CHAPTER 2

Simple Comparative Experiments

CHAPTER OUTLINE

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

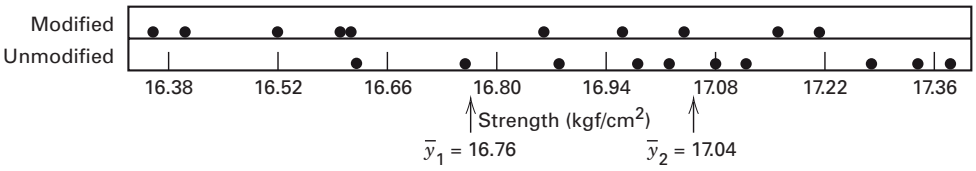
■ **TABLE 2.1**
Tension Bond Strength Data for the Portland Cement Formulation Experiment

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

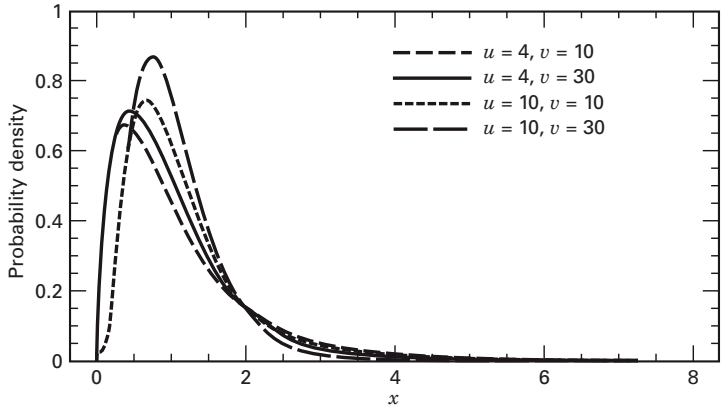
was completed, the experimenter did find a very large reduction in the cure time for the modified mortar formulation. Then he began to address the tension bond strength of the mortar. If the new mortar formulation has an adverse effect on bond strength, this could impact its usefulness.

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, $\bar{y}_1 = 16.76$ kgf/cm^2 for the modified mortar and $\bar{y}_2 = 17.04$ 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**



■ **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.

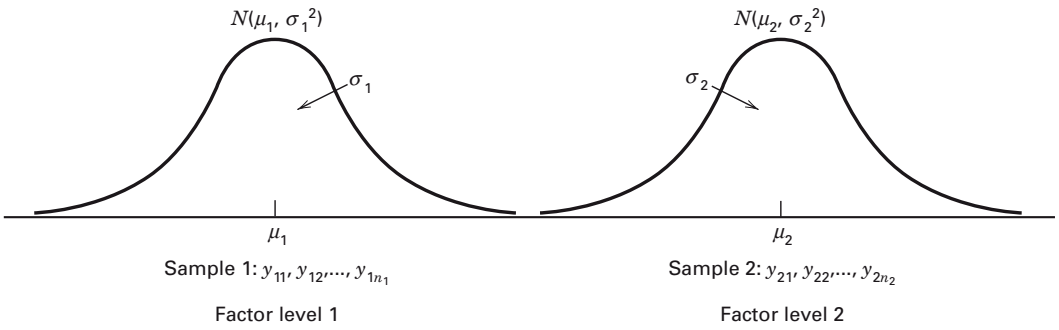
2.4.1 Hypothesis Testing

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 $y_{11}, y_{12}, \ldots, y_{1n_1}$ represent the n_1 observations from the first factor level and $y_{21}, y_{22}, \ldots, y_{2n_2}$ 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

$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2 \\ j = 1, 2, \ldots, n_i \end{cases} \tag{2.23}$$

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



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

that ϵ_{ij} are NID(0, σ_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 NID(μ_i, σ_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

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

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

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

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

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

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

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

and S_1^2 and S_2^2 are the two individual sample variances. The quality $S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 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

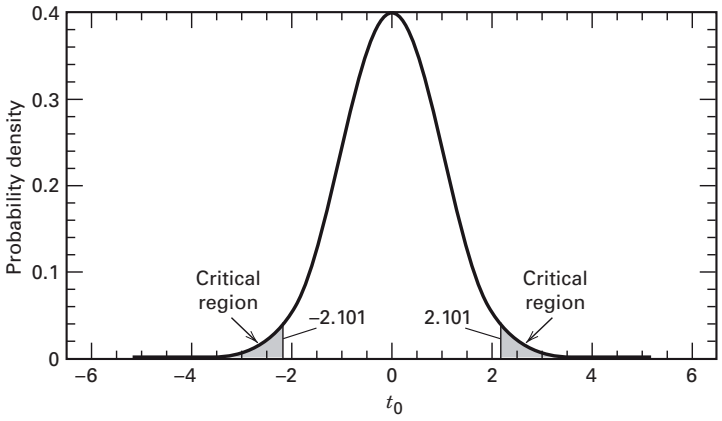
$$Z_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.26}$$

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
$\bar{y}_1 = 16.76$ kgf/cm^2	$\bar{y}_2 = 17.04$ 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

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

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.

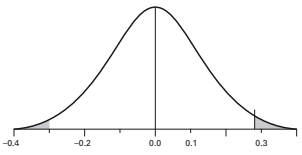
It is not always easy to compute the exact *P*-value for a test. However, most modern computer programs for statistical analysis report *P*-values, and they can be obtained on some handheld calculators. We will show how to approximate the *P*-value for the Portland cement mortar experiment. Because $|t_0| = 2.20 > t_{0.025,18} = 2.101$, 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
```

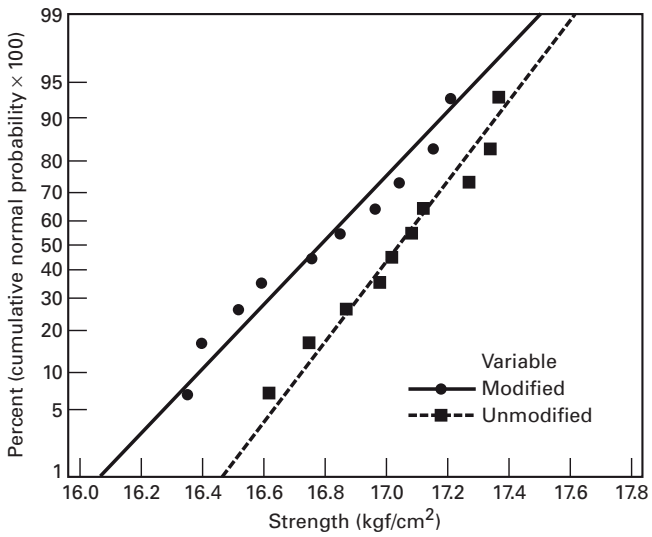


of the test statistic), and the *P*-value. Notice that the computed value of the *t* statistic differs slightly from our manually calculated value and that the *P*-value is reported to be $P = 0.042$. JMP also reports the *P*-values for the one-sided alternative hypothesis. Many software packages will 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 $y_1, y_2, \ldots, y_n$ is arranged as $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ 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

$$P(L \leqslant \theta \leqslant U) = 1 - \alpha \tag{2.27}$$

is true. The interval

$$L \leqslant \theta \leqslant U \tag{2.28}$$

is called a **100(1 − α) 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 − α) 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 − α 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 − α) percent of the time.

Suppose that we wish to find a 100(1 − α) 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,

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

or

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

Comparing Equations 2.29 and 2.27, we see that

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

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:

$$16.76 - 17.04 - (2.101)0.284\sqrt{\tfrac{1}{10}+\tfrac{1}{10}} \le \mu_1 - \mu_2 \le 16.76 - 17.04 + (2.101)0.284\sqrt{\tfrac{1}{10}+\tfrac{1}{10}}$$

$$-0.28 - 0.27 \le \mu_1 - \mu_2 \le -0.28 + 0.27$$

$$-0.55 \le \mu_1 - \mu_2 \le -0.01$$

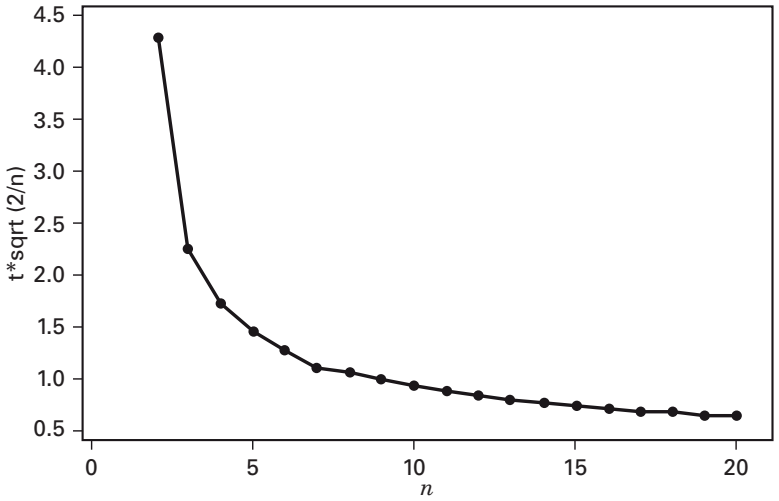
Thus, the 95 percent confidence interval estimate on the difference in means extends from -0.55 to -0.01 kgf/cm^2. Put another way, the confidence interval is $\mu_1 - \mu_2 = -0.28 \pm 0.27$ kgf/cm^2, or the difference in mean strengths is -0.28 kgf/cm^2, and the accuracy of this estimate is ± 0.27 kgf/cm^2. 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

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

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

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

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.

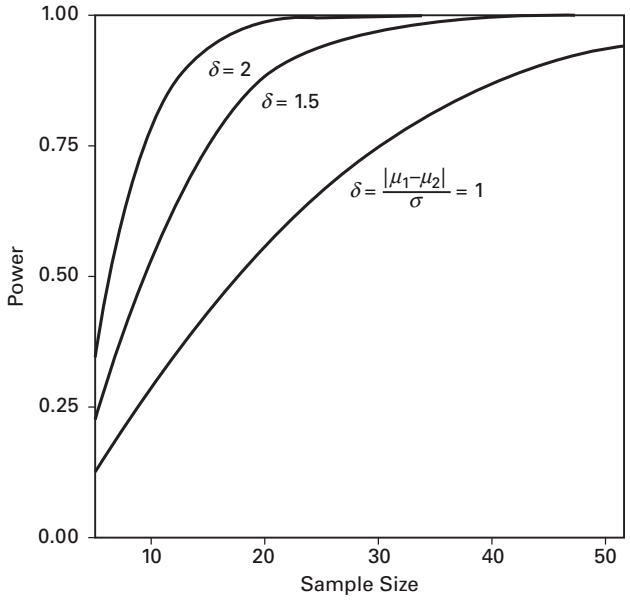
Operating curves and power curves are often helpful in selecting a sample size to use in an experiment. For example, consider the Portland cement mortar problem discussed previously. Suppose that a difference in mean strength of 0.5 kgf/cm^2 has practical impact on the use of the mortar, so if the difference in means is at least this large, we would like to detect it with a high probability. Thus, because $\mu_1 - \mu_2 = 0.5$ 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 $\sigma = 0.25$ 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 Varianes 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 = 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

                    Sample        Target        Actual
Difference            Size         Power         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

                    Sample
Difference            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

                    Sample        Target        Actual
Difference            Size         Power         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

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

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

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

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 then for muscle tissue. That is, if μ_δ is the mean normalized fluorescence for nerve tissue and 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

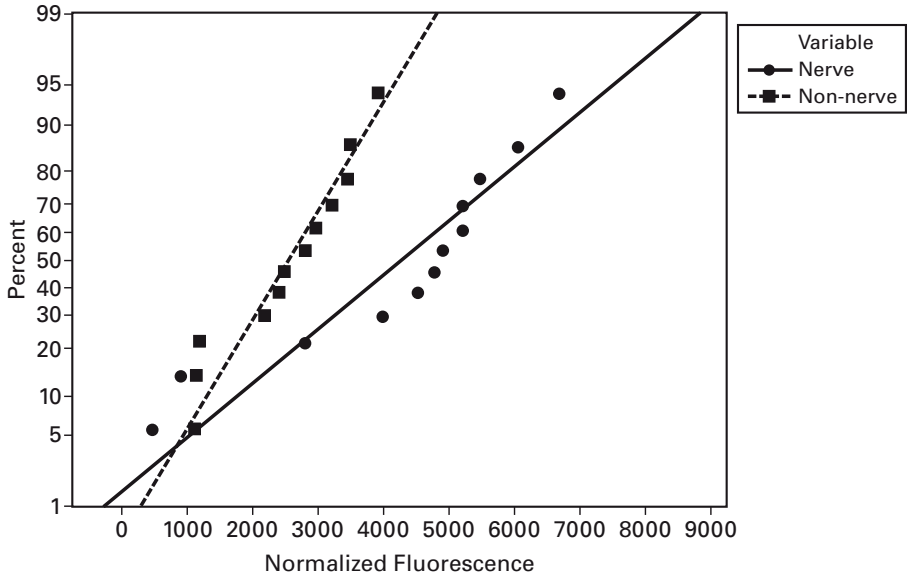
TABLE 2.3
Normalized Fluorescence After Two Hours

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

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

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

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

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

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

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

$$\bar{y} - Z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{y} + Z_{\alpha/2}\sigma/\sqrt{n} \tag{2.36}$$

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

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

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

$$\bar{y} - t_{\alpha/2,n-1}S/\sqrt{n} \le \mu \le \bar{y} + t_{\alpha/2,n-1}S/\sqrt{n} \tag{2.38}$$

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
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$Z_0 = \dfrac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$	$\lvert Z_0 \rvert > Z_{\alpha/2}$	$P = 2[1 - \Phi(\lvert Z_0 \rvert)]$
$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$		$Z_0 < -Z_\alpha$	$P = \Phi(Z_0)$
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$		$Z_0 > Z_\alpha$	$P = 1 - \Phi(Z_0)$
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$Z_0 = \dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$\lvert Z_0 \rvert > Z_{\alpha/2}$	$P = 2[1 - \Phi(\lvert Z_0 \rvert)]$
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$		$Z_0 < -Z_\alpha$	$P = \Phi(Z_0)$
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$		$Z_0 > Z_\alpha$	$P = 1 - \Phi(Z_0)$

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

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

$$y_{ij} = \mu_i + \beta_j + \epsilon_{ij} \begin{cases} i = 1, 2 \\ j = 1, 2, \ldots, 10 \end{cases} \tag{2.39}$$

■ **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 ith tip, β_j is an effect on hardness due to the jth 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 jth paired difference

$$d_j = y_{1j} - y_{2j} \qquad j = 1, 2, \ldots, 10 \tag{2.40}$$

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

$$t_0 = \frac{\bar{d}}{S_d/\sqrt{n}} \tag{2.41}$$

where

$$\bar{d} = \frac{1}{n}\sum_{j=1}^{n} d_j \tag{2.42}$$

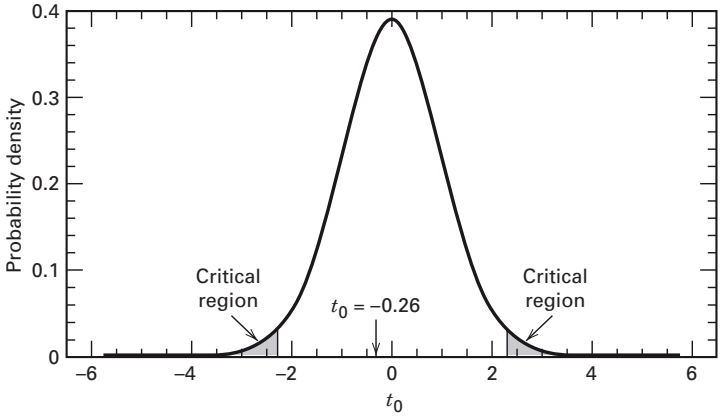
is the sample mean of the differences and

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

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 $P \simeq 0.80$, 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

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

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

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

The test statistic for Equation 2.44 is

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

where $SS = \Sigma_{i=1}^{n}(y_i - \bar{y})^2$ 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

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

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

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

is the ratio of the sample variances

$$F_0 = \frac{S_1^2}{S_2^2} \tag{2.48}$$

■ **TABLE 2.8**
Tests on Variances of Normal Distributions

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

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

$$F_{1-\alpha,v_1,v_2} = \frac{1}{F_{\alpha,v_2,v_1}} \tag{2.49}$$

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

$$\frac{S_1^2}{S_2^2} F_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} F_{\alpha/2,n_2-1,n_1-1} \tag{2.50}$$

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 $F_{0.975,9,11} = 1/F_{0.025,11,9} = 1/3.92 = 0.255$,

$$\frac{14.5}{10.8}(0.255) \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{14.5}{10.8}(3.59)$$

$$0.34 \le \frac{\sigma_1^2}{\sigma_2^2} \le 4.82$$

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.