

CHAPTER 2

Simple Comparative Experiments

CHAPTER OUTLINE

- 2.1 INTRODUCTION
 - 2.2 BASIC STATISTICAL CONCEPTS
 - 2.3 SAMPLING AND SAMPLING DISTRIBUTIONS
 - 2.4 INFERENCES ABOUT THE DIFFERENCES IN MEANS, RANDOMIZED DESIGNS
 - 2.4.1 Hypothesis Testing
 - 2.4.2 Confidence Intervals
 - 2.4.3 Choice of Sample Size
 - 2.4.4 The Case Where $\sigma_1^2 \neq \sigma_2^2$
 - 2.4.5 The Case Where σ_1^2 and σ_2^2 Are Known
 - 2.4.6 Comparing a Single Mean to a Specified Value
 - 2.4.7 Summary
 - 2.5 INFERENCES ABOUT THE DIFFERENCES IN MEANS, PAIRED COMPARISON DESIGNS
 - 2.5.1 The Paired Comparison Problem
 - 2.5.2 Advantages of the Paired Comparison Design
 - 2.6 INFERENCES ABOUT THE VARIANCES OF NORMAL DISTRIBUTIONS
- SUPPLEMENTAL MATERIAL FOR CHAPTER 2
- S2.1 Models for the Data and the t -Test
 - S2.2 Estimating the Model Parameters
 - S2.3 A Regression Model Approach to the t -Test
 - S2.4 Constructing Normal Probability Plots
 - S2.5 More about Checking Assumptions in the t -Test
 - S2.6 Some More Information about the Paired t -Test

The supplemental material is on the textbook website www.wiley.com/college/montgomery.

In this chapter, we consider experiments to compare two **conditions** (sometimes called **treatments**). These are often called **simple comparative experiments**. We begin with an example of an experiment performed to determine whether two different formulations of a product give equivalent results. The discussion leads to a review of several basic statistical concepts, such as random variables, probability distributions, random samples, sampling distributions, and tests of hypotheses.

2.1 Introduction

An engineer is studying the formulation of a Portland cement mortar. He has added a polymer latex emulsion during mixing to determine if this impacts the curing time and tension bond strength of the mortar. The experimenter prepared 10 samples of the original formulation and 10 samples of the modified formulation. We will refer to the two different formulations as two **treatments** or as two **levels** of the **factor** formulations. When the cure process

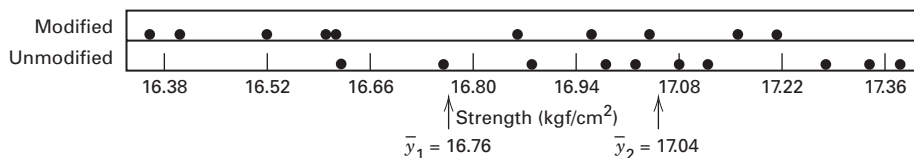
■ **TABLE 2.1**
**Tension Bond Strength Data for the Portland
 Cement Formulation Experiment**

| j | Modified Mortar y_{1j} | Unmodified Mortar y_{2j} |
|-----|--------------------------------|----------------------------------|
| 1 | 16.85 | 16.62 |
| 2 | 16.40 | 16.75 |
| 3 | 17.21 | 17.37 |
| 4 | 16.35 | 17.12 |
| 5 | 16.52 | 16.98 |
| 6 | 17.04 | 16.87 |
| 7 | 16.96 | 17.34 |
| 8 | 17.15 | 17.02 |
| 9 | 16.59 | 17.08 |
| 10 | 16.57 | 17.27 |

was completed, the experimenter did find a very large reduction in the cure time for the modified mortar formulation. Then he began to address the tension bond strength of the mortar. If the new mortar formulation has an adverse effect on bond strength, this could impact its usefulness.

The tension bond strength data from this experiment are shown in Table 2.1 and plotted in Figure 2.1. The graph is called a **dot diagram**. Visual examination of these data gives the impression that the strength of the unmodified mortar may be greater than the strength of the modified mortar. This impression is supported by comparing the *average* tension bond strengths, $\bar{y}_1 = 16.76 \text{ kgf/cm}^2$ for the modified mortar and $\bar{y}_2 = 17.04 \text{ kgf/cm}^2$ for the unmodified mortar. The average tension bond strengths in these two samples differ by what seems to be a modest amount. However, it is not obvious that this difference is large enough to imply that the two formulations really *are* different. Perhaps this observed difference in average strengths is the result of sampling fluctuation and the two formulations are really identical. Possibly another two samples would give opposite results, with the strength of the modified mortar exceeding that of the unmodified formulation.

A technique of statistical inference called **hypothesis testing** can be used to assist the experimenter in comparing these two formulations. Hypothesis testing allows the comparison of the two formulations to be made on **objective** terms, with knowledge of the risks associated with reaching the wrong conclusion. Before presenting procedures for hypothesis testing in simple comparative experiments, we will briefly summarize some elementary statistical concepts.



■ **FIGURE 2.1** Dot diagram for the tension bond strength data in Table 2.1

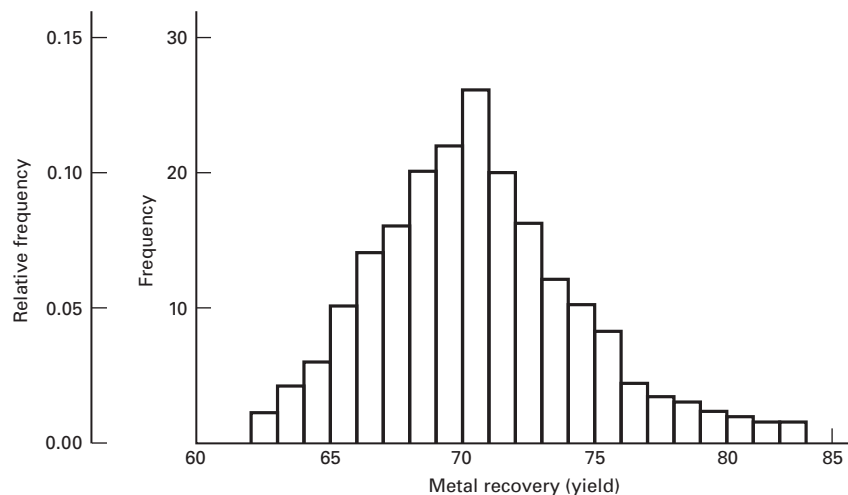
2.2 Basic Statistical Concepts

Each of the observations in the Portland cement experiment described above would be called a **run**. Notice that the individual runs differ, so there is fluctuation, or **noise**, in the observed bond strengths. This noise is usually called **experimental error** or simply **error**. It is a **statistical error**, meaning that it arises from variation that is uncontrolled and generally unavoidable. The presence of error or noise implies that the response variable, tension bond strength, is a **random variable**. A random variable may be either **discrete** or **continuous**. If the set of all possible values of the random variable is either finite or countably infinite, then the random variable is discrete, whereas if the set of all possible values of the random variable is an interval, then the random variable is continuous.

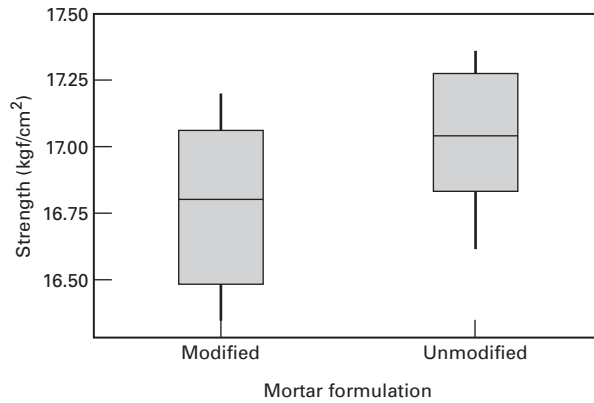
Graphical Description of Variability. We often use simple graphical methods to assist in analyzing the data from an experiment. The **dot diagram**, illustrated in Figure 2.1, is a very useful device for displaying a small body of data (say up to about 20 observations). The dot diagram enables the experimenter to see quickly the general **location** or **central tendency** of the observations and their **spread** or **variability**. For example, in the Portland cement tension bond experiment, the dot diagram reveals that the two formulations may differ in mean strength but that both formulations produce about the same variability in strength.

If the data are fairly numerous, the dots in a dot diagram become difficult to distinguish and a **histogram** may be preferable. Figure 2.2 presents a histogram for 200 observations on the metal recovery, or yield, from a smelting process. The histogram shows the central tendency, spread, and general shape of the distribution of the data. Recall that a histogram is constructed by dividing the horizontal axis into bins (usually of equal length) and drawing a rectangle over the j th bin with the area of the rectangle proportional to n_j , the number of observations that fall in that bin. The histogram is a large-sample tool. When the sample size is small the shape of the histogram can be very sensitive to the number of bins, the width of the bins, and the starting value for the first bin. Histograms should not be used with fewer than 75–100 observations.

The **box plot** (or **box-and-whisker plot**) is a very useful way to display data. A box plot displays the minimum, the maximum, the lower and upper quartiles (the 25th percentile and the 75th percentile, respectively), and the median (the 50th percentile) on a rectangular box aligned either horizontally or vertically. The box extends from the lower quartile to the



■ **FIGURE 2.2** Histogram for 200 observations on metal recovery (yield) from a smelting process



■ **FIGURE 2.3** Box plots for the Portland cement tension bond strength experiment

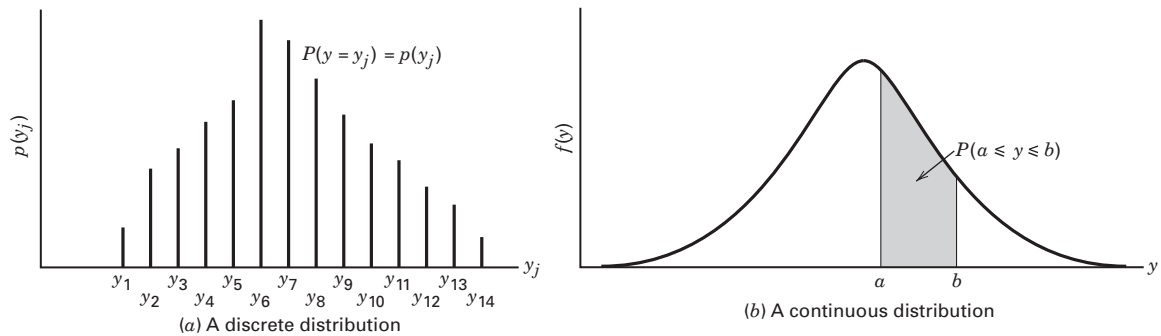
upper quartile, and a line is drawn through the box at the median. Lines (or whiskers) extend from the ends of the box to (typically) the minimum and maximum values. [There are several variations of box plots that have different rules for denoting the extreme sample points. See Montgomery and Runger (2011) for more details.]

Figure 2.3 presents the box plots for the two samples of tension bond strength in the Portland cement mortar experiment. This display indicates some difference in mean strength between the two formulations. It also indicates that both formulations produce reasonably symmetric distributions of strength with similar variability or spread.

Dot diagrams, histograms, and box plots are useful for summarizing the information in a **sample** of data. To describe the observations that might occur in a sample more completely, we use the concept of the probability distribution.

Probability Distributions. The probability structure of a random variable, say y , is described by its **probability distribution**. If y is discrete, we often call the probability distribution of y , say $p(y)$, the probability mass function of y . If y is continuous, the probability distribution of y , say $f(y)$, is often called the probability density function for y .

Figure 2.4 illustrates hypothetical discrete and continuous probability distributions. Notice that in the discrete probability distribution Fig. 2.4a, it is the height of the function $p(y_j)$ that represents probability, whereas in the continuous case Fig. 2.4b, it is the area under



■ **FIGURE 2.4** Discrete and continuous probability distributions

the curve $f(y)$ associated with a given interval that represents probability. The properties of probability distributions may be summarized quantitatively as follows:

$$\begin{aligned}
 \text{y discrete:} \quad & 0 \leq p(y_j) \leq 1 && \text{all values of } y_j \\
 & P(y = y_j) = p(y_j) && \text{all values of } y_j \\
 & \sum_{\substack{\text{all values} \\ \text{of } y_j}} p(y_j) = 1 \\
 \text{y continuous:} \quad & 0 \leq f(y) \\
 & P(a \leq y \leq b) = \int_a^b f(y) dy \\
 & \int_{-\infty}^{\infty} f(y) dy = 1
 \end{aligned}$$

Mean, Variance, and Expected Values. The **mean**, μ , of a probability distribution is a measure of its central tendency or location. Mathematically, we define the mean as

$$\mu = \begin{cases} \int_{-\infty}^{\infty} yf(y) dy & \text{y continuous} \\ \sum_{\text{all } y} yp(y) & \text{y discrete} \end{cases} \quad (2.1)$$

We may also express the mean in terms of the **expected value** or the long-run average value of the random variable y as

$$\mu = E(y) = \begin{cases} \int_{-\infty}^{\infty} yf(y) dy & \text{y continuous} \\ \sum_{\text{all } y} yp(y) & \text{y discrete} \end{cases} \quad (2.2)$$

where E denotes the **expected value operator**.

The variability or dispersion of a probability distribution can be measured by the **variance**, defined as

$$\sigma^2 = \begin{cases} \int_{-\infty}^{\infty} (y - \mu)^2 f(y) dy & \text{y continuous} \\ \sum_{\text{all } y} (y - \mu)^2 p(y) & \text{y discrete} \end{cases} \quad (2.3)$$

Note that the variance can be expressed entirely in terms of expectation because

$$\sigma^2 = E[(y - \mu)^2] \quad (2.4)$$

Finally, the variance is used so extensively that it is convenient to define a **variance operator** V such that

$$V(y) = E[(y - \mu)^2] = \sigma^2 \quad (2.5)$$

The concepts of expected value and variance are used extensively throughout this book, and it may be helpful to review several elementary results concerning these operators. If y is a random variable with mean μ and variance σ^2 and c is a constant, then

1. $E(c) = c$
2. $E(y) = \mu$

3. $E(cy) = cE(y) = c\mu$
4. $V(c) = 0$
5. $V(y) = \sigma^2$
6. $V(cy) = c^2V(y) = c^2\sigma^2$

If there are two random variables, say, y_1 with $E(y_1) = \mu_1$ and $V(y_1) = \sigma_1^2$ and y_2 with $E(y_2) = \mu_2$ and $V(y_2) = \sigma_2^2$, we have

$$7. E(y_1 + y_2) = E(y_1) + E(y_2) = \mu_1 + \mu_2$$

It is possible to show that

$$8. V(y_1 + y_2) = V(y_1) + V(y_2) + 2 \text{Cov}(y_1, y_2)$$

where

$$\text{Cov}(y_1, y_2) = E[(y_1 - \mu_1)(y_2 - \mu_2)] \quad (2.6)$$

is the **covariance** of the random variables y_1 and y_2 . The covariance is a measure of the linear association between y_1 and y_2 . More specifically, we may show that if y_1 and y_2 are independent,¹ then $\text{Cov}(y_1, y_2) = 0$. We may also show that

$$9. V(y_1 - y_2) = V(y_1) + V(y_2) - 2 \text{Cov}(y_1, y_2)$$

If y_1 and y_2 are **independent**, we have

$$10. V(y_1 \pm y_2) = V(y_1) + V(y_2) = \sigma_1^2 + \sigma_2^2$$

and

$$11. E(y_1 \cdot y_2) = E(y_1) \cdot E(y_2) = \mu_1 \cdot \mu_2$$

However, note that, in general

$$12. E\left(\frac{y_1}{y_2}\right) \neq \frac{E(y_1)}{E(y_2)}$$

regardless of whether or not y_1 and y_2 are independent.

2.3 Sampling and Sampling Distributions

Random Samples, Sample Mean, and Sample Variance. The objective of statistical inference is to draw conclusions about a population using a sample from that population. Most of the methods that we will study assume that **random samples** are used. A **random sample** is a sample that has been selected from the population in such a way that every possible sample has an equal probability of being selected. In practice, it is sometimes difficult to obtain random samples, and random numbers generated by a computer program may be helpful.

Statistical inference makes considerable use of quantities computed from the observations in the sample. We define a **statistic** as any function of the observations in a sample that

¹ Note that the converse of this is not necessarily so; that is, we may have $\text{Cov}(y_1, y_2) = 0$ and yet this does not imply independence. For an example, see Hines et al. (2003).

does not contain unknown parameters. For example, suppose that y_1, y_2, \dots, y_n represents a sample. Then the **sample mean**

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad (2.7)$$

and the **sample variance**

$$S^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1} \quad (2.8)$$

are both statistics. These quantities are measures of the central tendency and dispersion of the sample, respectively. Sometimes $S = \sqrt{S^2}$, called the **sample standard deviation**, is used as a measure of dispersion. Experimenters often prefer to use the standard deviation to measure dispersion because its units are the same as those for the variable of interest y .

Properties of the Sample Mean and Variance. The sample mean \bar{y} is a point estimator of the population mean μ , and the sample variance S^2 is a point estimator of the population variance σ^2 . In general, an **estimator** of an unknown parameter is a statistic that corresponds to that parameter. Note that a point estimator is a random variable. A particular numerical value of an estimator, computed from sample data, is called an **estimate**. For example, suppose we wish to estimate the mean and variance of the suspended solid material in the water of a lake. A random sample of $n = 25$ observation is tested, and the mg/l of suspended solid material is recorded for each. The sample mean and variance are computed according to Equations 2.7 and 2.8, respectively, and are $\bar{y} = 18.6$ and $S^2 = 1.20$. Therefore, the estimate of μ is $\bar{y} = 18.6$, and the estimate of σ^2 is $S^2 = 1.20$.

Several properties are required of good point estimators. Two of the most important are the following:

1. The point estimator should be **unbiased**. That is, the long-run average or expected value of the point estimator should be equal to the parameter that is being estimated. Although unbiasedness is desirable, this property alone does not always make an estimator a good one.
2. An unbiased estimator should have **minimum variance**. This property states that the minimum variance point estimator has a variance that is smaller than the variance of any other estimator of that parameter.

We may easily show that \bar{y} and S^2 are unbiased estimators of μ and σ^2 , respectively. First consider \bar{y} . Using the properties of expectation, we have

$$\begin{aligned} E(\bar{y}) &= E\left(\frac{\sum_{i=1}^n y_i}{n}\right) \\ &= \frac{1}{n} \sum_{i=1}^n E(y_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu \\ &= \mu \end{aligned}$$

because the expected value of each observation y_i is μ . Thus, \bar{y} is an unbiased estimator of μ .

Now consider the sample variance S^2 . We have

$$\begin{aligned} E(S^2) &= E\left[\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}\right] \\ &= \frac{1}{n-1} E\left[\sum_{i=1}^n (y_i - \bar{y})^2\right] \\ &= \frac{1}{n-1} E(SS) \end{aligned}$$

where $SS = \sum_{i=1}^n (y_i - \bar{y})^2$ is the **corrected sum of squares** of the observations y_i . Now

$$\begin{aligned} E(SS) &= E\left[\sum_{i=1}^n (y_i - \bar{y})^2\right] & (2.9) \\ &= E\left[\sum_{i=1}^n y_i^2 - n\bar{y}^2\right] \\ &= \sum_{i=1}^n (\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n) \\ &= (n-1)\sigma^2 & (2.10) \end{aligned}$$

Therefore,

$$E(S^2) = \frac{1}{n-1} E(SS) = \sigma^2$$

and we see that S^2 is an unbiased estimator of σ^2 .

Degrees of Freedom. The quantity $n-1$ in Equation 2.10 is called the **number of degrees of freedom** of the sum of squares SS . This is a very general result; that is, if y is a random variable with variance σ^2 and $SS = \sum (y_i - \bar{y})^2$ has ν degrees of freedom, then

$$E\left(\frac{SS}{\nu}\right) = \sigma^2 \quad (2.11)$$

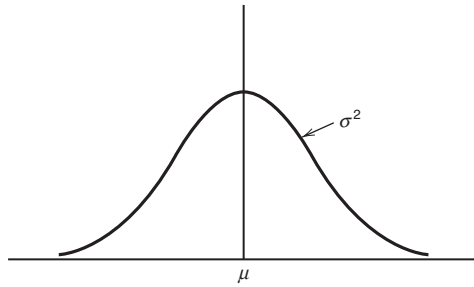
The number of degrees of freedom of a sum of squares is equal to the number of independent elements in that sum of squares. For example, $SS = \sum_{i=1}^n (y_i - \bar{y})^2$ in Equation 2.9 consists of the sum of squares of the n elements $y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y}$. These elements are not all independent because $\sum_{i=1}^n (y_i - \bar{y}) = 0$; in fact, only $n-1$ of them are independent, implying that SS has $n-1$ degrees of freedom.

The Normal and Other Sampling Distributions. Often we are able to determine the probability distribution of a particular statistic if we know the probability distribution of the population from which the sample was drawn. The probability distribution of a statistic is called a **sampling distribution**. We will now briefly discuss several useful sampling distributions.

One of the most important sampling distributions is the **normal distribution**. If y is a normal random variable, the probability distribution of y is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(y-\mu)/\sigma]^2} \quad -\infty < y < \infty \quad (2.12)$$

where $-\infty < \mu < \infty$ is the mean of the distribution and $\sigma^2 > 0$ is the variance. The normal distribution is shown in Figure 2.5.



■ **FIGURE 2.5** The normal distribution

Because sample runs that differ as a result of experimental error often are well described by the normal distribution, the normal plays a central role in the analysis of data from designed experiments. Many important sampling distributions may also be defined in terms of normal random variables. We often use the notation $y \sim N(\mu, \sigma^2)$ to denote that y is distributed normally with mean μ and variance σ^2 .

An important special case of the normal distribution is the **standard normal distribution**; that is, $\mu = 0$ and $\sigma^2 = 1$. We see that if $y \sim N(\mu, \sigma^2)$, the random variable

$$z = \frac{y - \mu}{\sigma} \quad (2.13)$$

follows the standard normal distribution, denoted $z \sim N(0, 1)$. The operation demonstrated in Equation 2.13 is often called **standardizing** the normal random variable y . The cumulative standard normal distribution is given in Table I of the Appendix.

Many statistical techniques assume that the random variable is normally distributed. The central limit theorem is often a justification of approximate normality.

THEOREM 2-1 The Central Limit Theorem

If y_1, y_2, \dots, y_n is a sequence of n independent and identically distributed random variables with $E(y_i) = \mu$ and $V(y_i) = \sigma^2$ (both finite) and $x = y_1 + y_2 + \dots + y_n$, then the limiting form of the distribution of

$$z_n = \frac{x - n\mu}{\sqrt{n\sigma^2}}$$

as $n \rightarrow \infty$, is the standard normal distribution.

This result states essentially that the sum of n independent and identically distributed random variables is approximately normally distributed. In many cases, this approximation is good for very small n , say $n < 10$, whereas in other cases large n is required, say $n > 100$. Frequently, we think of the error in an experiment as arising in an additive manner from several independent sources; consequently, the normal distribution becomes a plausible model for the combined experimental error.

An important sampling distribution that can be defined in terms of normal random variables is the **chi-square** or χ^2 **distribution**. If z_1, z_2, \dots, z_k are normally and independently distributed random variables with mean 0 and variance 1, abbreviated NID(0, 1), then the random variable

$$x = z_1^2 + z_2^2 + \dots + z_k^2$$

follows the **chi-square distribution with k degrees of freedom**. The density function of chi-square is

$$f(x) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{(k/2)-1} e^{-x/2} \quad x > 0 \quad (2.14)$$

Several chi-square distributions are shown in Figure 2.6. The distribution is asymmetric, or **skewed**, with mean and variance

$$\begin{aligned} \mu &= k \\ \sigma^2 &= 2k \end{aligned}$$

respectively. Percentage points of the chi-square distribution are given in Table III of the Appendix.

As an example of a random variable that follows the chi-square distribution, suppose that y_1, y_2, \dots, y_n is a random sample from an $N(\mu, \sigma^2)$ distribution. Then

$$\frac{SS}{\sigma^2} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sigma^2} \sim \chi_{n-1}^2 \quad (2.15)$$

That is, SS/σ^2 is distributed as chi-square with $n - 1$ degrees of freedom.

Many of the techniques used in this book involve the computation and manipulation of sums of squares. The result given in Equation 2.15 is extremely important and occurs repeatedly; a sum of squares in normal random variables when divided by σ^2 follows the chi-square distribution.

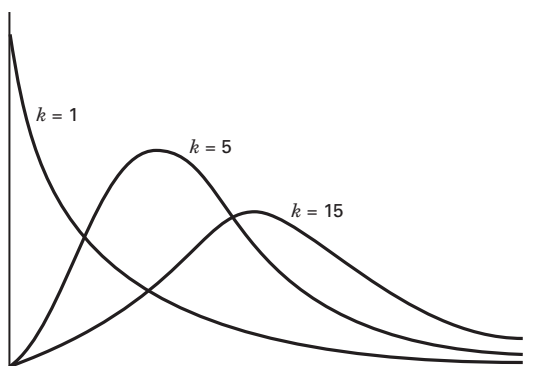
Examining Equation 2.8, we see that the sample variance can be written as

$$S^2 = \frac{SS}{n - 1} \quad (2.16)$$

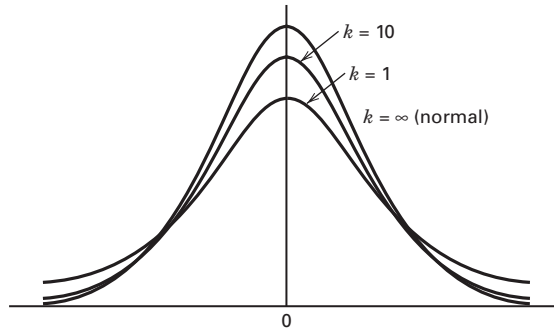
If the observations in the sample are $NID(\mu, \sigma^2)$, then the distribution of S^2 is $[\sigma^2/(n - 1)]\chi_{n-1}^2$. Thus, the sampling distribution of the sample variance is a constant times the chi-square distribution if the population is normally distributed.

If z and χ_k^2 are independent standard normal and chi-square random variables, respectively, the random variable

$$t_k = \frac{z}{\sqrt{\chi_k^2/k}} \quad (2.17)$$



■ FIGURE 2.6 Several Chi-square distributions



■ **FIGURE 2.7** Several t distributions

follows the **t distribution with k degrees of freedom**, denoted t_k . The density function of t is

$$f(t) = \frac{\Gamma[(k + 1)/2]}{\sqrt{k\pi}\Gamma(k/2)} \frac{1}{[(t^2/k) + 1]^{(k+1)/2}} \quad -\infty < t < \infty \quad (2.18)$$

and the mean and variance of t are $\mu = 0$ and $\sigma^2 = k/(k - 2)$ for $k > 2$, respectively. Several t distributions are shown in Figure 2.7. Note that if $k = \infty$, the t distribution becomes the standard normal distribution. The percentage points of the t distribution are given in Table II of the Appendix. If y_1, y_2, \dots, y_n is a random sample from the $N(\mu, \sigma^2)$ distribution, then the quantity

$$t = \frac{\bar{y} - \mu}{S/\sqrt{n}} \quad (2.19)$$

is distributed as t with $n - 1$ degrees of freedom.

The final sampling distribution that we will consider is the **F distribution**. If χ_u^2 and χ_v^2 are two independent chi-square random variables with u and v degrees of freedom, respectively, then the ratio

$$F_{u,v} = \frac{\chi_u^2/u}{\chi_v^2/v} \quad (2.20)$$

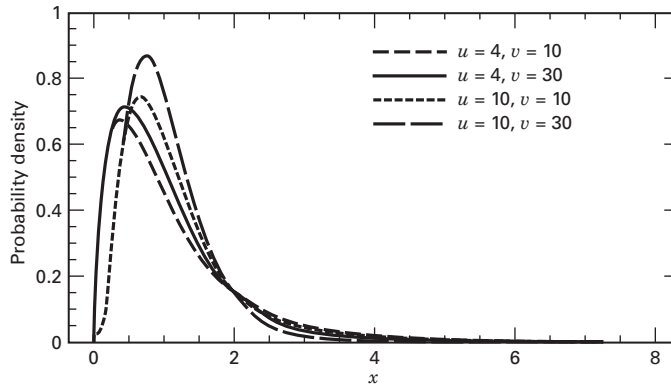
follows the **F distribution with u numerator degrees of freedom and v denominator degrees of freedom**. If x is an F random variable with u numerator and v denominator degrees of freedom, then the probability distribution of x is

$$h(x) = \frac{\Gamma\left(\frac{u + v}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x + 1\right]^{(u+v)/2}} \quad 0 < x < \infty \quad (2.21)$$

Several F distributions are shown in Figure 2.8. This distribution is very important in the statistical analysis of designed experiments. Percentage points of the F distribution are given in Table IV of the Appendix.

As an example of a statistic that is distributed as F , suppose we have two independent normal populations with common variance σ^2 . If $y_{11}, y_{12}, \dots, y_{1n_1}$ is a random sample of n_1 observations from the first population, and if $y_{21}, y_{22}, \dots, y_{2n_2}$ is a random sample of n_2 observations from the second, then

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1} \quad (2.22)$$



■ FIGURE 2.8 Several F distributions

where S_1^2 and S_2^2 are the two sample variances. This result follows directly from Equations 2.15 and 2.20.

2.4 Inferences About the Differences in Means, Randomized Designs

We are now ready to return to the Portland cement mortar problem posed in Section 2.1. Recall that two different formulations of mortar were being investigated to determine if they differ in tension bond strength. In this section we discuss how the data from this simple comparative experiment can be analyzed using **hypothesis testing** and **confidence interval** procedures for comparing two treatment means.

Throughout this section we assume that a **completely randomized experimental design** is used. In such a design, the data are usually viewed as if they were a random sample from a normal distribution.

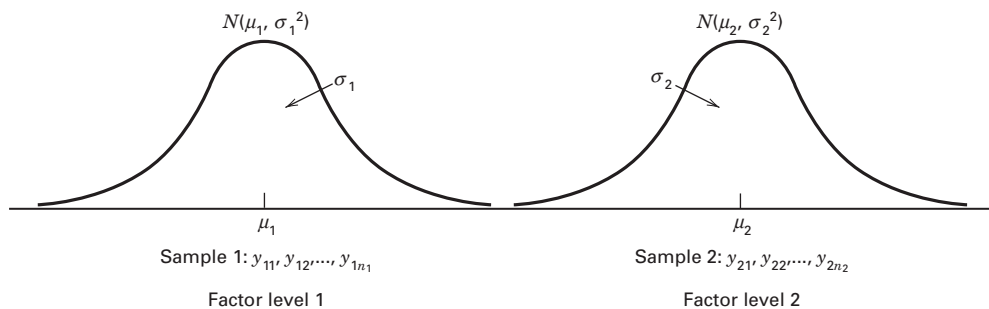
2.4.1 Hypothesis Testing

We now reconsider the Portland cement experiment introduced in Section 2.1. Recall that we are interested in comparing the strength of two different formulations: an unmodified mortar and a modified mortar. In general, we can think of these two formulations as two **levels of the factor** “formulations.” Let $y_{11}, y_{12}, \dots, y_{1n_1}$ represent the n_1 observations from the first factor level and $y_{21}, y_{22}, \dots, y_{2n_2}$ represent the n_2 observations from the second factor level. We assume that the samples are drawn at random from two independent normal populations. Figure 2.9 illustrates the situation.

A Model for the Data. We often describe the results of an experiment with a **model**. A simple statistical model that describes the data from an experiment such as we have just described is

$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2 \\ j = 1, 2, \dots, n_i \end{cases} \quad (2.23)$$

where y_{ij} is the j th observation from factor level i , μ_i is the mean of the response at the i th factor level, and ϵ_{ij} is a normal random variable associated with the ij th observation. We assume



■ **FIGURE 2.9** The sampling situation for the two-sample t -test

that ϵ_{ij} are $\text{NID}(0, \sigma_i^2)$, $i = 1, 2$. It is customary to refer to ϵ_{ij} as the **random error** component of the model. Because the means μ_1 and μ_2 are constants, we see directly from the model that y_{ij} are $\text{NID}(\mu_i, \sigma_i^2)$, $i = 1, 2$, just as we previously assumed. For more information about models for the data, refer to the supplemental text material.

Statistical Hypotheses. A **statistical hypothesis** is a statement either about the parameters of a probability distribution or the parameters of a model. The hypothesis reflects some **conjecture** about the problem situation. For example, in the Portland cement experiment, we may think that the mean tension bond strengths of the two mortar formulations are equal. This may be stated formally as

$$\begin{aligned} H_0: \mu_1 &= \mu_2 \\ H_1: \mu_1 &\neq \mu_2 \end{aligned}$$

where μ_1 is the mean tension bond strength of the modified mortar and μ_2 is the mean tension bond strength of the unmodified mortar. The statement $H_0: \mu_1 = \mu_2$ is called the **null hypothesis** and $H_1: \mu_1 \neq \mu_2$ is called the **alternative hypothesis**. The alternative hypothesis specified here is called a **two-sided alternative hypothesis** because it would be true if $\mu_1 < \mu_2$ or if $\mu_1 > \mu_2$.

To test a hypothesis, we devise a procedure for taking a random sample, computing an appropriate **test statistic**, and then rejecting or failing to reject the null hypothesis H_0 based on the computed value of the test statistic. Part of this procedure is specifying the set of values for the test statistic that leads to rejection of H_0 . This set of values is called the **critical region** or **rejection region** for the test.

Two kinds of errors may be committed when testing hypotheses. If the null hypothesis is rejected when it is true, a type I error has occurred. If the null hypothesis is *not* rejected when it is false, a type II error has been made. The probabilities of these two errors are given special symbols

$$\begin{aligned} \alpha &= P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true}) \\ \beta &= P(\text{type II error}) = P(\text{fail to reject } H_0 | H_0 \text{ is false}) \end{aligned}$$

Sometimes it is more convenient to work with the **power** of the test, where

$$\text{Power} = 1 - \beta = P(\text{reject } H_0 | H_0 \text{ is false})$$

The general procedure in hypothesis testing is to specify a value of the probability of type I error α , often called the **significance level** of the test, and then design the test procedure so that the probability of type II error β has a suitably small value.

The Two-Sample t -Test. Suppose that we could assume that the variances of tension bond strengths were identical for both mortar formulations. Then the appropriate test statistic to use for comparing two treatment means in the completely randomized design is

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2.24)$$

where \bar{y}_1 and \bar{y}_2 are the sample means, n_1 and n_2 are the sample sizes, S_p^2 is an estimate of the common variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$ computed from

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \quad (2.25)$$

and S_1^2 and S_2^2 are the two individual sample variances. The quantity $S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ in the denominator of Equation 2.24 is often called the **standard error** of the difference in means in the numerator, abbreviated *se* ($\bar{y}_1 - \bar{y}_2$). To determine whether to reject $H_0: \mu_1 = \mu_2$, we would compare t_0 to the t distribution with $n_1 + n_2 - 2$ degrees of freedom. If $|t_0| > t_{\alpha/2, n_1 + n_2 - 2}$, where $t_{\alpha/2, n_1 + n_2 - 2}$ is the upper $\alpha/2$ percentage point of the t distribution with $n_1 + n_2 - 2$ degrees of freedom, we would *reject* H_0 and conclude that the mean strengths of the two formulations of Portland cement mortar differ. This test procedure is usually called the **two-sample t -test**.

This procedure may be justified as follows. If we are sampling from independent normal distributions, then the distribution of $\bar{y}_1 - \bar{y}_2$ is $N[\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)]$. Thus, if σ^2 were known, and if $H_0: \mu_1 = \mu_2$ were true, the distribution of

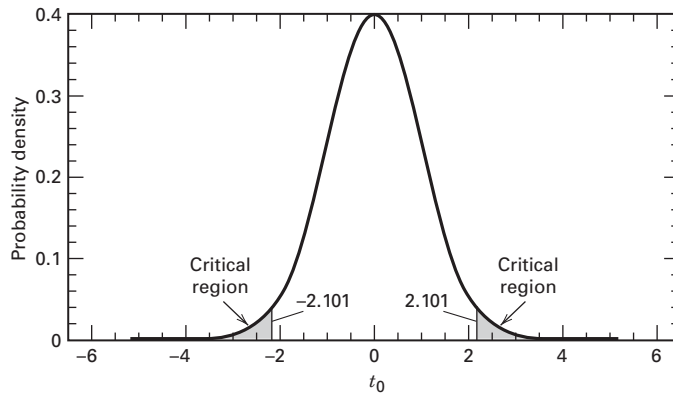
$$Z_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2.26)$$

would be $N(0, 1)$. However, in replacing σ in Equation 2.26 by S_p , the distribution of Z_0 changes from standard normal to t with $n_1 + n_2 - 2$ degrees of freedom. Now if H_0 is true, t_0 in Equation 2.24 is distributed as $t_{n_1 + n_2 - 2}$ and, consequently, we would expect $100(1 - \alpha)$ percent of the values of t_0 to fall between $-t_{\alpha/2, n_1 + n_2 - 2}$ and $t_{\alpha/2, n_1 + n_2 - 2}$. A sample producing a value of t_0 outside these limits would be unusual if the null hypothesis were true and is evidence that H_0 should be rejected. Thus the t distribution with $n_1 + n_2 - 2$ degrees of freedom is the appropriate **reference distribution** for the test statistic t_0 . That is, it describes the behavior of t_0 when the null hypothesis is true. Note that α is the probability of type I error for the test. Sometimes α is called the **significance level** of the test.

In some problems, one may wish to reject H_0 only if one mean is larger than the other. Thus, one would specify a **one-sided alternative hypothesis** $H_1: \mu_1 > \mu_2$ and would reject H_0 only if $t_0 > t_{\alpha, n_1 + n_2 - 2}$. If one wants to reject H_0 only if μ_1 is less than μ_2 , then the alternative hypothesis is $H_1: \mu_1 < \mu_2$, and one would reject H_0 if $t_0 < -t_{\alpha, n_1 + n_2 - 2}$.

To illustrate the procedure, consider the Portland cement data in Table 2.1. For these data, we find that

| Modified Mortar | Unmodified Mortar |
|--------------------------------------|--------------------------------------|
| $\bar{y}_1 = 16.76 \text{ kgf/cm}^2$ | $\bar{y}_2 = 17.04 \text{ kgf/cm}^2$ |
| $S_1^2 = 0.100$ | $S_2^2 = 0.061$ |
| $S_1 = 0.316$ | $S_2 = 0.248$ |
| $n_1 = 10$ | $n_2 = 10$ |



■ **FIGURE 2.10** The t distribution with 18 degrees of freedom with the critical region $\pm t_{0.025,18} = \pm 2.101$

Because the sample standard deviations are reasonably similar, it is not unreasonable to conclude that the population standard deviations (or variances) are equal. Therefore, we can use Equation 2.24 to test the hypotheses

$$\begin{aligned} H_0: \mu_1 &= \mu_2 \\ H_1: \mu_1 &\neq \mu_2 \end{aligned}$$

Furthermore, $n_1 + n_2 - 2 = 10 + 10 - 2 = 18$, and if we choose $\alpha = 0.05$, then we would reject $H_0: \mu_1 = \mu_2$ if the numerical value of the test statistic $t_0 > t_{0.025,18} = 2.101$, or if $t_0 < -t_{0.025,18} = -2.101$. These boundaries of the critical region are shown on the reference distribution (t with 18 degrees of freedom) in Figure 2.10.

Using Equation 2.25 we find that

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \\ &= \frac{9(0.100) + 9(0.061)}{10 + 10 - 2} = 0.081 \\ S_p &= 0.284 \end{aligned}$$

and the test statistic is

$$\begin{aligned} t_0 &= \frac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{16.76 - 17.04}{0.284 \sqrt{\frac{1}{10} + \frac{1}{10}}} \\ &= \frac{-0.28}{0.127} = -2.20 \end{aligned}$$

Because $t_0 = -2.20 < -t_{0.025,18} = -2.101$, we would reject H_0 and conclude that the mean tension bond strengths of the two formulations of Portland cement mortar are different. This is a potentially important engineering finding. The change in mortar formulation had the desired effect of reducing the cure time, but there is evidence that the change also affected the tension bond strength. One can conclude that the modified formulation reduces the bond strength (just because we conducted a two-sided test, this does not preclude drawing a one-sided conclusion when the null hypothesis is rejected). If the reduction in mean bond

strength is of practical importance (or has engineering significance in addition to statistical significance) then more development work and further experimentation will likely be required.

The Use of P -Values in Hypothesis Testing. One way to report the results of a hypothesis test is to state that the null hypothesis was or was not rejected at a specified α -value or **level of significance**. This is often called **fixed significance level testing**. For example, in the Portland cement mortar formulation above, we can say that $H_0 : \mu_1 = \mu_2$ was rejected at the 0.05 level of significance. This statement of conclusions is often inadequate because it gives the decision maker no idea about whether the computed value of the test statistic was just barely in the rejection region or whether it was very far into this region. Furthermore, stating the results this way imposes the predefined level of significance on other users of the information. This approach may be unsatisfactory because some decision makers might be uncomfortable with the risks implied by $\alpha = 0.05$.

To avoid these difficulties, the **P -value approach** has been adopted widely in practice. The P -value is the probability that the test statistic will take on a value that is at least as extreme as the observed value of the statistic when the null hypothesis H_0 is true. Thus, a P -value conveys much information about the weight of evidence against H_0 , and so a decision maker can draw a conclusion at *any* specified level of significance. More formally, we define the **P -value** as the smallest level of significance that would lead to rejection of the null hypothesis H_0 .

It is customary to call the test statistic (and the data) significant when the null hypothesis H_0 is rejected; therefore, we may think of the P -value as the smallest level α at which the data are significant. Once the P -value is known, the decision maker can determine how significant the data are without the data analyst formally imposing a preselected level of significance.

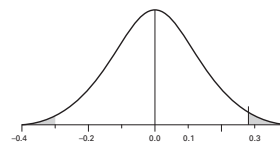
It is not always easy to compute the exact P -value for a test. However, most modern computer programs for statistical analysis report P -values, and they can be obtained on some handheld calculators. We will show how to approximate the P -value for the Portland cement mortar experiment. Because $|t_0| = 2.20 > t_{0.025,18} = 2.101$, we know that the P -value is less than 0.05. From Appendix Table II, for a t distribution with 18 degrees of freedom, and tail area probability 0.01 we find $t_{0.01,18} = 2.552$. Now $|t_0| = 2.20 < 2.552$, so because the alternative hypothesis is two sided, we know that the P -value must be between 0.05 and $2(0.01) = 0.02$. Some handheld calculators have the capability to calculate P -values. One such calculator is the HP-48. From this calculator, we obtain the P -value for the value $t_0 = -2.20$ in the Portland cement mortar formulation experiment as $P = 0.0411$. Thus the null hypothesis $H_0 : \mu_1 = \mu_2$ would be rejected at any level of significance $\alpha > 0.0411$.

Computer Solution. Many statistical software packages have capability for statistical hypothesis testing. The output from both the Minitab and the JMP two-sample t -test procedure applied to the Portland cement mortar formulation experiment is shown in Table 2.2. Notice that the output includes some summary statistics about the two samples (the abbreviation “SE mean” in the Minitab section of the table refers to the standard error of the mean, s/\sqrt{n}) as well as some information about confidence intervals on the difference in the two means (which we will discuss in the next section). The programs also test the hypothesis of interest, allowing the analyst to specify the nature of the alternative hypothesis (“not =” in the Minitab output implies $H_1 : \mu_1 \neq \mu_2$).

The output includes the computed value of t_0 , the value of the test statistic t_0 (JMP reports a positive value of t_0 because of how the sample means are subtracted in the numerator

■ **TABLE 2.2**
Computer Output for the Two-Sample t -Test

| Minitab | | | | |
|---|----------|-----------|-----------|---------|
| Two-sample T for Modified vs Unmodified | | | | |
| | N | Mean | Std. Dev. | SE Mean |
| Modified | 10 | 16.764 | 0.316 | 0.10 |
| Unmodified | 10 | 17.042 | 0.248 | 0.078 |
| Difference = mu (Modified) - mu (Unmodified) | | | | |
| Estimate for difference: -0.278000 | | | | |
| 95% CI for difference: (-0.545073, -0.010927) | | | | |
| T-Test of difference = 0 (vs not =): T-Value = -2.19 | | | | |
| P-Value = 0.042 DF = 18 | | | | |
| Both use Pooled Std. Dev. = 0.2843 | | | | |
| JMP t-test | | | | |
| Unmodified-Modified | | | | |
| Assuming equal variances | | | | |
| Difference | 0.278000 | t Ratio | 2.186876 | |
| Std Err Dif | 0.127122 | DF | 18 | |
| Upper CL Dif | 0.545073 | Prob > t | 0.0422 | |
| Lower CL Dif | 0.010927 | Prob > t | 0.0211 | |
| Confidence | 0.95 | Prob < t | 0.9789 | |

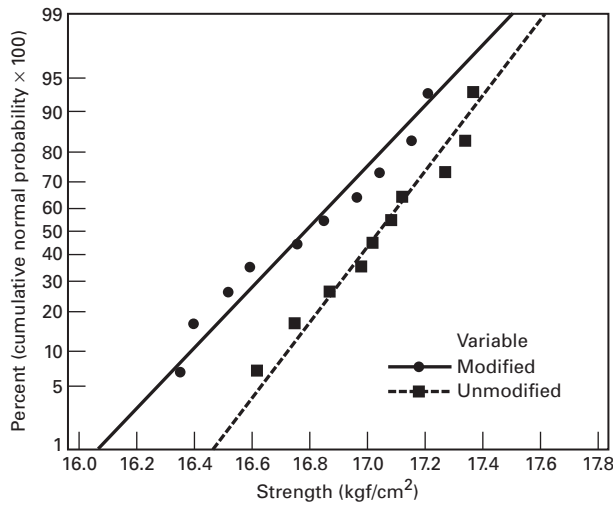


of the test statistic), and the P -value. Notice that the computed value of the t statistic differs slightly from our manually calculated value and that the P -value is reported to be $P = 0.042$. JMP also reports the P -values for the one-sided alternative hypothesis. Many software packages will not report an actual P -value less than some predetermined value such as 0.0001 and instead will return a “default” value such as “<0.001” or in some cases, zero.

Checking Assumptions in the t -Test. In using the t -test procedure we make the assumptions that both samples are random samples that are drawn from independent populations that can be described by a normal distribution, and that the standard deviation or variances of both populations are equal. The assumption of independence is critical, and if the run order is randomized (and, if appropriate, other experimental units and materials are selected at random), this assumption will usually be satisfied. The equal variance and normality assumptions are easy to check using a **normal probability plot**.

Generally, probability plotting is a graphical technique for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data. The general procedure is very simple and can be performed quickly with most statistics software packages. The **supplemental text material** discusses manual construction of normal probability plots.

To construct a probability plot, the observations in the sample are first ranked from smallest to largest. That is, the sample y_1, y_2, \dots, y_n is arranged as $y_{(1)}, y_{(2)}, \dots, y_{(n)}$ where $y_{(1)}$ is the smallest observation, $y_{(2)}$ is the second smallest observation, and so forth, with $y_{(n)}$ the largest. The ordered observations $y_{(j)}$ are then plotted against their observed cumulative frequency $(j - 0.5)/n$.



■ **FIGURE 2.11** Normal probability plots of tension bond strength in the Portland cement experiment

The cumulative frequency scale has been arranged so that if the hypothesized distribution adequately describes the data, the plotted points will fall approximately along a straight line; if the plotted points deviate significantly from a straight line, the hypothesized model is not appropriate. Usually, the determination of whether or not the data plot as a straight line is subjective.

To illustrate the procedure, suppose that we wish to check the assumption that tension bond strength in the Portland cement mortar formulation experiment is normally distributed. We initially consider only the observations from the unmodified mortar formulation. A computer-generated normal probability plot is shown in Figure 2.11. Most normal probability plots present $100(j - 0.5)/n$ on the left vertical scale (and occasionally $100[1 - (j - 0.5)/n]$ is plotted on the right vertical scale), with the variable value plotted on the horizontal scale. Some computer-generated normal probability plots convert the cumulative frequency to a standard normal z score. A straight line, chosen subjectively, has been drawn through the plotted points. In drawing the straight line, you should be influenced more by the points near the middle of the plot than by the extreme points. A good rule of thumb is to draw the line approximately between the 25th and 75th percentile points. This is how the lines in Figure 2.11 for each sample were determined. In assessing the “closeness” of the points to the straight line, imagine a fat pencil lying along the line. If all the points are covered by this imaginary pencil, a normal distribution adequately describes the data. Because the points for each sample in Figure 2.11 would pass the fat pencil test, we conclude that the normal distribution is an appropriate model for tension bond strength for both the modified and the unmodified mortar.

We can obtain an estimate of the mean and standard deviation directly from the normal probability plot. The mean is estimated as the 50th percentile on the probability plot, and the standard deviation is estimated as the difference between the 84th and 50th percentiles. This means that we can verify the assumption of equal population variances in the Portland cement experiment by simply comparing the slopes of the two straight lines in Figure 2.11. Both lines have very similar slopes, and so the assumption of equal variances is a reasonable one. If this assumption is violated, you should use the version of the t -test described in Section 2.4.4. The supplemental text material has more information about checking assumptions on the t -test.

When assumptions are badly violated, the performance of the t -test will be affected. Generally, small to moderate violations of assumptions are not a major concern, but *any* failure of the independence assumption and strong indications of nonnormality should not be ignored. Both the significance level of the test and the ability to detect differences between the means will be adversely affected by departures from assumptions. **Transformations** are one approach to dealing with this problem. We will discuss this in more detail in Chapter 3.

Nonparametric hypothesis testing procedures can also be used if the observations come from nonnormal populations. Refer to Montgomery and Runger (2011) for more details.

An Alternate Justification to the t -Test. The two-sample t -test we have just presented depends in theory on the underlying assumption that the two populations from which the samples were randomly selected are normal. Although the normality assumption is required to develop the test procedure formally, as we discussed above, moderate departures from normality will not seriously affect the results. It can be argued that the use of a randomized design enables one to test hypotheses without *any* assumptions regarding the form of the distribution. Briefly, the reasoning is as follows. If the treatments have no effect, all $[20!/(10!10!)] = 184,756$ possible ways that the 20 observations could occur are equally likely. Corresponding to each of these 184,756 possible arrangements is a value of t_0 . If the value of t_0 actually obtained from the data is unusually large or unusually small with reference to the set of 184,756 possible values, it is an indication that $\mu_1 \neq \mu_2$.

This type of procedure is called a **randomization test**. It can be shown that the t -test is a good approximation of the randomization test. Thus, we will use t -tests (and other procedures that can be regarded as approximations of randomization tests) without extensive concern about the assumption of normality. This is one reason a simple procedure such as normal probability plotting is adequate to check the assumption of normality.

2.4.2 Confidence Intervals

Although hypothesis testing is a useful procedure, it sometimes does not tell the entire story. It is often preferable to provide an interval within which the value of the parameter or parameters in question would be expected to lie. These interval statements are called **confidence intervals**. In many engineering and industrial experiments, the experimenter already knows that the means μ_1 and μ_2 differ; consequently, hypothesis testing on $\mu_1 = \mu_2$ is of little interest. The experimenter would usually be more interested in knowing how much the means differ. A confidence interval on the difference in means $\mu_1 - \mu_2$ is used in answering this question.

To define a confidence interval, suppose that θ is an unknown parameter. To obtain an interval estimate of θ , we need to find two statistics L and U such that the probability statement

$$P(L \leq \theta \leq U) = 1 - \alpha \quad (2.27)$$

is true. The interval

$$L \leq \theta \leq U \quad (2.28)$$

is called a **$100(1 - \alpha)$ percent confidence interval** for the parameter θ . The interpretation of this interval is that if, in repeated random samplings, a large number of such intervals are constructed, $100(1 - \alpha)$ percent of them will contain the true value of θ . The statistics L and U are called the **lower** and **upper confidence limits**, respectively, and $1 - \alpha$ is called the **confidence coefficient**. If $\alpha = 0.05$, Equation 2.28 is called a 95 percent confidence interval for θ . Note that confidence intervals have a frequency interpretation; that is, we do not know if the statement is true for this specific sample, but we do know that the *method* used to produce the confidence interval yields correct statements $100(1 - \alpha)$ percent of the time.

Suppose that we wish to find a $100(1 - \alpha)$ percent confidence interval on the true difference in means $\mu_1 - \mu_2$ for the Portland cement problem. The interval can be derived in the following way. The statistic

$$\frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is distributed as $t_{n_1+n_2-2}$. Thus,

$$P\left(-t_{\alpha/2, n_1+n_2-2} \leq \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha/2, n_1+n_2-2}\right) = 1 - \alpha$$

or

$$\begin{aligned} P\left(\bar{y}_1 - \bar{y}_2 - t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \right. \\ \left. \leq \bar{y}_1 - \bar{y}_2 + t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right) = 1 - \alpha \quad (2.29) \end{aligned}$$

Comparing Equations 2.29 and 2.27, we see that

$$\begin{aligned} \bar{y}_1 - \bar{y}_2 - t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \\ \leq \bar{y}_1 - \bar{y}_2 + t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (2.30) \end{aligned}$$

is a $100(1 - \alpha)$ percent confidence interval for $\mu_1 - \mu_2$.

The actual 95 percent confidence interval estimate for the difference in mean tension bond strength for the formulations of Portland cement mortar is found by substituting in Equation 2.30 as follows:

$$\begin{aligned} 16.76 - 17.04 - (2.101)0.284\sqrt{\frac{1}{10} + \frac{1}{10}} &\leq \mu_1 - \mu_2 \\ &\leq 16.76 - 17.04 + (2.101)0.284\sqrt{\frac{1}{10} + \frac{1}{10}} \\ -0.28 - 0.27 &\leq \mu_1 - \mu_2 \leq -0.28 + 0.27 \\ -0.55 &\leq \mu_1 - \mu_2 \leq -0.01 \end{aligned}$$

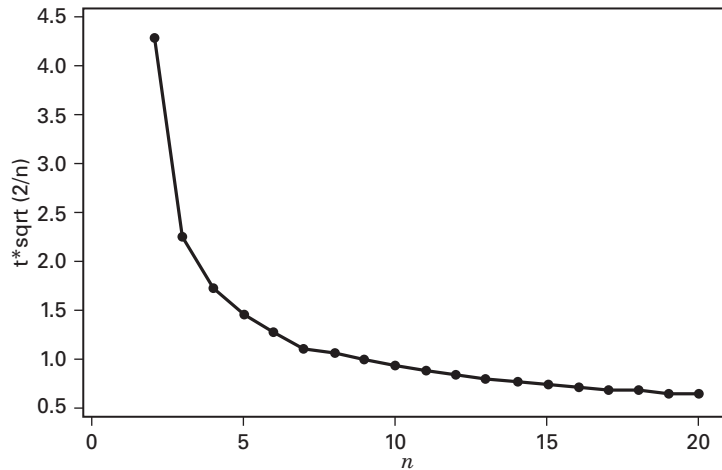
Thus, the 95 percent confidence interval estimate on the difference in means extends from -0.55 to -0.01 kgf/cm². Put another way, the confidence interval is $\mu_1 - \mu_2 = -0.28 \pm 0.27$ kgf/cm², or the difference in mean strengths is -0.28 kgf/cm², and the accuracy of this estimate is ± 0.27 kgf/cm². Note that because $\mu_1 - \mu_2 = 0$ is *not* included in this interval, the data do not support the hypothesis that $\mu_1 = \mu_2$ at the 5 percent level of significance (recall that the P -value for the two-sample t -test was 0.042, just slightly less than 0.05). It is likely that the mean strength of the unmodified formulation exceeds the mean strength of the modified formulation. Notice from Table 2.2 that both Minitab and JMP reported this confidence interval when the hypothesis testing procedure was conducted.

2.4.3 Choice of Sample Size

Selection of an appropriate sample size is one of the most important parts of any experimental design problem. One way to do this is to consider the impact of sample size on the estimate of the difference in two means. From Equation 2.30 we know that the $100(1 - \alpha)\%$ confidence interval on the difference in two means is a measure of the precision of estimation of the difference in the two means. The length of this interval is determined by

$$t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

We consider the case where the sample sizes from the two populations are equal, so that $n_1 = n_2 = n$. Then the length of the CI is determined by



■ **FIGURE 2.12** Plot of $t_{\alpha/2, 2n-2} \sqrt{2/n}$ versus sample size in each population n for $\alpha = 0.05$

$$t_{\alpha/2, 2n-2} S_p \sqrt{\frac{2}{n}}$$

Consequently the precision with which the difference in the two means is estimated depends on two quantities— S_p , over which we have no control, and $t_{\alpha/2, 2n-2} \sqrt{2/n}$, which we can control by choosing the sample size n . Figure 2.12 is a plot of $t_{\alpha/2, 2n-2} \sqrt{2/n}$ versus n for $\alpha = 0.05$. Notice that the curve descends rapidly as n increases up to about $n = 10$ and less rapidly beyond that. Since S_p is relatively constant and $t_{\alpha/2, 2n-2} \sqrt{2/n}$ isn't going to change much for sample sizes beyond $n = 10$ or 12 , we can conclude that choosing a sample size of $n = 10$ or 12 from each population in a two-sample 95% CI will result in a CI that results in about the best precision of estimation for the difference in the two means that is possible given the amount of inherent variability that is present in the two populations.

We can also use a hypothesis testing framework to determine sample size. The choice of sample size and the probability of type II error β are closely connected. Suppose that we are testing the hypotheses

$$\begin{aligned} H_0: \mu_1 &= \mu_2 \\ H_1: \mu_1 &\neq \mu_2 \end{aligned}$$

and that the means are *not* equal so that $\delta = \mu_1 - \mu_2$. Because $H_0: \mu_1 = \mu_2$ is not true, we are concerned about wrongly failing to reject H_0 . The probability of type II error depends on the true difference in means δ . A graph of β versus δ for a particular sample size is called the **operating characteristic curve**, or **O.C. curve** for the test. The β error is also a function of sample size. Generally, for a given value of δ , the β error decreases as the sample size increases. That is, a specified difference in means is easier to detect for larger sample sizes than for smaller ones.

An alternative to the OC curve is a **power curve**, which typically plots power or $1 - \beta$, versus sample size for a specified difference in the means. Some software packages perform power analysis and will plot power curves. A set of power curves constructed using JMP for the hypotheses

$$\begin{aligned} H_0: \mu_1 &= \mu_2 \\ H_1: \mu_1 &\neq \mu_2 \end{aligned}$$

is shown in Figure 2.13 for the case where the two population variances σ_1^2 and σ_2^2 are unknown but equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) and for a level of significance of $\alpha = 0.05$. These power

curves also assume that the sample sizes from the two populations are equal and that the sample size shown on the horizontal scale (say n) is the total sample size, so that the sample size in each population is $n/2$. Also notice that the difference in means is expressed as a ratio to the common standard deviation; that is

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$$

From examining these curves we observe the following:

1. The greater the difference in means $\mu_1 - \mu_2$, the higher the power (smaller type II error probability). That is, for a specified sample size and significance level α , the test will detect large differences in means more easily than small ones.
2. As the sample size get larger, the power of the test gets larger (the type II error probability gets smaller) for a given difference in means and significance level α . That is, to detect a specified difference in means we may make the test more powerful by increasing the sample size.

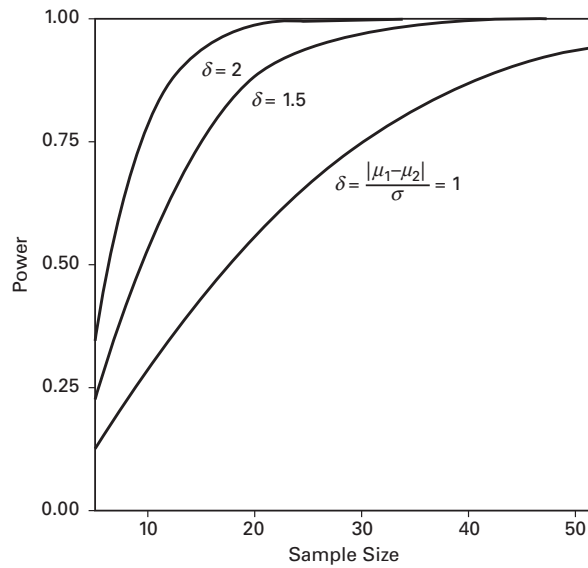
Operating curves and power curves are often helpful in selecting a sample size to use in an experiment. For example, consider the Portland cement mortar problem discussed previously. Suppose that a difference in mean strength of 0.5 kgf/cm^2 has practical impact on the use of the mortar, so if the difference in means is at least this large, we would like to detect it with a high probability. Thus, because $\mu_1 - \mu_2 = 0.5 \text{ kgf/cm}^2$ is the “critical” difference in means that we wish to detect, we find that the power curve parameter would be $\delta = 0.5/\sigma$. Unfortunately, δ involves the unknown standard deviation σ . However, suppose on the basis of past experience we think that it is very unlikely that the standard deviation will exceed 0.25 kgf/cm^2 . Then substituting $\sigma = 0.25 \text{ kgf/cm}^2$ into the expression for δ results in $\delta = 2$. If we wish to reject the null hypothesis when the difference in means $\mu_1 - \mu_2 = 0.5$ with probability at least 0.95 (power = 0.95) with $\alpha = 0.05$, then referring to Figure 2.13 we find that the required sample size on the horizontal axis is 16, approximately. This is the total sample size, so the sample size in each population should be

$$n = 16/2 = 8.$$

In our example, the experimenter actually used a sample size of 10. The experimenter could have elected to increase the sample size slightly to guard against the possibility that the prior estimate of the common standard deviation σ was too conservative and was likely to be somewhat larger than 0.25.

Operating characteristic curves often play an important role in the choice of sample size in experimental design problems. Their use in this respect is discussed in subsequent chapters. For a discussion of the uses of operating characteristic curves for other simple comparative experiments similar to the two-sample t -test, see Montgomery and Runger (2011).

Many statistics software packages can also assist the experimenter in performing power and sample size calculations. The following boxed display illustrates several computations for the Portland cement mortar problem from the power and sample size routine for the two-sample t test in Minitab. The first section of output repeats the analysis performed with the OC curves; find the sample size necessary for detecting the critical difference in means of 0.5 kgf/cm^2 , assuming that the standard deviation of strength is 0.25 kgf/cm^2 . Notice that the answer obtained from Minitab, $n_1 = n_2 = 8$, is identical to the value obtained from the OC curve analysis. The second section of the output computes the power for the case where the critical difference in means is much smaller; only 0.25 kgf/cm^2 . The power has dropped considerably, from over 0.95 to 0.562. The final section determines the sample sizes that would be necessary to detect an actual difference in means of 0.25 kgf/cm^2 with a power of at least 0.9. The required sample size turns out to be considerably larger, $n_1 = n_2 = 23$.



■ **FIGURE 2.13** Power Curves (from JMP) for the Two-Sample t -Test Assuming Equal Variances and $\alpha = 0.05$. The Sample Size on the Horizontal Axis is the Total sample Size, so the sample Size in Each population is $n = \text{sample size from graph}/2$.

Power and Sample Size

2-Sample t -Test

Testing mean 1 = mean 2 (versus not =)

Calculating power for mean 1 = mean 2 + difference

Alpha = 0.05 Sigma = 0.25

| Difference | Sample Size | Target Power | Actual Power |
|------------|-------------|--------------|--------------|
| 0.5 | 8 | 0.9500 | 0.9602 |

Power and Sample Size

2-Sample t -Test

Testing mean 1 = mean 2 (versus not =)

Calculating power for mean 1 = mean 2 + difference

Alpha = 0.05 Sigma = 0.25

| Difference | Sample Size | Power |
|------------|-------------|--------|
| 0.25 | 10 | 0.5620 |

Power and Sample Size

2-Sample t -Test

Testing mean 1 = mean 2 (versus not =)

Calculating power for mean 1 = mean 2 + difference

Alpha = 0.05 Sigma = 0.25

| Difference | Sample Size | Target Power | Actual Power |
|------------|-------------|--------------|--------------|
| 0.25 | 23 | 0.9000 | 0.9125 |

2.4.4 The Case Where $\sigma_1^2 \neq \sigma_2^2$

If we are testing

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

and cannot reasonably assume that the variances σ_1^2 and σ_2^2 are equal, then the two-sample t -test must be modified slightly. The test statistic becomes

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (2.31)$$

This statistic is not distributed exactly as t . However, the distribution of t_0 is well approximated by t if we use

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} \quad (2.32)$$

as the number of degrees of freedom. A strong indication of unequal variances on a normal probability plot would be a situation calling for this version of the t -test. You should be able to develop an equation for finding that confidence interval on the difference in mean for the unequal variances case easily.

EXAMPLE 2.1

Nerve preservation is important in surgery because accidental injury to the nerve can lead to post-surgical problems such as numbness, pain, or paralysis. Nerves are usually identified by their appearance and relationship to nearby structures or detected by local electrical stimulation (electromyography), but it is relatively easy to overlook them. An article in *Nature Biotechnology* (“Fluorescent Peptides

Highlight Peripheral Nerves During Surgery in Mice,” Vol. 29, 2011) describes the use of a fluorescently labeled peptide that binds to nerves to assist in identification. Table 2.3 shows the normalized fluorescence after two hours for nerve and muscle tissue for 12 mice (the data were read from a graph in the paper).

We would like to test the hypothesis that the mean normalized fluorescence after two hours is greater for nerve tissue than for muscle tissue. That is, if μ_δ is the mean normalized fluorescence for nerve tissue and μ_μ is the mean normalized fluorescence for muscle tissue, we want to test

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 > \mu_2$$

The descriptive statistics output from Minitab is shown below:

| Variable | N | Mean | StDev | Minimum | Median | Maximum |
|-----------|----|------|-------|---------|--------|---------|
| Nerve | 12 | 4228 | 1918 | 450 | 4825 | 6625 |
| Non-nerve | 12 | 2534 | 961 | 1130 | 2650 | 3900 |

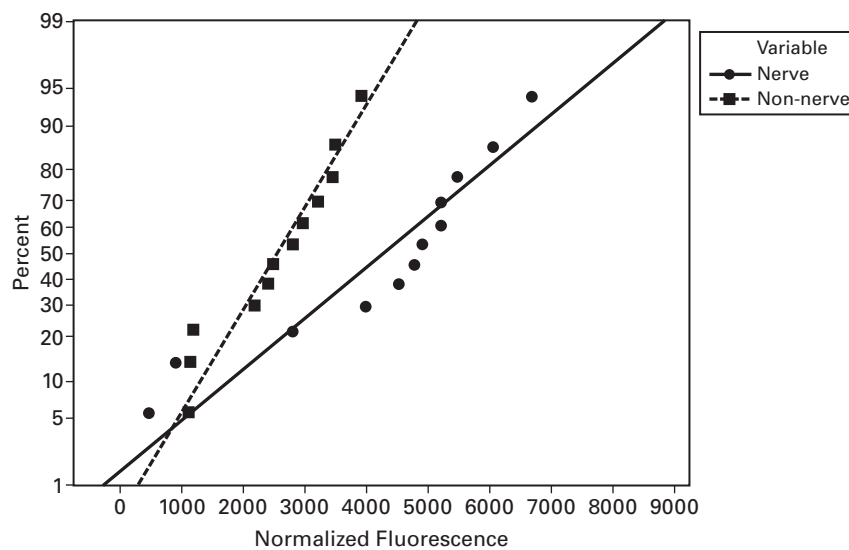
TABLE 2.3
Normalized Fluorescence After Two Hours

| Observation | Nerve | Muscle |
|-------------|-------|--------|
| 1 | 6625 | 3900 |
| 2 | 6000 | 3500 |
| 3 | 5450 | 3450 |
| 4 | 5200 | 3200 |
| 5 | 5175 | 2980 |
| 6 | 4900 | 2800 |
| 7 | 4750 | 2500 |
| 8 | 4500 | 2400 |
| 9 | 3985 | 2200 |
| 10 | 900 | 1200 |
| 11 | 450 | 1150 |
| 12 | 2800 | 1130 |

Notice that the two sample standard deviations are quite different, so the assumption of equal variances in the pooled t -test may not be appropriate. Figure 2.14 is the normal probability plot from Minitab for the two samples. This plot also indicates that the two population variances are probably not the same.

Because the equal variance assumption is not appropriate here, we will use the two-sample t -test described in this section to test the hypothesis of equal means. The test statistic, Equation 2.31, is

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{4228 - 2534}{\sqrt{\frac{(1918)^2}{12} + \frac{(961)^2}{12}}} = 2.7354$$



■ FIGURE 2.14 Normalized Fluorescence Data from Table 2.3

The number of degrees of freedom are calculated from Equation 2.32:

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}} = \frac{\left(\frac{(1918)^2}{12} + \frac{(961)^2}{12}\right)^2}{\frac{[(1918)^2/12]^2}{11} + \frac{[(961)^2/12]^2}{11}} = 16.1955$$

If we are going to find a P -value from a table of the t -distribution, we should round the degrees of freedom down to 16. Most computer programs interpolate to determine the P -value. The Minitab output for the two-sample t -test is shown below. Since the P -value reported is small (0.015), we would reject the null hypothesis and conclude that the mean normalized fluorescence for nerve tissue is greater than the mean normalized fluorescence for muscle tissue.

```
Difference = mu (Nerve) - mu (Non-nerve)
Estimate for difference: 1694
95% lower bound for difference: 613
T-Test of difference = 0 (vs >): T-Value = 2.74 P-Value = 0.007 DF = 16
```

2.4.5 The Case Where σ_1^2 and σ_2^2 Are Known

If the variances of both populations are **known**, then the hypotheses

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

may be tested using the statistic

$$Z_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (2.33)$$

If both populations are normal, or if the sample sizes are large enough so that the central limit theorem applies, the distribution of Z_0 is $N(0, 1)$ if the null hypothesis is true. Thus, the critical region would be found using the normal distribution rather than the t . Specifically, we would reject H_0 if $|Z_0| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution. This procedure is sometimes called the **two-sample Z-test**. A P -value approach can also be used with this test. The P -value would be found as $P = 2 [1 - \Phi(|Z_0|)]$, where $\Phi(x)$ is the cumulative standard normal distribution evaluated at the point x .

Unlike the t -test of the previous sections, the test on means with known variances does not require the assumption of sampling from normal populations. One can use the central limit theorem to justify an approximate normal distribution for the difference in sample means $\bar{y}_1 - \bar{y}_2$.

The $100(1 - \alpha)$ percent confidence interval on $\mu_1 - \mu_2$ where the variances are known is

$$\bar{y}_1 - \bar{y}_2 - Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{y}_1 - \bar{y}_2 + Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (2.34)$$

As noted previously, the confidence interval is often a useful supplement to the hypothesis testing procedure.

2.4.6 Comparing a Single Mean to a Specified Value

Some experiments involve comparing only one population mean μ to a specified value, say, μ_0 . The hypotheses are

$$H_0: \mu = \mu_0$$

If the population is normal with known variance, or if the population is nonnormal but the sample size is large enough so that the central limit theorem applies, then the hypothesis may be tested using a direct application of the normal distribution. The **one-sample Z-test** statistic is

$$Z_0 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \quad (2.35)$$

If $H_0: \mu = \mu_0$ is true, then the distribution of Z_0 is $N(0, 1)$. Therefore, the decision rule for $H_0: \mu = \mu_0$ is to reject the null hypothesis if $|Z_0| > Z_{\alpha/2}$. A P -value approach could also be used.

The value of the mean μ_0 specified in the null hypothesis is usually determined in one of three ways. It may result from past evidence, knowledge, or experimentation. It may be the result of some theory or model describing the situation under study. Finally, it may be the result of contractual specifications.

The $100(1 - \alpha)$ percent confidence interval on the true population mean is

$$\bar{y} - Z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{y} + Z_{\alpha/2}\sigma/\sqrt{n} \quad (2.36)$$

EXAMPLE 2.2

A supplier submits lots of fabric to a textile manufacturer. The customer wants to know if the lot average breaking strength exceeds 200 psi. If so, she wants to accept the lot. Past experience indicates that a reasonable value for the variance of breaking strength is $100(\text{psi})^2$. The hypotheses to be tested are

$$H_0: \mu = 200$$

$$H_1: \mu > 200$$

Note that this is a one-sided alternative hypothesis. Thus, we would accept the lot only if the null hypothesis $H_0: \mu = 200$ could be rejected (i.e., if $Z_0 > Z_\alpha$).

Four specimens are randomly selected, and the average breaking strength observed is $\bar{y} = 214$ psi. The value of the test statistic is

$$Z_0 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{214 - 200}{10/\sqrt{4}} = 2.80$$

If a type I error of $\alpha = 0.05$ is specified, we find $Z_\alpha = Z_{0.05} = 1.645$ from Appendix Table I. The P -value would be computed using only the area in the upper tail of the standard normal distribution, because the alternative hypothesis is one-sided. The P -value is $P = 1 - \Phi(2.80) = 1 - 0.99744 = 0.00256$. Thus H_0 is rejected, and we conclude that the lot average breaking strength exceeds 200 psi.

If the variance of the population is unknown, we must make the additional assumption that the population is normally distributed, although moderate departures from normality will not seriously affect the results.

To test $H_0: \mu = \mu_0$ in the variance unknown case, the sample variance S^2 is used to estimate σ^2 . Replacing σ with S in Equation 2.35, we have the **one-sample t-test** statistic

$$t_0 = \frac{\bar{y} - \mu_0}{S/\sqrt{n}} \quad (2.37)$$

The null hypothesis $H_0: \mu = \mu_0$ would be rejected if $|t_0| > t_{\alpha/2, n-1}$, where $t_{\alpha/2, n-1}$ denotes the upper $\alpha/2$ percentage point of the t distribution with $n - 1$ degrees of freedom. A P -value approach could also be used. The $100(1 - \alpha)$ percent confidence interval in this case is

$$\bar{y} - t_{\alpha/2, n-1}S/\sqrt{n} \leq \mu \leq \bar{y} + t_{\alpha/2, n-1}S/\sqrt{n} \quad (2.38)$$

2.4.7 Summary

Tables 2.4 and 2.5 summarize the t -test and z -test procedures discussed above for sample means. Critical regions are shown for both two-sided and one-sided alternative hypotheses.

■ **TABLE 2.4**
Tests on Means with Variance Known

| Hypothesis | Test Statistic | Fixed Significance Level Criteria for Rejection | P-Value |
|---|--|--|--------------------------|
| $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ | $Z_0 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$ | $ Z_0 > Z_{\alpha/2}$ | $P = 2[1 - \Phi(Z_0)]$ |
| $H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$ | | $Z_0 < -Z_\alpha$ | $P = \Phi(Z_0)$ |
| $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$ | | $Z_0 > Z_\alpha$ | $P = 1 - \Phi(Z_0)$ |
| $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ | $Z_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ | $ Z_0 > Z_{\alpha/2}$ | $P = 2[1 - \Phi(Z_0)]$ |
| $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$ | | $Z_0 < -Z_\alpha$ | $P = \Phi(Z_0)$ |
| $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$ | | $Z_0 > Z_\alpha$ | $P = 1 - \Phi(Z_0)$ |

■ **TABLE 2.5**
Tests on Means of Normal Distributions, Variance Unknown

| Hypothesis | Test Statistic | Fixed Significance Level Criteria for Rejection | P-Value |
|---|---|--|--|
| $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ | $t_0 = \frac{\bar{y} - \mu_0}{S/\sqrt{n}}$ | $ t_0 > t_{\alpha/2, n-1}$ | sum of the probability above t_0 and below $-t_0$ |
| $H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$ | | $t_0 < -t_{\alpha, n-1}$ | probability below t_0 |
| $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$ | | $t_0 > t_{\alpha, n-1}$ | probability above t_0 |
| $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$ | $t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $v = n_1 + n_2 - 2$ if $\sigma_1^2 = \sigma_2^2$ | $ t_0 > t_{\alpha/2, v}$ | sum of the probability above t_0 and below $-t_0$ |
| $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$ | $t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ | $t_0 < -t_{\alpha, v}$ | probability below t_0 |
| $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$ | $v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^2/n_1}{n_1 - 1} + \frac{S_2^2/n_2}{n_2 - 1}}$ | $t_0 > t_{\alpha, v}$ | probability above t_0 |

2.5 Inferences About the Differences in Means, Paired Comparison Designs

2.5.1 The Paired Comparison Problem

In some simple comparative experiments, we can greatly improve the precision by making comparisons within matched pairs of experimental material. For example, consider a hardness testing machine that presses a rod with a pointed tip into a metal specimen with a known force. By measuring the depth of the depression caused by the tip, the hardness of the specimen is determined. Two different tips are available for this machine, and although the precision (variability) of the measurements made by the two tips seems to be the same, it is suspected that one tip produces different mean hardness readings than the other.

An experiment could be performed as follows. A number of metal specimens (e.g., 20) could be randomly selected. Half of these specimens could be tested by tip 1 and the other half by tip 2. The exact assignment of specimens to tips would be randomly determined. Because this is a completely randomized design, the average hardness of the two samples could be compared using the t -test described in Section 2.4.

A little reflection will reveal a serious disadvantage in the completely randomized design for this problem. Suppose the metal specimens were cut from different bar stock that were produced in different heats or that were not exactly homogeneous in some other way that might affect the hardness. This lack of homogeneity between specimens will contribute to the variability of the hardness measurements and will tend to inflate the experimental error, thus making a true difference between tips harder to detect.

To protect against this possibility, consider an alternative experimental design. Assume that each specimen is large enough so that *two* hardness determinations may be made on it. This alternative design would consist of dividing each specimen into two parts, then randomly assigning one tip to one-half of each specimen and the other tip to the remaining half. The order in which the tips are tested for a particular specimen would also be randomly selected. The experiment, when performed according to this design with 10 specimens, produced the (coded) data shown in Table 2.6.

We may write a **statistical model** that describes the data from this experiment as

$$y_{ij} = \mu_i + \beta_j + \epsilon_{ij} \begin{cases} i = 1, 2 \\ j = 1, 2, \dots, 10 \end{cases} \quad (2.39)$$

■ **TABLE 2.6**
Data for the Hardness Testing Experiment

| Specimen | Tip 1 | Tip 2 |
|----------|-------|-------|
| 1 | 7 | 6 |
| 2 | 3 | 3 |
| 3 | 3 | 5 |
| 4 | 4 | 3 |
| 5 | 8 | 8 |
| 6 | 3 | 2 |
| 7 | 2 | 4 |
| 8 | 9 | 9 |
| 9 | 5 | 4 |
| 10 | 4 | 5 |

where y_{ij} is the observation on hardness for tip i on specimen j , μ_i is the true mean hardness of the i th tip, β_j is an effect on hardness due to the j th specimen, and ϵ_{ij} is a random experimental error with mean zero and variance σ_i^2 . That is, σ_1^2 is the variance of the hardness measurements from tip 1, and σ_2^2 is the variance of the hardness measurements from tip 2.

Note that if we compute the j th paired difference

$$d_j = y_{1j} - y_{2j} \quad j = 1, 2, \dots, 10 \quad (2.40)$$

the expected value of this difference is

$$\begin{aligned} \mu_d &= E(d_j) \\ &= E(y_{1j} - y_{2j}) \\ &= E(y_{1j}) - E(y_{2j}) \\ &= \mu_1 + \beta_j - (\mu_2 + \beta_j) \\ &= \mu_1 - \mu_2 \end{aligned}$$

That is, we may make inferences about the difference in the mean hardness readings of the two tips $\mu_1 - \mu_2$ by making inferences about the mean of the differences μ_d . Notice that the additive effect of the specimens β_j cancels out when the observations are paired in this manner.

Testing $H_0: \mu_1 = \mu_2$ is equivalent to testing

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

This is a single-sample t -test. The test statistic for this hypothesis is

$$t_0 = \frac{\bar{d}}{S_d/\sqrt{n}} \quad (2.41)$$

where

$$\bar{d} = \frac{1}{n} \sum_{j=1}^n d_j \quad (2.42)$$

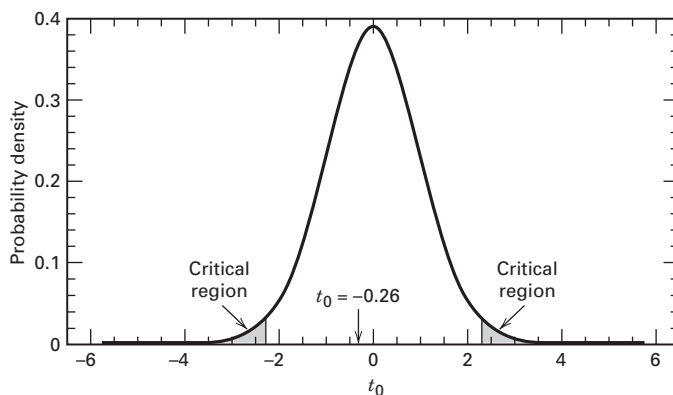
is the sample mean of the differences and

$$S_d = \left[\frac{\sum_{j=1}^n (d_j - \bar{d})^2}{n-1} \right]^{1/2} = \left[\frac{\sum_{j=1}^n d_j^2 - \frac{1}{n} \left(\sum_{j=1}^n d_j \right)^2}{n-1} \right]^{1/2} \quad (2.43)$$

is the sample standard deviation of the differences. $H_0: \mu_d = 0$ would be rejected if $|t_0| > t_{\alpha/2, n-1}$. A P -value approach could also be used. Because the observations from the factor levels are “paired” on each experimental unit, this procedure is usually called the **paired t -test**.

For the data in Table 2.6, we find

$$\begin{aligned} d_1 &= 7 - 6 = 1 & d_6 &= 3 - 2 = 1 \\ d_2 &= 3 - 3 = 0 & d_7 &= 2 - 4 = -2 \\ d_3 &= 3 - 5 = -2 & d_8 &= 9 - 9 = 0 \\ d_4 &= 4 - 3 = 1 & d_9 &= 5 - 4 = 1 \\ d_5 &= 8 - 8 = 0 & d_{10} &= 4 - 5 = -1 \end{aligned}$$



■ **FIGURE 2.15** The reference distribution (t with 9 degrees of freedom) for the hardness testing problem

Thus,

$$\bar{d} = \frac{1}{n} \sum_{j=1}^n d_j = \frac{1}{10} (-1) = -0.10$$

$$S_d = \left[\frac{\sum_{j=1}^n d_j^2 - \frac{1}{n} \left(\sum_{j=1}^n d_j \right)^2}{n-1} \right]^{1/2} = \left[\frac{13 - \frac{1}{10}(-1)^2}{10-1} \right]^{1/2} = 1.20$$

Suppose we choose $\alpha = 0.05$. Now to make a decision, we would compute t_0 and reject H_0 if $|t_0| > t_{0.025,9} = 2.262$.

The computed value of the paired t -test statistic is

$$t_0 = \frac{\bar{d}}{S_d/\sqrt{n}} = \frac{-0.10}{1.20/\sqrt{10}} = -0.26$$

and because $|t_0| = 0.26 \not> t_{0.025,9} = 2.262$, we cannot reject the hypothesis $H_0: \mu_d = 0$. That is, there is no evidence to indicate that the two tips produce different hardness readings. Figure 2.15 shows the t_0 distribution with 9 degrees of freedom, the reference distribution for this test, with the value of t_0 shown relative to the critical region.

Table 2.7 shows the computer output from the Minitab paired t -test procedure for this problem. Notice that the P -value for this test is $P \approx 0.80$, implying that we cannot reject the null hypothesis at *any* reasonable level of significance.

■ **TABLE 2.7**

Minitab Paired t -Test Results for the Hardness Testing Example

| Paired T for Tip 1–Tip 2 | | | | |
|--------------------------|----|--------|-----------|---------|
| | N | Mean | Std. Dev. | SE Mean |
| Tip 1 | 10 | 4.800 | 2.394 | 0.757 |
| Tip 2 | 10 | 4.900 | 2.234 | 0.706 |
| Difference | 10 | -0.100 | 1.197 | 0.379 |

95% CI for mean difference: (-0.956, 0.756)
 t-Test of mean difference = 0 (vs not = 0):
 T-Value = -0.26 P-Value = 0.798

2.5.2 Advantages of the Paired Comparison Design

The design actually used for this experiment is called the **paired comparison design**, and it illustrates the blocking principle discussed in Section 1.3. Actually, it is a special case of a more general type of design called the **randomized block design**. The term *block* refers to a relatively homogeneous experimental unit (in our case, the metal specimens are the blocks), and the block represents a restriction on complete randomization because the treatment combinations are only randomized within the block. We look at designs of this type in Chapter 4. In that chapter the mathematical model for the design, Equation 2.39, is written in a slightly different form.

Before leaving this experiment, several points should be made. Note that, although $2n = 2(10) = 20$ observations have been taken, only $n - 1 = 9$ degrees of freedom are available for the t statistic. (We know that as the degrees of freedom for t increase, the test becomes more sensitive.) By blocking or pairing we have effectively “lost” $n - 1$ degrees of freedom, but we hope we have gained a better knowledge of the situation by eliminating an additional source of variability (the difference between specimens).

We may obtain an indication of the quality of information produced from the paired design by comparing the standard deviation of the differences S_d with the pooled standard deviation S_p that would have resulted had the experiment been conducted in a completely randomized manner and the data of Table 2.5 been obtained. Using the data in Table 2.5 as two independent samples, we compute the pooled standard deviation from Equation 2.25 to be $S_p = 2.32$. Comparing this value to $S_d = 1.20$, we see that blocking or pairing has reduced the estimate of variability by nearly 50 percent.

Generally, when we don’t block (or pair the observations) when we really should have, S_p will always be larger than S_d . It is easy to show this formally. If we pair the observations, it is easy to show that S_d^2 is an unbiased estimator of the variance of the differences d_j under the model in Equation 2.39 because the block effects (the β_j) cancel out when the differences are computed. However, if we don’t block (or pair) and treat the observations as two independent samples, then S_p^2 is not an unbiased estimator of σ^2 under the model in Equation 2.39. In fact, assuming that both population variances are equal,

$$E(S_p^2) = \sigma^2 + \sum_{j=1}^n \beta_j^2$$

That is, the block effects β_j inflate the variance estimate. This is why blocking serves as a **noise reduction** design technique.

We may also express the results of this experiment in terms of a confidence interval on $\mu_1 - \mu_2$. Using the paired data, a 95 percent confidence interval on $\mu_1 - \mu_2$ is

$$\begin{aligned} \bar{d} \pm t_{0.025,9} S_d / \sqrt{n} \\ -0.10 \pm (2.262)(1.20) / \sqrt{10} \\ -0.10 \pm 0.86 \end{aligned}$$

Conversely, using the pooled or independent analysis, a 95 percent confidence interval on $\mu_1 - \mu_2$ is

$$\begin{aligned} \bar{y}_1 - \bar{y}_2 \pm t_{0.025,18} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ 4.80 - 4.90 \pm (2.101)(2.32) \sqrt{\frac{1}{10} + \frac{1}{10}} \\ -0.10 \pm 2.18 \end{aligned}$$

The confidence interval based on the paired analysis is much narrower than the confidence interval from the independent analysis. This again illustrates the **noise reduction** property of blocking.

Blocking is not always the best design strategy. If the within-block variability is the same as the between-block variability, the variance of $\bar{y}_1 - \bar{y}_2$ will be the same regardless of which design is used. Actually, blocking in this situation would be a poor choice of design because blocking results in the loss of $n - 1$ degrees of freedom and will actually lead to a wider confidence interval on $\mu_1 - \mu_2$. A further discussion of blocking is given in Chapter 4.

2.6 Inferences About the Variances of Normal Distributions

In many experiments, we are interested in possible differences in the mean response for two treatments. However, in some experiments it is the comparison of variability in the data that is important. In the food and beverage industry, for example, it is important that the variability of filling equipment be small so that all packages have close to the nominal net weight or volume of content. In chemical laboratories, we may wish to compare the variability of two analytical methods. We now briefly examine tests of hypotheses and confidence intervals for variances of normal distributions. Unlike the tests on means, the procedures for tests on variances are rather sensitive to the normality assumption. A good discussion of the normality assumption is in Appendix 2A of Davies (1956).

Suppose we wish to test the hypothesis that the variance of a normal population equals a constant, for example, σ_0^2 . Stated formally, we wish to test

$$\begin{aligned} H_0: \sigma^2 &= \sigma_0^2 \\ H_1: \sigma^2 &\neq \sigma_0^2 \end{aligned} \quad (2.44)$$

The test statistic for Equation 2.44 is

$$\chi_0^2 = \frac{SS}{\sigma_0^2} = \frac{(n-1)S^2}{\sigma_0^2} \quad (2.45)$$

where $SS = \sum_{i=1}^n (y_i - \bar{y})^2$ is the corrected sum of squares of the sample observations. The appropriate reference distribution for χ_0^2 is the chi-square distribution with $n - 1$ degrees of freedom. The null hypothesis is rejected if $\chi_0^2 > \chi_{\alpha/2, n-1}^2$ or if $\chi_0^2 < \chi_{1-(\alpha/2), n-1}^2$, where $\chi_{\alpha/2, n-1}^2$ and $\chi_{1-(\alpha/2), n-1}^2$ are the upper $\alpha/2$ and lower $1 - (\alpha/2)$ percentage points of the chi-square distribution with $n - 1$ degrees of freedom, respectively. Table 2.8 gives the critical regions for the one-sided alternative hypotheses. The $100(1 - \alpha)$ percent confidence interval on σ^2 is

$$\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-(\alpha/2), n-1}^2} \quad (2.46)$$

Now consider testing the equality of the variances of two normal populations. If independent random samples of size n_1 and n_2 are taken from populations 1 and 2, respectively, the test statistic for

$$\begin{aligned} H_0: \sigma_1^2 &= \sigma_2^2 \\ H_1: \sigma_1^2 &\neq \sigma_2^2 \end{aligned} \quad (2.47)$$

is the ratio of the sample variances

$$F_0 = \frac{S_1^2}{S_2^2} \quad (2.48)$$

■ TABLE 2.8
Tests on Variances of Normal Distributions

| Hypothesis | Test Statistic | Fixed Significance Level Criteria for Rejection |
|---|--|---|
| $H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$ | | $\chi_0^2 > \chi_{\alpha/2, n-1}^2$ or $\chi_0^2 < \chi_{1-\alpha/2, n-1}^2$ |
| $H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$ | $\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$ | $\chi_0^2 < \chi_{1-\alpha, n-1}^2$ |
| $H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$ | | $\chi_0^2 > \chi_{\alpha, n-1}^2$ |
| $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$ | $F_0 = \frac{S_1^2}{S_2^2}$ | $F_0 > F_{\alpha/2, n_1-1, n_2-1}$ or $F_0 < F_{1-\alpha/2, n_1-1, n_2-1}$ |
| $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$ | $F_0 = \frac{S_2^2}{S_1^2}$ | $F_0 > F_{\alpha, n_2-1, n_1-1}$ |
| $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$ | $F_0 = \frac{S_1^2}{S_2^2}$ | $F_0 > F_{\alpha, n_1-1, n_2-1}$ |

The appropriate reference distribution for F_0 is the F distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom. The null hypothesis would be rejected if $F_0 > F_{\alpha/2, n_1-1, n_2-1}$ or if $F_0 < F_{1-(\alpha/2), n_1-1, n_2-1}$, where $F_{\alpha/2, n_1-1, n_2-1}$ and $F_{1-(\alpha/2), n_1-1, n_2-1}$ denote the upper $\alpha/2$ and lower $1 - (\alpha/2)$ percentage points of the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Table IV of the Appendix gives only upper-tail percentage points of F ; however, the upper- and lower-tail points are related by

$$F_{1-\alpha, v_1, v_2} = \frac{1}{F_{\alpha, v_2, v_1}} \quad (2.49)$$

Critical values for the one-sided alternative hypothesis are given in Table 2.8. Test procedures for more than two variances are discussed in Section 3.4.3. We will also discuss the use of the variance or standard deviation as a response variable in more general experimental settings.

EXAMPLE 2.3

A chemical engineer is investigating the inherent variability of two types of test equipment that can be used to monitor the output of a production process. He suspects that the old equipment, type 1, has a larger variance than the new one. Thus, he wishes to test the hypothesis

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

Two random samples of $n_1 = 12$ and $n_2 = 10$ observations are taken, and the sample variances are $S_1^2 = 14.5$ and $S_2^2 =$

10.8. The test statistic is

$$F_0 = \frac{S_1^2}{S_2^2} = \frac{14.5}{10.8} = 1.34$$

From Appendix Table IV we find that $F_{0.05, 11, 9} = 3.10$, so the null hypothesis cannot be rejected. That is, we have found insufficient statistical evidence to conclude that the variance of the old equipment is greater than the variance of the new equipment.

The $100(1 - \alpha)$ confidence interval for the ratio of the population variances σ_1^2/σ_2^2 is

$$\frac{S_1^2}{S_2^2} F_{1-\alpha/2, n_2-1, n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} F_{\alpha/2, n_2-1, n_1-1} \quad (2.50)$$

To illustrate the use of Equation 2.50, the 95 percent confidence interval for the ratio of variances σ_1^2/σ_2^2 in Example 2.2 is, using $F_{0.025, 9, 11} = 3.59$ and $F_{0.975, 9, 11} = 1/F_{0.025, 11, 9} = 1/3.92 = 0.255$,

$$\begin{aligned} \frac{14.5}{10.8} (0.255) &\leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{14.5}{10.8} (3.59) \\ 0.34 &\leq \frac{\sigma_1^2}{\sigma_2^2} \leq 4.82 \end{aligned}$$

2.7 Problems

2.1. Computer output for a random sample of data is shown below. Some of the quantities are missing. Compute the values of the missing quantities.

| Variable | N | Mean | SE Mean | Std. Dev. | Variance | Minimum | Maximum |
|----------|---|-------|---------|-----------|----------|---------|---------|
| Y | 9 | 19.96 | ? | 3.12 | ? | 15.94 | 27.16 |

2.2. Computer output for a random sample of data is shown below. Some of the quantities are missing. Compute the values of the missing quantities.

| Variable | N | Mean | SE Mean | Std. Dev. | Sum |
|----------|----|------|---------|-----------|---------|
| Y | 16 | ? | 0.159 | ? | 399.851 |

2.3. Suppose that we are testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Calculate the P -value for the following observed values of the test statistic:

- (a) $Z_0 = 2.25$ (b) $Z_0 = 1.55$ (c) $Z_0 = 2.10$
 (d) $Z_0 = 1.95$ (e) $Z_0 = -0.10$

2.4. Suppose that we are testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$. Calculate the P -value for the following observed values of the test statistic:

- (a) $Z_0 = 2.45$ (b) $Z_0 = -1.53$ (c) $Z_0 = 2.15$
 (d) $Z_0 = 1.95$ (e) $Z_0 = -0.25$

2.5. Consider the computer output shown below.

| One-Sample Z | | | | | | |
|--------------------------------------|---------|---------|--------------------|---|---|--|
| Test of $\mu = 30$ vs not = 30 | | | | | | |
| The assumed standard deviation = 1.2 | | | | | | |
| N | Mean | SE Mean | 95% CI | Z | P | |
| 16 | 31.2000 | 0.3000 | (30.6120, 31.7880) | ? | ? | |

- (a) Fill in the missing values in the output. What conclusion would you draw?
 (b) Is this a one-sided or two-sided test?

- (c) Use the output and the normal table to find a 99 percent CI on the mean.
 (d) What is the P -value if the alternative hypothesis is $H_1 : \mu > 30$?

2.6. Suppose that we are testing $H_0 : \mu_1 = \mu_2$ versus $H_0 : \mu_1 \neq \mu_2$ where the two sample sizes are $n_1 = n_2 = 12$. Both sample variances are unknown but assumed equal. Find bounds on the P -value for the following observed values of the test statistic.

- (a) $t_0 = 2.30$ (b) $t_0 = 3.41$ (c) $t_0 = 1.95$ (d) $t_0 = -2.45$

2.7. Suppose that we are testing $H_0 : \mu_1 = \mu_2$ versus $H_0 : \mu_1 > \mu_2$ where the two sample sizes are $n_1 = n_2 = 10$. Both sample variances are unknown but assumed equal. Find bounds on the P -value for the following observed values of the test statistic.

- (a) $t_0 = 2.31$ (b) $t_0 = 3.60$ (c) $t_0 = 1.95$ (d) $t_0 = 2.19$

2.8. Consider the following sample data: 9.37, 13.04, 11.69, 8.21, 11.18, 10.41, 13.15, 11.51, 13.21, and 7.75. Is it reasonable to assume that this data is a sample from a normal distribution? Is there evidence to support a claim that the mean of the population is 10?

2.9. A computer program has produced the following output for a hypothesis-testing problem:

| | |
|---|--------------|
| Difference in sample means: | 2.35 |
| Degrees of freedom: | 18 |
| Standard error of the difference in sample means: | ? |
| Test statistic: | $t_0 = 2.01$ |
| P -value: | 0.0298 |

- (a) What is the missing value for the standard error?
 (b) Is this a two-sided or a one-sided test?
 (c) If $\alpha = 0.05$, what are your conclusions?
 (d) Find a 90% two-sided CI on the difference in means.

2.10. A computer program has produced the following output for a hypothesis-testing problem:

```
Difference in sample means: 11.5
Degrees of freedom: 24
Standard error of the difference in sample means: ?
Test statistic:  $t_0 = -1.88$ 
P-value: 0.0723
```

- What is the missing value for the standard error?
- Is this a two-sided or a one-sided test?
- If $\alpha = 0.05$, what are your conclusions?
- Find a 95% two-sided CI on the difference in means.

2.11. Suppose that we are testing $H_0: \mu = \mu_0$ versus $H_1: \mu > \mu_0$ with a sample size of $n = 15$. Calculate bounds on the P -value for the following observed values of the test statistic:

- (a) $t_0 = 2.35$ (b) $t_0 = 3.55$ (c) $t_0 = 2.00$ (d) $t_0 = 1.55$

2.12. Suppose that we are testing $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$ with a sample size of $n = 10$. Calculate bounds on the P -value for the following observed values of the test statistic:

- (a) $t_0 = 2.48$ (b) $t_0 = -3.95$ (c) $t_0 = 2.69$
 (d) $t_0 = 1.88$ (e) $t_0 = -1.25$

2.13. Consider the computer output shown below.

```
One-Sample T: Y
Test of mu=91 vs. not=91

Variable N Mean Std. Dev. SE Mean 95% CI T P
Y 25 92.5805 ? 0.4673 (91.6160, ?) 3.38 0.002
```

- Fill in the missing values in the output. Can the null hypothesis be rejected at the 0.05 level? Why?
- Is this a one-sided or a two-sided test?
- If the hypotheses had been $H_0: \mu = 90$ versus $H_1: \mu \neq 90$ would you reject the null hypothesis at the 0.05 level?
- Use the output and the t table to find a 99 percent two-sided CI on the mean.
- What is the P -value if the alternative hypothesis is $H_1: \mu > 91$?

2.14. Consider the computer output shown below.

```
One-Sample T: Y
Test of mu=25 vs > 25

Variable N Mean Std. Dev. SE Mean 95% Lower Bound T P
Y 12 25.6818 ? 0.3360 ? ? 0.034
```

- How many degrees of freedom are there on the t -test statistic?
- Fill in the missing information.

2.15. Consider the computer output shown below.

```
Two-Sample T-Test and CI: Y1, Y2

Two-sample T for Y1 vs Y2

      N      Mean      Std. Dev.      SE Mean
Y1    20    50.19      1.71          0.38
Y2    20    52.52      2.48          0.55

Difference = mu (X1) - mu (X2)
Estimate for difference: -2.33341
95% CI for difference: (-3.69547, -0.97135)
T-Test of difference=0 (vs not =): T-Value=-3.47
P-Value=0.001 DF=38
Both use Pooled Std. Dev.=2.1277
```

- Can the null hypothesis be rejected at the 0.05 level? Why?
- Is this a one-sided or a two-sided test?
- If the hypotheses had been $H_0: \mu_1 - \mu_2 = 2$ versus $H_1: \mu_1 - \mu_2 \neq 2$ would you reject the null hypothesis at the 0.05 level?
- If the hypotheses had been $H_0: \mu_1 - \mu_2 = 2$ versus $H_1: \mu_1 - \mu_2 < 2$ would you reject the null hypothesis at the 0.05 level? Can you answer this question without doing any additional calculations? Why?
- Use the output and the t table to find a 95 percent upper confidence bound on the difference in means.
- What is the P -value if the hypotheses are $H_0: \mu_1 - \mu_2 = 2$ versus $H_1: \mu_1 - \mu_2 \neq 2$?

2.16. The breaking strength of a fiber is required to be at least 150 psi. Past experience has indicated that the standard deviation of breaking strength is $\sigma = 3$ psi. A random sample of four specimens is tested, and the results are $y_1 = 145$, $y_2 = 153$, $y_3 = 150$, and $y_4 = 147$.

- State the hypotheses that you think should be tested in this experiment.
- Test these hypotheses using $\alpha = 0.05$. What are your conclusions?
- Find the P -value for the test in part (b).
- Construct a 95 percent confidence interval on the mean breaking strength.


2.17. The viscosity of a liquid detergent is supposed to average 800 centistokes at 25°C. A random sample of 16 batches of detergent is collected, and the average viscosity is 812. Suppose we know that the standard deviation of viscosity is $\sigma = 25$ centistokes.

- State the hypotheses that should be tested.
- Test these hypotheses using $\alpha = 0.05$. What are your conclusions?
- What is the P -value for the test?
- Find a 95 percent confidence interval on the mean.

2.18. The diameters of steel shafts produced by a certain manufacturing process should have a mean diameter of 0.255 inches. The diameter is known to have a standard deviation of $\sigma = 0.0001$ inch. A random sample of 10 shafts has an average diameter of 0.2545 inch.

- (a) Set up appropriate hypotheses on the mean μ .
- (b) Test these hypotheses using $\alpha = 0.05$. What are your conclusions?
- (c) Find the P -value for this test.
- (d) Construct a 95 percent confidence interval on the mean shaft diameter.

2.19. A normally distributed random variable has an unknown mean μ and a known variance $\sigma^2 = 9$. Find the sample size required to construct a 95 percent confidence interval on the mean that has total length of 1.0.

 **2.20.** The shelf life of a carbonated beverage is of interest. Ten bottles are randomly selected and tested, and the following results are obtained:

| Days | |
|------|-----|
| 108 | 138 |
| 124 | 163 |
| 124 | 159 |
| 106 | 134 |
| 115 | 139 |

- (a) We would like to demonstrate that the mean shelf life exceeds 120 days. Set up appropriate hypotheses for investigating this claim.
- (b) Test these hypotheses using $\alpha = 0.01$. What are your conclusions?
- (c) Find the P -value for the test in part (b).
- (d) Construct a 99 percent confidence interval on the mean shelf life.

2.21. Consider the shelf life data in Problem 2.20. Can shelf life be described or modeled adequately by a normal distribution? What effect would the violation of this assumption have on the test procedure you used in solving Problem 2.15?

2.22. The time to repair an electronic instrument is a normally distributed random variable measured in hours. The repair times for 16 such instruments chosen at random are as follows:

| Hours | | | |
|-------|-----|-----|-----|
| 159 | 280 | 101 | 212 |
| 224 | 379 | 179 | 264 |
| 222 | 362 | 168 | 250 |
| 149 | 260 | 485 | 170 |

- (a) You wish to know if the mean repair time exceeds 225 hours. Set up appropriate hypotheses for investigating this issue.
- (b) Test the hypotheses you formulated in part (a). What are your conclusions? Use $\alpha = 0.05$.
- (c) Find the P -value for the test.
- (d) Construct a 95 percent confidence interval on mean repair time.


2.23. Reconsider the repair time data in Problem 2.22. Can repair time, in your opinion, be adequately modeled by a normal distribution?

2.24. Two machines are used for filling plastic bottles with a net volume of 16.0 ounces. The filling processes can be assumed to be normal, with standard deviations of $\sigma_1 = 0.015$ and $\sigma_2 = 0.018$. The quality engineering department suspects that both machines fill to the same net volume, whether or not this volume is 16.0 ounces. An experiment is performed by taking a random sample from the output of each machine.

| Machine 1 | | Machine 2 | |
|-----------|-------|-----------|-------|
| 16.03 | 16.01 | 16.02 | 16.03 |
| 16.04 | 15.96 | 15.97 | 16.04 |
| 16.05 | 15.98 | 15.96 | 16.02 |
| 16.05 | 16.02 | 16.01 | 16.01 |
| 16.02 | 15.99 | 15.99 | 16.00 |

- (a) State the hypotheses that should be tested in this experiment.
- (b) Test these hypotheses using $\alpha = 0.05$. What are your conclusions?
- (c) Find the P -value for this test.
- (d) Find a 95 percent confidence interval on the difference in mean fill volume for the two machines.


2.25. Two types of plastic are suitable for use by an electronic calculator manufacturer. The breaking strength of this plastic is important. It is known that $\sigma_1 = \sigma_2 = 1.0$ psi. From random samples of $n_1 = 10$ and $n_2 = 12$ we obtain $\bar{y}_1 = 162.5$ and $\bar{y}_2 = 155.0$. The company will not adopt plastic 1 unless its breaking strength exceeds that of plastic 2 by at least 10 psi. Based on the sample information, should they use plastic 1? In answering this question, set up and test appropriate hypotheses using $\alpha = 0.01$. Construct a 99 percent confidence interval on the true mean difference in breaking strength.

2.26. The following are the burning times (in minutes) of chemical flares of two different formulations. The design engineers are interested in both the mean and variance of the burning times. 

| Type 1 | | Type 2 | |
|--------|----|--------|----|
| 65 | 82 | 64 | 56 |
| 81 | 67 | 71 | 69 |
| 57 | 59 | 83 | 74 |
| 66 | 75 | 59 | 82 |
| 82 | 70 | 65 | 79 |

- (a) Test the hypothesis that the two variances are equal. Use $\alpha = 0.05$.
- (b) Using the results of (a), test the hypothesis that the mean burning times are equal. Use $\alpha = 0.05$. What is the P -value for this test?

- (c) Discuss the role of the normality assumption in this problem. Check the assumption of normality for both types of flares.

 **2.27.** An article in *Solid State Technology*, “Orthogonal Design for Process Optimization and Its Application to Plasma Etching” by G. Z. Yin and D. W. Jillie (May 1987) describes an experiment to determine the effect of the C_2F_6 flow rate on the uniformity of the etch on a silicon wafer used in integrated circuit manufacturing. All of the runs were made in random order. Data for two flow rates are as follows:

| C_2F_6 Flow (SCCM) | Uniformity Observation | | | | | |
|-------------------------|------------------------|-----|-----|-----|-----|-----|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 125 | 2.7 | 4.6 | 2.6 | 3.0 | 3.2 | 3.8 |
| 200 | 4.6 | 3.4 | 2.9 | 3.5 | 4.1 | 5.1 |

- (a) Does the C_2F_6 flow rate affect average etch uniformity? Use $\alpha = 0.05$.
- (b) What is the P -value for the test in part (a)?
- (c) Does the C_2F_6 flow rate affect the wafer-to-wafer variability in etch uniformity? Use $\alpha = 0.05$.
- (d) Draw box plots to assist in the interpretation of the data from this experiment.

2.28. A new filtering device is installed in a chemical unit. Before its installation, a random sample yielded the following information about the percentage of impurity: $\bar{y}_1 = 12.5$, $S_1^2 = 101.17$, and $n_1 = 8$. After installation, a random sample yielded $\bar{y}_2 = 10.2$, $S_2^2 = 94.73$, $n_2 = 9$.

- (a) Can you conclude that the two variances are equal? Use $\alpha = 0.05$.
- (b) Has the filtering device reduced the percentage of impurity significantly? Use $\alpha = 0.05$.

2.29. Photoresist is a light-sensitive material applied to semiconductor wafers so that the circuit pattern can be imaged on to the wafer. After application, the coated wafers are baked to remove the solvent in the photoresist mixture and to harden the resist. Here are measurements of photoresist thickness (in kA) for eight wafers baked at two different temperatures. Assume that all of the runs were made in random order.

| 95 °C | 100 °C |
|--------|--------|
| 11.176 | 5.263 |
| 7.089 | 6.748 |
| 8.097 | 7.461 |
| 11.739 | 7.015 |
| 11.291 | 8.133 |
| 10.759 | 7.418 |
| 6.467 | 3.772 |
| 8.315 | 8.963 |

- (a) Is there evidence to support the claim that the higher baking temperature results in wafers with a lower mean photoresist thickness? Use $\alpha = 0.05$.
- (b) What is the P -value for the test conducted in part (a)?
- (c) Find a 95 percent confidence interval on the difference in means. Provide a practical interpretation of this interval.
- (d) Draw dot diagrams to assist in interpreting the results from this experiment.
- (e) Check the assumption of normality of the photoresist thickness.
- (f) Find the power of this test for detecting an actual difference in means of 2.5 kA.
- (g) What sample size would be necessary to detect an actual difference in means of 1.5 kA with a power of at least 0.9?

2.30. Front housings for cell phones are manufactured in an injection molding process. The time the part is allowed to cool in the mold before removal is thought to influence the occurrence of a particularly troublesome cosmetic defect, flow lines, in the finished housing. After manufacturing, the housings are inspected visually and assigned a score between 1 and 10 based on their appearance, with 10 corresponding to a perfect part and 1 corresponding to a completely defective part. An experiment was conducted using two cool-down times, 10 and 20 seconds, and 20 housings were evaluated at each level of cool-down time. All 40 observations in this experiment were run in random order. The data are as follows.

| 10 seconds | | 20 seconds | |
|------------|---|------------|---|
| 1 | 3 | 7 | 6 |
| 2 | 6 | 8 | 9 |
| 1 | 5 | 5 | 5 |
| 3 | 3 | 9 | 7 |
| 5 | 2 | 5 | 4 |
| 1 | 1 | 8 | 6 |
| 5 | 6 | 6 | 8 |
| 2 | 8 | 4 | 5 |
| 3 | 2 | 6 | 8 |
| 5 | 3 | 7 | 7 |

- (a) Is there evidence to support the claim that the longer cool-down time results in fewer appearance defects? Use $\alpha = 0.05$.
- (b) What is the P -value for the test conducted in part (a)?
- (c) Find a 95 percent confidence interval on the difference in means. Provide a practical interpretation of this interval.
- (d) Draw dot diagrams to assist in interpreting the results from this experiment.
- (e) Check the assumption of normality for the data from this experiment.

2.31. Twenty observations on etch uniformity on silicon wafers are taken during a qualification experiment for a plasma etcher. The data are as follows:

| | | | | |
|------|------|------|------|------|
| 5.34 | 6.65 | 4.76 | 5.98 | 7.25 |
| 6.00 | 7.55 | 5.54 | 5.62 | 6.21 |
| 5.97 | 7.35 | 5.44 | 4.39 | 4.98 |
| 5.25 | 6.35 | 4.61 | 6.00 | 5.32 |

- (a) Construct a 95 percent confidence interval estimate of σ^2 .
- (b) Test the hypothesis that $\sigma^2 = 1.0$. Use $\alpha = 0.05$. What are your conclusions?
- (c) Discuss the normality assumption and its role in this problem.
- (d) Check normality by constructing a normal probability plot. What are your conclusions?

2.32. The diameter of a ball bearing was measured by 12 inspectors, each using two different kinds of calipers. The results were

| Inspector | Caliper 1 | Caliper 2 |
|-----------|-----------|-----------|
| 1 | 0.265 | 0.264 |
| 2 | 0.265 | 0.265 |
| 3 | 0.266 | 0.264 |
| 4 | 0.267 | 0.266 |
| 5 | 0.267 | 0.267 |
| 6 | 0.265 | 0.268 |
| 7 | 0.267 | 0.264 |
| 8 | 0.267 | 0.265 |
| 9 | 0.265 | 0.265 |
| 10 | 0.268 | 0.267 |
| 11 | 0.268 | 0.268 |
| 12 | 0.265 | 0.269 |

- (a) Is there a significant difference between the means of the population of measurements from which the two samples were selected? Use $\alpha = 0.05$.
- (b) Find the P -value for the test in part (a).
- (c) Construct a 95 percent confidence interval on the difference in mean diameter measurements for the two types of calipers.

2.33. An article in the journal *Neurology* (1998, Vol. 50, pp. 1246–1252) observed that monozygotic twins share numerous physical, psychological, and pathological traits. The investigators measured an intelligence score of 10 pairs of twins. The data obtained are as follows:

| Pair | Birth Order: 1 | Birth Order: 2 |
|------|----------------|----------------|
| 1 | 6.08 | 5.73 |
| 2 | 6.22 | 5.80 |

| | | |
|----|------|------|
| 3 | 7.99 | 8.42 |
| 4 | 7.44 | 6.84 |
| 5 | 6.48 | 6.43 |
| 6 | 7.99 | 8.76 |
| 7 | 6.32 | 6.32 |
| 8 | 7.60 | 7.62 |
| 9 | 6.03 | 6.59 |
| 10 | 7.52 | 7.67 |

- (a) Is the assumption that the difference in score is normally distributed reasonable?
- (b) Find a 95% confidence interval on the difference in mean score. Is there any evidence that mean score depends on birth order?
- (c) Test an appropriate set of hypotheses indicating that the mean score does not depend on birth order.

2.34. An article in the *Journal of Strain Analysis* (vol. 18, no. 2, 1983) compares several procedures for predicting the shear strength for steel plate girders. Data for nine girders in the form of the ratio of predicted to observed load for two of these procedures, the Karlsruhe and Lehigh methods, are as follows:

| Girder | Karlsruhe Method | Lehigh Method |
|--------|------------------|---------------|
| S1/1 | 1.186 | 1.061 |
| S2/1 | 1.151 | 0.992 |
| S3/1 | 1.322 | 1.063 |
| S4/1 | 1.339 | 1.062 |
| S5/1 | 1.200 | 1.065 |
| S2/1 | 1.402 | 1.178 |
| S2/2 | 1.365 | 1.037 |
| S2/3 | 1.537 | 1.086 |
| S2/4 | 1.559 | 1.052 |


- (a) Is there any evidence to support a claim that there is a difference in mean performance between the two methods? Use $\alpha = 0.05$.
- (b) What is the P -value for the test in part (a)?
- (c) Construct a 95 percent confidence interval for the difference in mean predicted to observed load.
- (d) Investigate the normality assumption for both samples.
- (e) Investigate the normality assumption for the difference in ratios for the two methods.
- (f) Discuss the role of the normality assumption in the paired t -test.

2.35. The deflection temperature under load for two different formulations of ABS plastic pipe is being studied. Two samples of 12 observations each are prepared using each formulation and the deflection temperatures (in °F) are reported below:

| Formulation 1 | | | Formulation 2 | | |
|---------------|-----|-----|---------------|-----|-----|
| 206 | 193 | 192 | 177 | 176 | 198 |
| 188 | 207 | 210 | 197 | 185 | 188 |
| 205 | 185 | 194 | 206 | 200 | 189 |
| 187 | 189 | 178 | 201 | 197 | 203 |

- (a) Construct normal probability plots for both samples. Do these plots support assumptions of normality and equal variance for both samples?
- (b) Do the data support the claim that the mean deflection temperature under load for formulation 1 exceeds that of formulation 2? Use $\alpha = 0.05$.
- (c) What is the P -value for the test in part (a)?

2.36. Refer to the data in Problem 2.35. Do the data support a claim that the mean deflection temperature under load for formulation 1 exceeds that of formulation 2 by at least 3°F ?

 **2.37.** In semiconductor manufacturing wet chemical etching is often used to remove silicon from the backs of wafers prior to metalization. The etch rate is an important characteristic of this process. Two different etching solutions are being evaluated. Eight randomly selected wafers have been etched in each solution, and the observed etch rates (in mils/min) are as follows.

| Solution 1 | | Solution 2 | |
|------------|------|------------|------|
| 9.9 | 10.6 | 10.2 | 10.6 |
| 9.4 | 10.3 | 10.0 | 10.2 |
| 10.0 | 9.3 | 10.7 | 10.4 |
| 10.3 | 9.8 | 10.5 | 10.3 |

- (a) Do the data indicate that the claim that both solutions have the same mean etch rate is valid? Use $\alpha = 0.05$ and assume equal variances.
- (b) Find a 95 percent confidence interval on the difference in mean etch rates.
- (c) Use normal probability plots to investigate the adequacy of the assumptions of normality and equal variances.
- 2.38.** Two popular pain medications are being compared on the basis of the speed of absorption by the body. Specifically, tablet 1 is claimed to be absorbed twice as fast as tablet 2. Assume that σ_1^2 and σ_2^2 are known. Develop a test statistic for

$$H_0: 2\mu_1 = \mu_2$$

$$H_1: 2\mu_1 \neq \mu_2$$

2.39. Continuation of Problem 2.38. An article in *Nature* (1972, pp. 225–226) reported on the levels of monoamine oxidase in blood platelets for a sample of 43 schizophrenic

patients resulting in $\bar{y}_1 = 2.69$ and $s_1 = 2.30$ while for a sample of 45 normal patients the results were $\bar{y}_2 = 6.35$ and $s_2 = 4.03$. The units are nm/mg protein/h. Use the results of the previous problem to test the claim that the mean monoamine oxidase level for normal patients is at least twice the mean level for schizophrenic patients. Assume that the sample sizes are large enough to use the sample standard deviations as the true parameter values.

2.40. Suppose we are testing

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

where $\sigma_1^2 > \sigma_2^2$ are known. Our sampling resources are constrained such that $n_1 + n_2 = N$. Show that an allocation of the observation n_1 n_2 to the two samp that lead the most powerful test is in the ratio $n_1/n_2 = \sigma_1/\sigma_2$.


2.41. Continuation of Problem 2.40. Suppose that we want to construct a 95% two-sided confidence interval on the difference in two means where the two sample standard deviations are known to be $\sigma_1 = 4$ and $\sigma_2 = 8$. The total sample size is restricted to $N = 30$. What is the length of the 95% CI if the sample sizes used by the experimenter are $n_1 = n_2 = 15$? How much shorter would the 95% CI have been if the experimenter had used an optimal sample size allocation?

2.42. Develop Equation 2.46 for a $100(1 - \alpha)$ percent confidence interval for the variance of a normal distribution.

2.43. Develop Equation 2.50 for a $100(1 - \alpha)$ percent confidence interval for the ratio σ_1^2/σ_2^2 , where σ_1^2 and σ_2^2 are the variances of two normal distributions.

2.44. Develop an equation for finding a $100(1 - \alpha)$ percent confidence interval on the difference in the means of two normal distributions where $\sigma_1^2 \geq \sigma_2^2$. Apply your equation to the Portland cement experiment data, and find a 95 percent confidence interval.

2.45. Construct a data set for which the paired t -test statistic is very large, but for which the usual two-sample or pooled t -test statistic is small. In general, describe how you created the data. Does this give you any insight regarding how the paired t -test works?

2.46. Consider the experiment described in Problem 2.26. If the mean burning times of the two flares differ by as much as 2 minutes, find the power of the test. What sample size would be required to detect an actual difference in mean burning time of 1 minute with a power of at least 0.90? 

2.47. Reconsider the bottle filling experiment described in Problem 2.24. Rework this problem assuming that the two population variances are unknown but equal.

2.48. Consider the data from Problem 2.24. If the mean fill volume of the two machines differ by as much as 0.25 ounces, what is the power of the test used in Problem 2.19? What sample size would result in a power of at least 0.9 if the actual difference in mean fill volume is 0.25 ounces?