correlation coefficient will be obtained by the methods discussed under the regression model.

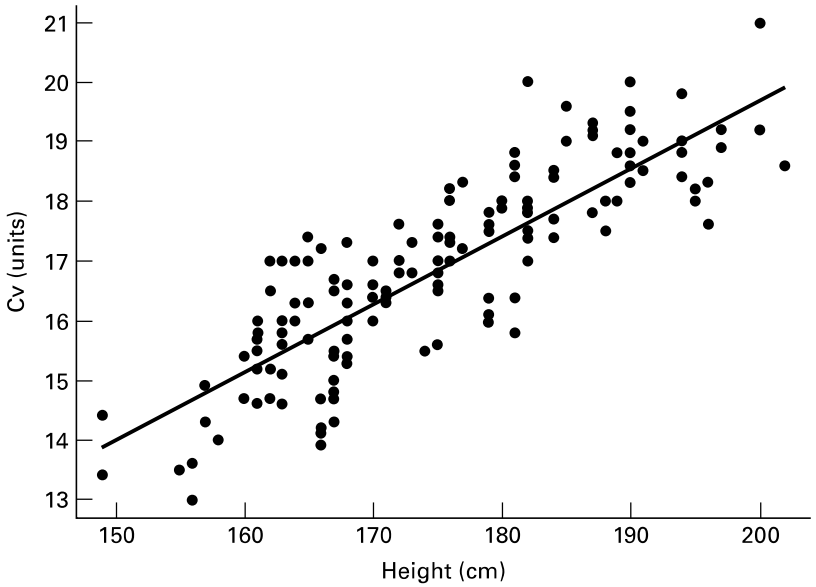

FIGURE 9.7.2 Height and cervical (spine) potentials in digit I stimulation for the data described in Example 9.7.1.

9.1 INTRODUCTION

In analyzing data for the health sciences disciplines, we find that it is frequently desirable to learn something about the relationship between two numeric variables. We may, for example, be interested in studying the relationship between blood pressure and age, height and weight, the concentration of an injected drug and heart rate, the consumption level of some nutrient and weight gain, the intensity of a stimulus and reaction time, or total family income and medical care expenditures. The nature and strength of the relationships between variables such as these may be examined using linear models such as *regression* and *correlation* analysis, two statistical techniques that, although related, serve different purposes.

Regression Regression analysis is helpful in assessing specific forms of the relationship between variables, and the ultimate objective when this method of analysis is employed usually is to *predict* or *estimate* the value of one variable corresponding to a given value of another variable. The ideas of regression were first elucidated by the English scientist Sir Francis Galton (1822–1911) in reports of his research on heredity—first in sweet peas and later in human stature. He described a tendency of adult offspring, having either short or tall parents, to revert back toward the average height of the general population. He first used the word *reversion*, and later *regression*, to refer to this phenomenon.

Correlation Correlation analysis, on the other hand, is concerned with measuring the strength of the relationship between variables. When we compute measures of correlation from a set of data, we are interested in the degree of the *correlation* between variables. Again, the concepts and terminology of correlation analysis originated with Galton, who first used the word *correlation* in 1888.

In this chapter our discussion is limited to the exploration of the linear relationship between two variables. The concepts and methods of regression are covered first, beginning in the next section. In Section 9.6 the ideas and techniques of correlation are introduced. In the next chapter we consider the case where there is an interest in the relationships among three or more variables.

Regression and correlation analysis are areas in which the speed and accuracy of a computer are most appreciated. The data for the exercises of this chapter, therefore, are presented in a way that makes them suitable for computer processing. As is always the case, the input requirements and output features of the particular programs and software packages to be used should be studied carefully.

9.2 THE REGRESSION MODEL

In the typical regression problem, as in most problems in applied statistics, researchers have available for analysis a sample of observations from some real or hypothetical population. Based on the results of their analysis of the sample data, they are interested in reaching decisions about the population from which the sample is presumed to have been drawn. It is important, therefore, that the researchers understand the nature of the population in which they are interested. They should know enough about the population to be able either to construct a mathematical model for its representation or to determine if it reasonably fits

some established model. A researcher about to analyze a set of data by the methods of simple linear regression, for example, should be secure in the knowledge that the simple linear regression model is, at least, an approximate representation of the population. It is unlikely that the model will be a perfect portrait of the real situation, since this characteristic is seldom found in models of practical value. A model constructed so that it corresponds precisely with the details of the situation is usually too complex to yield any information of value. On the other hand, the results obtained from the analysis of data that have been forced into a model that does not fit are also worthless. Fortunately, however, a perfectly fitting model is not a requirement for obtaining useful results. Researchers, then, should be able to distinguish between the occasion when their chosen models and the data are sufficiently compatible for them to proceed and the case where their chosen model must be abandoned.

Assumptions Underlying Simple Linear Regression In the simple linear regression model two variables, usually labeled X and Y, are of interest. The letter X is usually used to designate a variable referred to as the *independent variable*, since frequently it is controlled by the investigator; that is, values of X may be selected by the investigator and, corresponding to each preselected value of X, one or more values of another variable, labeled Y, are obtained. The variable, Y, accordingly, is called the *dependent variable*, and we speak of the regression of Y on X. The following are the assumptions underlying the simple linear regression model.

1. Values of the independent variable X are said to be "fixed." This means that the values of X are preselected by the investigator so that in the collection of the data they are not allowed to vary from these preselected values. In this model, X is referred to by some writers as a *nonrandom* variable and by others as a *mathematical* variable. It should be pointed out at this time that the statement of this assumption classifies our model as the *classical regression model*. Regression analysis also can be carried out on data in which X is a random variable.
2. The variable X is measured without error. Since no measuring procedure is perfect, this means that the magnitude of the measurement error in X is negligible.
3. For each value of X there is a subpopulation of Y values. For the usual inferential procedures of estimation and hypothesis testing to be valid, these subpopulations must be normally distributed. In order that these procedures may be presented it will be assumed that the Y values are normally distributed in the examples and exercises that follow.
4. The variances of the subpopulations of Y are all equal and denoted by σ^2.
5. The means of the subpopulations of Y all lie on the same straight line. This is known as the *assumption of linearity*. This assumption may be expressed symbolically as

$$\mu_{y|x} = \beta_0 + \beta_1 x \tag{9.2.1}$$

where $\mu_{y|x}$ is the mean of the subpopulation of Y values for a particular value of X, and β_0 and β_1 are called the population regression coefficients. Geometrically, β_0 and β_1 represent the y-intercept and slope, respectively, of the line on which all of the means are assumed to lie.

6. The Y values are statistically independent. In other words, in drawing the sample, it is assumed that the values of Y chosen at one value of X in no way depend on the values of Y chosen at another value of X.

These assumptions may be summarized by means of the following equation, which is called the simple linear regression model:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{9.2.2}$$

where y is a typical value from one of the subpopulations of Y, β_0 and β_1 are as defined for Equation 9.2.1, and ϵ is called the error term. If we solve 9.2.2 for ϵ, we have

$$\begin{aligned} \epsilon &= y - (\beta_0 + \beta_1 x) \\ &= y - \mu_{y|x} \end{aligned} \tag{9.2.3}$$

and we see that ϵ shows the amount by which y deviates from the mean of the subpopulation of Y values from which it is drawn. As a consequence of the assumption that the subpopulations of Y values are normally distributed with equal variances, the ϵ's for each subpopulation are normally distributed with a variance equal to the common variance of the subpopulations of Y values.

The following acronym will help the reader remember most of the assumptions necessary for inference in linear regression analysis:

LINE [Linear (assumption 5), Independent (assumption 6), Normal (assumption 3), Equal variances (assumption 4)]

A graphical representation of the regression model is given in Figure 9.2.1.

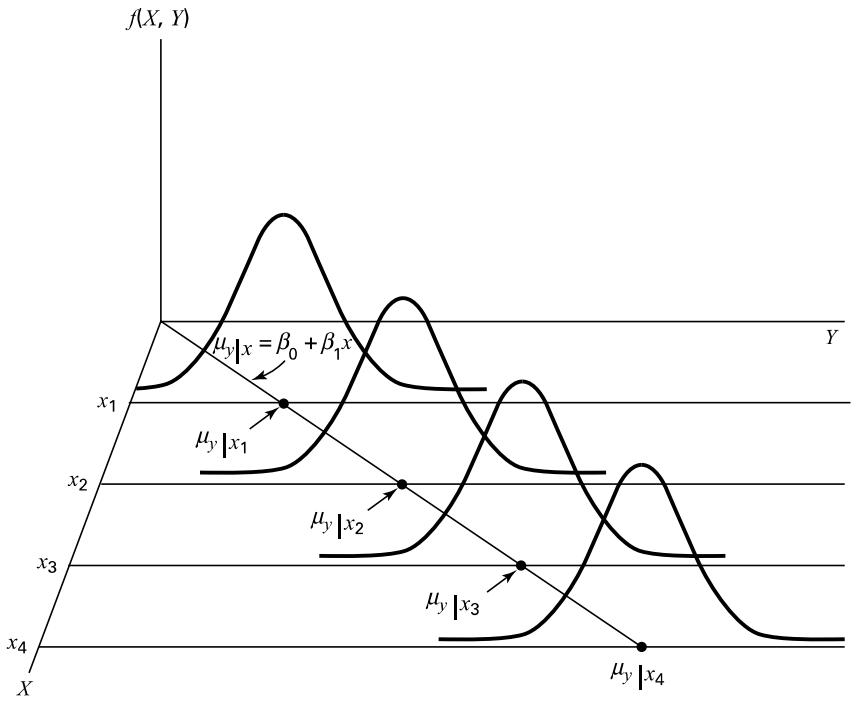

FIGURE 9.2.1 Representation of the simple linear regression model.

9.3 THE SAMPLE REGRESSION EQUATION

In simple linear regression the object of the researcher's interest is the population regression equation—the equation that describes the true relationship between the dependent variable Y and the independent variable X. The variable designated by Y is sometimes called the *response variable* and X is sometimes called the *predictor variable*.

In an effort to reach a decision regarding the likely form of this relationship, the researcher draws a sample from the population of interest and using the resulting data, computes a sample regression equation that forms the basis for reaching conclusions regarding the unknown population regression equation.

Steps in Regression Analysis In the absence of extensive information regarding the nature of the variables of interest, a frequently employed strategy is to assume initially that they are linearly related. Subsequent analysis, then, involves the following steps.

1. Determine whether or not the assumptions underlying a linear relationship are met in the data available for analysis.
2. Obtain the equation for the line that best fits the sample data.
3. Evaluate the equation to obtain some idea of the strength of the relationship and the usefulness of the equation for predicting and estimating.
4. If the data appear to conform satisfactorily to the linear model, use the equation obtained from the sample data to predict and to estimate.

When we use the regression equation to *predict*, we will be predicting the value Y is likely to have when X has a given value. When we use the equation to *estimate*, we will be estimating the mean of the subpopulation of Y values assumed to exist at a given value of X. Note that the sample data used to obtain the regression equation consist of known values of both X and Y. When the equation is used to predict and to estimate Y, only the corresponding values of X will be known. We illustrate the steps involved in simple linear regression analysis by means of the following example.

EXAMPLE 9.3.1

Després et al. (A-1) point out that the topography of adipose tissue (AT) is associated with metabolic complications considered as risk factors for cardiovascular disease. It is important, they state, to measure the amount of intraabdominal AT as part of the evaluation of the cardiovascular-disease risk of an individual. Computed tomography (CT), the only available technique that precisely and reliably measures the amount of deep abdominal AT, however, is costly and requires irradiation of the subject. In addition, the technique is not available to many physicians. Després and his colleagues conducted a study to develop equations to predict the amount of deep abdominal AT from simple anthropometric measurements. Their subjects were men between the ages of 18 and 42 years who were free from metabolic disease that would require treatment. Among the measurements taken on each subject were deep abdominal AT obtained by CT and waist circumference as

shown in Table 9.3.1. A question of interest is how well one can predict and estimate deep abdominal AT from knowledge of the waist circumference. This question is typical of those that can be answered by means of regression analysis. Since deep abdominal AT is the variable about which we wish to make predictions and estimations, it is the dependent variable. The variable waist measurement, knowledge of which will be used to make the predictions and estimations, is the independent variable.

TABLE 9.3.1 Waist Circumference (cm), *X*, and Deep Abdominal AT, *Y*, of 109 Men

Subject	*X*	*Y*	Subject	*X*	*Y*	Subject	*X*	*Y*
1	74.75	25.72	38	103.00	129.00	75	108.00	217.00
2	72.60	25.89	39	80.00	74.02	76	100.00	140.00
3	81.80	42.60	40	79.00	55.48	77	103.00	109.00
4	83.95	42.80	41	83.50	73.13	78	104.00	127.00
5	74.65	29.84	42	76.00	50.50	79	106.00	112.00
6	71.85	21.68	43	80.50	50.88	80	109.00	192.00
7	80.90	29.08	44	86.50	140.00	81	103.50	132.00
8	83.40	32.98	45	83.00	96.54	82	110.00	126.00
9	63.50	11.44	46	107.10	118.00	83	110.00	153.00
10	73.20	32.22	47	94.30	107.00	84	112.00	158.00
11	71.90	28.32	48	94.50	123.00	85	108.50	183.00
12	75.00	43.86	49	79.70	65.92	86	104.00	184.00
13	73.10	38.21	50	79.30	81.29	87	111.00	121.00
14	79.00	42.48	51	89.80	111.00	88	108.50	159.00
15	77.00	30.96	52	83.80	90.73	89	121.00	245.00
16	68.85	55.78	53	85.20	133.00	90	109.00	137.00
17	75.95	43.78	54	75.50	41.90	91	97.50	165.00
18	74.15	33.41	55	78.40	41.71	92	105.50	152.00
19	73.80	43.35	56	78.60	58.16	93	98.00	181.00
20	75.90	29.31	57	87.80	88.85	94	94.50	80.95
21	76.85	36.60	58	86.30	155.00	95	97.00	137.00
22	80.90	40.25	59	85.50	70.77	96	105.00	125.00
23	79.90	35.43	60	83.70	75.08	97	106.00	241.00
24	89.20	60.09	61	77.60	57.05	98	99.00	134.00
25	82.00	45.84	62	84.90	99.73	99	91.00	150.00
26	92.00	70.40	63	79.80	27.96	100	102.50	198.00
27	86.60	83.45	64	108.30	123.00	101	106.00	151.00
28	80.50	84.30	65	119.60	90.41	102	109.10	229.00
29	86.00	78.89	66	119.90	106.00	103	115.00	253.00
30	82.50	64.75	67	96.50	144.00	104	101.00	188.00
31	83.50	72.56	68	105.50	121.00	105	100.10	124.00
32	88.10	89.31	69	105.00	97.13	106	93.30	62.20
33	90.80	78.94	70	107.00	166.00	107	101.80	133.00
34	89.40	83.55	71	107.00	87.99	108	107.90	208.00
35	102.00	127.00	72	101.00	154.00	109	108.50	208.00
36	94.50	121.00	73	97.00	100.00			
37	91.00	107.00	74	100.00	123.00			

Source: Data provided courtesy of Jean-Pierre Després, Ph.D.

The Scatter Diagram

A first step that is usually useful in studying the relationship between two variables is to prepare a *scatter diagram* of the data such as is shown in Figure 9.3.1. The points are plotted by assigning values of the independent variable X to the horizontal axis and values of the dependent variable Y to the vertical axis.

The pattern made by the points plotted on the scatter diagram usually suggests the basic nature and strength of the relationship between two variables. As we look at Figure 9.3.1, for example, the points seem to be scattered around an invisible straight line. The scatter diagram also shows that, in general, subjects with large waist circumferences also have larger amounts of deep abdominal AT. These impressions suggest that the relationship between the two variables may be described by a straight line crossing the Y-axis below the origin and making approximately a 45-degree angle with the X-axis. It looks as if it would be simple to draw, freehand, through the data points the line that describes the relationship between X and Y. It is highly unlikely, however, that the lines drawn by any two people would be exactly the same. In other words, for every person drawing such a line by eye, or freehand, we would expect a slightly different line. The question then arises as to which line best describes the relationship between the two variables. We cannot obtain an answer to this question by inspecting the lines. In fact, it is not likely that any freehand line

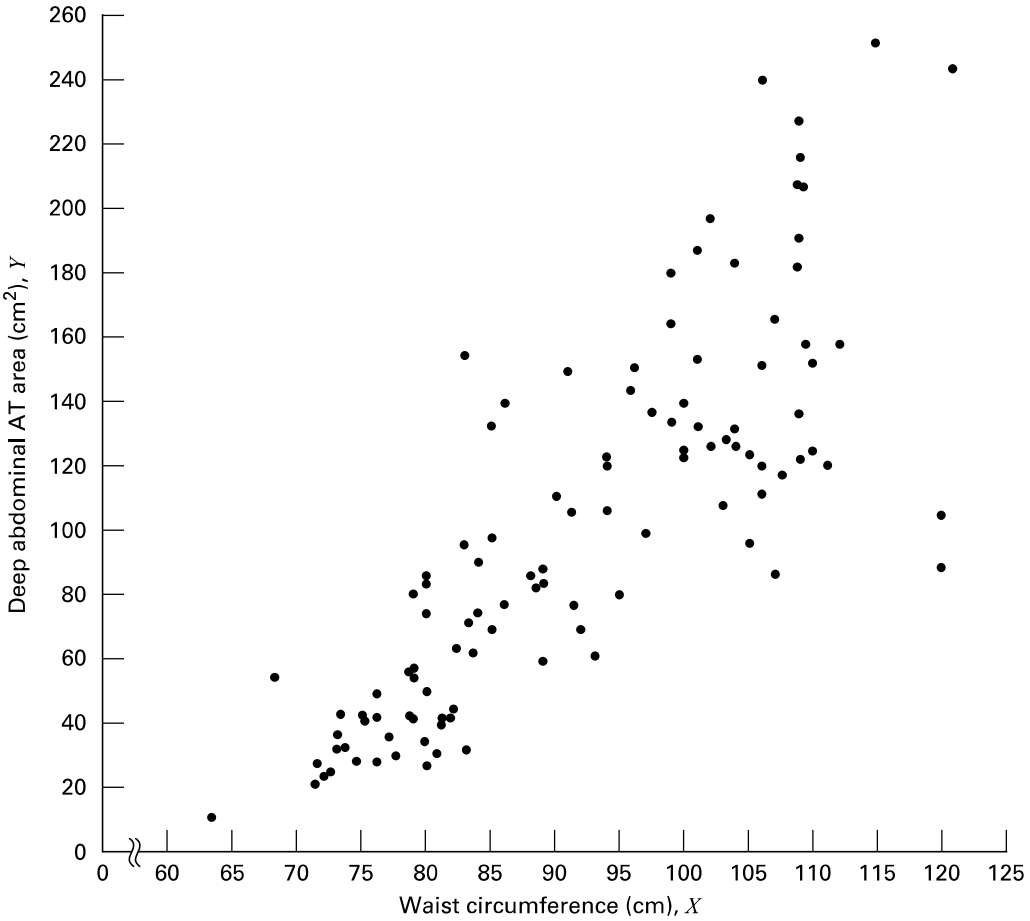

FIGURE 9.3.1 Scatter diagram of data shown in Table 9.3.1.

drawn through the data will be the line that best describes the relationship between X and Y, since freehand lines will reflect any defects of vision or judgment of the person drawing the line. Similarly, when judging which of two lines best describes the relationship, subjective evaluation is liable to the same deficiencies.

What is needed for obtaining the desired line is some method that is not fraught with these difficulties.

The Least-Squares Line

The method commonly employed for obtaining the desired line is known as the *method of least squares*, and the resulting line is called the *least-squares line*. The reason for calling the method by this name will be explained in the discussion that follows.

We recall from algebra that the general equation for a straight line may be written as

$$y = a + bx \tag{9.3.1}$$

where y is a value on the vertical axis, x is a value on the horizontal axis, a is the point where the line crosses the vertical axis, and b shows the amount by which y changes for each unit change in x. We refer to a as the y-intercept and b as the *slope* of the line. To draw a line based on Equation 9.3.1, we need the numerical values of the constants a and b. Given these constants, we may substitute various values of x into the equation to obtain corresponding values of y. The resulting points may be plotted. Since any two such coordinates determine a straight line, we may select any two, locate them on a graph, and connect them to obtain the line corresponding to the equation.

Obtaining the Least-Square Line

The least-squares regression line equation may be obtained from sample data by simple arithmetic calculations that may be carried out by hand using the following equations

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{9.3.2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{9.3.3}$$

where x_i and y_i are the corresponding values of each data point (X, Y), $\bar{x}$ and $\bar{y}$ are the means of the X and Y sample data values, respectively, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the intercept β_0 and slope β_1, respectively, of the population regression line. Since the necessary hand calculations are time consuming, tedious, and subject to error, the regression line equation is best obtained through the use of a computer software package. Although the typical researcher need not be concerned with the arithmetic involved, the interested reader will find them discussed in references listed at the end of this chapter.

For the data in Table 9.3.1 we obtain the least-squares regression equation by means of MINITAB. After entering the X values in Column 1 and the Y values in Column 2 we proceed as shown in Figure 9.3.2.

For now, the only information from the output in Figure 9.3.2 that we are interested in is the regression equation. Other information in the output will be discussed later.

12.1 INTRODUCTION

In the chapters on estimation and hypothesis testing, brief mention is made of the chi-square distribution in the construction of confidence intervals for, and the testing of, hypotheses concerning a population variance. This distribution, which is one of the most widely used distributions in statistical applications, has many other uses. Some of the more common ones are presented in this chapter along with a more complete description of the distribution itself, which follows in the next section.

The chi-square distribution is the most frequently employed statistical technique for the analysis of count or frequency data. For example, we may know for a sample of hospitalized patients how many are male and how many are female. For the same sample we may also know how many have private insurance coverage, how many have Medicare insurance, and how many are on Medicaid assistance. We may wish to know, for the population from which the sample was drawn, if the type of insurance coverage differs according to gender. For another sample of patients, we may have frequencies for each diagnostic category represented and for each geographic area represented. We might want to know if, in the population from which the same was drawn, there is a relationship between area of residence and diagnosis. We will learn how to use chi-square analysis to answer these types of questions.

There are other statistical techniques that may be used to analyze frequency data in an effort to answer other types of questions. In this chapter we will also learn about these techniques.

12.2 THE MATHEMATICAL PROPERTIES OF THE CHI-SQUARE DISTRIBUTION

The chi-square distribution may be derived from normal distributions. Suppose that from a normally distributed random variable Y with mean μ and variance σ^2 we randomly and independently select samples of size $n = 1$. Each value selected may be transformed to the standard normal variable z by the familiar formula

$$z_i = \frac{y_i - \mu}{\sigma} \tag{12.2.1}$$

Each value of z may be squared to obtain z^2. When we investigate the sampling distribution of z^2, we find that it follows a chi-square distribution with 1 degree of freedom. That is,

$$\chi^2_{(1)} = \left(\frac{y - \mu}{\sigma}\right)^2 = z^2$$

Now suppose that we randomly and independently select samples of size $n = 2$ from the normally distributed population of Y values. Within each sample we may transform each

value of y to the standard normal variable z and square as before. If the resulting values of z^2 for each sample are added, we may designate this sum by

$$\chi^2_{(2)} = \left(\frac{y_1 - \mu}{\sigma}\right)^2 + \left(\frac{y_2 - \mu}{\sigma}\right)^2 = z_1^2 + z_2^2$$

since it follows the chi-square distribution with 2 degrees of freedom, the number of independent squared terms that are added together.

The procedure may be repeated for any sample size n. The sum of the resulting z^2 values in each case will be distributed as chi-square with n degrees of freedom. In general, then,

$$\chi^2_{(n)} = z_1^2 + z_2^2 + \cdots + z_n^2 \tag{12.2.2}$$

follows the chi-square distribution with n degrees of freedom. The mathematical form of the chi-square distribution is as follows:

$$f(u) = \frac{1}{\left(\frac{k}{2} - 1\right)!} \frac{1}{2^{k/2}} u^{(k/2)-1} e^{-(u/2)}, \quad u > 0 \tag{12.2.3}$$

where e is the irrational number 2.71828 . . . and k is the number of degrees of freedom. The variate u is usually designated by the Greek letter chi (χ) and, hence, the distribution is called the chi-square distribution. As we pointed out in Chapter 6, the chi-square distribution has been tabulated in Appendix Table F. Further use of the table is demonstrated as the need arises in succeeding sections.

The mean and variance of the chi-square distribution are k and $2k$, respectively. The modal value of the distribution is $k - 2$ for values of k greater than or equal to 2 and is zero for $k = 1$.

The shapes of the chi-square distributions for several values of k are shown in Figure 6.9.1. We observe in this figure that the shapes for $k = 1$ and $k = 2$ are quite different from the general shape of the distribution for $k > 2$. We also see from this figure that chi-square assumes values between 0 and infinity. It cannot take on negative values, since it is the sum of values that have been squared. A final characteristic of the chi-square distribution worth noting is that the sum of two or more independent chi-square variables also follows a chi-square distribution.

Types of Chi-Square Tests As already noted, we make use of the chi-square distribution in this chapter in testing hypotheses where the data available for analysis are in the form of frequencies. These hypothesis testing procedures are discussed under the topics of *tests of goodness-of-fit, tests of independence,* and *tests of homogeneity.* We will discover that, in a sense, all of the chi-square tests that we employ may be thought of as goodness-of-fit tests, in that they test the goodness-of-fit of observed frequencies to frequencies that one would expect if the data were generated under some particular theory or hypothesis. We, however, reserve the phrase "goodness-of-fit" for use in a more

restricted sense. We use it to refer to a comparison of a sample distribution to some theoretical distribution that it is assumed describes the population from which the sample came. The justification of our use of the distribution in these situations is due to Karl Pearson (1), who showed that the chi-square distribution may be used as a test of the agreement between observation and hypothesis whenever the data are in the form of frequencies. An extensive treatment of the chi-square distribution is to be found in the book by Lancaster (2). Nikulin and Greenwood (3) offer practical advice for conducting chi-square tests.

Observed Versus Expected Frequencies The chi-square statistic is most appropriate for use with categorical variables, such as marital status, whose values are the categories married, single, widowed, and divorced. The quantitative data used in the computation of the test statistic are the frequencies associated with each category of the one or more variables under study. There are two sets of frequencies with which we are concerned, *observed frequencies* and *expected frequencies*. The observed frequencies are the number of subjects or objects in our sample that fall into the various categories of the variable of interest. For example, if we have a sample of 100 hospital patients, we may observe that 50 are married, 30 are single, 15 are widowed, and 5 are divorced. Expected frequencies are the number of subjects or objects in our sample that we would expect to observe if some null hypothesis about the variable is true. For example, our null hypothesis might be that the four categories of marital status are equally represented in the population from which we drew our sample. In that case we would expect our sample to contain 25 married, 25 single, 25 widowed, and 25 divorced patients.

The Chi-Square Test Statistic The test statistic for the chi-square tests we discuss in this chapter is

$$X^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right] \tag{12.2.4}$$

When the null hypothesis is true, X^2 is distributed approximately as χ^2 with $k - r$ degrees of freedom. In determining the degrees of freedom, k is equal to the number of groups for which observed and expected frequencies are available, and r is the number of restrictions or constraints imposed on the given comparison. A restriction is imposed when we force the sum of the expected frequencies to equal the sum of the observed frequencies, and an additional restriction is imposed for each parameter that is estimated from the sample.

In Equation 12.2.4, O_i is the observed frequency for the ith category of the variable of interest, and E_i is the expected frequency (given that H_0 is true) for the ith category.

The quantity X^2 is a measure of the extent to which, in a given situation, pairs of observed and expected frequencies agree. As we will see, the nature of X^2 is such that when there is close agreement between observed and expected frequencies it is small, and when the agreement is poor it is large. Consequently, only a sufficiently large value of X^2 will cause rejection of the null hypothesis.

If there is perfect agreement between the observed frequencies and the frequencies that one would expect, given that H_0 is true, the term $O_i - E_i$ in Equation 12.2.4 will be

equal to zero for each pair of observed and expected frequencies. Such a result would yield a value of X^2 equal to zero, and we would be unable to reject H_0.

When there is disagreement between observed frequencies and the frequencies one would expect given that H_0 is true, at least one of the $O_i - E_i$ terms in Equation 12.2.4 will be a nonzero number. In general, the poorer the agreement between the O_i and the E_i, the greater or the more frequent will be these nonzero values. As noted previously, if the agreement between the O_i and the E_i is sufficiently poor (resulting in a sufficiently large X^2 value,) we will be able to reject H_0.

When there is disagreement between a pair of observed and expected frequencies, the difference may be either positive or negative, depending on which of the two frequencies is the larger. Since the measure of agreement, X^2, is a sum of component quantities whose magnitudes depend on the difference $O_i - E_i$, positive and negative differences must be given equal weight. This is achieved by squaring each $O_i - E_i$ difference. Dividing the squared differences by the appropriate expected frequency converts the quantity to a term that is measured in original units. Adding these individual $(O_i - E_i)^2/E_i$ terms yields X^2, a summary statistic that reflects the extent of the overall agreement between observed and expected frequencies.

The Decision Rule The quantity $\sum[(O_i - E_i)^2/E_i]$ will be small if the observed and expected frequencies are close together and will be large if the differences are large.

The computed value of X^2 is compared with the tabulated value of χ^2 with $k - r$ degrees of freedom. The decision rule, then, is: Reject H_0 if X^2 is greater than or equal to the tabulated χ^2 for the chosen value of α.

Small Expected Frequencies Frequently in applications of the chi-square test the expected frequency for one or more categories will be small, perhaps much less than 1. In the literature the point is frequently made that the approximation of X^2 to χ^2 is not strictly valid when some of the expected frequencies are small. There is disagreement among writers, however, over what size expected frequencies are allowable before making some adjustment or abandoning χ^2 in favor of some alternative test. Some writers, especially the earlier ones, suggest lower limits of 10, whereas others suggest that all expected frequencies should be no less than 5. Cochran (4,5), suggests that for goodness-of-fit tests of unimodal distributions (such as the normal), the minimum expected frequency can be as low as 1. If, in practice, one encounters one or more expected frequencies less than 1, adjacent categories may be combined to achieve the suggested minimum. Combining reduces the number of categories and, therefore, the number of degrees of freedom. Cochran's suggestions appear to have been followed extensively by practitioners in recent years.

12.3 TESTS OF GOODNESS-OF-FIT

As we have pointed out, a goodness-of-fit test is appropriate when one wishes to decide if an observed distribution of frequencies is incompatible with some preconceived or hypothesized distribution.

We may, for example, wish to determine whether or not a sample of observed values of some random variable is compatible with the hypothesis that it was drawn from a population of values that is normally distributed. The procedure for reaching a decision consists of placing the values into mutually exclusive categories or class intervals and noting the frequency of occurrence of values in each category. We then make use of our knowledge of normal distributions to determine the frequencies for each category that one could expect if the sample had come from a normal distribution. If the discrepancy is of such magnitude that it could have come about due to chance, we conclude that the sample may have come from a normal distribution. In a similar manner, tests of goodness-of-fit may be carried out in cases where the hypothesized distribution is the binomial, the Poisson, or any other distribution. Let us illustrate in more detail with some examples of tests of hypotheses of goodness-of-fit.

EXAMPLE 12.3.1 *The Normal Distribution*

Cranor and Christensen (A-1) conducted a study to assess short-term clinical, economic, and humanistic outcomes of pharmaceutical care services for patients with diabetes in community pharmacies. For 47 of the subjects in the study, cholesterol levels are summarized in Table 12.3.1.

We wish to know whether these data provide sufficient evidence to indicate that the sample did not come from a normally distributed population. Let $\alpha = .05$

Solution:

1. **Data.** See Table 12.3.1.
2. **Assumptions.** We assume that the sample available for analysis is a simple random sample.

TABLE 12.3.1 Cholesterol Levels as Described in Example 12.3.1

Cholesterol Level (mg/dl)	Number of Subjects
100.0–124.9	1
125.0–149.9	3
150.0–174.9	8
175.0–199.9	18
200.0–224.9	6
225.0–249.9	4
250.0–274.9	4
275.0–299.9	3

Source: Data provided courtesy of Carole W. Cranor, and Dale B. Christensen, "The Asheville Project: Short-Term Outcomes of a Community Pharmacy Diabetes Care Program," *Journal of the American Pharmaceutical Association*, 43 (2003), 149–159.

3. **Hypotheses.**
   H_0: In the population from which the sample was drawn, cholesterol levels are normally distributed.
   H_A: The sampled population is not normally distributed.
4. **Test statistic.** The test statistic is

$$X^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

5. **Distribution of test statistic.** If H_0 is true, the test statistic is distributed approximately as chi-square with $k - r$ degrees of freedom. The values of k and r will be determined later.
6. **Decision rule.** We will reject H_0 if the computed value of X^2 is equal to or greater than the critical value of chi-square.
7. **Calculation of test statistic.** Since the mean and variance of the hypothesized distribution are not specified, the sample data must be used to estimate them. These parameters, or their estimates, will be needed to compute the frequency that would be expected in each class interval when the null hypothesis is true. The mean and standard deviation computed from the grouped data of Table 12.3.1 are

$$\bar{x} = 198.67$$
$$s = 41.31$$

As the next step in the analysis, we must obtain for each class interval the frequency of occurrence of values that we would expect when the null hypothesis is true, that is, if the sample were, in fact, drawn from a normally distributed population of values. To do this, we first determine the expected relative frequency of occurrence of values for each class interval and then multiply these expected relative frequencies by the total number of values to obtain the expected number of values for each interval.

The Expected Relative Frequencies

It will be recalled from our study of the normal distribution that the relative frequency of occurrence of values equal to or less than some specified value, say, x_0, of the normally distributed random variable X is equivalent to the area under the curve and to the left of x_0 as represented by the shaded area in Figure 12.3.1. We obtain the numerical value of this area by converting x_0 to a standard normal deviation by the formula $z_0 = (x_0 - \mu)/\sigma$ and finding the appropriate value in Appendix Table D. We use this procedure to obtain the expected relative frequencies corresponding to each of the class intervals in Table 12.3.1. We estimate μ and σ with $\bar{x}$ and s as computed from the grouped sample data. The first step consists of obtaining z values corresponding to the lower limit of each class interval. The area between two successive z values will give the expected relative frequency of occurrence of values for the corresponding class interval.

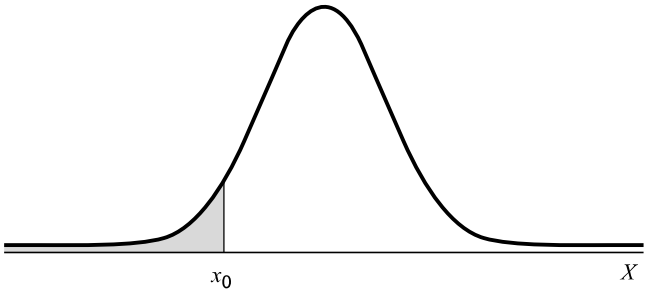

FIGURE 12.3.1 A normal distribution showing the relative frequency of occurrence of values less than or equal to x_0. The shaded area represents the relative frequency of occurrence of values equal to or less than x_0.

For example, to obtain the expected relative frequency of occurrence of values in the interval 100.0 to 124.9 we proceed as follows:

$$\text{The } z \text{ value corresponding to } X = 100.0 \text{ is } z = \frac{100.0 - 198.67}{41.31} = -2.39$$

$$\text{The } z \text{ value corresponding to } X = 125.0 \text{ is } z = \frac{125.0 - 198.67}{41.31} = -1.78$$

In Appendix Table D we find that the area to the left of -2.39 is .0084, and the area to the left of -1.78 is .0375. The area between -1.78 and -2.39 is equal to $.0375 - .0084 = .0291$, which is equal to the expected relative frequency of occurrence of cholesterol levels within the interval 100.0 to 124.9. This tells us that if the null hypothesis is true, that is, if the cholesterol levels are normally distributed, we should expect 2.91 percent of the values in our sample to be between 100.0 and 124.9. When we multiply our total sample size, 47, by .0291 we find the expected frequency for the interval to be 1.4. Similar calculations will give the expected frequencies for the other intervals as shown in Table 12.3.2.

TABLE 12.3.2 Class Intervals and Expected Frequencies for Example 12.3.1

Class Interval	$z(x_i - \bar{x})/s$ At Lower Limit of Interval	Expected Relative Frequency	Expected Frequency	
< 100		.0084	.4	1.8
100.0–124.9	−2.39	.0291	1.4	
125.0–149.9	−1.78	.0815	3.8	
150.0–174.9	−1.18	.1653	7.8	
175.0–199.9	−.57	.2277	10.7	
200.0–224.9	.03	.2269	10.7	
225.0–249.9	.64	.1536	7.2	
250.0–274.9	1.24	.0753	3.5	
275.0–299.9	1.85	.0251	1.2	1.5
300.0 and greater	2.45	.0071	.3	

Comparing Observed and Expected Frequencies

We are now interested in examining the magnitudes of the discrepancies between the observed frequencies and the expected frequencies, since we note that the two sets of frequencies do not agree. We know that even if our sample were drawn from a normal distribution of values, sampling variability alone would make it highly unlikely that the observed and expected frequencies would agree perfectly. We wonder, then, if the discrepancies between the observed and expected frequencies are small enough that we feel it reasonable that they could have occurred by chance alone, when the null hypothesis is true. If they are of this magnitude, we will be unwilling to reject the null hypothesis that the sample came from a normally distributed population.

If the discrepancies are so large that it does not seem reasonable that they could have occurred by chance alone when the null hypothesis is true, we will want to reject the null hypothesis. The criterion against which we judge whether the discrepancies are "large" or "small" is provided by the chi-square distribution.

The observed and expected frequencies along with each value of $(O_i - E_i)^2/E_i$ are shown in Table 12.3.3. The first entry in the last column, for example, is computed from $(1 - 1.8)^2/1.8 = .356$. The other values of $(O_i - E_i)^2/E_i$ are computed in a similar manner.

From Table 12.3.3 we see that $X^2 = \sum[(O_i - E_i)^2/E_i] = 10.566$. The appropriate degrees of freedom are 8 (the number of groups or class intervals) -3 (for the three restrictions: making $\sum E_i = \sum O_i$, and estimating μ and σ from the sample data) $= 5$.

8. **Statistical decision.** When we compare $X^2 = 10.566$ with values of χ^2 in Appendix Table F, we see that it is less than $\chi^2_{.95} = 11.070$, so that, at the .05 level of significance, we cannot reject the null hypothesis that the sample came from a normally distributed population.

TABLE 12.3.3 Observed and Expected Frequencies and $(O_i - E_i)^2/E_i$ for Example 12.3.1

Class Interval	Observed Frequency (O_i)	Expected Frequency (E_i)	$(O_i - E_i)^2/E_i$
< 100	0	.4 } 1.8	.356
100.0–124.9	1	1.4 }	
125.0–149.9	3	3.8	.168
150.0–174.9	8	7.8	.005
175.0–199.9	18	10.7	4.980
200.0–224.9	6	10.7	2.064
225.0–249.9	4	7.2	1.422
250.0–274.9	4	3.5	.071
275.0–299.9	3	1.2 } 1.5	1.500
300.0 and greater	0	.3 }	
Total	47	47	10.566

9. **Conclusion.** We conclude that in the sampled population, cholesterol levels may follow a normal distribution.
10. ***p* value.** Since $11.070 > 10.566 > 9.236$, $.05 < p < .10$. In other words, the probability of obtaining a value of X^2 as large as 10.566, when the null hypothesis is true, is between .05 and .10. Thus we conclude that such an event is not sufficiently rare to reject the null hypothesis that the data come from a normal distribution. ■

Sometimes the parameters are specified in the null hypothesis. It should be noted that had the mean and variance of the population been specified as part of the null hypothesis in Example 12.3.1, we would not have had to estimate them from the sample and our degrees of freedom would have been $8 - 1 = 7$.

Alternatives Although one frequently encounters in the literature the use of chi-square to test for normality, it is not the most appropriate test to use when the hypothesized distribution is continuous. The Kolmogorov–Smirnov test, described in Chapter 13, was especially designed for goodness-of-fit tests involving continuous distributions.

EXAMPLE 12.3.2 *The Binomial Distribution*

In a study designed to determine patient acceptance of a new pain reliever, 100 physicians each selected a sample of 25 patients to participate in the study. Each patient, after trying the new pain reliever for a specified period of time, was asked whether it was preferable to the pain reliever used regularly in the past.

The results of the study are shown in Table 12.3.4.

TABLE 12.3.4 Results of Study Described in Example 12.3.2

Number of Patients Out of 25 Preferring New Pain Reliever	Number of Doctors Reporting this Number	Total Number of Patients Preferring New Pain Reliever by Doctor
0	5	0
1	6	6
2	8	16
3	10	30
4	10	40
5	15	75
6	17	102
7	10	70
8	10	80
9	9	81
10 or more	0	0
Total	100	500

8.1 INTRODUCTION

In the preceding chapters the basic concepts of statistics have been examined, and they provide a foundation for this and the next several chapters. In this chapter and the three that follow, we provide an overview of two of the most commonly employed analytical tools used by applied statisticians, analysis of variance and linear regression. The conceptual foundations of these analytical tools are statistical models that provide useful representations of the relationships among several variables simultaneously.

Linear Models A statistical model is a mathematical representation of the relationships among variables. More specifically for the purposes of this book, a statistical model is most often used to describe how random variables are related to one another in a context in which the value of one *outcome variable,* often referred to with the letter "*y*," can be modeled as a function of one or more *explanatory variables,* often referred to with the letter "*x*." In this way, we are interested in determining how much variability in outcomes can be explained by random variables that were measured or controlled as part of an experiment. The linear model can be expanded easily to the more generalized form, in which we include multiple outcome variables simultaneously. These models are referred to as General Linear Models, and can be found in more advanced statistics books.

DEFINITION

An *outcome variable* is represented by the set of measured values that result from an experiment or some other statistical process. An *explanatory variable,* on the other hand, is a variable that is useful for predicting the value of the outcome variable.

A linear model is any model that is linear in the parameters that define the model. We can represent such models generically in the form:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj} + \epsilon_j \tag{8.1.1}$$

In this equation, β_j represents the coefficients in the model and ϵ_j represents random error. Therefore, any model that can be represented in this form, where the coefficients are constants and the algebraic order of the model is one, is considered a linear model. Though at first glance this equation may seem daunting, it actually is generally easy to find values for the parameters using basic algebra or calculus, as we shall see as the chapter progresses.

We will see many representations of linear models in this and other forms in the next several chapters. In particular, we will focus on the use of linear models for analyzing data using the analysis of variance for testing differences among means, regression for making predictions, and correlation for understanding associations among variables. In the context of analysis of variance, the predictor variables are *classification variables* used to define factors of interest (e.g., differentiating between a control group and a treatment group), and in the context of correlation and linear regression the predictor variables are most often continuous variables, or at least variables at a higher level than nominal classes. Though the underlying purposes of these tasks may seem quite different, studying these techniques and

the structure of the models used to represent them will prove to be valuable for understanding some of the most commonly used inferential statistics.

Analysis of Variance This chapter is concerned with *analysis of variance*, which may be defined as *a technique whereby the total variation present in a set of data is partitioned into two or more components. Associated with each of these components is a specific source of variation, so that in the analysis it is possible to ascertain the magnitude of the contributions of each of these sources to the total variation.*

The development of analysis of variance (ANOVA) is due mainly to the work of R. A. Fisher (1), whose contributions to statistics, spanning the years 1912 to 1962, have had a tremendous influence on modern statistical thought (2,3).

Applications Analysis of variance finds its widest application in the analysis of data derived from experiments. The principles of the design of experiments are well covered in many books, including those by Hinkelmann and Kempthorne (4), Montgomery (5), and Myers and Well (6). We do not study this topic in detail, since to do it justice would require a minimum of an additional chapter. Some of the important concepts in experimental design, however, will become apparent as we discuss analysis of variance.

Analysis of variance is used for two different purposes: (1) to estimate and test hypotheses about population variances, and (2) to estimate and test hypotheses about population means. We are concerned here with the latter use. However, as we will see, our conclusions regarding the means will depend on the magnitudes of the observed variances.

The concepts and techniques that we cover under the heading of analysis of variance are extensions of the concepts and techniques covered in Chapter 7. In Chapter 7 we learned to test the null hypothesis that two means are equal. In this chapter we learn to test the null hypothesis that three or more means are equal. Whereas, for example, what we learned in Chapter 7 enables us to determine if we can conclude that two treatments differ in effectiveness, what we learn in this chapter enables us to determine if we can conclude that three or more treatments differ in effectiveness. The following example illustrates some basic ideas involved in the application of analysis of variance. These will be extended and elaborated on later in this chapter.

EXAMPLE 8.1.1

Suppose we wish to know if three drugs differ in their effectiveness in lowering serum cholesterol in human subjects. Some subjects receive drug A, some drug B, and some drug C. After a specified period of time, measurements are taken to determine the extent to which serum cholesterol was reduced in each subject. We find that the amount by which serum cholesterol was lowered is not the same in all subjects. In other words, there is *variability among* the measurements. Why, we ask ourselves, are the measurements not all the same? Presumably, one reason they are not the same is that the subjects received different drugs. We now look at the measurements of those subjects who received drug A. We find that the amount by which serum cholesterol was lowered is not the same among these subjects. We find this to be the case when we look at the measurements for subjects

who received drug B and those subjects who received drug C. We see that there is *variability* among the measurements *within* the treatment groups. Why, we ask ourselves again, are these measurements not the same? Among the reasons that come to mind are differences in the genetic makeup of the subjects and differences in their diets. Through an analysis of the *variability* that we have observed, we will be able to reach a conclusion regarding the equivalence of the effectiveness of the three drugs. To do this we employ the techniques and concepts of analysis of variance. ■

Variables In our example we allude to three kinds of variables. We find these variables to be present in all situations in which the use of analysis of variance is appropriate. First, we have the *treatment variable*, which in our example was "drug." We had three "values" of this variable, drug A, drug B, and drug C. The second kind of variable we refer to is the *response variable*. In the example it is change in serum cholesterol. The response variable is the variable that we expect to exhibit different values when different "values" of the treatment variable are employed. Finally, we have the other variables that we mention—genetic composition and diet. These are called *extraneous variables*. These variables may have an effect on the response variable, but they are not the focus of our attention in the experiment. The treatment variable is the variable of primary concern, and the question to be answered is: Do the different "values" of the treatment variable result in differences, on the average, in the response variable?

Assumptions Underlying the valid use of analysis of variance as a tool of statistical inference is a set of fundamental assumptions. Although an experimenter must not expect to find all the assumptions met to perfection, it is important that the user of analysis of variance techniques be aware of the underlying assumptions and be able to recognize when they are substantially unsatisfied. Because experiments in which all the assumptions are perfectly met are rare, analysis of variance results should be considered as approximate rather than exact. These assumptions are pointed out at appropriate points in the following sections.

We discuss analysis of variance as it is used to analyze the results of two different experimental designs, the completely randomized and the randomized complete block designs. In addition to these, the concept of a factorial experiment is given through its use in a completely randomized design. These do not exhaust the possibilities. A discussion of additional designs may be found in the references (4–6).

The ANOVA Procedure In our presentation of the analysis of variance for the different designs, we follow the ten-step procedure presented in Chapter 7. The following is a restatement of the steps of the procedure, including some new concepts necessary for its adaptation to analysis of variance.

1. **Description of data.** In addition to describing the data in the usual way, we display the sample data in tabular form.
2. **Assumptions.** Along with the assumptions underlying the analysis, we present the model for each design we discuss. The model consists of a symbolic representation of a typical value from the data being analyzed.
3. **Hypotheses.**

4. **Test statistic.**
5. **Distribution of test statistic.**
6. **Decision rule.**
7. **Calculation of test statistic.** The results of the arithmetic calculations will be summarized in a table called the analysis of variance (ANOVA) table. The entries in the table make it easy to evaluate the results of the analysis.
8. **Statistical decision.**
9. **Conclusion.**
10. **Determination of p value.**

We discuss these steps in greater detail in Section 8.2.

The Use of Computers The calculations required by analysis of variance are lengthier and more complicated than those we have encountered in preceding chapters. For this reason the computer assumes an important role in analysis of variance. All the exercises appearing in this chapter are suitable for computer analysis and may be solved with the statistical packages mentioned in Chapter 1. The output of the statistical packages may vary slightly from that presented in this chapter, but this should pose no major problem to those who use a computer to analyze the data of the exercises. The basic concepts of analysis of variance that we present here should provide the necessary background for understanding the description of the programs and their output in any of the statistical packages.

8.2 THE COMPLETELY RANDOMIZED DESIGN

We saw in Chapter 7 how it is possible to test the null hypothesis of no difference between two population means. It is not unusual for the investigator to be interested in testing the null hypothesis of no difference among several population means. The student first encountering this problem might be inclined to suggest that all possible pairs of sample means be tested separately by means of the Student t test. Suppose there are five populations involved. The number of possible pairs of sample means is ${}_5C_2 = 10$. As the amount of work involved in carrying out this many t tests is substantial, it would be worthwhile if a more efficient alternative for analysis were available. A more important consequence of performing all possible t tests, however, is that it is very likely to lead to a false conclusion.

Suppose we draw five samples from populations having equal means. As we have seen, there would be 10 tests if we were to do each of the possible tests separately. If we select a significance level of $\alpha = .05$ for each test, the probability of failing to reject a hypothesis of no difference in each case would be .95. By the multiplication rule of probability, if the tests were independent of one another, the probability of failing to reject a hypothesis of no difference in all 10 cases would be $(.95)^{10} = .5987$. The probability of rejecting at least one hypothesis of no difference, then, would be $1 - .5987 = .4013$. Since we know that the null hypothesis is true in every case in this illustrative example, rejecting the null hypothesis constitutes the committing of a type I error. In the long run, then, in

testing all possible pairs of means from five samples, we would commit a type I error 40 percent of the time. The problem becomes even more complicated in practice, since three or more t tests based on the same data would not be independent of one another.

It becomes clear, then, that some other method for testing for a significant difference among several means is needed. Analysis of variance provides such a method.

One-Way ANOVA The simplest type of analysis of variance is that known as *one-way analysis of variance*, in which only one source of variation, or *factor*, is investigated. It is an extension to three or more samples of the t test procedure (discussed in Chapter 7) for use with two independent samples. Stated another way, we can say that the t test for use with two independent samples is a special case of one-way analysis of variance.

In a typical situation we want to use one-way analysis of variance to test the null hypothesis that three or more treatments are equally effective. The necessary experiment is designed in such a way that the treatments of interest are assigned completely at random to the subjects or objects on which the measurements to determine treatment effectiveness are to be made. For this reason the design is called the *completely randomized experimental design*.

We may randomly allocate subjects to treatments as follows. Suppose we have 16 subjects available to participate in an experiment in which we wish to compare four drugs. We number the subjects from 01 through 16. We then go to a table of random numbers and select 16 consecutive, unduplicated numbers between 01 and 16. To illustrate, let us use Appendix Table A and a random starting point that, say, is at the intersection of Row 4 and Columns 11 and 12. The two-digit number at this intersection is 98. The succeeding (moving downward) 16 consecutive two-digit numbers between 01 and 16 are 16, 09, 06, 15, 14, 11, 02, 04, 10, 07, 05, 13, 03, 12, 01, and 08. We allocate subjects 16, 09, 06, and 15 to drug A; subjects 14, 11, 02, and 04 to drug B; subjects 10, 07, 05, and 13 to drug C; and subjects 03, 12, 01, and 08 to drug D. We emphasize that the number of subjects in each treatment group does not have to be the same. Figure 8.2.1 illustrates the scheme of random allocation.

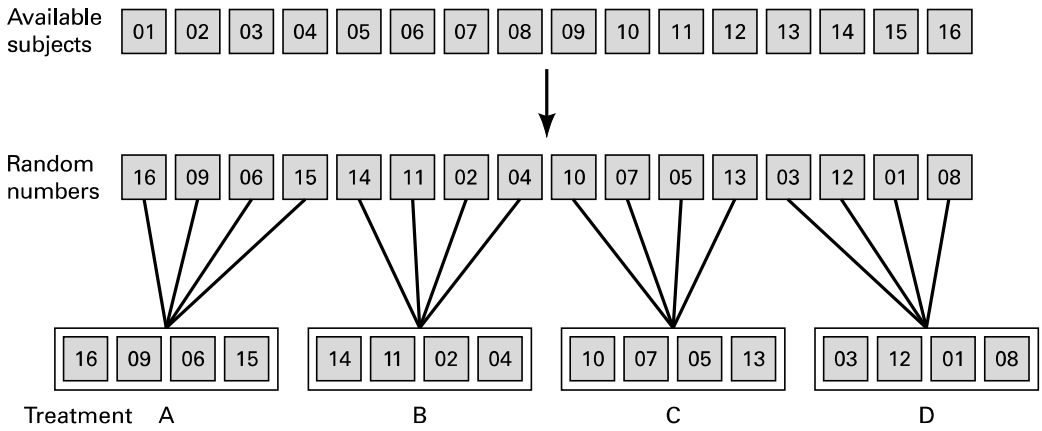

FIGURE 8.2.1 Allocation of subjects to treatments, completely randomized design.

TABLE 8.2.1 Table of Sample Values for the Completely Randomized Design

	Treatment					
	1	2	3	...	*k*	
	x_{11}	x_{12}	x_{13}	...	x_{1k}	
	x_{21}	x_{22}	x_{23}	...	x_{2k}	
	x_{31}	x_{32}	x_{33}	...	x_{3k}	
	⋮	⋮	⋮	⋮	⋮	
	$x_{n_1 1}$	$x_{n_2 2}$	$x_{n_3 3}$	...	$x_{n_k k}$	
Total	$T_{\cdot 1}$	$T_{\cdot 2}$	$T_{\cdot 3}$	...	$T_{\cdot k}$	$T_{\cdot\cdot}$
Mean	$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	$\bar{x}_{\cdot 3}$	...	$\bar{x}_{\cdot k}$	$\bar{x}_{\cdot\cdot}$

Hypothesis Testing Steps Once we decide that the completely randomized design is the appropriate design, we may proceed with the hypothesis testing steps. We discuss these in detail first, and follow with an example.

1. **Description of data.** The measurements (or observations) resulting from a completely randomized experimental design, along with the means and totals that can be computed from them, may be displayed for convenience as in Table 8.2.1. The symbols used in Table 8.2.1 are defined as follows:

$$
\begin{aligned}
x_{ij} &= \text{the } i\text{th observation resulting from the } j\text{th treatment} \\
&\quad (\text{there are a total of } k \text{ treatments}) \\
i &= 1, 2, \ldots, n_j, \quad j = 1, 2, \ldots, k \\
T_{\cdot j} &= \sum_{i=1}^{n_j} x_{ij} = \text{total of the } j\text{th treatment} \\
\bar{x}_{\cdot j} &= \frac{T_{\cdot j}}{n_j} = \text{mean of the } j\text{th treatment} \\
T_{\cdot\cdot} &= \sum_{j=1}^{k} T_{\cdot j} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij} = \text{total of all observations} \\
\bar{x}_{\cdot\cdot} &= \frac{T_{\cdot\cdot}}{N}, \quad N = \sum_{j=1}^{k} n_j
\end{aligned}
$$

2. **Assumptions.** Before stating the assumptions, let us specify the model for the experiment described here.

The Model As already noted, a model is a symbolic representation of a typical value of a data set. To write down the model for the completely randomized experimental design, let us begin by identifying a typical value from the set of data represented by the sample displayed in Table 8.2.1. We use the symbol x_{ij} to represent this typical value.

The one-way analysis of variance model may be written as follows:

$$x_{ij} = \mu + \tau_j + \epsilon_{ij}; \qquad i = 1, 2, \ldots, n_j, \qquad j = 1, 2, \ldots, k \tag{8.2.1}$$

The terms in this model are defined as follows:

1. μ represents the mean of all k population means and is called the *grand mean.*
2. τ_j represents the difference between the mean of the jth population and the grand mean and is called the *treatment effect.*
3. ϵ_{ij} represents the amount by which an individual measurement differs from the mean of the population to which it belongs and is called the *error term.*

Components of the Model By looking at our model we can see that a typical observation from the total set of data under study is composed of (1) the grand mean, (2) a treatment effect, and (3) an error term representing the deviation of the observation from its group mean.

In most situations we are interested only in the k treatments represented in our experiment. Any inferences that we make apply only to these treatments. We do not wish to extend our inference to any larger collection of treatments. When we place such a restriction on our inference goals, we refer to our model as the *fixed-effects model*, or *model 1*. The discussion in this book is limited to this model.

Assumptions of the Model The assumptions for the fixed-effects model are as follows:

(a) The k sets of observed data constitute k independent random samples from the respective populations.

(b) Each of the populations from which the samples come is normally distributed with mean μ_j and variance σ_j^2.

(c) Each of the populations has the same variance. That is, $\sigma_1^2 = \sigma_2^2 = \ldots \sigma_k^2 = \sigma^2$ the common variance.

(d) The τ_j are unknown constants and $\sum \tau_j = 0$ since the sum of all deviations of the μ_j from their mean, μ, is zero.

(e) The ϵ_{ij} have a mean of 0, since the mean of x_{ij} is μ_j.

(f) The ϵ_{ij} have a variance equal to the variance of the x_{ij}, since the ϵ_{ij} and x_{ij} differ only by a constant; that is, the error variance is equal to σ^2, the common variance specified in assumption c.

(g) The ϵ_{ij} are normally (and independently) distributed.

3. **Hypotheses.** We test the null hypothesis that all population or treatment means are equal against the alternative that the members of at least one pair are not equal. We may state the hypotheses formally as follows:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_A : \text{not all } \mu_j \text{ are equal}$$

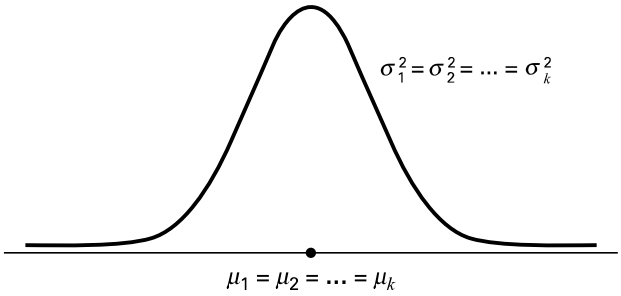

FIGURE 8.2.2 Picture of the populations represented in a completely randomized design when H_0 is true and the assumptions are met.

If the population means are equal, each treatment effect is equal to zero, so that, alternatively, the hypotheses may be stated as

$$H_0 : \tau_j = 0, \qquad j = 1, 2, \ldots, k$$
$$H_A : \text{not all } \tau_j = 0$$

If H_0 is true and the assumptions of equal variances and normally distributed populations are met, a picture of the populations will look like Figure 8.2.2. When H_0 is true the population means are all equal, and the populations are centered at the same point (the common mean) on the horizontal axis. If the populations are all normally distributed with equal variances the distributions will be identical, so that in drawing their pictures each is superimposed on each of the others, and a single picture sufficiently represents them all.

When H_0 is false it may be false because one of the population means is different from the others, which are all equal. Or, perhaps, all the population means are different. These are only two of the possibilities when H_0 is false. There are many other possible combinations of equal and unequal means. Figure 8.2.3 shows a picture of the populations when the assumptions are met, but H_0 is false because no two population means are equal.

4. **Test statistic.** The test statistic for one-way analysis of variance is a computed variance ratio, which we designate by V.R. as we did in Chapter 7. The two

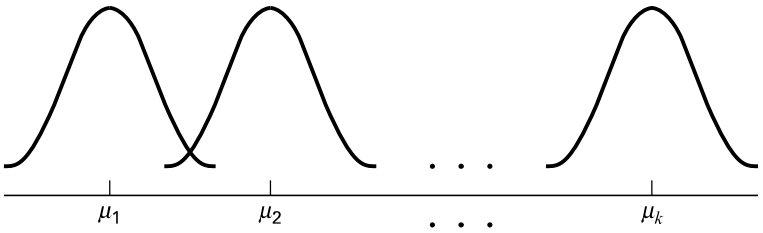

FIGURE 8.2.3 Picture of the populations represented in a completely randomized design when the assumptions of equal variances and normally distributed populations are met, but H_0 is false because none of the population means are equal.

variances from which V.R. is calculated are themselves computed from the sample data. The methods by which they are calculated will be given in the discussion that follows.

5. **Distribution of test statistic.** As discussed in Section 7.8, V.R. is distributed as the F distribution when H_0 is true and the assumptions are met.
6. **Decision rule.** In general, the decision rule is: reject the null hypothesis if the computed value of V.R. is equal to or greater than the critical value of F for the chosen α level.
7. **Calculation of test statistic.** We have defined analysis of variance as a process whereby the total variation present in a set of data is partitioned into components that are attributable to different sources. The term *variation* used in this context refers to the *sum of squared deviations of observations from their mean*, or *sum of squares* for short.

The initial computations performed in one-way ANOVA consist of the partitioning of the total variation present in the observed data into its basic components, each of which is attributable to an identifiable source.

Those who use a computer for calculations may wish to skip the following discussion of the computations involved in obtaining the test statistic.

The Total Sum of Squares Before we can do any partitioning, we must first obtain the total sum of squares. The total sum of squares is the sum of the squares of the deviations of individual observations from the mean of all the observations taken together. This *total sum of squares* is defined as

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_{..}\right)^2 \tag{8.2.2}$$

where $\Sigma_{i=1}^{n_j}$ tells us to sum the squared deviations for each treatment group, and $\Sigma_{j=1}^{k}$ tells us to add the k group totals obtained by applying $\Sigma_{i=1}^{n_j}$. The reader will recognize Equation 8.2.2 as the numerator of the variance that may be computed from the complete set of observations taken together.

The Within Groups Sum of Squares Now let us show how to compute the first of the two components of the total sum of squares.

The first step in the computation calls for performing certain calculations *within* each group. These calculations involve computing within each group the sum of the squared deviations of the individual observations from their mean. When these calculations have been performed within each group, we obtain the sum of the individual group results. This component of variation is called the *within groups sum of squares* and may be designated *SSW*. This quantity is sometimes referred to as the *residual* or *error* sum of squares. The expression for these calculations is written as follows:

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_{.j}\right)^2 \tag{8.2.3}$$

The Among Groups Sum of Squares To obtain the second component of the total sum of squares, we compute for each group the squared deviation of the group mean from the grand mean and multiply the result by the size of the group. Finally, we add these results over all groups. This quantity is a measure of the variation among groups and is referred to as the *sum of squares among groups* or *SSA*. The formula for calculating this quantity is as follows:

$$SSA = \sum_{j=1}^{k} n_j \left(\bar{x}_{.j} - \bar{x}_{..}\right)^2 \tag{8.2.4}$$

In summary, then, we have found that the total sum of squares is equal to the sum of the among and the within sum of squares. We express this relationship as follows:

$$SST = SSA + SSW$$

From the sums of squares that we have now learned to compute, it is possible to obtain two estimates of the common population variance, σ^2. It can be shown that when the assumptions are met and the population means are all equal, both the among sum of squares and the within sum of squares, when divided by their respective degrees of freedom, yield independent and unbiased estimates of σ^2.

The First Estimate of σ^2 Within any sample,

$$\frac{\sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_{.j}\right)^2}{n_j - 1}$$

provides an unbiased estimate of the true variance of the population from which the sample came. Under the assumption that the population variances are all equal, we may pool the k estimates to obtain

$$\text{MSW} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_{.j}\right)^2}{\sum_{j=1}^{k} \left(n_j - 1\right)} \tag{8.2.5}$$

This is our first estimate of σ^2 and may be called the *within groups variance*, since it is the within groups sum of squares of Equation 8.2.3 divided by the appropriate degrees of freedom. The student will recognize this as an extension to k samples of the pooling of variances procedure encountered in Chapters 6 and 7 when the variances from two samples were pooled in order to use the t distribution. The quantity in Equation 8.2.5 is customarily referred to as the within groups *mean square* rather than the within groups variance.

The within groups mean square is a valid estimate of σ^2 only if the population variances are equal. It is not necessary, however, for H_0 to be true in order for the within groups mean square to be a valid estimate of σ^2; that is, the within groups mean square estimates σ^2 regardless of whether H_0 is true or false, as long as the population variances are equal.

The Second Estimate of σ^2 The second estimate of σ^2 may be obtained from the familiar formula for the variance of sample means, $\sigma_{\bar{x}}^2 = \sigma^2/n$. If we solve this equation for σ^2, the variance of the population from which the samples were drawn, we have

$$\sigma^2 = n\sigma_{\bar{x}}^2 \tag{8.2.6}$$

An unbiased estimate of $\sigma_{\bar{x}}^2$ computed from sample data is provided by

$$\frac{\sum_{j=1}^{k} \left(\bar{x}_{.j} - \bar{x}_{..}\right)^2}{k-1}$$

If we substitute this quantity into Equation 8.2.6, we obtain the desired estimate of σ^2,

$$\text{MSA} = \frac{n\sum_{j=1}^{k} \left(\bar{x}_{.j} - \bar{x}_{..}\right)^2}{k-1} \tag{8.2.7}$$

The reader will recognize the numerator of Equation 8.2.7 as the among groups sum of squares for the special case when all sample sizes are equal. This sum of squares when divided by the associated degrees of freedom $k-1$ is referred to as the *among groups mean square*.

When the sample sizes are not all equal, an estimate of σ^2 based on the variability among sample means is provided by

$$\text{MSA} = \frac{\sum_{j=1}^{k} n_j\left(\bar{x}_{.j} - \bar{x}_{..}\right)^2}{k-1} \tag{8.2.8}$$

If, indeed, the null hypothesis is true we would expect these two estimates of σ^2 to be fairly close in magnitude. If the null hypothesis is false, that is, if all population means are not equal, we would expect the among groups mean square, which is computed by using the squared deviations of the sample means from the overall mean, to be larger than the within groups mean square.

In order to understand analysis of variance we must realize that the among groups mean square provides a valid estimate of σ^2 when the assumption of equal population

variances is met *and when* H_0 is true. Both conditions, a true null hypothesis and equal population variances, must be met in order for the among groups mean square to be a valid estimate of σ^2.

The Variance Ratio What we need to do now is to compare these two estimates of σ^2, and we do this by computing the following variance ratio, which is the desired test statistic:

$$\text{V.R.} = \frac{\text{among groups mean square}}{\text{within groups mean square}} = \frac{\text{MSA}}{\text{MSW}}$$

If the two estimates are about equal, V.R. will be close to 1. A ratio close to 1 tends to support the hypothesis of equal population means. If, on the other hand, the among groups mean square is considerably larger than the within groups mean square, V.R. will be considerably greater than 1. A value of V.R. sufficiently greater than 1 will cast doubt on the hypothesis of equal population means.

We know that because of the vagaries of sampling, even when the null hypothesis is true, it is unlikely that the among and within groups mean squares will be equal. We must decide, then, how big the observed difference must be before we can conclude that the difference is due to something other than sampling fluctuation. In other words, how large a value of V.R. is required for us to be willing to conclude that the observed difference between our two estimates of σ^2 is not the result of chance alone?

The *F* Test To answer the question just posed, we must consider the sampling distribution of the ratio of two sample variances. In Chapter 6 we learned that the quantity $\left(s_1^2/\sigma_1^2\right)/\left(s_2^2/\sigma_2^2\right)$ follows a distribution known as the F distribution when the sample variances are computed from random and independently drawn samples from normal populations. The F distribution, introduced by R. A. Fisher in the early 1920s, has become one of the most widely used distributions in modern statistics. We have already become acquainted with its use in constructing confidence intervals for, and testing hypotheses about, population variances. In this chapter, we will see that it is the distribution fundamental to analysis of variance. For this reason the ratio that we designate V.R. is frequently referred to as F, and the testing procedure is frequently called the F test. It is of interest to note that the F distribution is the ratio of two Chi-square distributions.

In Chapter 7 we learned that when the population variances are the same, they cancel in the expression $\left(s_1^2/\sigma_1^2\right)/\left(s_2^2/\sigma_2^2\right)$, leaving s_1^2/s_2^2, which is itself distributed as F. The F distribution is really a family of distributions, and the particular F distribution we use in a given situation depends on the number of degrees of freedom associated with the sample variance in the numerator (*numerator degrees of freedom*) and the number of degrees of freedom associated with the sample variance in the denominator (*denominator degrees of freedom*).

Once the appropriate F distribution has been determined, the size of the observed V.R. that will cause rejection of the hypothesis of equal population variances depends on the significance level chosen. The significance level chosen determines the critical value of F, the value that separates the nonrejection region from the rejection region.

TABLE 8.2.2 Analysis of Variance Table for the Completely Randomized Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio
Among samples	$SSA = \sum_{j=1}^{k} n_j(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2$	$k-1$	$MSA = SSA/(k-1)$	V.R. $= \dfrac{MSA}{MSW}$
Within samples	$SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{\cdot j})^2$	$N-k$	$MSW = SSW/(N-k)$	
Total	$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{\cdot\cdot})^2$	$N-1$		

As we have seen, we compute V.R. in situations of this type by placing the among groups mean square in the numerator and the within groups mean square in the denominator, so that the numerator degrees of freedom is equal to $(k-1)$, the number of groups minus 1, and the denominator degrees of freedom value is equal to

$$\sum_{j=1}^{k}(n_j - 1) = \left(\sum_{j=1}^{k} n_j\right) - k = N - k$$

The ANOVA Table The calculations that we perform may be summarized and displayed in a table such as Table 8.2.2 , which is called the ANOVA table.

8. **Statistical decision.** To reach a decision we must compare our computed V.R. with the critical value of F, which we obtain by entering Appendix Table G with $k-1$ numerator degrees of freedom and $N-k$ denominator degrees of freedom.

If the computed V.R. is equal to or greater than the critical value of F, we reject the null hypothesis. If the computed value of V.R. is smaller than the critical value of F, we do not reject the null hypothesis.

Explaining a Rejected Null Hypothesis There are two possible explanations for a rejected null hypothesis. If the null hypothesis is true, that is, if the two sample variances are estimates of a common variance, we know that the probability of getting a value of V.R. as large as or larger than the critical F is equal to our chosen level of significance. When we reject H_0 we may, if we wish, conclude that the null hypothesis is true and assume that because of chance we got a set of data that gave rise to a rare event. On the other hand, we may prefer to take the position that our large computed V.R. value does not represent a rare event brought about by chance but, instead, reflects the fact that something other than chance is operative. We then conclude that we have a false null hypothesis.

It is this latter explanation that we usually give for computed values of V.R. that exceed the critical value of F. In other words, if the computed value of V.R. is greater than the critical value of F, we reject the null hypothesis.

It will be recalled that the original hypothesis we set out to test was

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

Does rejection of the hypothesis about variances imply a rejection of the hypothesis of equal population means? The answer is yes. A large value of V.R. resulted from the fact that the among groups mean square was considerably larger than the within groups mean square. Since the among groups mean square is based on the dispersion of the sample means about their mean (called the grand mean), this quantity will be large when there is a large discrepancy among the sizes of the sample means. Because of this, then, a significant value of V.R. tells us to reject the null hypothesis that all population means are equal.

9. **Conclusion.** When we reject H_0, we conclude that not all population means are equal. When we fail to reject H_0, we conclude that the population means are not significantly different from each other.
10. **Determination of *p* value.**

EXAMPLE 8.2.1

Game meats, including those from white-tailed deer and eastern gray squirrels, are used as food by families, hunters, and other individuals for health, cultural, or personal reasons. A study by David Holben (A-1) assessed the selenium content of meat from free-roaming white-tailed deer (venison) and gray squirrel (squirrel) obtained from a low selenium region of the United States. These selenium content values were also compared to those of beef produced within and outside the same region. We want to know if the selenium levels are different among the four meat groups.

Solution:

1. **Description of data.** Selenium content of raw venison (VEN), squirrel meat (SQU), region-raised beef (RRB), and nonregion-raised beef (NRB), in μg/100 g of dry weight, are shown in Table 8.2.3. A graph of the data in the form of a *dotplot* is shown in Figure 8.2.4. Such a graph highlights the main features of the data and brings into clear focus differences in selenium levels among the different meats.

TABLE 8.2.3 Selenium Content, in μg/100 g, of Four Different Meat Types

Meat Type						
VEN		SQU		RRB		NRB
26.72	14.86	37.42	37.57	11.23	15.82	44. 33
28.58	16.47	56.46	25.71	29.63	27.74	76.86
29.71	25.19	51.91	23.97	20.42	22.35	4.45
26.95	37.45	62.73	13.82	10.12	34.78	55.01
10.97	45.08	4.55	42.21	39.91	35.09	58.21
21.97	25.22	39.17	35.88	32.66	32.60	74.72

(*Continued*)

Meat Type						
VEN		SQU		RRB		NRB
14.35	22.11	38.44	10.54	38.38	37.03	11.84
32.21	33.01	40.92	27.97	36.21	27.00	139.09
19.19	31.20	58.93	41.89	16.39	44.20	69.01
30.92	26.50	61.88	23.94	27.44	13.09	94.61
10.42	32.77	49.54	49.81	17.29	33.03	48.35
35.49	8.70	64.35	30.71	56.20	9.69	37.65
36.84	25.90	82.49	50.00	28.94	32.45	66.36
25.03	29.80	38.54	87.50	20.11	37.38	72.48
33.59	37.63	39.53	68.99	25.35	34.91	87.09
33.74	21.69			21.77	27.99	26.34
18.02	21.49			31.62	22.36	71.24
22.27	18.11			32.63	22.68	90.38
26.10	31.50			30.31	26.52	50.86
20.89	27.36			46.16	46.01	
29.44	21.33			56.61	38.04	
				24.47	30.88	
				29.39	30.04	
				40.71	25.91	
				18.52	18.54	
				27.80	25.51	
				19.49		

Source: Data provided courtesy of David H. Holben, Ph.D.

2. **Assumptions.** We assume that the four sets of data constitute independent simple random samples from the four indicated populations. We assume that the four populations of measurements are normally distributed with equal variances.

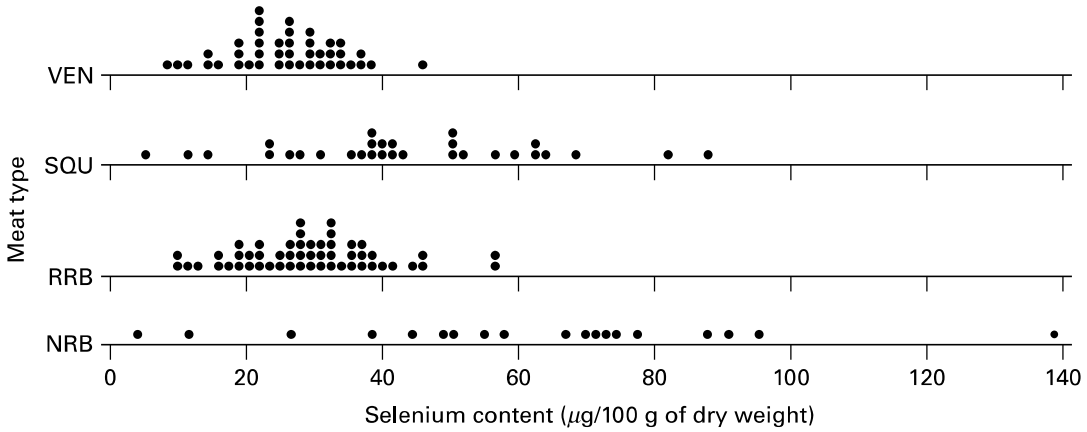

FIGURE 8.2.4 Selenium content of four meat types. VEN = venison, SQU = squirrel, RRB = region-raised beef, and NRB = nonregion-raised beef.

TABLE 8.2.4 ANOVA Table for Example 8.2.1

Source	SS	df	MS	F
Among samples	21261.82886	3	7087.27629	27.00
Within samples	36747.22674	140	262.48019	
Total	58009.05560	143		

3. **Hypotheses.** $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ (On average the four meats have the same selenium content.)

   H_A: Not all μ's are equal (At least one meat yields an average selenium content different from the average selenium content of at least one other meat.)

4. **Test statistic.** The test statistic is V.R. $= MSA/MSW$.
5. **Distribution of test statistic.** If H_0 is true and the assumptions are met, the V.R. follows the F distribution with $4 - 1 = 3$ numerator degrees of freedom and $144 - 4 = 140$ denominator degrees of freedom.
6. **Decision rule.** Suppose we let $\alpha = .01$. The critical value of F from Appendix Table G is < 3.95. The decision rule, then, is reject H_0 if the computed V.R. statistic is equal to or greater than 3.95.
7. **Calculation of test statistic.** By Equation 8.2.2 we compute

   $$SST = 58009.05560$$

   By Equation 8.2.4 we compute

   $$SSA = 21261.82886$$

   $$SSW = 58009.05560 - 21261.82886 = 36747.22674$$

   The results of our calculations are displayed in Table 8.2.4.
8. **Statistical decision.** Since our computed F of 27.00 is greater than 3.95 we reject H_0.
9. **Conclusion.** Since we reject H_0, we conclude that the alternative hypothesis is true. That is, we conclude that the four meat types do not all have the same average selenium content.
10. ***p* value.** Since $27.00 > 3.95$, $p < .01$ for this test. ■

A Word of Caution The completely randomized design is simple and, therefore, widely used. It should be used, however, only when the units receiving the treatments are homogeneous. If the experimental units are not homogeneous, the researcher should consider an alternative design such as one of those to be discussed later in this chapter.

In our illustrative example the treatments are treatments in the usual sense of the word. This is not always the case, however, as the term "treatment" as used in experimental design is quite general. We might, for example, wish to study the response to the same

CHAPTER 7

HYPOTHESIS TESTING

CHAPTER OVERVIEW

This chapter covers hypothesis testing, the second of two general areas of statistical inference. Hypothesis testing is a topic with which you as a student are likely to have some familiarity. Interval estimation, discussed in the preceding chapter, and hypothesis testing are based on similar concepts. In fact, confidence intervals may be used to arrive at the same conclusions that are reached through the use of hypothesis tests. This chapter provides a format, followed throughout the remainder of this book, for conducting a hypothesis test.

TOPICS

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand how to correctly state a null and alternative hypothesis and carry out a structured hypothesis test.
2. understand the concepts of type I error, type II error, and the power of a test.
3. be able to calculate and interpret z, t, F, and chi-square test statistics for making statistical inferences.
4. understand how to calculate and interpret p values.

7.1 INTRODUCTION

One type of statistical inference, estimation, is discussed in the preceding chapter. The other type, hypothesis testing, is the subject of this chapter. As is true with estimation, the *purpose of hypothesis testing is to aid the clinician, researcher, or administrator in reaching a conclusion concerning a population by examining a sample from that population*. Estimation and hypothesis testing are not as different as they are made to appear by the fact that most textbooks devote a separate chapter to each. As we will explain later, one may use confidence intervals to arrive at the same conclusions that are reached by using the hypothesis testing procedures discussed in this chapter.

Basic Concepts In this section some of the basic concepts essential to an understanding of hypothesis testing are presented. The specific details of particular tests will be given in succeeding sections.

DEFINITION

A *hypothesis* may be defined simply as *a statement about one or more populations.*

The hypothesis is frequently concerned with the parameters of the populations about which the statement is made. A hospital administrator may hypothesize that the average length of stay of patients admitted to the hospital is 5 days; a public health nurse may hypothesize that a particular educational program will result in improved communication between nurse and patient; a physician may hypothesize that a certain drug will be effective in 90 percent of the cases for which it is used. By means of hypothesis testing one determines whether or not such statements are compatible with the available data.

Types of Hypotheses Researchers are concerned with two types of hypotheses—research hypotheses and statistical hypotheses.

DEFINITION

The *research hypothesis* is the conjecture or supposition that motivates the research.

It may be the result of years of observation on the part of the researcher. A public health nurse, for example, may have noted that certain clients responded more readily to a particular type of health education program. A physician may recall numerous instances in which certain combinations of therapeutic measures were more effective than any one of them alone. Research projects often result from the desire of such health practitioners to determine whether or not their theories or suspicions can be supported when subjected to the rigors of scientific investigation.

Research hypotheses lead directly to statistical hypotheses.

DEFINITION

***Statistical hypotheses* are hypotheses that are stated in such a way that they may be evaluated by appropriate statistical techniques.**

In this book the hypotheses that we will focus on are statistical hypotheses. We will assume that the research hypotheses for the examples and exercises have already been considered.

Hypothesis Testing Steps For convenience, hypothesis testing will be presented as a ten-step procedure. There is nothing magical or sacred about this particular format. It merely breaks the process down into a logical sequence of actions and decisions.

1. **Data.** The nature of the data that form the basis of the testing procedures must be understood, since this determines the particular test to be employed. Whether the data consist of counts or measurements, for example, must be determined.
2. **Assumptions.** As we learned in the chapter on estimation, different assumptions lead to modifications of confidence intervals. The same is true in hypothesis testing: A general procedure is modified depending on the assumptions. In fact, the same assumptions that are of importance in estimation are important in hypothesis testing. We have seen that these include assumptions about the normality of the population distribution, equality of variances, and independence of samples.
3. **Hypotheses.** There are two statistical hypotheses involved in hypothesis testing, and these should be stated explicitly. The *null hypothesis* is the *hypothesis to be tested*. It is designated by the symbol H_0. The null hypothesis is sometimes referred to as a *hypothesis of no difference,* since it is a statement of agreement with (or no difference from) conditions presumed to be true in the population of interest. In general, the null hypothesis is set up for the express purpose of being discredited. Consequently, the complement of the conclusion that the researcher is seeking to reach becomes the statement of the null hypothesis. In the testing process the null hypothesis either is rejected or is not rejected. If the null hypothesis is not rejected, we will say that the data on which the test is based do not provide sufficient evidence to cause rejection. If the testing procedure leads to rejection, we will say that the data at hand are not compatible with the null hypothesis, but are supportive of some other hypothesis. The *alternative hypothesis* is a statement of what we will believe is true if our sample data cause us to reject the null hypothesis. Usually the alternative hypothesis and the research hypothesis are the same, and in fact the two terms are used interchangeably. We shall designate the alternative hypothesis by the symbol H_A.

Rules for Stating Statistical Hypotheses When hypotheses are of the type considered in this chapter an indication of equality (either $=, \leq$, or $\geq$) must appear in the null hypothesis. Suppose, for example, that we want to answer the

question: Can we conclude that a certain population mean is not 50? The null hypothesis is

$$H_0: \mu = 50$$

and the alternative is

$$H_A: \mu \neq 50$$

Suppose we want to know if we can conclude that the population mean is greater than 50. Our hypotheses are

$$H_0: \mu \leq 50 \qquad H_A: \mu > 50$$

If we want to know if we can conclude that the population mean is less than 50, the hypotheses are

$$H_0: \mu \geq 50 \qquad H_A: \mu < 50$$

In summary, we may state the following rules of thumb for deciding what statement goes in the null hypothesis and what statement goes in the alternative hypothesis:

(a) What you hope or expect to be able to conclude as a result of the test usually should be placed in the alternative hypothesis.

(b) The null hypothesis should contain a statement of equality, either $=, \leq,$ or $\geq$.

(c) The null hypothesis is the hypothesis that is tested.

(d) The null and alternative hypotheses are complementary. That is, the two together exhaust all possibilities regarding the value that the hypothesized parameter can assume.

A Precaution It should be pointed out that neither hypothesis testing nor statistical inference, in general, leads to the proof of a hypothesis; it merely indicates whether the hypothesis is supported or is not supported by the available data. When we fail to reject a null hypothesis, therefore, we do not say that it is true, but that it may be true. When we speak of accepting a null hypothesis, we have this limitation in mind and do not wish to convey the idea that accepting implies proof.

4. **Test statistic.** The test statistic is some statistic that may be computed from the data of the sample. As a rule, there are many possible values that the test statistic may assume, the particular value observed depending on the particular sample drawn. As we will see, the test statistic serves as a decision maker, since the decision to reject or not to reject the null hypothesis depends on the magnitude of the test statistic. An example of a test statistic is the quantity

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{7.1.1}$$

where μ_0 is a hypothesized value of a population mean. This test statistic is related to the statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{7.1.2}$$

with which we are already familiar.

General Formula for Test Statistic The following is a general formula for a test statistic that will be applicable in many of the hypothesis tests discussed in this book:

$$\text{test statistic} = \frac{\text{relevant statistic} - \text{hypothesized parameter}}{\text{standard error of the relevant statistic}}$$

In Equation 7.1.1, $\bar{x}$ is the relevant statistic, μ_0 is the hypothesized parameter, and $\sigma/\sqrt{n}$ is the standard error of $\bar{x}$, the relevant statistic.

5. **Distribution of test statistic.** It has been pointed out that the key to statistical inference is the sampling distribution. We are reminded of this again when it becomes necessary to specify the probability distribution of the test statistic. The distribution of the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

for example, follows the standard normal distribution if the null hypothesis is true and the assumptions are met.

6. **Decision rule.** All possible values that the test statistic can assume are points on the horizontal axis of the graph of the distribution of the test statistic and are divided into two groups; one group constitutes what is known as the *rejection region* and the other group makes up the *nonrejection region*. The values of the test statistic forming the rejection region are those values that are less likely to occur if the null hypothesis is true, while the values making up the acceptance region are more likely to occur if the null hypothesis is true. *The decision rule tells us to reject the null hypothesis if the value of the test statistic that we compute from our sample is one of the values in the rejection region and to not reject the null hypothesis if the computed value of the test statistic is one of the values in the nonrejection region.*

Significance Level The decision as to which values go into the rejection region and which ones go into the nonrejection region is made on the basis of the desired *level of significance,* designated by α. The term *level of significance* reflects the fact that hypothesis tests are sometimes called significance tests, and a computed value of the test statistic that falls in the rejection region is said to be *significant*. The level of significance, α, specifies the area under the curve of the distribution of the test statistic that is above the values on the horizontal axis constituting the rejection region.

DEFINITION

The *level of significance* α is a probability and, in fact, is the *probability of rejecting a true null hypothesis*.

Since to reject a true null hypothesis would constitute an error, it seems only reasonable that we should make the probability of rejecting a true null hypothesis small and, in fact, that is what is done. We select a small value of α in order to make the probability of rejecting a true null hypothesis small. The more frequently encountered values of α are .01, .05, and .10.

Types of Errors The error committed when a true null hypothesis is rejected is called the *type I error*. The *type II error* is the error committed when a false null hypothesis is not rejected. The probability of committing a type II error is designated by β.

Whenever we reject a null hypothesis there is always the concomitant risk of committing a type I error, rejecting a true null hypothesis. Whenever we fail to reject a null hypothesis the risk of failing to reject a false null hypothesis is always present. We make α small, but we generally exercise no control over β, although we know that in most practical situations it is larger than α.

We never know whether we have committed one of these errors when we reject or fail to reject a null hypothesis, since the true state of affairs is unknown. If the testing procedure leads to rejection of the null hypothesis, we can take comfort from the fact that we made α small and, therefore, the probability of committing a type I error was small. If we fail to reject the null hypothesis, we do not know the concurrent risk of committing a type II error, since β is usually unknown but, as has been pointed out, we do know that, in most practical situations, it is larger than α.

Figure 7.1.1 shows for various conditions of a hypothesis test the possible actions that an investigator may take and the conditions under which each of the two types of error will be made. The table shown in this figure is an example of what is generally referred to as a *confusion matrix*.

7. **Calculation of test statistic.** From the data contained in the sample we compute a value of the test statistic and compare it with the rejection and nonrejection regions that have already been specified.
8. **Statistical decision.** The statistical decision consists of rejecting or of not rejecting the null hypothesis. It is rejected if the computed value of the test statistic falls in the

		Condition of Null Hypothesis	
		True	False
Possible Action	Fail to reject H_0	Correct action	Type II error
	Reject H_0	Type I error	Correct action

FIGURE 7.1.1 Conditions under which type I and type II errors may be committed.

rejection region, and it is not rejected if the computed value of the test statistic falls in the nonrejection region.

9. **Conclusion.** If H_0 is rejected, we conclude that H_A is true. If H_0 is not rejected, we conclude that H_0 may be true.
10. ***p* values.** The p value is a number that tells us how unusual our sample results are, given that the null hypothesis is true. A p value indicating that the sample results are not likely to have occurred, if the null hypothesis is true, provides justification for doubting the truth of the null hypothesis.

DEFINITION

A *p value* is the probability that the computed value of a test statistic is at least as extreme as a specified value of the test statistic when the null hypothesis is true. Thus, the *p value* is the smallest value of α for which we can reject a null hypothesis.

We emphasize that when the null hypothesis is not rejected one should not say that the null hypothesis is accepted. We should say that the null hypothesis is "not rejected." We avoid using the word "accept" in this case because we may have committed a type II error. Since, frequently, the probability of committing a type II error can be quite high, we do not wish to commit ourselves to accepting the null hypothesis.

Figure 7.1.2 is a flowchart of the steps that we follow when we perform a hypothesis test.

Purpose of Hypothesis Testing The purpose of hypothesis testing is to assist administrators and clinicians in making decisions. The administrative or clinical decision usually depends on the statistical decision. If the null hypothesis is rejected, the administrative or clinical decision usually reflects this, in that the decision is compatible with the alternative hypothesis. The reverse is usually true if the null hypothesis is not rejected. The administrative or clinical decision, however, may take other forms, such as a decision to gather more data.

We also emphasize that the hypothesis testing procedures highlighted in the remainder of this chapter generally examine the case of normally distributed data or cases where the procedures are appropriate because the central limit theorem applies. In practice, it is not uncommon for samples to be small relative to the size of the population, or to have samples that are highly skewed, and hence the assumption of normality is violated. Methods to handle this situation, that is *distribution-free* or *nonparametric methods,* are examined in detail in Chapter 13. Most computer packages include an analytical procedure (for example, the Shapiro-Wilk or Anderson-Darling test) for testing normality. It is important that such tests are carried out prior to analysis of data. Further, when testing two samples, there is an implicit assumption that the variances are equal. Tests for this assumption are provided in Section 7.8. Finally, it should be noted that hypothesis tests, just like confidence intervals, are relatively

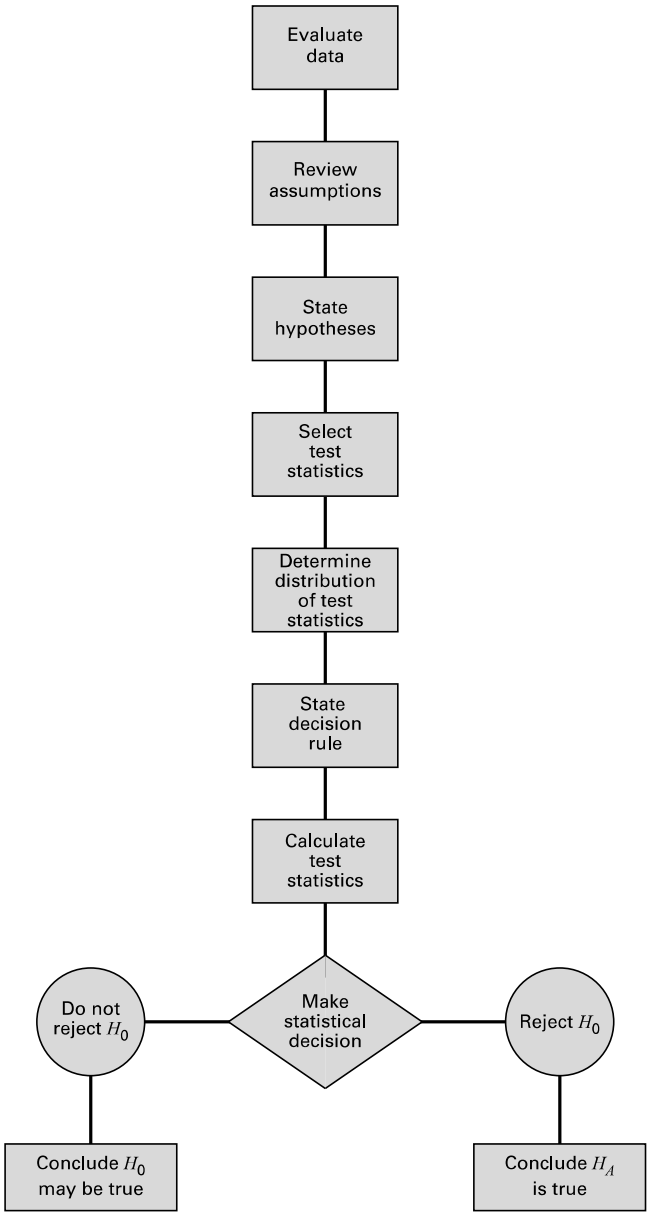

FIGURE 7.1.2 Steps in the hypothesis testing procedure.

sensitive to the size of the samples being tested, and caution should be taken when interpreting results involving very small sample sizes.

We must emphasize at this point, however, that the outcome of the statistical test is only one piece of evidence that influences the administrative or clinical decision. The statistical decision should not be interpreted as definitive but should be considered along with all the other relevant information available to the experimenter.

With these general comments as background, we now discuss specific hypothesis tests.

7.2 HYPOTHESIS TESTING: A SINGLE POPULATION MEAN

In this section we consider the testing of a hypothesis about a population mean under three different conditions: (1) when sampling is from a normally distributed population of values with known variance; (2) when sampling is from a normally distributed population with unknown variance, and (3) when sampling is from a population that is not normally distributed. Although the theory for conditions 1 and 2 depends on normally distributed populations, it is common practice to make use of the theory when relevant populations are only approximately normally distributed. This is satisfactory as long as the departure from normality is not drastic. When sampling is from a normally distributed population and the population variance is known, the test statistic for testing H_0: $\mu = \mu_0$ is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{7.2.1}$$

which, when H_0 is true, is distributed as the standard normal. Examples 7.2.1 and 7.2.2 illustrate hypothesis testing under these conditions.

Sampling from Normally Distributed Populations: Population Variances Known As we did in Chapter 6, we again emphasize that situations in which the variable of interest is normally distributed with a known variance are rare. The following example, however, will serve to illustrate the procedure.

EXAMPLE 7.2.1

Researchers are interested in the mean age of a certain population. Let us say that they are asking the following question: Can we conclude that the mean age of this population is different from 30 years?

Solution: Based on our knowledge of hypothesis testing, we reply that they can conclude that the mean age is different from 30 if they can reject the null hypothesis that the mean is equal to 30. Let us use the ten-step hypothesis testing procedure given in the previous section to help the researchers reach a conclusion.

1. **Data.** The data available to the researchers are the ages of a simple random sample of 10 individuals drawn from the population of interest. From this sample a mean of $\bar{x} = 27$ has been computed.
2. **Assumptions.** It is assumed that the sample comes from a population whose ages are approximately normally distributed. Let us also assume that the population has a known variance of $\sigma^2 = 20$.
3. **Hypotheses.** The hypothesis to be tested, or null hypothesis, is that the mean age of the population is equal to 30. The alternative hypothesis is

that the mean age of the population is not equal to 30. Note that we are identifying with the alternative hypothesis the conclusion the researchers wish to reach, so that if the data permit rejection of the null hypothesis, the researchers' conclusion will carry more weight, since the accompanying probability of rejecting a true null hypothesis will be small. We will make sure of this by assigning a small value to α, the probability of committing a type I error. We may present the relevant hypotheses in compact form as follows:

$$H_0\text{: } \mu = 30$$
$$H_A\text{: } \mu \neq 30$$

4. **Test statistic.** Since we are testing a hypothesis about a population mean, since we assume that the population is normally distributed, and since the population variance is known, our test statistic is given by Equation 7.2.1.
5. **Distribution of test statistic.** Based on our knowledge of sampling distributions and the normal distribution, we know that the test statistic is normally distributed with a mean of 0 and a variance of 1, if H_0 is true. There are many possible values of the test statistic that the present situation can generate; one for every possible sample of size 10 that can be drawn from the population. Since we draw only one sample, we have only one of these possible values on which to base a decision.
6. **Decision rule.** The decision rule tells us to reject H_0 if the computed value of the test statistic falls in the rejection region and to fail to reject H_0 if it falls in the nonrejection region. We must now specify the rejection and nonrejection regions. We can begin by asking ourselves what magnitude of values of the test statistic will cause rejection of H_0. If the null hypothesis is false, it may be so either because the population mean is less than 30 or because the population mean is greater than 30. Therefore, either sufficiently small values or sufficiently large values of the test statistic will cause rejection of the null hypothesis. We want these extreme values to constitute the rejection region. How extreme must a possible value of the test statistic be to qualify for the rejection region? The answer depends on the significance level we choose, that is, the size of the probability of committing a type I error. Let us say that we want the probability of rejecting a true null hypothesis to be $\alpha = .05$. Since our rejection region is to consist of two parts, sufficiently small values and sufficiently large values of the test statistic, part of α will have to be associated with the large values and part with the small values. It seems reasonable that we should divide α equally and let $\alpha/2 = .025$ be associated with small values and $\alpha/2 = .025$ be associated with large values.

Critical Value of Test Statistic

What value of the test statistic is so large that, when the null hypothesis is true, the probability of obtaining a value this large or larger is .025? In other words, what is the value of z to the right of which lies .025 of the area under the standard normal distribution? The value of z to the right of which lies .025 of the area is the same value that has .975 of the area between it and $-\infty$. We look in the body of Appendix Table D until we find .975 or its closest value and read the corresponding marginal entries to obtain our z value. In the present example the value of z is 1.96. Similar reasoning will lead us to find -1.96 as the value of the test statistic so small that when the null hypothesis is true, the probability of obtaining a value this small or smaller is .025. Our rejection region, then, consists of all values of the test statistic equal to or greater than 1.96 and less than or equal to -1.96. The nonrejection region consists of all values in between. We may state the decision rule for this test as follows: *reject H_0 if the computed value of the test statistic is either ≥ 1.96 or ≤ -1.96.* Otherwise, do not reject H_0. The rejection and nonrejection regions are shown in Figure 7.2.1. The values of the test statistic that separate the rejection and nonrejection regions are called *critical values* of the test statistic, and the rejection region is sometimes referred to as the *critical region.*

The decision rule tells us to compute a value of the test statistic from the data of our sample and to reject H_0 if we get a value that is either equal to or greater than 1.96 or equal to or less than -1.96 and to fail to reject H_0 if we get any other value. The value of α and, hence, the decision rule should be decided on before gathering the data. This prevents our being accused of allowing the sample results to influence our choice of α. This condition of objectivity is highly desirable and should be preserved in all tests.

7. **Calculation of test statistic.** From our sample we compute

$$z = \frac{27 - 30}{\sqrt{20/10}} = \frac{-3}{1.4142} = -2.12$$

8. **Statistical decision.** Abiding by the decision rule, we are able to reject the null hypothesis since -2.12 is in the rejection region. We

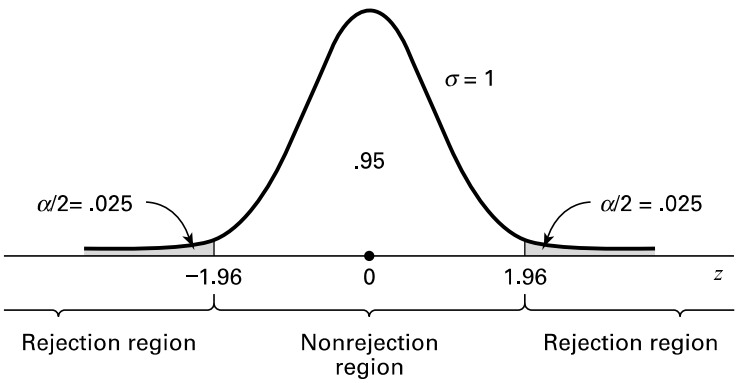

FIGURE 7.2.1 Rejection and nonrejection regions for Example 7.2.1.

can say that the computed value of the test statistic is significant at the .05 level.

9. **Conclusion.** We conclude that μ is not equal to 30 and let our administrative or clinical actions be in accordance with this conclusion.

10. ***p* values.** Instead of saying that an observed value of the test statistic is significant or is not significant, most writers in the research literature prefer to report the exact probability of getting a value as extreme as or more extreme than that observed if the null hypothesis is true. In the present instance these writers would give the computed value of the test statistic along with the statement $p = .0340$. The statement $p = .0340$ means that the probability of getting a value as extreme as 2.12 in either direction, when the null hypothesis is true, is .0340. The value .0340 is obtained from Appendix Table D and is the probability of observing a $z \geq 2.12$ or a $z \leq -2.12$ when the null hypothesis is true. That is, when H_0 is true, the probability of obtaining a value of z as large as or larger than 2.12 is .0170, and the probability of observing a value of z as small as or smaller than -2.12 is .0170. The probability of one or the other of these events occurring, when H_0 is true, is equal to the sum of the two individual probabilities, and hence, in the present example, we say that $p = .0170 + .0170 = .0340$.

Recall that the p value for a test may be defined also as the smallest value of α for which the null hypothesis can be rejected. Since, in Example 7.2.1, our p value is .0340, we know that we could have chosen an α value as small as .0340 and still have rejected the null hypothesis. If we had chosen an α smaller than .0340, we would not have been able to reject the null hypothesis. A general rule worth remembering, then, is this: *if the p value is less than or equal to* α, *we reject the null hypothesis; if the p value is greater than* α, *we do not reject the null hypothesis.*

The reporting of p values as part of the results of an investigation is more informative to the reader than such statements as "the null hypothesis is rejected at the .05 level of significance" or "the results were not significant at the .05 level." Reporting the p value associated with a test lets the reader know just how common or how rare is the computed value of the test statistic given that H_0 is true. ■

Testing H_0 by Means of a Confidence Interval Earlier, we stated that one can use confidence intervals to test hypotheses. In Example 7.2.1 we used a hypothesis testing procedure to test H_0: $\mu = 30$ against the alternative, H_A: $\mu \neq 30$. We were able to reject H_0 because the computed value of the test statistic fell in the rejection region.

Let us see how we might have arrived at this same conclusion by using a $100(1-\alpha)$ percent confidence interval. The 95 percent confidence interval for μ is

$$27 \pm 1.96\sqrt{20/10}$$
$$27 \pm 1.96(1.414)$$
$$27 \pm 2.7714$$
$$(24.2286,\ 29.7714)$$

Since this interval does not include 30, we say 30 is not a candidate for the mean we are estimating and, therefore, μ is not equal to 30 and H_0 is rejected. This is the same conclusion reached by means of the hypothesis testing procedure.

If the hypothesized parameter, 30, had been within the 95 percent confidence interval, we would have said that H_0 is not rejected at the .05 level of significance. In general, *when testing a null hypothesis by means of a two-sided confidence interval, we reject H_0 at the α level of significance if the hypothesized parameter is not contained within the $100(1-\alpha)$ percent confidence interval. If the hypothesized parameter is contained within the interval, H_0 cannot be rejected at the α level of significance.*

One-Sided Hypothesis Tests The hypothesis test illustrated by Example 7.2.1 is an example of a *two-sided test,* so called because the rejection region is split between the two sides or tails of the distribution of the test statistic. A hypothesis test may be *one-sided,* in which case all the rejection region is in one or the other tail of the distribution. Whether a one-sided or a two-sided test is used depends on the nature of the question being asked by the researcher.

If both large and small values will cause rejection of the null hypothesis, a two-sided test is indicated. When either sufficiently "small" values only or sufficiently "large" values only will cause rejection of the null hypothesis, a one-sided test is indicated.

EXAMPLE 7.2.2

Refer to Example 7.2.1. Suppose, instead of asking if they could conclude that $\mu \neq 30$, the researchers had asked: Can we conclude that $\mu < 30$? To this question we would reply that they can so conclude if they can reject the null hypothesis that $\mu \geq 30$.

Solution: Let us go through the ten-step procedure to reach a decision based on a one-sided test.

1. **Data.** See the previous example.
2. **Assumptions.** See the previous example.
3. **Hypotheses.**

$$H_0\colon \mu \geq 30$$
$$H_A\colon \mu < 30$$

The inequality in the null hypothesis implies that the null hypothesis consists of an infinite number of hypotheses. The test will be made only

at the point of equality, since it can be shown that if H_0 is rejected when the test is made at the point of equality it would be rejected if the test were done for any other value of μ indicated in the null hypothesis.

4. **Test statistic.**

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

5. **Distribution of test statistic.** See the previous example.

6. **Decision rule.** Let us again use $\alpha = .05$. To determine where to place the rejection region, let us ask ourselves what magnitude of values would cause rejection of the null hypothesis. If we look at the hypotheses, we see that sufficiently small values would cause rejection and that large values would tend to reinforce the null hypothesis. We will want our rejection region to be where the small values are—at the lower tail of the distribution. This time, since we have a one-sided test, all of α will go in the one tail of the distribution. By consulting Appendix Table D, we find that the value of z to the left of which lies .05 of the area under the standard normal curve is -1.645 after interpolating. Our rejection and nonrejection regions are now specified and are shown in Figure 7.2.2.

   Our decision rule tells us to reject H_0 if the computed value of the test statistic is less than or equal to -1.645.

7. **Calculation of test statistic.** From our data we compute

$$z = \frac{27 - 30}{\sqrt{20/10}} = -2.12$$

8. **Statistical decision.** We are able to reject the null hypothesis since $-2.12 < -1.645$.

9. **Conclusion.** We conclude that the population mean is smaller than 30 and act accordingly.

10. ***p* value.** The p value for this test is .0170, since $P(z \leq -2.12)$, when H_0 is true, is .0170 as given by Appendix Table D when we determine the

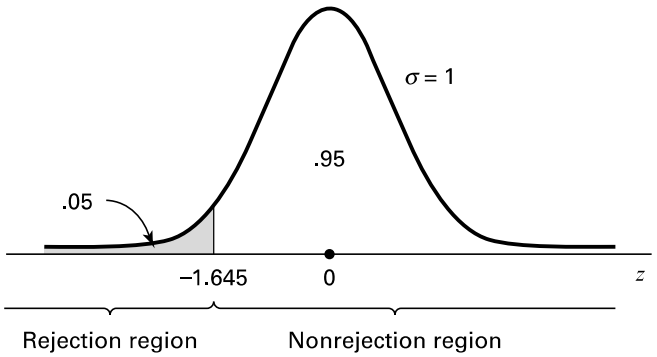

FIGURE 7.2.2 Rejection and nonrejection regions for Example 7.2.2.

magnitude of the area to the left of -2.12 under the standard normal curve. One can test a one-sided null hypothesis by means of a one-sided confidence interval. However, we will not cover the construction and interpretation of this type of confidence interval in this book.

If the researcher's question had been, "Can we conclude that the mean is greater than 30?", following the above ten-step procedure would have led to a one-sided test with all the rejection region at the upper tail of the distribution of the test statistic and a critical value of $+1.645$. ■

Sampling from a Normally Distributed Population: Population Variance Unknown As we have already noted, the population variance is usually unknown in actual situations involving statistical inference about a population mean. When sampling is from an approximately normal population with an unknown variance, the test statistic for testing H_0: $\mu = \mu_0$ is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{7.2.2}$$

which, when H_0 is true, is distributed as Student's t with $n - 1$ degrees of freedom. The following example illustrates the hypothesis testing procedure when the population is assumed to be normally distributed and its variance is unknown. This is the usual situation encountered in practice.

EXAMPLE 7.2.3

Nakamura et al. (A-1) studied subjects with medial collateral ligament (MCL) and anterior cruciate ligament (ACL) tears. Between February 1995 and December 1997, 17 consecutive patients with combined acute ACL and grade III MCL injuries were treated by the same physician at the research center. One of the variables of interest was the length of time in days between the occurrence of the injury and the first magnetic resonance imaging (MRI). The data are shown in Table 7.2.1. We wish to know if we can conclude that the mean number of days between injury and initial MRI is not 15 days in a population presumed to be represented by these sample data.

TABLE 7.2.1 Number of Days Until MRI for Subjects with MCL and ACL Tears

Subject	Days	Subject	Days	Subject	Days	Subject	Days
1	14	6	0	11	28	16	14
2	9	7	10	12	24	17	9
3	18	8	4	13	24		
4	26	9	8	14	2		
5	12	10	21	15	3		

Source: Norimasa Nakamura, Shuji Horibe, Yukyoshi Toritsuka, Tomoki Mitsuoka, Hideki Yoshikawa, and Konsei Shino, "Acute Grade III Medial Collateral Ligament Injury of the Knee Associated with Anterior Cruciate Ligament Tear," *American Journal of Sports Medicine, 31* (2003), 261–267.

Solution: We will be able to conclude that the mean number of days for the population is not 15 if we can reject the null hypothesis that the population mean is equal to 15.

1. **Data.** The data consist of number of days until MRI on 17 subjects as previously described.
2. **Assumptions.** The 17 subjects constitute a simple random sample from a population of similar subjects. We assume that the number of days until MRI in this population is approximately normally distributed.
3. **Hypotheses.**

$$H_0: \mu = 15$$
$$H_A: \mu \neq 15$$

4. **Test statistic.** Since the population variance is unknown, our test statistic is given by Equation 7.2.2.
5. **Distribution of test statistic.** Our test statistic is distributed as Student's t with $n - 1 = 17 - 1 = 16$ degrees of freedom if H_0 is true.
6. **Decision rule.** Let $\alpha = .05$. Since we have a two-sided test, we put $\alpha/2 = .025$ in each tail of the distribution of our test statistic. The t values to the right and left of which .025 of the area lies are 2.1199 and -2.1199. These values are obtained from Appendix Table E. The rejection and nonrejection regions are shown in Figure 7.2.3.

   The decision rule tells us to compute a value of the test statistic and reject H_0 if the computed t is either greater than or equal to 2.1199 or less than or equal to -2.1199.
7. **Calculation of test statistic.** From our sample data we compute a sample mean of 13.2941 and a sample standard deviation of 8.88654. Substituting these statistics into Equation 7.2.2 gives

$$t = \frac{13.2941 - 15}{8.88654/\sqrt{17}} = \frac{-1.7059}{2.1553} = -.791$$

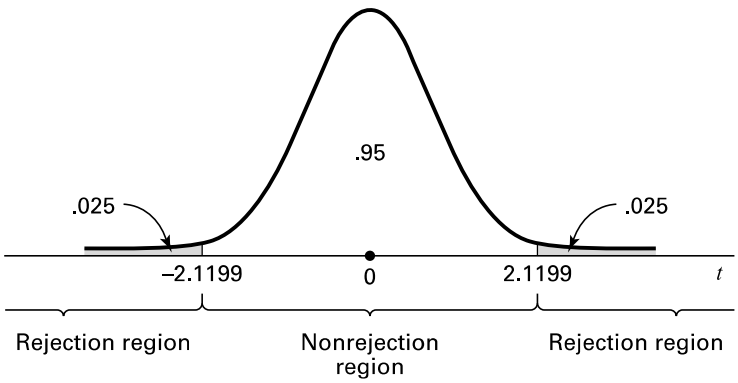

FIGURE 7.2.3 Rejection and nonrejection regions for Example 7.2.3.

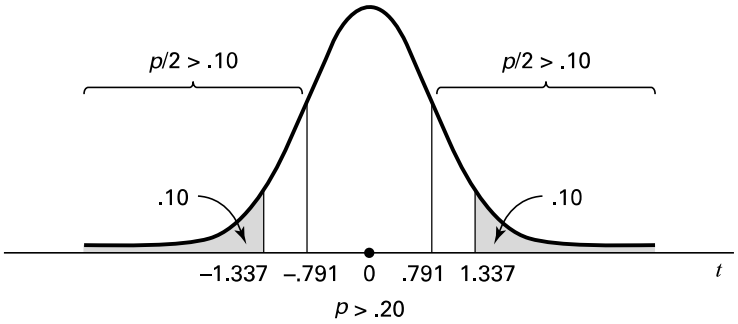

FIGURE 7.2.4 Determination of p value for Example 7.2.3.

8. **Statistical decision.** Do not reject H_0, since $-.791$ falls in the nonrejection region.
9. **Conclusion.** Our conclusion, based on these data, is that the mean of the population from which the sample came may be 15.
10. ***p value.*** The exact p value for this test cannot be obtained from Appendix Table E since it gives t values only for selected percentiles. The p value can be stated as an interval, however. We find that $-.791$ is less than -1.337, the value of t to the left of which lies .10 of the area under the t with 16 degrees of freedom. Consequently, when H_0 is true, the probability of obtaining a value of t as small as or smaller than $-.791$ is greater than .10. That is $P(t \leq -.791) > .10$. Since the test was two-sided, we must allow for the possibility of a computed value of the test statistic as large in the opposite direction as that observed. Appendix Table E reveals that $P(t \geq .791) > .10$ also. The p value, then, is $p > .20$. In fact, Excel calculates the p value to be .4403. Figure 7.2.4 shows the p value for this example.

If in the previous example the hypotheses had been

$$H_0: \mu \geq 15$$
$$H_A: \mu < 15$$

the testing procedure would have led to a one-sided test with all the rejection region at the lower tail of the distribution, and if the hypotheses had been

$$H_0: \mu \leq 15$$
$$H_A: \mu > 15$$

we would have had a one-sided test with all the rejection region at the upper tail of the distribution. ■

Sampling from a Population That Is Not Normally Distributed

If, as is frequently the case, the sample on which we base our hypothesis test about a population mean comes from a population that is not normally distributed, we may, if our sample is large (greater than or equal to 30), take advantage of the central limit theorem and use $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ as the test statistic. If the population standard deviation is not

Student's t-test

Definition:

The t-test is a statistical method used to determine if there is a significant difference between the means of two groups. It helps in hypothesis testing when sample sizes are small and population variances are unknown.

Properties:

- **Parametric Test:** Assumes that the data is normally distributed.
- **Types:**
    - **Independent (Unpaired) t-test** – Compares means of two independent groups.
    - **Paired t-test** – Compares means of the same group under two different conditions.
- **Assumptions:**
    - The sample data should be randomly selected.
    - Data should follow a normal distribution.
    - Variances of the two groups should be approximately equal (for independent t-test).
    - Observations should be independent.

Uses:

- Comparing two treatment groups in clinical trials.
- Evaluating pre- and post-treatment differences in the same group.
- Comparing sample means in biological and healthcare research.

Formula:

1. Independent Samples t-test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\bar{X}_1, \bar{X}_2$ = sample means of groups 1 and 2
- s_1, s_2 = sample standard deviations
- n_1, n_2 = sample sizes

2. Paired t-test:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where:

- $\bar{d}$ = mean of the differences between paired observations
- s_d = standard deviation of differences
- n = number of pairs

Classification in Biostatistics:

- **Inferential Statistics:** Helps in drawing conclusions about a population based on sample data.
- **Parametric Test:** Used when data follows a normal distribution.
- **Hypothesis Testing:** Determines if differences in means are statistically significant.

References:

1. Daniel, W. W., & Cross, C. L. (2018). *Biostatistics: A Foundation for Analysis in the Health Sciences* (11th ed.). Wiley.
2. Rosner, B. (2015). *Fundamentals of Biostatistics* (8th ed.). Cengage Learning.
3. Pagano, M., & Gauvreau, K. (2018). *Principles of Biostatistics* (2nd ed.). CRC Press.
4. Kirkwood, B. R., & Sterne, J. A. C. (2003). *Essential Medical Statistics* (2nd ed.). Blackwell Science.

knowledge of test results (positive or negative) and/or the status of presenting symptoms (present or absent). Also of interest is information regarding the likelihood of positive and negative test results and the likelihood of the presence or absence of a particular symptom in patients with and without a particular disease.

In our consideration of screening tests, we must be aware of the fact that they are not always infallible. That is, a testing procedure may yield a *false positive* or a *false negative.*

DEFINITION

1. **A false positive results when a test indicates a positive status when the true status is negative.**
2. **A false negative results when a test indicates a negative status when the true status is positive.**

In summary, the following questions must be answered in order to evaluate the usefulness of test results and symptom status in determining whether or not a subject has some disease:

1. Given that a subject has the disease, what is the probability of a positive test result (or the presence of a symptom)?
2. Given that a subject does not have the disease, what is the probability of a negative test result (or the absence of a symptom)?
3. Given a positive screening test (or the presence of a symptom), what is the probability that the subject has the disease?
4. Given a negative screening test result (or the absence of a symptom), what is the probability that the subject does not have the disease?

Suppose we have for a sample of n subjects (where n is a large number) the information shown in Table 3.5.1. The table shows for these n subjects their status with regard to a disease and results from a screening test designed to identify subjects with the disease. The cell entries represent the number of subjects falling into the categories defined by the row and column headings. For example, a is the number of subjects who have the disease and whose screening test result was positive.

As we have learned, a variety of probability estimates may be computed from the information displayed in a two-way table such as Table 3.5.1. For example, we may

TABLE 3.5.1 Sample of n Subjects (Where n Is Large) Cross-Classified According to Disease Status and Screening Test Result

	Disease		
Test Result	**Present (D)**	**Absent ($\bar{D}$)**	**Total**
Positive (T)	a	b	$a+b$
Negative ($\bar{T}$)	c	d	$c+d$
Total	$a+c$	$b+d$	n

compute the conditional probability estimate $P(T \mid D) = a/(a + c)$. This ratio is an estimate of the *sensitivity* of the screening test.

DEFINITION

The sensitivity of a test (or symptom) is the probability of a positive test result (or presence of the symptom) given the presence of the disease.

We may also compute the conditional probability estimate $P(\bar{T} \mid \bar{D}) = d/(b + d)$. This ratio is an estimate of the *specificity* of the screening test.

DEFINITION

The specificity of a test (or symptom) is the probability of a negative test result (or absence of the symptom) given the absence of the disease.

From the data in Table 3.5.1 we answer Question 3 by computing the conditional probability estimate $P(D \mid T)$. This ratio is an estimate of a probability called the *predictive value positive* of a screening test (or symptom).

DEFINITION

The predictive value positive of a screening test (or symptom) is the probability that a subject has the disease given that the subject has a positive screening test result (or has the symptom).

Similarly, the ratio $P(\bar{D} \mid \bar{T})$ is an estimate of the conditional probability that a subject does not have the disease given that the subject has a negative screening test result (or does not have the symptom). The probability estimated by this ratio is called the *predictive value negative* of the screening test or symptom.

DEFINITION

The predictive value negative of a screening test (or symptom) is the probability that a subject does not have the disease, given that the subject has a negative screening test result (or does not have the symptom).

Estimates of the predictive value positive and predictive value negative of a test (or symptom) may be obtained from knowledge of a test's (or symptom's) sensitivity and specificity and the probability of the relevant disease in the general population. To obtain these predictive value estimates, we make use of Bayes's theorem. The following statement of Bayes's theorem, employing the notation established in Table 3.5.1, gives the predictive value positive of a screening test (or symptom):

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid \bar{D})P(\bar{D})} \tag{3.5.1}$$

It is instructive to examine the composition of Equation 3.5.1. We recall from Equation 3.4.2 that the conditional probability $P(D \mid T)$ is equal to $P(D \cap T)/P(T)$. To understand the logic of Bayes's theorem, we must recognize that the numerator of Equation 3.5.1 represents $P(D \cap T)$ and that the denominator represents $P(T)$. We know from the multiplication rule of probability given in Equation 3.4.1 that the numerator of Equation 3.5.1, $P(T \mid D)\,P(D)$, is equal to $P(D \cap T)$.

Now let us show that the denominator of Equation 3.5.1 is equal to $P(T)$. We know that event T is the result of a subject's being classified as positive with respect to a screening test (or classified as having the symptom). A subject classified as positive may have the disease or may not have the disease. Therefore, the occurrence of T is the result of a subject having the disease and being positive $[P(D \cap T)]$ or not having the disease and being positive $[P(\bar{D} \cap T)]$. These two events are mutually exclusive (their intersection is zero), and consequently, by the addition rule given by Equation 3.4.3, we may write

$$P(T) = P(D \cap T) + P(\bar{D} \cap T) \tag{3.5.2}$$

Since, by the multiplication rule, $P(D \cap T) = P(T \mid D)\,P(D)$ and $P(\bar{D} \cap T) = P(T \mid \bar{D})\,P(\bar{D})$, we may rewrite Equation 3.5.2 as

$$P(T) = P(T \mid D)P(D) + P(T \mid \bar{D})P(\bar{D}) \tag{3.5.3}$$

which is the denominator of Equation 3.5.1.

Note, also, that the numerator of Equation 3.5.1 is equal to the sensitivity times the rate (prevalence) of the disease and the denominator is equal to the sensitivity times the rate of the disease plus the term *1 minus the sensitivity* times the term *1 minus the rate of the disease*. Thus, we see that the predictive value positive can be calculated from knowledge of the sensitivity, specificity, and the rate of the disease.

Evaluation of Equation 3.5.1 answers Question 3. To answer Question 4 we follow a now familiar line of reasoning to arrive at the following statement of Bayes's theorem:

$$P(\bar{D} \mid \bar{T}) = \frac{P(\bar{T} \mid \bar{D})P(\bar{D})}{P(\bar{T} \mid \bar{D})P(\bar{D}) + P(\bar{T} \mid D)P(D)} \tag{3.5.4}$$

Equation 3.5.4 allows us to compute an estimate of the probability that a subject who is negative on the test (or has no symptom) does not have the disease, which is the predictive value negative of a screening test or symptom.

We illustrate the use of Bayes' theorem for calculating a predictive value positive with the following example.

EXAMPLE 3.5.1

A medical research team wished to evaluate a proposed screening test for Alzheimer's disease. The test was given to a random sample of 450 patients with Alzheimer's disease and an independent random sample of 500 patients without symptoms of the disease.

The two samples were drawn from populations of subjects who were 65 years of age or older. The results are as follows:

	Alzheimer's Diagnosis?		
Test Result	**Yes (D)**	**No ($\bar{D}$)**	**Total**
Positive (T)	436	5	441
Negative ($\bar{T}$)	14	495	509
Total	450	500	950

Using these data we estimate the sensitivity of the test to be $P(T \mid D) = 436/450 = .97$. The specificity of the test is estimated to be $P(\bar{T} \mid \bar{D}) = 495/500 = .99$. We now use the results of the study to compute the predictive value positive of the test. That is, we wish to estimate the probability that a subject who is positive on the test has Alzheimer's disease. From the tabulated data we compute $P(T \mid D) = 436/450 = .9689$ and $P(T \mid \bar{D}) = 5/500 = .01$. Substitution of these results into Equation 3.5.1 gives

$$P(D \mid T) = \frac{(.9689)P(D)}{(.9689)P(D) + (.01)P(\bar{D})} \tag{3.5.5}$$

We see that the predictive value positive of the test depends on the rate of the disease in the relevant population in general. In this case the relevant population consists of subjects who are 65 years of age or older. We emphasize that the rate of disease in the relevant general population, $P(D)$, cannot be computed from the sample data, since two independent samples were drawn from two different populations. We must look elsewhere for an estimate of $P(D)$. Evans et al. (A-5) estimated that 11.3 percent of the U.S. population aged 65 and over have Alzheimer's disease. When we substitute this estimate of $P(D)$ into Equation 3.5.5 we obtain

$$P(D \mid T) = \frac{(.9689)(.113)}{(.9689)(.113) + (.01)(1 - .113)} = .93$$

As we see, in this case, the predictive value of the test is very high.

Similarly, let us now consider the predictive value negative of the test. We have already calculated all entries necessary except for $P(\bar{T} \mid D) = 14/450 = .0311$. Using the values previously obtained and our new value, we find

$$P(\bar{D} \mid T) = \frac{(.99)(1 - .113)}{(.99)(1 - .113) + (.0311)(.113)} = .996$$

As we see, the predictive value negative is also quite high. ■

pursue this subject should refer to the many books on probability available in most college and university libraries. The books by Gut (1), Isaac (2), and Larson (3) are recommended. The objectives of this chapter are to help students gain some mathematical ability in the area of probability and to assist them in developing an understanding of the more important concepts. Progress along these lines will contribute immensely to their success in understanding the statistical inference procedures presented later in this book.

The concept of probability is not foreign to health workers and is frequently encountered in everyday communication. For example, we may hear a physician say that a patient has a 50–50 chance of surviving a certain operation. Another physician may say that she is 95 percent certain that a patient has a particular disease. A public health nurse may say that nine times out of ten a certain client will break an appointment. As these examples suggest, most people express probabilities in terms of percentages. In dealing with probabilities mathematically, it is more convenient to express probabilities as fractions. (Percentages result from multiplying the fractions by 100.) Thus, we measure the probability of the occurrence of some event by a number between zero and one. The more likely the event, the closer the number is to one; and the more unlikely the event, the closer the number is to zero. An event that cannot occur has a probability of zero, and an event that is certain to occur has a probability of one.

Health sciences researchers continually ask themselves if the results of their efforts could have occurred by chance alone or if some other force was operating to produce the observed effects. For example, suppose six out of ten patients suffering from some disease are cured after receiving a certain treatment. Is such a cure rate likely to have occurred if the patients had not received the treatment, or is it evidence of a true curative effect on the part of the treatment? We shall see that questions such as these can be answered through the application of the concepts and laws of probability.

3.2 TWO VIEWS OF PROBABILITY: OBJECTIVE AND SUBJECTIVE

Until fairly recently, probability was thought of by statisticians and mathematicians only as an *objective* phenomenon derived from objective processes.

The concept of *objective probability* may be categorized further under the headings of (1) *classical,* or *a priori, probability*, and (2) the *relative frequency,* or *a posteriori,* concept of probability.

Classical Probability The classical treatment of probability dates back to the 17th century and the work of two mathematicians, Pascal and Fermat. Much of this theory developed out of attempts to solve problems related to games of chance, such as those involving the rolling of dice. Examples from games of chance illustrate very well the principles involved in classical probability. For example, if a fair six-sided die is rolled, the probability that a 1 will be observed is equal to $1/6$ and is the same for the other five faces. If a card is picked at random from a well-shuffled deck of ordinary playing cards, the probability of picking a heart is $13/52$. Probabilities such as these are calculated by the processes of abstract reasoning. It is not necessary to roll a die or draw a card to compute

these probabilities. In the rolling of the die, we say that each of the six sides is *equally likely* to be observed if there is no reason to favor any one of the six sides. Similarly, if there is no reason to favor the drawing of a particular card from a deck of cards, we say that each of the 52 cards is equally likely to be drawn. We may define probability in the classical sense as follows:

DEFINITION

If an event can occur in N mutually exclusive and equally likely ways, and if m of these possess a trait E, the probability of the occurrence of E is equal to m/N.

If we read $P(E)$ as "the probability of E," we may express this definition as

$$P(E) = \frac{m}{N} \tag{3.2.1}$$

Relative Frequency Probability The relative frequency approach to probability depends on the repeatability of some process and the ability to count the number of repetitions, as well as the number of times that some event of interest occurs. In this context we may define the probability of observing some characteristic, E, of an event as follows:

DEFINITION

If some process is repeated a large number of times, n, and if some resulting event with the characteristic E occurs m times, the relative frequency of occurrence of E, m/n, will be approximately equal to the probability of E.

To express this definition in compact form, we write

$$P(E) = \frac{m}{n} \tag{3.2.2}$$

We must keep in mind, however, that, strictly speaking, m/n is only an estimate of $P(E)$.

Subjective Probability In the early 1950s, L. J. Savage (4) gave considerable impetus to what is called the "personalistic" or subjective concept of probability. This view holds that probability measures the confidence that a particular individual has in the truth of a particular proposition. This concept does not rely on the repeatability of any process. In fact, by applying this concept of probability, one may evaluate the probability of an event that can only happen once, for example, the probability that a cure for cancer will be discovered within the next 10 years.

Although the subjective view of probability has enjoyed increased attention over the years, it has not been fully accepted by statisticians who have traditional orientations.

Bayesian Methods Bayesian methods are named in honor of the Reverend Thomas Bayes (1702–1761), an English clergyman who had an interest in mathematics. Bayesian methods are an example of subjective probability, since it takes into consideration the degree of belief that one has in the chance that an event will occur. While probabilities based on classical or relative frequency concepts are designed to allow for decisions to be made solely on the basis of collected data, Bayesian methods make use of what are known as *prior probabilities* and *posterior probabilities*.

DEFINITION

The *prior probability* of an event is a probability based on prior knowledge, prior experience, or results derived from prior data collection activity.

DEFINITION

The *posterior probability* of an event is a probability obtained by using new information to update or revise a prior probability.

As more data are gathered, the more is likely to be known about the "true" probability of the event under consideration. Although the idea of updating probabilities based on new information is in direct contrast to the philosophy behind frequency-of-occurrence probability, Bayesian concepts are widely used. For example, Bayesian techniques have found recent application in the construction of e-mail spam filters. Typically, the application of Bayesian concepts makes use of a mathematical formula called *Bayes' theorem*. In Section 3.5 we employ Bayes' theorem in the evaluation of diagnostic screening test data.

3.3 ELEMENTARY PROPERTIES OF PROBABILITY

In 1933 the axiomatic approach to probability was formalized by the Russian mathematician A. N. Kolmogorov (5). The basis of this approach is embodied in three properties from which a whole system of probability theory is constructed through the use of mathematical logic. The three properties are as follows.

1. Given some process (or experiment) with n mutually exclusive outcomes (called events), $E_1, E_2, \ldots, E_n$, the probability of any event E_i is assigned a nonnegative number. That is,

$$P(E_i) \geq 0 \tag{3.3.1}$$

In other words, all events must have a probability greater than or equal to zero, a reasonable requirement in view of the difficulty of conceiving of negative probability. A key concept in the statement of this property is the concept of *mutually exclusive* outcomes. Two events are said to be mutually exclusive if they cannot occur simultaneously.

2. The sum of the probabilities of the mutually exclusive outcomes is equal to 1.

$$P(E_1) + P(E_2) + \cdots + P(E_n) = 1 \tag{3.3.2}$$

This is the property of *exhaustiveness* and refers to the fact that the observer of a probabilistic process must allow for all possible events, and when all are taken together, their total probability is 1. The requirement that the events be mutually exclusive is specifying that the events $E_1, E_2, \ldots, E_n$ do not overlap; that is, no two of them can occur at the same time.

3. Consider any two mutually exclusive events, E_i and E_j. The probability of the occurrence of either E_i or E_j is equal to the sum of their individual probabilities.

$$P(E_i + E_j) = P(E_i) + P(E_j) \tag{3.3.3}$$

Suppose the two events were not mutually exclusive; that is, suppose they could occur at the same time. In attempting to compute the probability of the occurrence of either E_i or E_j the problem of overlapping would be discovered, and the procedure could become quite complicated. This concept will be discusses further in the next section.

3.4 CALCULATING THE PROBABILITY OF AN EVENT

We now make use of the concepts and techniques of the previous sections in calculating the probabilities of specific events. Additional ideas will be introduced as needed.

EXAMPLE 3.4.1

The primary aim of a study by Carter et al. (A-1) was to investigate the effect of the age at onset of bipolar disorder on the course of the illness. One of the variables investigated was family history of mood disorders. Table 3.4.1 shows the frequency of a family history of

TABLE 3.4.1 Frequency of Family History of Mood Disorder by Age Group among Bipolar Subjects

Family History of Mood Disorders	Early = 18(*E*)	Later > 18(*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

mood disorders in the two groups of interest (Early age at onset defined to be 18 years or younger and Later age at onset defined to be later than 18 years). Suppose we pick a person at random from this sample. What is the probability that this person will be 18 years old or younger?

Solution: For purposes of illustrating the calculation of probabilities we consider this group of 318 subjects to be the largest group for which we have an interest. In other words, for this example, we consider the 318 subjects as a population. We assume that Early and Later are mutually exclusive categories and that the likelihood of selecting any one person is equal to the likelihood of selecting any other person. We define the desired probability as the number of subjects with the characteristic of interest (Early) divided by the total number of subjects. We may write the result in probability notation as follows:

$$\begin{aligned} P(E) &= \text{number of Early subjects/total number of subjects} \\ &= 141/318 = .4434 \end{aligned}$$

■

Conditional Probability On occasion, the set of "all possible outcomes" may constitute a subset of the total group. In other words, the size of the group of interest may be reduced by conditions not applicable to the total group. When probabilities are calculated with a subset of the total group as the denominator, the result is a *conditional probability.*

The probability computed in Example 3.4.1, for example, may be thought of as an unconditional probability, since the size of the total group served as the denominator. No conditions were imposed to restrict the size of the denominator. We may also think of this probability as a *marginal probability* since one of the marginal totals was used as the numerator.

We may illustrate the concept of conditional probability by referring again to Table 3.4.1.

EXAMPLE 3.4.2

Suppose we pick a subject at random from the 318 subjects and find that he is 18 years or younger (E). What is the probability that this subject will be one who has no family history of mood disorders (A)?

Solution: The total number of subjects is no longer of interest, since, with the selection of an Early subject, the Later subjects are eliminated. We may define the desired probability, then, as follows: What is the probability that a subject has no family history of mood disorders (A), given that the selected subject is Early (E)? This is a conditional probability and is written as $P(A \mid E)$ in which the vertical line is read "given." The 141 Early subjects become the denominator of this conditional probability, and 28, the number of Early subjects with no family history of mood disorders, becomes the numerator. Our desired probability, then, is

$$P(A \mid E) = 28/141 = .1986$$

■

Joint Probability Sometimes we want to find the probability that a subject picked at random from a group of subjects possesses two characteristics at the same time. Such a probability is referred to as a *joint probability.* We illustrate the calculation of a joint probability with the following example.

EXAMPLE 3.4.3

Let us refer again to Table 3.4.1. What is the probability that a person picked at random from the 318 subjects will be Early (E) *and* will be a person who has no family history of mood disorders (A)?

Solution: The probability we are seeking may be written in symbolic notation as $P(E \cap A)$ in which the symbol $\cap$ is read either as "intersection" or "and." The statement $E \cap A$ indicates the joint occurrence of conditions E and A. The number of subjects satisfying both of the desired conditions is found in Table 3.4.1 at the intersection of the column labeled E and the row labeled A and is seen to be 28. Since the selection will be made from the total set of subjects, the denominator is 318. Thus, we may write the joint probability as

$$P(E \cap A) = 28/318 = .0881$$

■

The Multiplication Rule A probability may be computed from other probabilities. For example, a joint probability may be computed as the product of an appropriate marginal probability and an appropriate conditional probability. This relationship is known as the *multiplication rule* of probability. We illustrate with the following example.

EXAMPLE 3.4.4

We wish to compute the joint probability of Early age at onset (E) and a negative family history of mood disorders (A) from a knowledge of an appropriate marginal probability and an appropriate conditional probability.

Solution: The probability we seek is $P(E \cap A)$. We have already computed a marginal probability, $P(E) = 141/318 = .4434$, and a conditional probability, $P(A|E) = 28/141 = .1986$. It so happens that these are appropriate marginal and conditional probabilities for computing the desired joint probability. We may now compute $P(E \cap A) = P(E)P(A \mid E) = (.4434)(.1986) = .0881$. This, we note, is, as expected, the same result we obtained earlier for $P(E \cap A)$. ■

We may state the multiplication rule in general terms as follows: For any two events A and B,

$$P(A \cap B) = P(B)P(A \mid B), \qquad \text{if } P(B) \neq 0 \tag{3.4.1}$$

For the same two events A and B, the multiplication rule may also be written as $P(A \cap B) = P(A)P(B \mid A)$, if $P(A) \neq 0$.

We see that through algebraic manipulation the multiplication rule as stated in Equation 3.4.1 may be used to find any one of the three probabilities in its statement if the other two are known. We may, for example, find the conditional probability $P(A \mid B)$ by

dividing $P(A \cap B)$ by $P(B)$. This relationship allows us to formally define conditional probability as follows.

DEFINITION

The *conditional probability* of A given B is equal to the probability of $A \cap B$ divided by the probability of B, provided the probability of B is not zero.

That is,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) \neq 0 \tag{3.4.2}$$

We illustrate the use of the multiplication rule to compute a conditional probability with the following example.

EXAMPLE 3.4.5

We wish to use Equation 3.4.2 and the data in Table 3.4.1 to find the conditional probability, $P(A \mid E)$

Solution: According to Equation 3.4.2,

$$P(A \mid E) = P(A \cap E)/P(E)$$ ■

Earlier we found $P(E \cap A) = P(A \cap E) = 28/318 = .0881$. We have also determined that $P(E) = 141/318 = .4434$. Using these results we are able to compute $P(A \mid E) = .0881/.4434 = .1987$, which, as expected, is the same result we obtained by using the frequencies directly from Table 3.4.1. (The slight discrepancy is due to rounding.)

The Addition Rule The third property of probability given previously states that the probability of the occurrence of either one or the other of two mutually exclusive events is equal to the sum of their individual probabilities. Suppose, for example, that we pick a person at random from the 318 represented in Table 3.4.1. What is the probability that this person will be Early age at onset (E) or Later age at onset (L)? We state this probability in symbols as $P(E \cup L)$, where the symbol $\cup$ is read either as "union" or "or." Since the two age conditions are mutually exclusive, $P(E \cap L) = (141/318) + (177/318) = .4434 + .5566 = 1$.

What if two events are not mutually exclusive? This case is covered by what is known as the *addition rule,* which may be stated as follows:

DEFINITION

Given two events A and B, the probability that event A, or event B, or both occur is equal to the probability that event A occurs, plus the probability that event B occurs, minus the probability that the events occur simultaneously.

The addition rule may be written

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3.4.3}$$

When events A and B cannot occur simultaneously, $P(A \cap B)$ is sometimes called "exclusive or," and $P(A \cup B) = 0$. When events A and B can occur simultaneously, $P(A \cup B)$ is sometimes called "inclusive or," and we use the addition rule to calculate $P(A \cup B)$. Let us illustrate the use of the addition rule by means of an example.

EXAMPLE 3.4.6

If we select a person at random from the 318 subjects represented in Table 3.4.1, what is the probability that this person will be an Early age of onset subject (E) or will have no family history of mood disorders (A) or both?

Solution: The probability we seek is $P(E \cup A)$. By the addition rule as expressed by Equation 3.4.3, this probability may be written as $P(E \cup A) = P(E) + P(A) - P(E \cap A)$. We have already found that $P(E) = 141/318 = .4434$ and $P(E \cap A) = 28/318 = .0881$. From the information in Table 3.4.1 we calculate $P(A) = 63/318 = .1981$. Substituting these results into the equation for $P(E \cup A)$ we have $P(E \cup A) = .4434 + .1981 - .0881 = .5534$. ■

Note that the 28 subjects who are *both* Early *and* have no family history of mood disorders are included in the 141 who are Early as well as in the 63 who have no family history of mood disorders. Since, in computing the probability, these 28 have been added into the numerator twice, they have to be subtracted out once to overcome the effect of duplication, or overlapping.

Independent Events Suppose that, in Equation 3.4.2, we are told that event B has occurred, but that this fact has no effect on the probability of A. That is, suppose that the probability of event A is the same regardless of whether or not B occurs. In this situation, $P(A \mid B) = P(A)$. In such cases we say that A and B are *independent events.* The multiplication rule for two independent events, then, may be written as

$$P(A \cap B) = P(A)P(B); \qquad P(A) \neq 0, \qquad P(B) \neq 0 \tag{3.4.4}$$

Thus, we see that if two events are independent, the probability of their joint occurrence is equal to the product of the probabilities of their individual occurrences.

Note that when two events with nonzero probabilities are independent, each of the following statements is true:

$$P(A \mid B) = P(A), \qquad P(B|A) = P(B), \qquad P(A \cap B) = P(A)P(B)$$

Two events are not independent unless all these statements are true. It is important to be aware that the terms *independent* and *mutually exclusive* do not mean the same thing.

Let us illustrate the concept of independence by means of the following example.

EXAMPLE 3.4.7

In a certain high school class, consisting of 60 girls and 40 boys, it is observed that 24 girls and 16 boys wear eyeglasses. If a student is picked at random from this class, the probability that the student wears eyeglasses, $P(E)$, is $40/100$, or .4.

(a) What is the probability that a student picked at random wears eyeglasses, given that the student is a boy?

Solution: By using the formula for computing a conditional probability, we find this to be

$$P(E \mid B) = \frac{P(E \cap B)}{P(B)} = \frac{16/100}{40/100} = .4$$

Thus the additional information that a student is a boy does not alter the probability that the student wears eyeglasses, and $P(E) = P(E \mid B)$. We say that the events being a boy and wearing eyeglasses for this group are independent. We may also show that the event of wearing eyeglasses, E, and *not* being a boy, $\bar{B}$ are also independent as follows:

$$P(E \mid \bar{B}) = \frac{P(E \cap \bar{B})}{P(\bar{B})} = \frac{24/100}{60/100} = \frac{24}{60} = .4$$

(b) What is the probability of the joint occurrence of the events of wearing eyeglasses and being a boy?

Solution: Using the rule given in Equation 3.4.1, we have

$$P(E \cap B) = P(B)P(E \mid B)$$

but, since we have shown that events E and B are independent we may replace $P(E \mid B)$ by $P(E)$ to obtain, by Equation 3.4.4,

$$\begin{aligned} P(E \cap B) &= P(B)P(E) \\ &= \left(\frac{40}{100}\right)\left(\frac{40}{100}\right) \\ &= .16 \end{aligned}$$

■

Complementary Events Earlier, using the data in Table 3.4.1, we computed the probability that a person picked at random from the 318 subjects will be an Early age of onset subject as $P(E) = 141/318 = .4434$. We found the probability of a Later age at onset to be $P(L) = 177/318 = .5566$. The sum of these two probabilities we found to be equal to 1. This is true because the events being Early age at onset and being Later age at onset are *complementary events.* In general, we may make the following statement about complementary events. The probability of an event A is equal to 1 minus the probability of its

complement, which is written $\bar{A}$ and

$$P(\bar{A}) = 1 - P(A) \tag{3.4.5}$$

This follows from the third property of probability since the event, A, and its complement, $\bar{A}$ are mutually exclusive.

EXAMPLE 3.4.8

Suppose that of 1200 admissions to a general hospital during a certain period of time, 750 are private admissions. If we designate these as set A, then $\bar{A}$ is equal to 1200 minus 750, or 450. We may compute

$$P(A) = 750/1200 = .625$$

and

$$P(\bar{A}) = 450/1200 = .375$$

and see that

$$\begin{aligned} P(\bar{A}) &= 1 - P(A) \\ .375 &= 1 - .625 \\ .375 &= .375 \end{aligned}$$

■

Marginal Probability Earlier we used the term *marginal probability* to refer to a probability in which the numerator of the probability is a marginal total from a table such as Table 3.4.1. For example, when we compute the probability that a person picked at random from the 318 persons represented in Table 3.4.1 is an Early age of onset subject, the numerator of the probability is the total number of Early subjects, 141. Thus, $P(E) = 141/318 = .4434$. We may define marginal probability more generally as follows:

DEFINITION

Given some variable that can be broken down into m categories designated by $A_1, A_2, \ldots, A_i, \ldots, A_m$ and another jointly occurring variable that is broken down into n categories designated by $B_1, B_2, \ldots, B_j, \ldots, B_n$, the marginal probability of A_i, $P(A_i)$, is equal to the sum of the joint probabilities of A_i with all the categories of B. That is,

$$P(A_i) = \Sigma P(A_i \cap B_j), \qquad \text{for all values of } j \tag{3.4.6}$$

The following example illustrates the use of Equation 3.4.6 in the calculation of a marginal probability.

EXAMPLE 3.4.9

We wish to use Equation 3.4.6 and the data in Table 3.4.1 to compute the marginal probability $P(E)$.

Solution: The variable age at onset is broken down into two categories, Early for onset 18 years or younger (E) and Later for onset occurring at an age over 18 years (L). The variable family history of mood disorders is broken down into four categories: negative family history (A), bipolar disorder only (B), unipolar disorder only (C), and subjects with a history of both unipolar and bipolar disorder (D). The category Early occurs jointly with all four categories of the variable family history of mood disorders. The four joint probabilities that may be computed are

$$P(E \cap A) = 28/318 = .0881$$
$$P(E \cap B) = 19/318 = .0597$$
$$P(E \cap C) = 41/318 = .1289$$
$$P(E \cap D) = 53/318 = .1667$$

We obtain the marginal probability $P(E)$ by adding these four joint probabilities as follows:

$$\begin{aligned} P(E) &= P(E \cap A) + P(E \cap B) + P(E \cap C) + P(E \cap D) \\ &= .0881 + .0597 + .1289 + .1667 \\ &= .4434 \end{aligned}$$

■

The result, as expected, is the same as the one obtained by using the marginal total for Early as the numerator and the total number of subjects as the denominator.

EXERCISES

3.4.1 In a study of violent victimization of women and men, Porcerelli et al. (A-2) collected information from 679 women and 345 men aged 18 to 64 years at several family practice centers in the metropolitan Detroit area. Patients filled out a health history questionnaire that included a question about victimization. The following table shows the sample subjects cross-classified by sex and the type of violent victimization reported. The victimization categories are defined as no victimization, partner victimization (and not by others), victimization by persons other than partners (friends, family members, or strangers), and those who reported multiple victimization.

	No Victimization	Partners	Nonpartners	Multiple Victimization	Total
Women	611	34	16	18	679
Men	308	10	17	10	345
Total	919	44	33	28	1024

Source: Data provided courtesy of John H. Porcerelli, Ph.D., Rosemary Cogan, Ph.D.

(a) Suppose we pick a subject at random from this group. What is the probability that this subject will be a woman?

(b) What do we call the probability calculated in part a?

(c) Show how to calculate the probability asked for in part a by two additional methods.

2.3 GROUPED DATA: THE FREQUENCY DISTRIBUTION

Although a set of observations can be made more comprehensible and meaningful by means of an ordered array, further useful summarization may be achieved by grouping the data. Before the days of computers one of the main objectives in grouping large data sets was to facilitate the calculation of various descriptive measures such as percentages and averages. Because computers can perform these calculations on large data sets without first grouping the data, the main purpose in grouping data now is summarization. One must bear in mind that data contain information and that summarization is a way of making it easier to determine the nature of this information. One must also be aware that reducing a large quantity of information in order to summarize the data succinctly carries with it the potential to inadvertently lose some amount of specificity with regard to the underlying data set. Therefore, it is important to group the data sufficiently such that the vast amounts of information are reduced into understandable summaries. At the same time data should be summarized to the extent that useful intricacies in the data are not readily obvious.

To group a set of observations we select a set of contiguous, nonoverlapping intervals such that each value in the set of observations can be placed in one, and only one, of the intervals. These intervals are usually referred to as *class intervals*.

One of the first considerations when data are to be grouped is how many intervals to include. Too few intervals are undesirable because of the resulting loss of information. On the other hand, if too many intervals are used, the objective of summarization will not be met. The best guide to this, as well as to other decisions to be made in grouping data, is your knowledge of the data. It may be that class intervals have been determined by precedent, as in the case of annual tabulations, when the class intervals of previous years are maintained for comparative purposes. A commonly followed rule of thumb states that there should be no fewer than five intervals and no more than 15. If there are fewer than five intervals, the data have been summarized too much and the information they contain has been lost. If there are more than 15 intervals, the data have not been summarized enough.

Those who need more specific guidance in the matter of deciding how many class intervals to employ may use a formula given by Sturges (1). This formula gives $k = 1 + 3.322(\log_{10} n)$, where k stands for the number of class intervals and n is the number of values in the data set under consideration. The answer obtained by applying *Sturges's rule* should not be regarded as final, but should be considered as a guide only. The number of class intervals specified by the rule should be increased or decreased for convenience and clear presentation.

Suppose, for example, that we have a sample of 275 observations that we want to group. The logarithm to the base 10 of 275 is 2.4393. Applying Sturges's formula gives $k = 1 + 3.322(2.4393) \simeq 9$. In practice, other considerations might cause us to use eight or fewer or perhaps 10 or more class intervals.

Another question that must be decided regards the width of the class intervals. Class intervals generally should be of the same width, although this is sometimes impossible to accomplish. This width may be determined by dividing the range by k, the number of class intervals. Symbolically, the class interval width is given by

$$w = \frac{R}{k} \tag{2.3.1}$$

where R (the range) is the difference between the smallest and the largest observation in the data set, and k is defined as above. As a rule this procedure yields a width that is inconvenient for use. Again, we may exercise our good judgment and select a width (usually close to one given by Equation 2.3.1) that is more convenient.

There are other rules of thumb that are helpful in setting up useful class intervals. When the nature of the data makes them appropriate, class interval widths of 5 units, 10 units, and widths that are multiples of 10 tend to make the summarization more comprehensible. When these widths are employed it is generally good practice to have the lower limit of each interval end in a zero or 5. Usually class intervals are ordered from smallest to largest; that is, the first class interval contains the smaller measurements and the last class interval contains the larger measurements. When this is the case, the lower limit of the first class interval should be equal to or smaller than the smallest measurement in the data set, and the upper limit of the last class interval should be equal to or greater than the largest measurement.

Most statistical packages allow users to interactively change the number of class intervals and/or the class widths, so that several visualizations of the data can be obtained quickly. This feature allows users to exercise their judgment in deciding which data display is most appropriate for a given purpose. Let us use the 189 ages shown in Table 1.4.1 and arrayed in Table 2.2.1 to illustrate the construction of a frequency distribution.

EXAMPLE 2.3.1

We wish to know how many class intervals to have in the frequency distribution of the data. We also want to know how wide the intervals should be.

Solution: To get an idea as to the number of class intervals to use, we can apply Sturges's rule to obtain

$$\begin{aligned} k &= 1 + 3.322(\log 189) \\ &= 1 + 3.322(2.2764618) \\ &\approx 9 \end{aligned}$$

Now let us divide the range by 9 to get some idea about the class interval width. We have

$$\frac{R}{k} = \frac{82 - 30}{9} = \frac{52}{9} = 5.778$$

It is apparent that a class interval width of 5 or 10 will be more convenient to use, as well as more meaningful to the reader. Suppose we decide on 10. We may now construct our intervals. Since the smallest value in Table 2.2.1 is 30 and the largest value is 82, we may begin our intervals with 30 and end with 89. This gives the following intervals:

30–39
40–49
50–59
60–69

70–79
80–89

We see that there are six of these intervals, three fewer than the number suggested by Sturges's rule.

It is sometimes useful to refer to the center, called the *midpoint,* of a class interval. The midpoint of a class interval is determined by obtaining the sum of the upper and lower limits of the class interval and dividing by 2. Thus, for example, the midpoint of the class interval 30–39 is found to be $(30 + 39)/2 = 34.5$. ■

When we group data manually, determining the number of values falling into each class interval is merely a matter of looking at the ordered array and counting the number of observations falling in the various intervals. When we do this for our example, we have Table 2.3.1.

A table such as Table 2.3.1 is called a *frequency distribution*. This table shows the way in which the values of the variable are distributed among the specified class intervals. By consulting it, we can determine the frequency of occurrence of values within any one of the class intervals shown.

Relative Frequencies It may be useful at times to know the proportion, rather than the number, of values falling within a particular class interval. We obtain this information by dividing the number of values in the particular class interval by the total number of values. If, in our example, we wish to know the proportion of values between 50 and 59, inclusive, we divide 70 by 189, obtaining .3704. Thus we say that 70 out of 189, or 70/189ths, or .3704, of the values are between 50 and 59. Multiplying .3704 by 100 gives us the percentage of values between 50 and 59. We can say, then, that 37.04 percent of the subjects are between 50 and 59 years of age. We may refer to the proportion of values falling within a class interval as the *relative frequency of occurrence* of values in that interval. In Section 3.2 we shall see that a relative frequency may be interpreted also as the probability of occurrence within the given interval. This probability of occurrence is also called the *experimental probability* or the *empirical probability.*

TABLE 2.3.1 Frequency Distribution of Ages of 189 Subjects Shown in Tables 1.4.1 and 2.2.1

Class Interval	Frequency
30–39	11
40–49	46
50–59	70
60–69	45
70–79	16
80–89	1
Total	189

TABLE 2.3.2 Frequency, Cumulative Frequency, Relative Frequency, and Cumulative Relative Frequency Distributions of the Ages of Subjects Described in Example 1.4.1

Class Interval	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
30–39	11	11	.0582	.0582
40–49	46	57	.2434	.3016
50–59	70	127	.3704	.6720
60–69	45	172	.2381	.9101
70–79	16	188	.0847	.9948
80–89	1	189	.0053	1.0001
Total	189		1.0001	

Note: Frequencies do not add to 1.0000 exactly because of rounding.

In determining the frequency of values falling within two or more class intervals, we obtain the sum of the number of values falling within the class intervals of interest. Similarly, if we want to know the relative frequency of occurrence of values falling within two or more class intervals, we add the respective relative frequencies. We may sum, or *cumulate,* the frequencies and relative frequencies to facilitate obtaining information regarding the frequency or relative frequency of values within two or more contiguous class intervals. Table 2.3.2 shows the data of Table 2.3.1 along with the *cumulative frequencies,* the *relative frequencies,* and *cumulative relative frequencies*.

Suppose that we are interested in the relative frequency of values between 50 and 79. We use the cumulative relative frequency column of Table 2.3.2 and subtract .3016 from .9948, obtaining .6932.

We may use a statistical package to obtain a table similar to that shown in Table 2.3.2. Tables obtained from both MINITAB and SPSS software are shown in Figure 2.3.1.

The Histogram We may display a frequency distribution (or a relative frequency distribution) graphically in the form of a *histogram,* which is a special type of bar graph.

When we construct a histogram the values of the variable under consideration are represented by the horizontal axis, while the vertical axis has as its scale the frequency (or relative frequency if desired) of occurrence. Above each class interval on the horizontal axis a rectangular bar, or cell, as it is sometimes called, is erected so that the height corresponds to the respective frequency when the class intervals are of equal width. The cells of a histogram must be joined and, to accomplish this, we must take into account the true boundaries of the class intervals to prevent gaps from occurring between the cells of our graph.

The level of precision observed in reported data that are measured on a continuous scale indicates some order of rounding. The order of rounding reflects either the reporter's personal preference or the limitations of the measuring instrument employed. When a frequency distribution is constructed from the data, the class interval limits usually reflect the degree of precision of the raw data. This has been done in our illustrative example.

Dialog box:

Stat ➤ Tables ➤ Tally Individual Variables

Type *C2* in **Variables.** Check **Counts, Percents, Cumulative counts,** and **Cumulative percents** in **Display.** Click **OK.**

Session command:

```
MTB > Tally  C2;
SUBC>   Counts;
SUBC>   CumCounts;
SUBC>   Percents;
SUBC>   CumPercents;
```

Output:

Tally for Discrete Variables: C2

MINITAB Output

C2	Count	CumCnt	Percent	CumPct
0	11	11	5.82	5.82
1	46	57	24.34	30.16
2	70	127	37.04	67.20
3	45	172	23.81	91.01
4	16	188	8.47	99.47
5	1	189	0.53	100.00
N=	189			

SPSS Output

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	30-39	11	5.8	5.8	5.8
	40-49	46	24.3	24.3	30.2
	50-59	70	37.0	37.0	67.2
	60-69	45	23.8	23.8	91.0
	70-79	16	8.5	8.5	99.5
	80-89	1	.5	.5	100.0
	Total	189	100.0	100.0	

FIGURE 2.3.1 Frequency, cumulative frequencies, percent, and cumulative percent distribution of the ages of subjects described in Example 1.4.1 as constructed by MINITAB and SPSS.

We know, however, that some of the values falling in the second class interval, for example, when measured precisely, would probably be a little less than 40 and some would be a little greater than 49. Considering the underlying continuity of our variable, and assuming that the data were rounded to the nearest whole number, we find it convenient to think of 39.5 and 49.5 as the true limits of this second interval. The true limits for each of the class intervals, then, we take to be as shown in Table 2.3.3.

If we construct a graph using these class limits as the base of our rectangles, no gaps will result, and we will have the histogram shown in Figure 2.3.2. We used MINITAB to construct this histogram, as shown in Figure 2.3.3.

We refer to the space enclosed by the boundaries of the histogram as the *area* of the histogram. Each observation is allotted one unit of this area. Since we have 189 observations, the histogram consists of a total of 189 units. Each cell contains a certain proportion of the total area, depending on the frequency. The second cell, for example, contains 46/189 of the area. This, as we have learned, is the relative frequency of occurrence of values between 39.5 and 49.5. From this we see that subareas of the histogram defined by the cells correspond to the frequencies of occurrence of values between the horizontal scale boundaries of the areas. The ratio of a particular subarea to the total area of the histogram is equal to the relative frequency of occurrence of values between the corresponding points on the horizontal axis.

TABLE 2.3.3 The Data of Table 2.3.1 Showing True Class Limits

True Class Limits	Frequency
29.5–39.5	11
39.5–49.5	46
49.5–59.5	70
59.5–69.5	45
69.5–79.5	16
79.5–89.5	1
Total	189

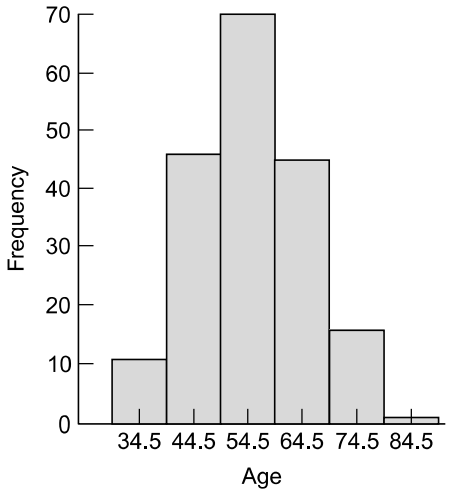

FIGURE 2.3.2 Histogram of ages of 189 subjects from Table 2.3.1.

The Frequency Polygon A frequency distribution can be portrayed graphically in yet another way by means of a *frequency polygon,* which is a special kind of line graph. To draw a frequency polygon we first place a dot above the midpoint of each class interval represented on the horizontal axis of a graph like the one shown in Figure 2.3.2. The height of a given dot above the horizontal axis corresponds to the frequency of the relevant class interval. Connecting the dots by straight lines produces the frequency polygon. Figure 2.3.4 is the frequency polygon for the age data in Table 2.2.1.

Note that the polygon is brought down to the horizontal axis at the ends at points that would be the midpoints if there were an additional cell at each end of the corresponding histogram. This allows for the total area to be enclosed. The total area under the frequency polygon is equal to the area under the histogram. Figure 2.3.5 shows the frequency polygon of Figure 2.3.4 superimposed on the histogram of Figure 2.3.2. This figure allows you to see, for the same set of data, the relationship between the two graphic forms.

Dialog box:

Graph ➤ Histogram ➤ Simple ➤ OK

Type *Age* in **Graph Variables:** Click **OK.**

Now double click the histogram and click **Binning** Tab.
Type 34.5:84.5/10 in **MidPoint/CutPoint positions:**
Click **OK.**

Session command:

```
MTB > Histogram 'Age';
SUBC>   MidPoint 34.5:84.5/10;
SUBC>   Bar.
```

FIGURE 2.3.3 MINITAB dialog box and session command for constructing histogram from data on ages in Example 1.4.1.

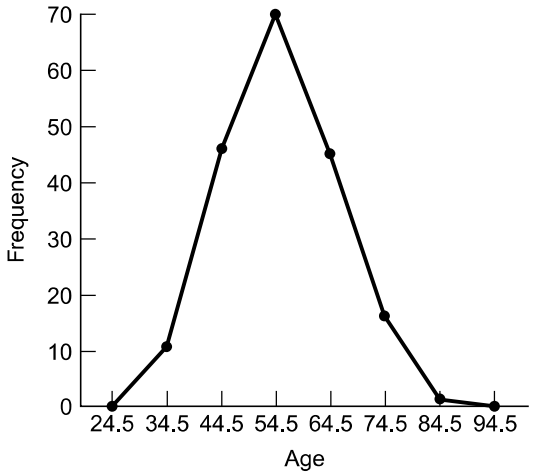

FIGURE 2.3.4 Frequency polygon for the ages of 189 subjects shown in Table 2.2.1.

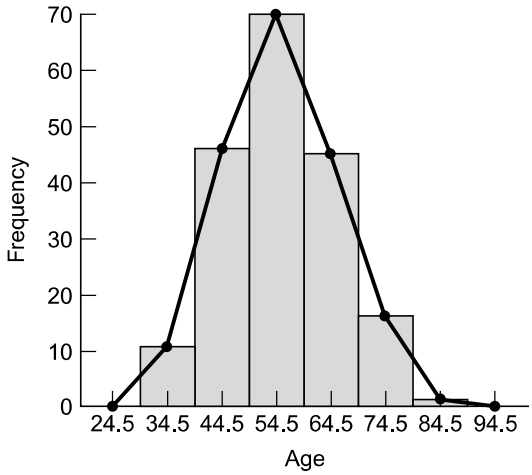

FIGURE 2.3.5 Histogram and frequency polygon for the ages of 189 subjects shown in Table 2.2.1.

Stem-and-Leaf Displays Another graphical device that is useful for representing quantitative data sets is the *stem-and-leaf display.* A stem-and-leaf display bears a strong resemblance to a histogram and serves the same purpose. A properly constructed stem-and-leaf display, like a histogram, provides information regarding the range of the data set, shows the location of the highest concentration of measurements, and reveals the presence or absence of symmetry. An advantage of the stem-and-leaf display over the histogram is the fact that it preserves the information contained in the individual measurements. Such information is lost when measurements are assigned to the class intervals of a histogram. As will become apparent, another advantage of stem-and-leaf displays is the fact that they can be constructed during the tallying process, so the intermediate step of preparing an ordered array is eliminated.

To construct a stem-and-leaf display we partition each measurement into two parts. The first part is called the *stem*, and the second part is called the *leaf*. The stem consists of one or more of the initial digits of the measurement, and the leaf is composed of one or more of the remaining digits. All partitioned numbers are shown together in a single display; the stems form an ordered column with the smallest stem at the top and the largest at the bottom. We include in the stem column all stems within the range of the data even when a measurement with that stem is not in the data set. The rows of the display contain the leaves, ordered and listed to the right of their respective stems. When leaves consist of more than one digit, all digits after the first may be deleted. Decimals when present in the original data are omitted in the stem-and-leaf display. The stems are separated from their leaves by a vertical line. Thus we see that a stem-and-leaf display is also an ordered array of the data.

Stem-and-leaf displays are most effective with relatively small data sets. As a rule they are not suitable for use in annual reports or other communications aimed at the general public. They are primarily of value in helping researchers and decision makers understand the nature of their data. Histograms are more appropriate for externally circulated publications. The following example illustrates the construction of a stem-and-leaf display.

```
Stem   Leaf
   3 | 04577888899
   4 | 00223333334444444555666666777777888888899999999
   5 | 000000001111222222333333333333333334444444444455566666677777788999999
   6 | 000011111111111222222233444444556666667888999
   7 | 0111111123567888
   8 | 2
```

FIGURE 2.3.6 Stem-and-leaf display of ages of 189 subjects shown in Table 2.2.1 (stem unit = 10, leaf unit = 1).

EXAMPLE 2.3.2

Let us use the age data shown in Table 2.2.1 to construct a stem-and-leaf display.

Solution: Since the measurements are all two-digit numbers, we will have one-digit stems and one-digit leaves. For example, the measurement 30 has a stem of 3 and a leaf of 0. Figure 2.3.6 shows the stem-and-leaf display for the data.

The MINITAB statistical software package may be used to construct stem-and-leaf displays. The MINITAB procedure and output are as shown in Figure 2.3.7. The increment subcommand specifies the distance from one stem to the next. The numbers in the leftmost output column of Figure 2.3.7

Dialog box:

Graph ➤ Stem-and-Leaf

Type *Age* in **Graph Variables.** Type *10* in **Increment.** Click **OK.**

Session command:

```
MTB > Stem-and-Leaf 'Age';
SUBC>   Increment 10.
```

Output:

Stem-and-Leaf Display: Age

```
Stem-and-leaf of Age    N = 189
Leaf Unit = 1.0

 11   3 04577888899
 57   4 00223333334444444555666666777777888888899999999
(70)  5 000000001111222222333333333333333334444444444455566666677777777889+
 62   6 000011111111111222222233444444556666667888999
 17   7 0111111123567888
  1   8 2
```

FIGURE 2.3.7 Stem-and-leaf display prepared by MINITAB from the data on subjects' ages shown in Table 2.2.1.

```
Stem-and-leaf of Age          N = 189
Leaf Unit = 1.0
  2     3 04
 11     3 577888899
 28     4 00223333334444444
 57     4 55566666677777788888889999999
(46)    5 00000000111122222233333333333333333344444444444
 86     5 55566666677777788999999
 62     6 00001111111111122222233444444
 32     6 556666667888999
 17     7 0111111123
  7     7 567888
  1     8 2
```

FIGURE 2.3.8 Stem-and-leaf display prepared by MINITAB from the data on subjects' ages shown in Table 2.2.1; class interval width = 5.

provide information regarding the number of observations (leaves) on a given line and above or the number of observations on a given line and below. For example, the number 57 on the second line shows that there are 57 observations (or leaves) on that line and the one above it. The number 62 on the fourth line from the top tells us that there are 62 observations on that line and all the ones below. The number in parentheses tells us that there are 70 observations on that line. The parentheses mark the line containing the middle observation if the total number of observations is odd or the two middle observations if the total number of observations is even.

The + at the end of the third line in Figure 2.3.7 indicates that the frequency for that line (age group 50 through 59) exceeds the line capacity, and that there is at least one additional leaf that is not shown. In this case, the frequency for the 50–59 age group was 70. The line contains only 65 leaves, so the + indicates that there are five more leaves, the number 9, that are not shown. ■

One way to avoid exceeding the capacity of a line is to have more lines. This is accomplished by making the distance between lines shorter, that is, by decreasing the widths of the class intervals. For the present example, we may use class interval widths of 5, so that the distance between lines is 5. Figure 2.3.8 shows the result when MINITAB is used to produce the stem-and-leaf display.

EXERCISES

2.3.1 In a study of the oral home care practice and reasons for seeking dental care among individuals on renal dialysis, Atassi (A-1) studied 90 subjects on renal dialysis. The oral hygiene status of all subjects was examined using a plaque index with a range of 0 to 3 (0 = no soft plaque deposits,

3 = an abundance of soft plaque deposits). The following table shows the plaque index scores for all 90 subjects.

1.17	2.50	2.00	2.33	1.67	1.33
1.17	2.17	2.17	1.33	2.17	2.00
2.17	1.17	2.50	2.00	1.50	1.50
1.00	2.17	2.17	1.67	2.00	2.00
1.33	2.17	2.83	1.50	2.50	2.33
0.33	2.17	1.83	2.00	2.17	2.00
1.00	2.17	2.17	1.33	2.17	2.50
0.83	1.17	2.17	2.50	2.00	2.50
0.50	1.50	2.00	2.00	2.00	2.00
1.17	1.33	1.67	2.17	1.50	2.00
1.67	0.33	1.50	2.17	2.33	2.33
1.17	0.00	1.50	2.33	1.83	2.67
0.83	1.17	1.50	2.17	2.67	1.50
2.00	2.17	1.33	2.00	2.33	2.00
2.17	2.17	2.00	2.17	2.00	2.17

Source: Data provided courtesy of Farhad Atassi, DDS, MSc, FICOI.

(a) Use these data to prepare:

A frequency distribution
A relative frequency distribution
A cumulative frequency distribution
A cumulative relative frequency distribution
A histogram
A frequency polygon

(b) What percentage of the measurements are less than 2.00?

(c) What proportion of the subjects have measurements greater than or equal to 1.50?

(d) What percentage of the measurements are between 1.50 and 1.99 inclusive?

(e) How many of the measurements are greater than 2.49?

(f) What proportion of the measurements are either less than 1.0 or greater than 2.49?

(g) Someone picks a measurement at random from this data set and asks you to guess the value. What would be your answer? Why?

(h) Frequency distributions and their histograms may be described in a number of ways depending on their shape. For example, they may be symmetric (the left half is at least approximately a mirror image of the right half), skewed to the left (the frequencies tend to increase as the measurements increase in size), skewed to the right (the frequencies tend to decrease as the measurements increase in size), or U-shaped (the frequencies are high at each end of the distribution and small in the center). How would you describe the present distribution?

2.3.2 Janardhan et al. (A-2) conducted a study in which they measured incidental intracranial aneurysms (IIAs) in 125 patients. The researchers examined postprocedural complications and concluded that IIAs can be safely treated without causing mortality and with a lower complications rate than previously reported. The following are the sizes (in millimeters) of the 159 IIAs in the sample.

8.1	10.0	5.0	7.0	10.0	3.0
20.0	4.0	4.0	6.0	6.0	7.0

(*Continued*)

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand the importance and basic principles of estimation.
2. be able to calculate interval estimates for a variety of parameters.
3. be able to interpret a confidence interval from both a practical and a probabilistic viewpoint.
4. understand the basic properties and uses of the *t* distribution, chi-square distribution, and *F* distribution.

6.1 INTRODUCTION

We come now to a consideration of *estimation,* the first of the two general areas of statistical inference. The second general area, *hypothesis testing,* is examined in the next chapter.

We learned in Chapter 1 that inferential statistics is defined as follows.

DEFINITION

***Statistical inference* is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample drawn from that population.**

The process of estimation entails calculating, from the data of a sample, some statistic that is offered as an approximation of the corresponding parameter of the population from which the sample was drawn.

The rationale behind estimation in the health sciences field rests on the assumption that workers in this field have an interest in the parameters, such as means and proportions, of various populations. If this is the case, there is a good reason why one must rely on estimating procedures to obtain information regarding these parameters. Many populations of interest, although finite, are so large that a 100 percent examination would be prohibitive from the standpoint of cost.

Suppose the administrator of a large hospital is interested in the mean age of patients admitted to his hospital during a given year. He may consider it too expensive to go through the records of all patients admitted during that particular year and, consequently, elect to examine a sample of the records from which he can compute an estimate of the mean age of patients admitted that year.

A physician in general practice may be interested in knowing what proportion of a certain type of individual, treated with a particular drug, suffers undesirable side effects. No doubt, her concept of the population consists of all those persons who ever have been or ever will be treated with this drug. Deferring a conclusion until the entire population has been observed could have an adverse effect on her practice.

These two examples have implied an interest in estimating, respectively, a population mean and a population proportion. Other parameters, the estimation of which we will cover in this chapter, are the difference between two means, the difference between two proportions, the population variance, and the ratio of two variances.

We will find that for each of the parameters we discuss, we can compute two types of estimate: a point estimate and an interval estimate.

DEFINITION

A *point estimate* is a single numerical value used to estimate the corresponding population parameter.

DEFINITION

An *interval estimate* consists of two numerical values defining a range of values that, with a specified degree of confidence, most likely includes the parameter being estimated.

These concepts will be elaborated on in the succeeding sections.

Choosing an Appropriate Estimator Note that a single computed value has been referred to as an *estimate*. The rule that tells us how to compute this value, or estimate, is referred to as an *estimator*. Estimators are usually presented as formulas. For example,

$$\bar{x} = \frac{\sum x_i}{n}$$

is an estimator of the population mean, μ. The single numerical value that results from evaluating this formula is called an estimate of the parameter μ.

In many cases, a parameter may be estimated by more than one estimator. For example, we could use the sample median to estimate the population mean. How then do we decide which estimator to use for estimating a given parameter? The decision is based on an objective measure or set of criteria that reflect some desired property of a particular estimator. When measured against these criteria, some estimators are better than others. One of these criteria is the property of *unbiasedness*.

DEFINITION

An estimator, say, T, of the parameter θ is said to be an unbiased *estimator* of θ if $E(T) = \theta$.

$E(T)$ is read, "the expected value of T." For a finite population, $E(T)$ is obtained by taking the average value of T computed from all possible samples of a given size that may be drawn from the population. That is, $E(T) = \mu_T$. For an infinite population, $E(T)$ is defined in terms of calculus.

In the previous chapter we have seen that the sample mean, the sample proportion, the difference between two sample means, and the difference between two sample proportions are each unbiased estimates of their corresponding parameters. This property was implied when the parameters were said to be the means of the respective sampling distributions. For example, since the mean of the sampling distribution of $\bar{x}$ is equal to μ, we know that $\bar{x}$ is an unbiased estimator of μ. The other criteria of good estimators will not

Thus, the probability of a continuous random variable to assume values between a and b is denoted by $P(a < X < b)$.

4.6 THE NORMAL DISTRIBUTION

We come now to the most important distribution in all of statistics—the *normal distribution*. The formula for this distribution was first published by Abraham De Moivre (1667–1754) on November 12, 1733. Many other mathematicians figure prominently in the history of the normal distribution, including Carl Friedrich Gauss (1777–1855). The distribution is frequently called the *Gaussian distribution* in recognition of his contributions.

The normal density is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty \tag{4.6.1}$$

In Equation 4.6.1, π and e are the familiar constants, 3.14159 . . . and 2.71828 . . . , respectively, which are frequently encountered in mathematics. The two parameters of the distribution are μ, the mean, and σ, the standard deviation. For our purposes we may think of μ and σ of a normal distribution, respectively, as measures of central tendency and dispersion as discussed in Chapter 2. Since, however, a normally distributed random variable is continuous and takes on values between $-\infty$ and $+\infty$, its mean and standard deviation may be more rigorously defined; but such definitions cannot be given without using calculus. The graph of the normal distribution produces the familiar bell-shaped curve shown in Figure 4.6.1.

Characteristics of the Normal Distribution The following are some important characteristics of the normal distribution.

1. It is symmetrical about its mean, μ. As is shown in Figure 4.6.1, the curve on either side of μ is a mirror image of the other side.
2. The mean, the median, and the mode are all equal.
3. The total area under the curve above the x-axis is one square unit. This characteristic follows from the fact that the normal distribution is a probability distribution. Because of the symmetry already mentioned, 50 percent of the area is to the right of a perpendicular erected at the mean, and 50 percent is to the left.

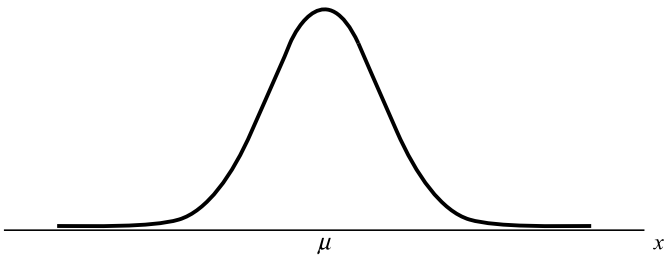

FIGURE 4.6.1 Graph of a normal distribution.

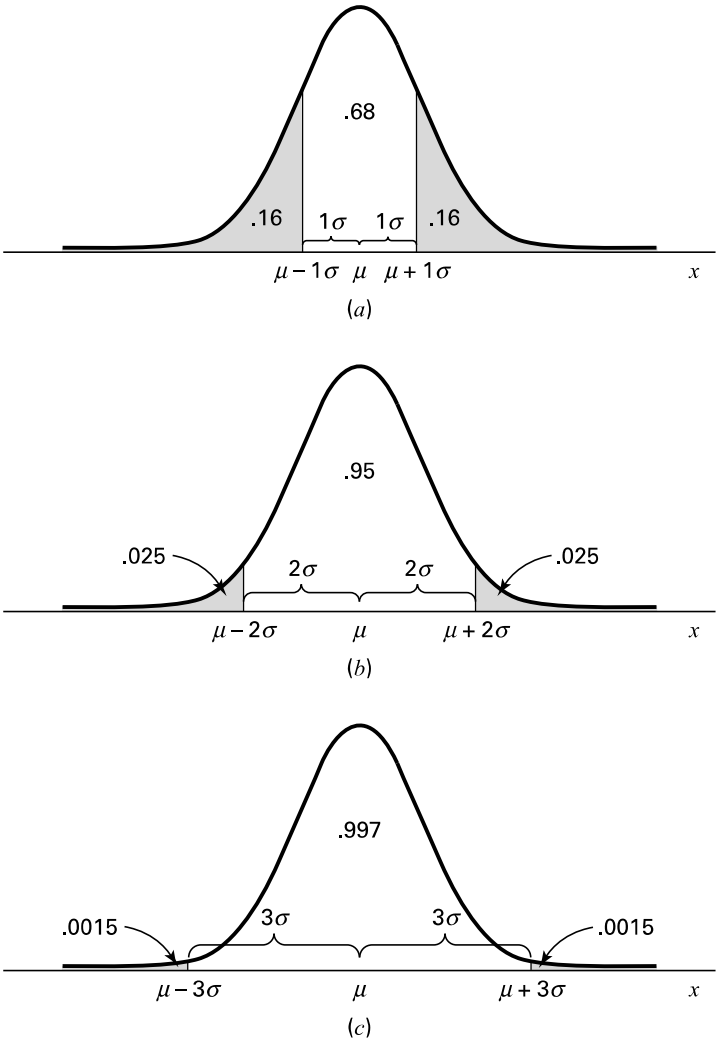

FIGURE 4.6.2 Subdivision of the area under the normal curve (areas are approximate).

4. If we erect perpendiculars a distance of 1 standard deviation from the mean in both directions, the area enclosed by these perpendiculars, the x-axis, and the curve will be approximately 68 percent of the total area. If we extend these lateral boundaries a distance of two standard deviations on either side of the mean, approximately 95 percent of the area will be enclosed, and extending them a distance of three standard deviations will cause approximately 99.7 percent of the total area to be enclosed. These approximate areas are illustrated in Figure 4.6.2.
5. The normal distribution is completely determined by the parameters μ and σ. In other words, a different normal distribution is specified for each different value of μ and σ. Different values of μ shift the graph of the distribution along the x-axis as is shown in Figure 4.6.3. Different values of σ determine the degree of flatness or peakedness of the graph of the distribution as is shown in Figure 4.6.4. Because of the characteristics of these two parameters, μ is often referred to as a *location parameter* and σ is often referred to as a *shape parameter.*

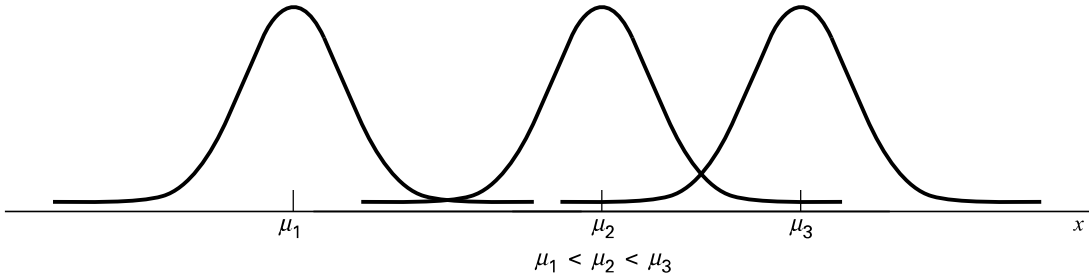

FIGURE 4.6.3 Three normal distributions with different means but the same amount of variability.

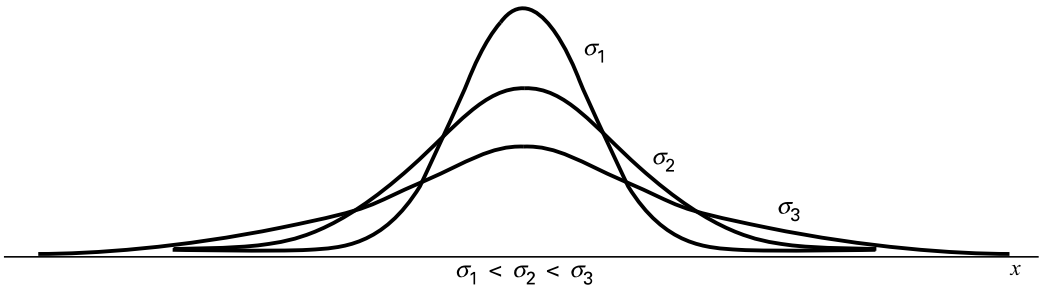

FIGURE 4.6.4 Three normal distributions with different standard deviations but the same mean.

The Standard Normal Distribution The last-mentioned characteristic of the normal distribution implies that the normal distribution is really a family of distributions in which one member is distinguished from another on the basis of the values of μ and σ. The most important member of this family is the *standard normal distribution* or *unit normal distribution,* as it is sometimes called, because it has a mean of 0 and a standard deviation of 1. It may be obtained from Equation 4.6.1 by creating a random variable.

$$z = (x - \mu)/\sigma \tag{4.6.2}$$

The equation for the standard normal distribution is written

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty \tag{4.6.3}$$

The graph of the standard normal distribution is shown in Figure 4.6.5.

The z-transformation will prove to be useful in the examples and applications that follow. This value of z denotes, for a value of a random variable, the number of standard deviations that value falls above $(+z)$ or below $(-z)$ the mean, which in this case is 0. For example, a z-transformation that yields a value of $z = 1$ indicates that the value of x used in the transformation is 1 standard deviation above 0. A value of $z = -1$ indicates that the value of x used in the transformation is 1 standard deviation below 0. This property is illustrated in the examples of Section 4.7.

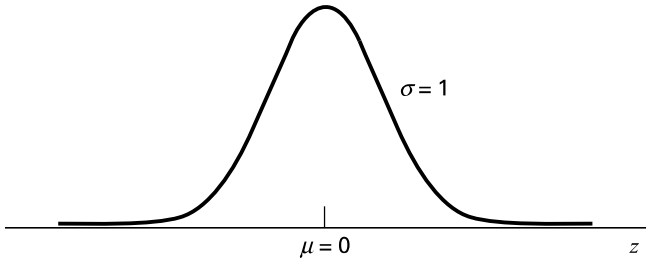

FIGURE 4.6.5 The standard normal distribution.

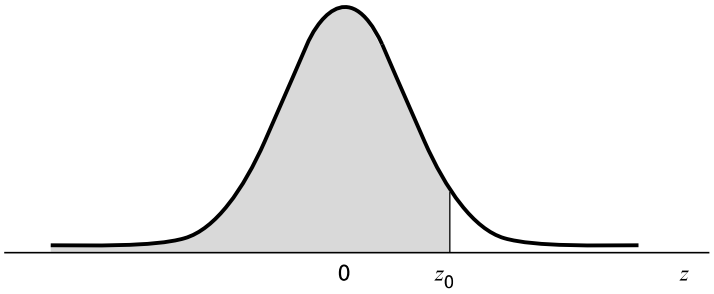

FIGURE 4.6.6 Area given by Appendix Table D.

To find the probability that z takes on a value between any two points on the z-axis, say, z_0 and z_1, we must find the area bounded by perpendiculars erected at these points, the curve, and the horizontal axis. As we mentioned previously, areas under the curve of a continuous distribution are found by integrating the function between two values of the variable. In the case of the standard normal, then, to find the area between z_0 and z_1 directly, we would need to evaluate the following integral:

$$\int_{z_0}^{z_1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

Although a closed-form solution for the integral does not exist, we can use numerical methods of calculus to approximate the desired areas beneath the curve to a desired accuracy. Fortunately, we do not have to concern ourselves with such matters, since there are tables available that provide the results of any integration in which we might be interested. Table D in the Appendix is an example of these tables. In the body of Table D are found the areas under the curve between $-\infty$ and the values of z shown in the leftmost column of the table. The shaded area of Figure 4.6.6 represents the area listed in the table as being between $-\infty$ and z_0, where z_0 is the specified value of z.

We now illustrate the use of Table D by several examples.

EXAMPLE 4.6.1

Given the standard normal distribution, find the area under the curve, above the z-axis between $z = -\infty$ and $z = 2$.

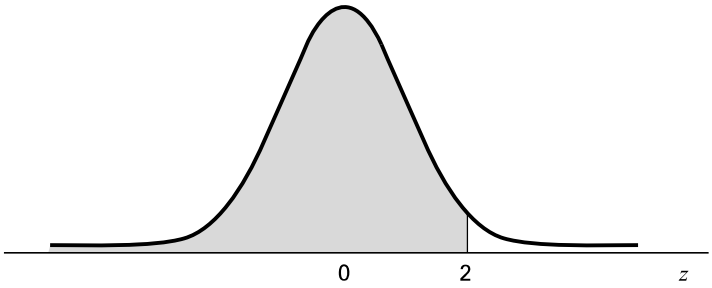

FIGURE 4.6.7 The standard normal distribution showing area between $z = -\infty$ and $z = 2$.

Solution: It will be helpful to draw a picture of the standard normal distribution and shade the desired area, as in Figure 4.6.7. If we locate $z = 2$ in Table D and read the corresponding entry in the body of the table, we find the desired area to be .9772. We may interpret this area in several ways. We may interpret it as the probability that a z picked at random from the population of z's will have a value between $-\infty$ and 2. We may also interpret it as the relative frequency of occurrence (or proportion) of values of z between $-\infty$ and 2, or we may say that 97.72 percent of the z's have a value between $-\infty$ and 2. ■

EXAMPLE 4.6.2

What is the probability that a z picked at random from the population of z's will have a value between -2.55 and $+2.55$?

Solution: Figure 4.6.8 shows the area desired. Table D gives us the area between $-\infty$ and 2.55, which is found by locating 2.5 in the leftmost column of the table and then moving across until we come to the entry in the column headed by 0.05. We find this area to be .9946. If we look at the picture we draw, we see that this is more area than is desired. We need to subtract from .9946 the area to the left of -2.55. Reference to Table D shows that the area to the left of -2.55 is .0054. Thus the desired probability is

$$P(-2.55 < z < 2.55) = .9946 - .0054 = .9892$$

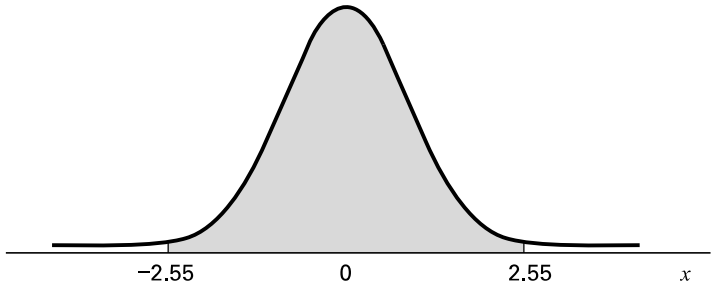

FIGURE 4.6.8 Standard normal curve showing $P(-2.55 < z < 2.55)$.

■

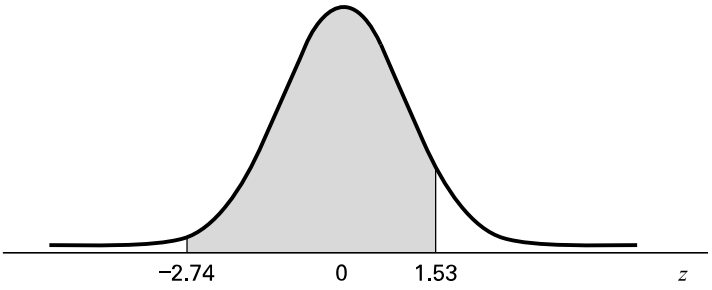

FIGURE 4.6.9 Standard normal curve showing proportion of z values between $z = -2.74$ and $z = 1.53$.

Suppose we had been asked to find the probability that z is between -2.55 and 2.55 inclusive. The desired probability is expressed as $P(-2.55 \leq z \leq 2.55)$. Since, as we noted in Section 4.5, $P(z = z_0) = 0, P(-2.55 \leq z \leq 2.55) = P(-2.55 < z < 2.55) = .9892$.

EXAMPLE 4.6.3

What proportion of z values are between -2.74 and 1.53?

Solution: Figure 4.6.9 shows the area desired. We find in Table D that the area between $-\infty$ and 1.53 is .9370, and the area between $-\infty$ and -2.74 is .0031. To obtain the desired probability we subtract .0031 from .9370. That is,

$$P(-2.74 \leq z \leq 1.53) = .9370 - .0031 = .9339$$

■

EXAMPLE 4.6.4

Given the standard normal distribution, find $P(z \geq 2.71)$.

Solution: The area desired is shown in Figure 4.6.10. We obtain the area to the right of $z = 2.71$ by subtracting the area between $-\infty$ and 2.71 from 1. Thus,

$$\begin{aligned} P(z \geq 2.71) &= 1 - P(z \leq 2.71) \\ &= 1 - .9966 \\ &= .0034 \end{aligned}$$

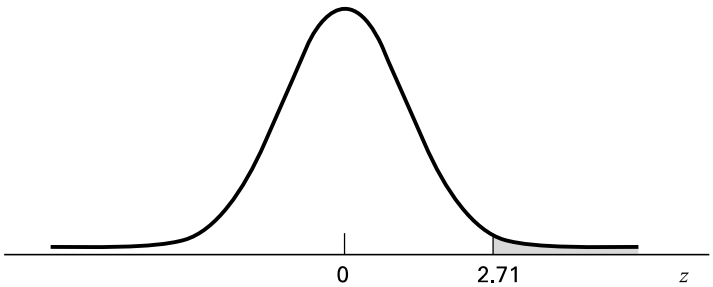

FIGURE 4.6.10 Standard normal distribution showing $P(z \geq 2.71)$.

■

randomly selected material, very little research of this type would be conducted. Again, nonstatistical considerations must play a part in the generalization process. Researchers may contend that the samples actually used are equivalent to simple random samples, since there is no reason to believe that the material actually used is not representative of the population about which inferences are desired.

In many health research projects, samples of convenience, rather than random samples, are employed. Researchers may have to rely on volunteer subjects or on readily available subjects such as students in their classes. Samples obtained from such sources are examples of *convenience samples*. Again, generalizations must be made on the basis of nonstatistical considerations. The consequences of such generalizations, however, may be useful or they may range from misleading to disastrous.

In some situations it is possible to introduce randomization into an experiment even though available subjects are not randomly selected from some well-defined population. In comparing two treatments, for example, each subject may be randomly assigned to one or the other of the treatments. Inferences in such cases apply to the treatments and not the subjects, and hence the inferences are valid.

6.2 CONFIDENCE INTERVAL FOR A POPULATION MEAN

Suppose researchers wish to estimate the mean of some normally distributed population. They draw a random sample of size n from the population and compute $\bar{x}$, which they use as a point estimate of μ. Although this estimator of μ possesses all the qualities of a good estimator, we know that because random sampling inherently involves chance, $\bar{x}$ cannot be expected to be equal to μ.

It would be much more meaningful, therefore, to estimate μ by an interval that somehow communicates information regarding the probable magnitude of μ.

Sampling Distributions and Estimation To obtain an interval estimate, we must draw on our knowledge of sampling distributions. In the present case, because we are concerned with the sample mean as an estimator of a population mean, we must recall what we know about the sampling distribution of the sample mean.

In the previous chapter we learned that if sampling is from a normally distributed population, the sampling distribution of the sample mean will be normally distributed with a mean $\mu_{\bar{x}}$ equal to the population mean μ, and a variance $\sigma_{\bar{x}}^2$ equal to σ^2/n. We could plot the sampling distribution if we only knew where to locate it on the $\bar{x}$-axis. From our knowledge of normal distributions, in general, we know even more about the distribution of $\bar{x}$ in this case. We know, for example, that regardless of where the distribution of $\bar{x}$ is located, approximately 95 percent of the possible values of $\bar{x}$ constituting the distribution are within two standard deviations of the mean. The two points that are two standard deviations from the mean are $\mu - 2\sigma_{\bar{x}}$ and $\mu + 2\sigma_{\bar{x}}$, so that the interval $\mu \pm 2\sigma_{\bar{x}}$ will contain approximately 95 percent of the possible values of $\bar{x}$. We know that μ and, hence $\mu_{\bar{x}}$, are unknown, but we may arbitrarily place the sampling distribution of $\bar{x}$ on the $\bar{x}$-axis.

Since we do not know the value of μ, not a great deal is accomplished by the expression $\mu \pm 2\sigma_{\bar{x}}$. We do, however, have a point estimate of μ, which is $\bar{x}$. Would it be

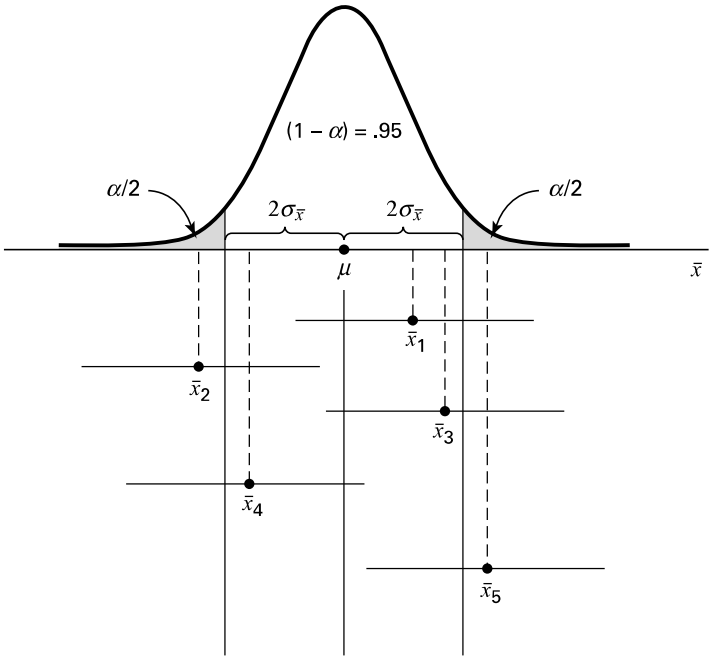

FIGURE 6.2.1 The 95 percent confidence interval for μ.

useful to construct an interval about this point estimate of μ? The answer is yes. Suppose we constructed intervals about every possible value of $\bar{x}$ computed from all possible samples of size n from the population of interest. We would have a large number of intervals of the form $\bar{x} \pm 2\sigma_{\bar{x}}$ with widths all equal to the width of the interval about the unknown μ. Approximately 95 percent of these intervals would have centers falling within the $\pm 2\sigma_{\bar{x}}$ interval about μ. Each of the intervals whose centers fall within $2\sigma_{\bar{x}}$ of μ would contain μ. These concepts are illustrated in Figure 6.2.1, in which we see that $\bar{x}, \bar{x}_3$, and $\bar{x}_4$ all fall within the interval about μ, and, consequently, the $2\sigma_{\bar{x}}$ intervals about these sample means include the value of μ. The sample means $\bar{x}_2$ and $\bar{x}_5$ do not fall within the $2\sigma_{\bar{x}}$ interval about μ, and the $2\sigma_{\bar{x}}$ intervals about them do not include μ.

EXAMPLE 6.2.1

Suppose a researcher, interested in obtaining an estimate of the average level of some enzyme in a certain human population, takes a sample of 10 individuals, determines the level of the enzyme in each, and computes a sample mean of $\bar{x} = 22$. Suppose further it is known that the variable of interest is approximately normally distributed with a variance of 45. We wish to estimate μ.

Solution: An approximate 95 percent confidence interval for μ is given by

$$\bar{x} \pm 2\sigma_{\bar{x}}$$

$$22 \pm 2\sqrt{\frac{45}{10}}$$

$$22 \pm 2(2.1213)$$

$$17.76, 26.24$$

■

Interval Estimate Components Let us examine the composition of the interval estimate constructed in Example 6.2.1. It contains in its center the point estimate of μ. The 2 we recognize as a value from the standard normal distribution that tells us within how many standard errors lie approximately 95 percent of the possible values of $\bar{x}$. This value of z is referred to as the *reliability coefficient*. The last component, $\sigma_{\bar{x}}$, is the standard error, or standard deviation of the sampling distribution of $\bar{x}$. In general, then, an interval estimate may be expressed as follows:

$$\text{estimator} \pm (\text{reliability coefficient}) \times (\text{standard error}) \tag{6.2.1}$$

In particular, when sampling is from a normal distribution with known variance, an interval estimate for μ may be expressed as

$$\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}} \tag{6.2.2}$$

where $z_{(1-\alpha/2)}$ is the value of z to the left of which lies $1 - \alpha/2$ and to the right of which lies $\alpha/2$ of the area under its curve.

Interpreting Confidence Intervals How do we interpret the interval given by Expression 6.2.2? In the present example, where the reliability coefficient is equal to 2, we say that in repeated sampling approximately 95 percent of the intervals constructed by Expression 6.2.2 will include the population mean. This interpretation is based on the probability of occurrence of different values of $\bar{x}$. We may generalize this interpretation if we designate the total area under the curve of $\bar{x}$ that is outside the interval $\mu \pm 2\sigma_{\bar{x}}$ as α and the area within the interval as $1 - \alpha$ and give the following *probabilistic interpretation* of Expression 6.2.2.

Probabilistic Interpretation

In repeated sampling, from a normally distributed population with a known standard deviation, $100(1 - \alpha)$ *percent of all intervals of the form* $\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$ *will in the long run include the population mean* μ.

The quantity $1 - \alpha$, in this case .95, is called the *confidence coefficient* (or confidence level), and the interval $\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$ is called a *confidence interval* for μ. When $(1 - \alpha) = .95$, the interval is called the 95 percent confidence interval for μ. In the present example we say that we are 95 percent confident that the population mean is between 17.76 and 26.24. This is called the *practical interpretation* of Expression 6.2.2. In general, it may be expressed as follows.

Practical Interpretation

When sampling is from a normally distributed population with known standard deviation, we are $100(1 - \alpha)$ *percent confident that the single computed interval,* $\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$, *contains the population mean* μ.

In the example given here we might prefer, rather than 2, the more exact value of z, 1.96, corresponding to a confidence coefficient of .95. Researchers may use any confidence coefficient they wish; the most frequently used values are .90, .95, and .99, which have associated reliability factors, respectively, of 1.645, 1.96, and 2.58.

Precision The quantity obtained by multiplying the reliability factor by the standard error of the mean is called the *precision* of the estimate. This quantity is also called the *margin of error.*

EXAMPLE 6.2.2

A physical therapist wished to estimate, with 99 percent confidence, the mean maximal strength of a particular muscle in a certain group of individuals. He is willing to assume that strength scores are approximately normally distributed with a variance of 144. A sample of 15 subjects who participated in the experiment yielded a mean of 84.3.

Solution: The z value corresponding to a confidence coefficient of .99 is found in Appendix Table D to be 2.58. This is our reliability coefficient. The standard error is $\sigma_{\bar{x}} = 12/\sqrt{15} = 3.0984$. Our 99 percent confidence interval for μ, then, is

$$84.3 \pm 2.58(3.0984)$$

$$84.3 \pm 8.0$$

$$76.3, 92.3$$

We say we are 99 percent confident that the population mean is between 76.3 and 92.3 since, in repeated sampling, 99 percent of all intervals that could be constructed in the manner just described would include the population mean. ■

Situations in which the variable of interest is approximately normally distributed with a known variance are quite rare. The purpose of the preceding examples, which assumed that these ideal conditions existed, was to establish the theoretical background for constructing confidence intervals for population means. In most practical situations either the variables are not approximately normally distributed or the population variances are not known or both. Example 6.2.3 and Section 6.3 explain the procedures that are available for use in the less than ideal, but more common, situations.

Sampling from Nonnormal Populations As noted, it will not always be possible or prudent to assume that the population of interest is normally distributed. Thanks to the central limit theorem, this will not deter us if we are able to select a large enough sample. We have learned that for large samples, the sampling distribution of $\bar{x}$ is approximately normally distributed regardless of how the parent population is distributed.

EXAMPLE 6.2.3

Punctuality of patients in keeping appointments is of interest to a research team. In a study of patient flow through the offices of general practitioners, it was found that a sample of 35 patients was 17.2 minutes late for appointments, on the average. Previous research had shown the standard deviation to be about 8 minutes. The population distribution was felt to be nonnormal. What is the 90 percent confidence interval for μ, the true mean amount of time late for appointments?