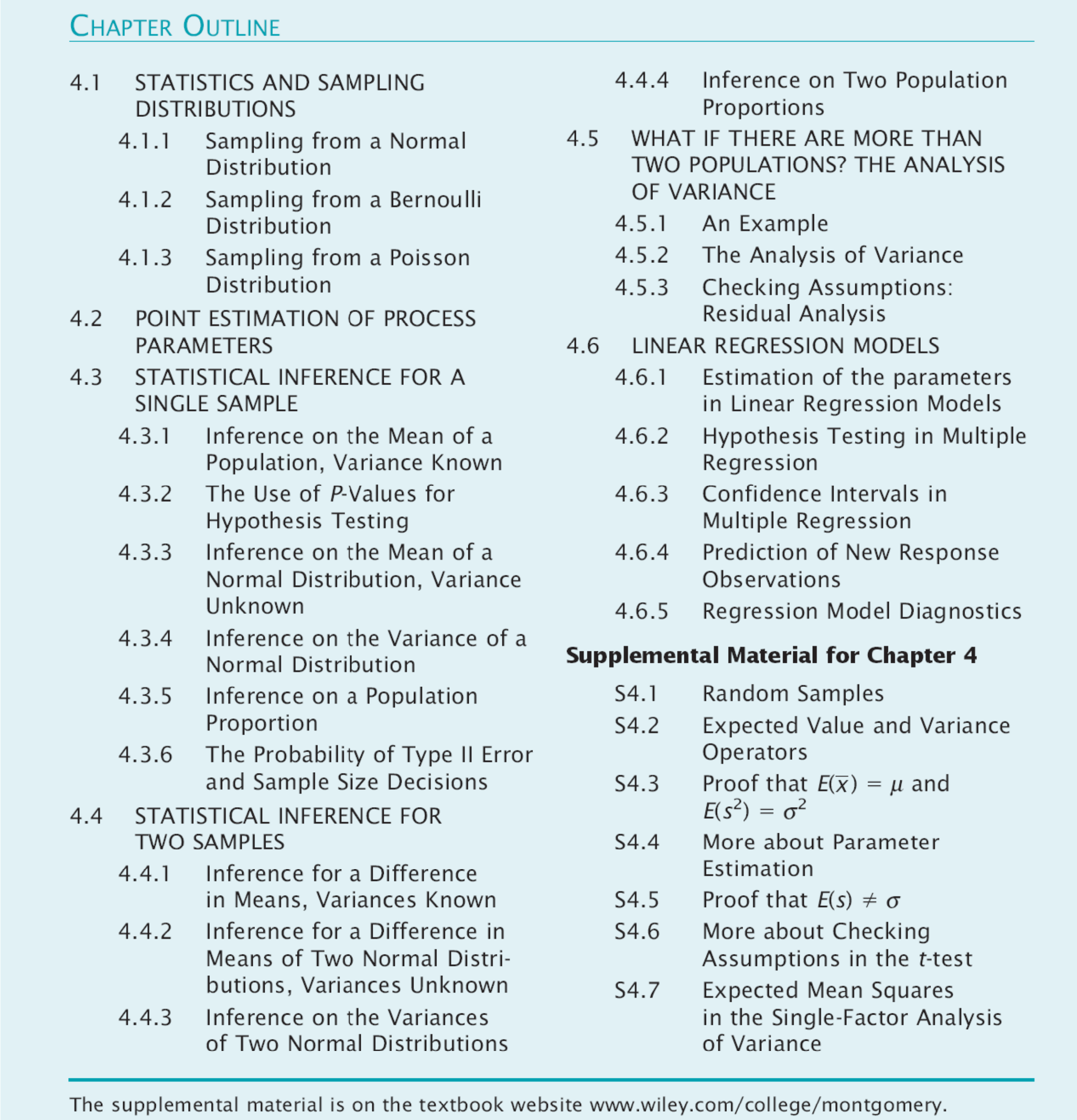
4 Inferences about Process Quality

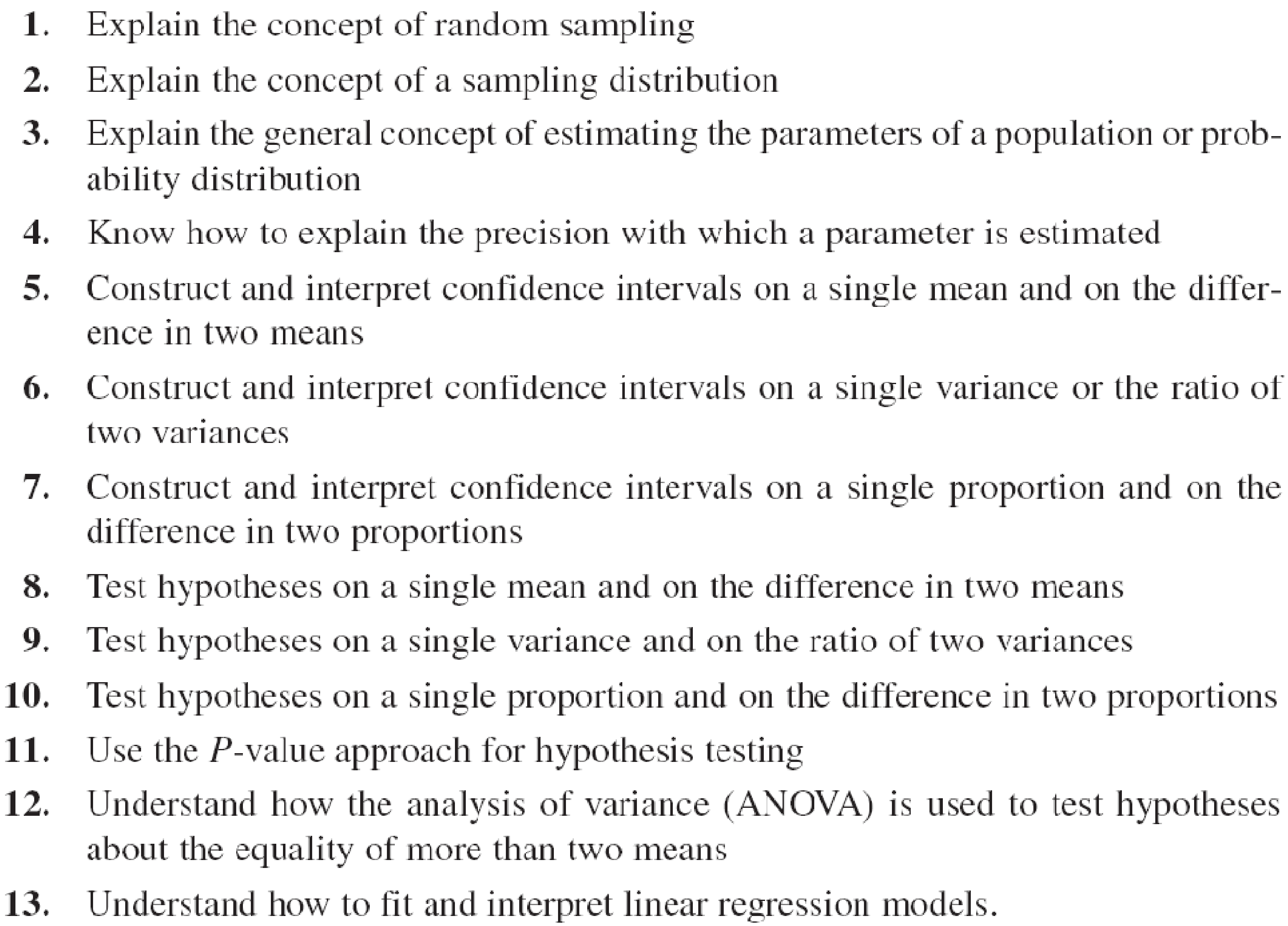


Chapter Outline

Supplemental Material for Chapter 4

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

Learning Objectives



1. Explain the concept of random sampling
2. Explain the concept of a sampling distribution
3. Explain the general concept of estimating the parameters of a population or probability distribution
4. Know how to explain the precision with which a parameter is estimated
5. Construct and interpret confidence intervals on a single mean and on the difference in two means
6. Construct and interpret confidence intervals on a single variance or the ratio of two variances
7. Construct and interpret confidence intervals on a single proportion and on the difference in two proportions
8. Test hypotheses on a single mean and on the difference in two means
9. Test hypotheses on a single variance and on the ratio of two variances
10. Test hypotheses on a single proportion and on the difference in two proportions
11. Use the *P*-value approach for hypothesis testing
12. Understand how the analysis of variance (ANOVA) is used to test hypotheses about the equality of more than two means
13. Understand how to fit and interpret linear regression models.

4.1 Statistics and Sampling Distributions

- Statistical inference is concerned with drawing conclusions about populations (or processes) based on sample data from that system
- Random sample – a sample that is selected so that the observations are independent – a random sample has the property that it has the same probability of selection as any other sample.
- Statistic – any function of the observations in a sample that doesn't contain unknown parameters
- The sample mean, the sample variance, and the sample standard deviation are all statistics

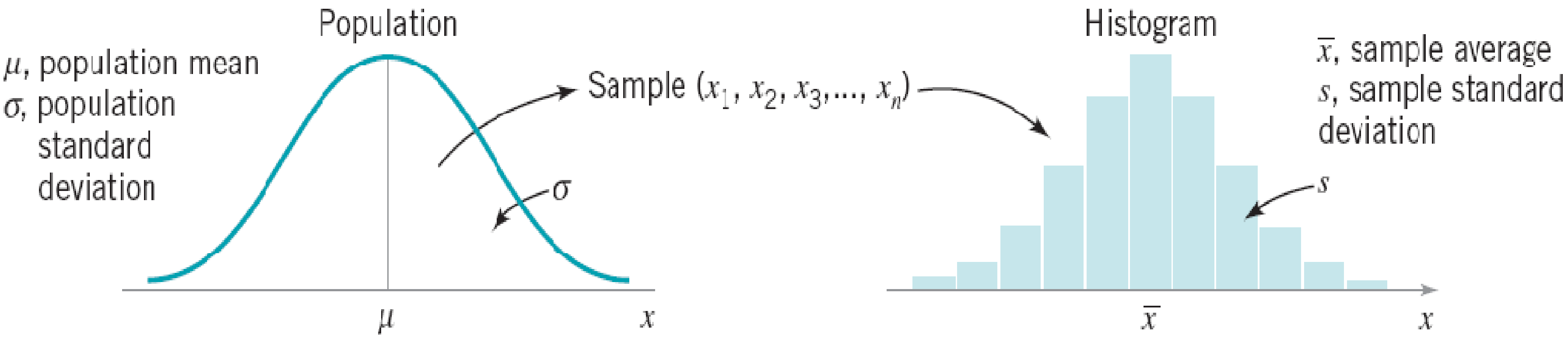

Population
μ, population mean
σ, population standard deviation
Sample ($x_1, x_2, x_3, \ldots, x_n$)
σ
μ
x
Histogram
$\bar{x}$, sample average
s, sample standard deviation
s
$\bar{x}$
x

FIGURE 4.1 Relationship between a population and a sample.

Observations in a sample are used to draw conclusions about the population

Sampling Distributions

- A statistic is a random variable, because a different sample with produce a different observed value of the statistic
- Every statistic has a probability distribution
- The probability distribution of a statistic is called a **sampling distribution**

Sampling from a Normal Distribution

Suppose that x is a normally distributed random variable with mean μ and variance σ^2. If $x_1, x_2, \ldots, x_n$ is a random sample of size n from this process, then the distribution of the sample mean $\bar{x}$ is $N(\mu, \sigma^2/n)$. This follows directly from the results on the distribution of linear combinations of normal random variables in Section 3.3.1.

This property of the sample mean is not restricted exclusively to the case of sampling from normal populations. Note that we may write

$$\left(\frac{\bar{x}-\mu}{\sigma}\right)\sqrt{n} = \frac{\sum_{i=1}^{n} x_i - n\mu}{\sigma\sqrt{n}}$$

From the central limit theorem we know that, regardless of the distribution of the population, the distribution of $\Sigma_{i=1}^{u} x_i$ is approximately normal with mean $n\mu$ and variance $n\sigma^2$. Therefore, regardless of the distribution of the population, the sampling distribution of the sample mean is approximately

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

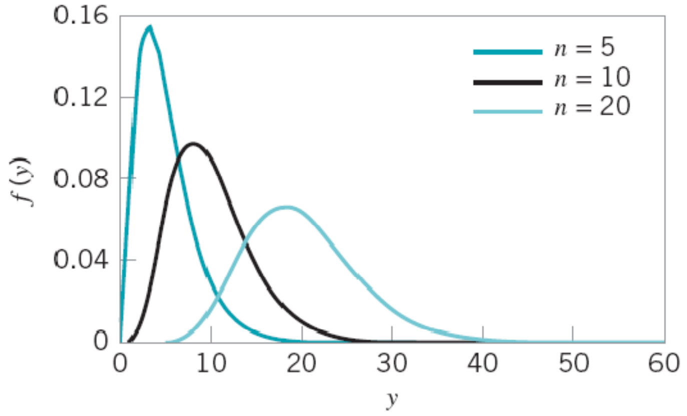
An important sampling distribution defined in terms of the normal distribution is the **chi-square** or χ^2 **distribution.** If $x_1, x_2, \ldots, x_n$ are normally and independently distributed random variables with mean zero and variance one, then the random variable

$$y = x_1^2 + x_2^2 + \cdots + x_n^2$$

is distributed as chi-square with n degrees of freedom. The chi-square probability distribution with n degrees of freedom is

$$f(y) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} y^{(n/2)-1} e^{-y/2} \qquad y > 0 \tag{4.4}$$

Several chi-square distributions are shown in Fig. 4.2. The distribution is skewed with mean $\mu = n$ and variance $\sigma^2 = 2n$. A table of the percentage points of the chi-square distribution is given in Appendix Table III.



■ FIGURE 4.2 Chi-square distribution for selected values of n (number of degrees of freedom).

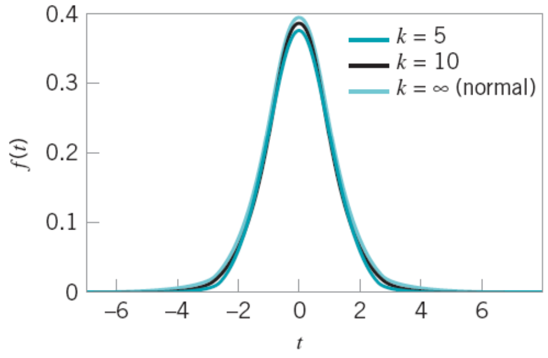
Another useful sampling distribution is the ***t* distribution.** If x is a standard normal random variable and if y is a chi-square random variable with k degrees of freedom, and if x and y are independent, then the random variable

$$t = \frac{x}{\sqrt{y/k}} \tag{4.6}$$

is distributed as t with k degrees of freedom. The probability distribution of t is

$$f(t) = \frac{\Gamma[(k+1)/2]}{\sqrt{k\pi}\Gamma(k/2)}\left(\frac{t^2}{k}+1\right)^{-(k+1)/2} \qquad -\infty < t < \infty \tag{4.7}$$

and the mean and variance of t are $\mu = 0$ and $\sigma^2 = k/(k-2)$ for $k > 2$, respectively. The degrees of freedom for t are the degrees of freedom associated with the chi-square random variable in the denominator of equation (4.6). Several t distributions are shown in Fig. 4.3. Note that if $k = \infty$, the t distribution reduces to the standard normal distribution. A table of percentage points of the t distribution is given in Appendix Table IV.



■ FIGURE 4.3 The *t* distribution for selected values of *k* (number of degrees of freedom).

The last sampling distribution based on the normal process that we will consider is the ***F* distribution.** If w and y are two independent chi-square random variables with u and v degrees of freedom, respectively, then the ratio

$$F_{u,v} = \frac{w/u}{y/v} \tag{4.9}$$

is distributed as F 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 distribution is

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)} \frac{x^{(u/2)-1}}{\left[\left(\frac{u}{v}\right)x+1\right]^{(u+v)/2}} \qquad 0 < x < \infty \tag{4.10}$$

Several F distributions are shown in Fig. 4.4. A table of percentage points of the F distribution is given in Appendix Table V.

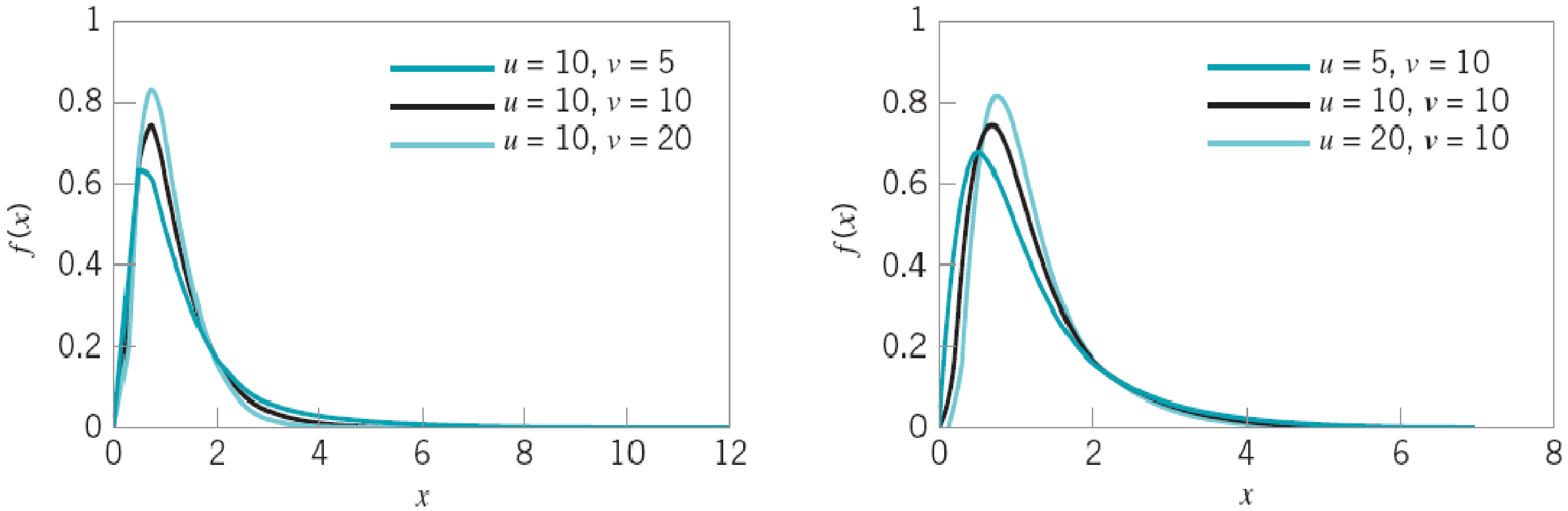
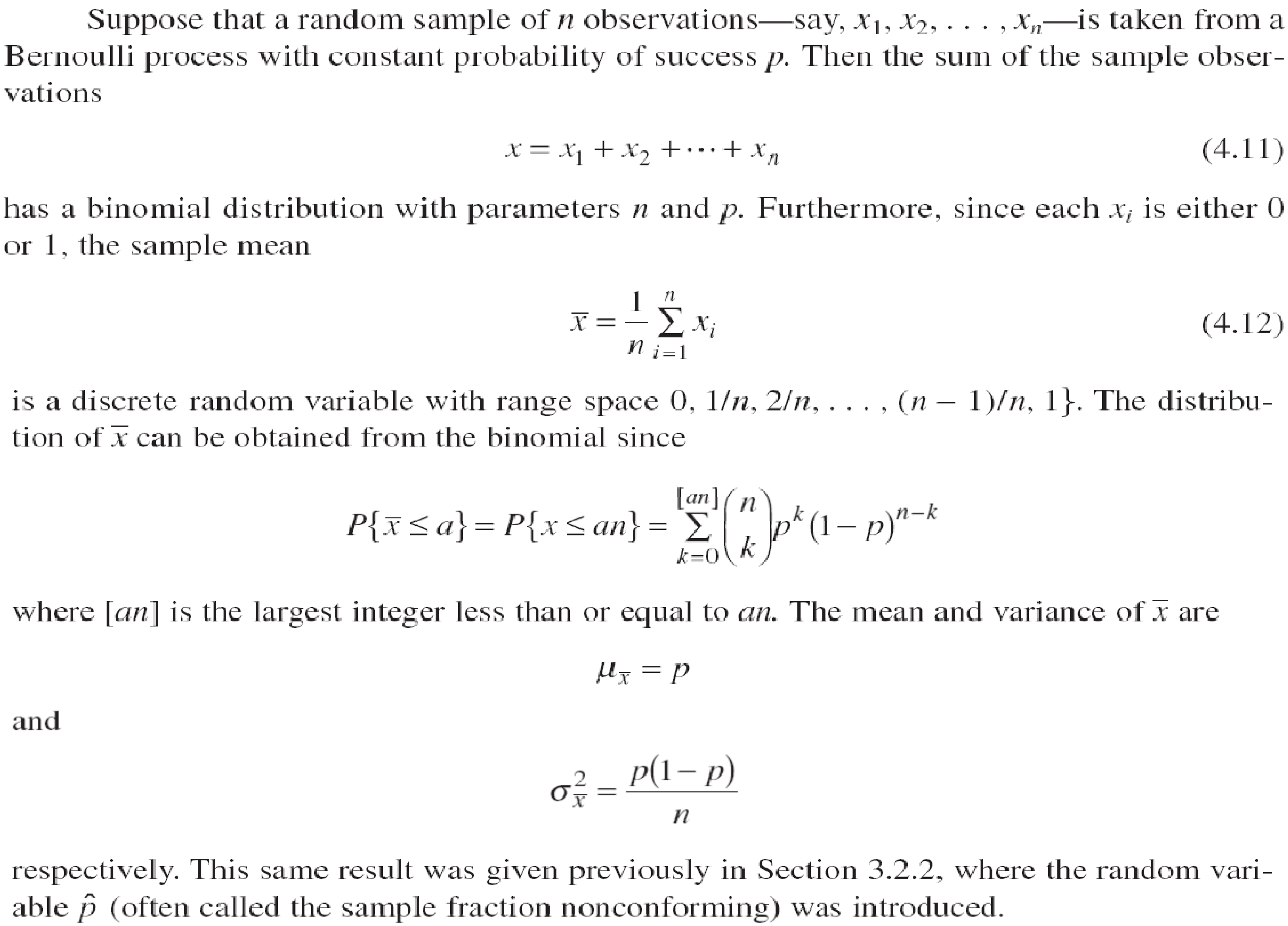


FIGURE 4.4 The *F* distribution for selected values of *u* (numerator degrees of freedom) and *v* (denominator degrees of freedom).

Sampling from a Bernoulli Distribution



Suppose that a random sample of n observations—say, $x_1, x_2, \ldots, x_n$—is taken from a Bernoulli process with constant probability of success p. Then the sum of the sample observations

$$x = x_1 + x_2 + \cdots + x_n \qquad (4.11)$$

has a binomial distribution with parameters n and p. Furthermore, since each x_i is either 0 or 1, the sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (4.12)$$

is a discrete random variable with range space 0, $1/n$, $2/n$, . . . , $(n-1)/n$, 1}. The distribution of $\overline{x}$ can be obtained from the binomial since

$$P\{\overline{x} \le a\} = P\{x \le an\} = \sum_{k=0}^{[an]} \binom{n}{k} p^k (1-p)^{n-k}$$

where $[an]$ is the largest integer less than or equal to an. The mean and variance of $\overline{x}$ are

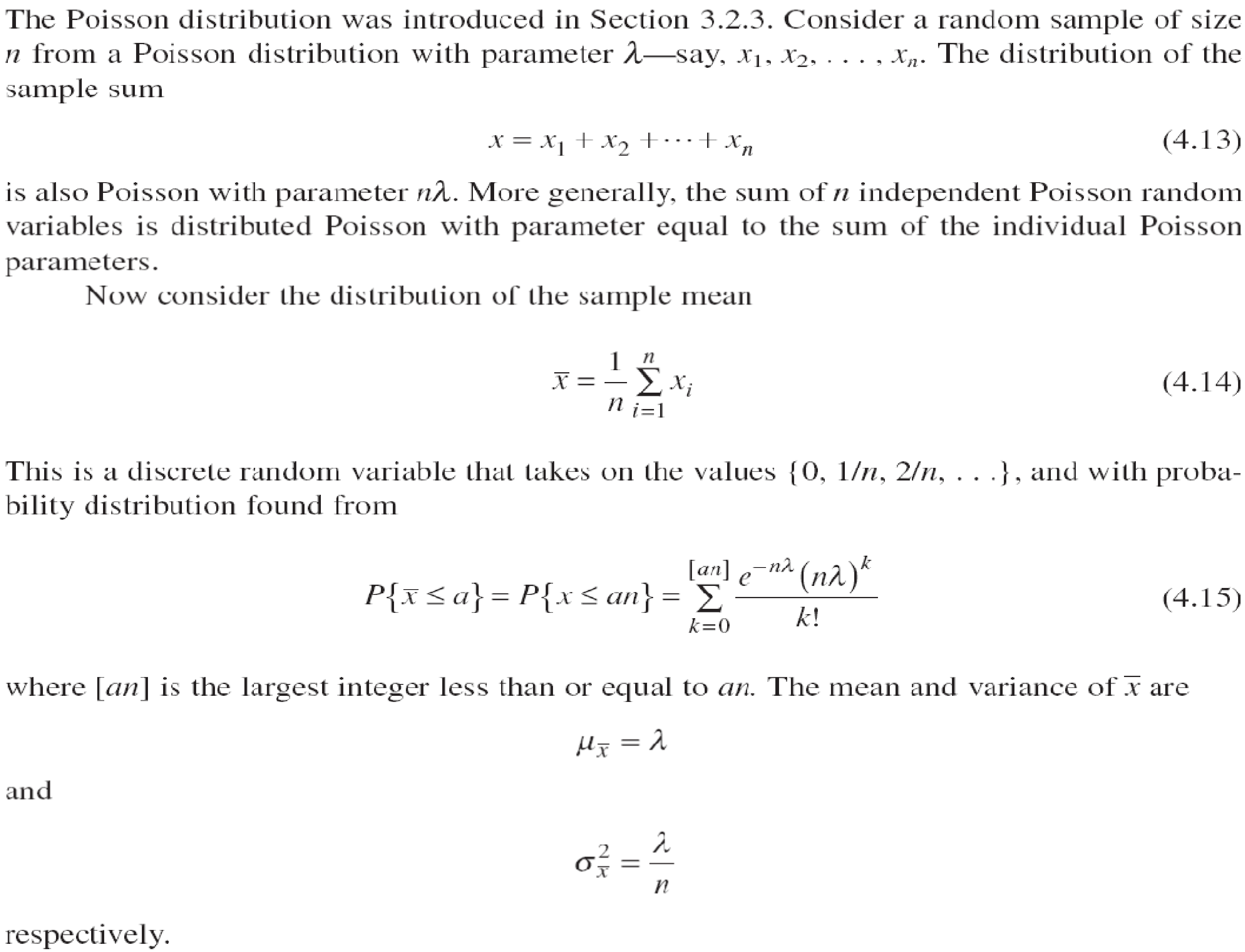
$$\mu_{\overline{x}} = p$$

and

$$\sigma_{\overline{x}}^2 = \frac{p(1-p)}{n}$$

respectively. This same result was given previously in Section 3.2.2, where the random variable $\hat{p}$ (often called the sample fraction nonconforming) was introduced.

Sampling from a Poisson Distribution



The Poisson distribution was introduced in Section 3.2.3. Consider a random sample of size n from a Poisson distribution with parameter λ—say, $x_1, x_2, \ldots, x_n$. The distribution of the sample sum

$$x = x_1 + x_2 + \cdots + x_n \tag{4.13}$$

is also Poisson with parameter $n\lambda$. More generally, the sum of n independent Poisson random variables is distributed Poisson with parameter equal to the sum of the individual Poisson parameters.

Now consider the distribution of the sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.14}$$

This is a discrete random variable that takes on the values $\{0, 1/n, 2/n, \ldots\}$, and with probability distribution found from

$$P\{\overline{x} \le a\} = P\{x \le an\} = \sum_{k=0}^{[an]} \frac{e^{-n\lambda}(n\lambda)^k}{k!} \tag{4.15}$$

where $[an]$ is the largest integer less than or equal to an. The mean and variance of $\overline{x}$ are

$$\mu_{\overline{x}} = \lambda$$

and

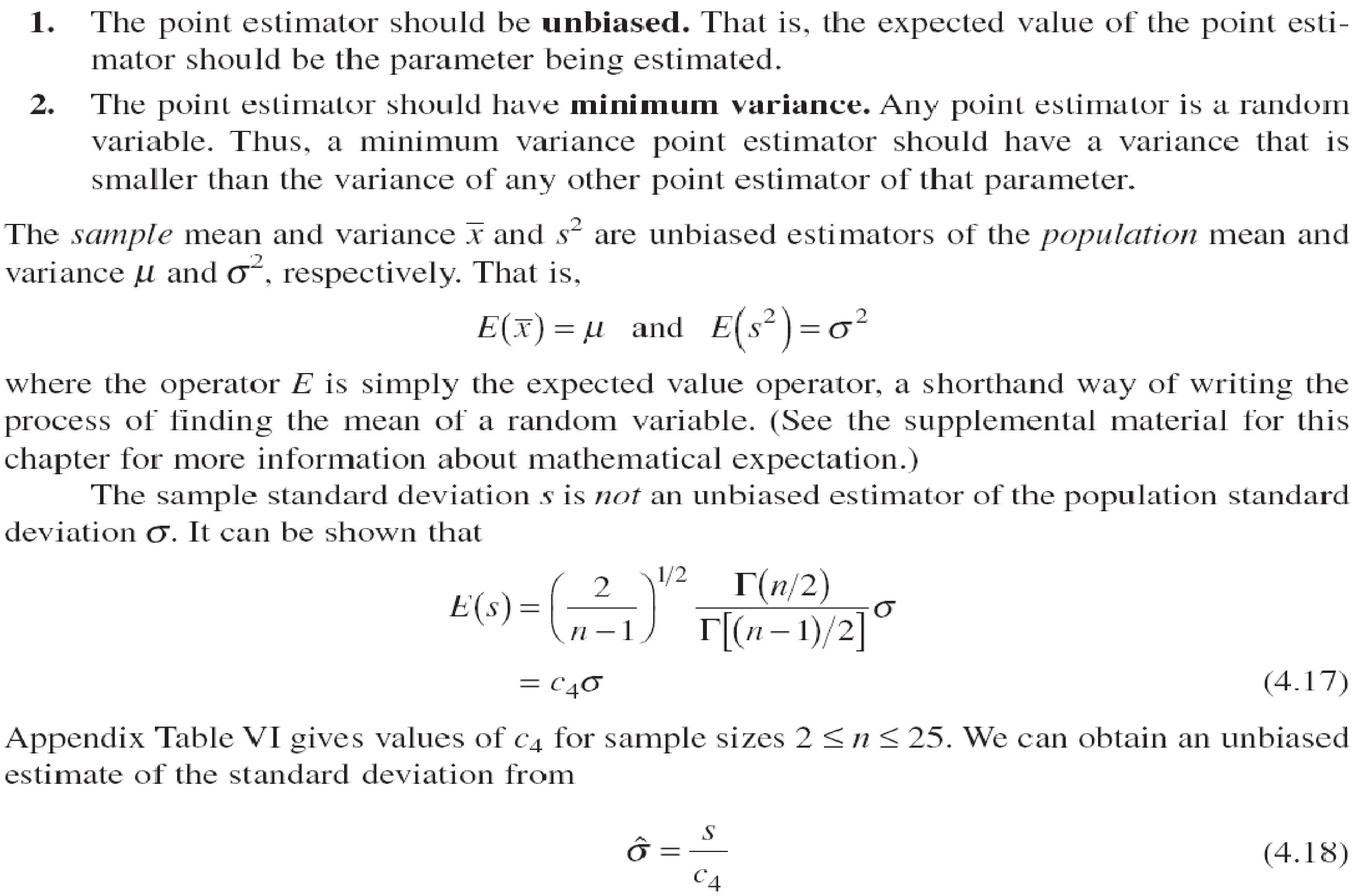
$$\sigma_{\overline{x}}^2 = \frac{\lambda}{n}$$

respectively.

4.2 Point Estimation of Process Parameters

- Distributions are described by their parameters
- Parameters are generally unknown and must be estimated
- Point estimator – a statistic that a single numerical value that is the **estimate** of the parameter
- Examples, page 110 & 111

Properties of Point Estimators



1. The point estimator should be **unbiased.** That is, the expected value of the point estimator should be the parameter being estimated.
2. The point estimator should have **minimum variance.** Any point estimator is a random variable. Thus, a minimum variance point estimator should have a variance that is smaller than the variance of any other point estimator of that parameter.

The *sample* mean and variance $\bar{x}$ and s^2 are unbiased estimators of the *population* mean and variance μ and σ^2, respectively. That is,

$$E(\bar{x}) = \mu \quad \text{and} \quad E\left(s^2\right) = \sigma^2$$

where the operator E is simply the expected value operator, a shorthand way of writing the process of finding the mean of a random variable. (See the supplemental material for this chapter for more information about mathematical expectation.)

The sample standard deviation s is *not* an unbiased estimator of the population standard deviation σ. It can be shown that

$$\begin{aligned} E(s) &= \left(\frac{2}{n-1}\right)^{1/2} \frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}\sigma \\ &= c_4\sigma \end{aligned} \qquad (4.17)$$

Appendix Table VI gives values of c_4 for sample sizes $2 \le n \le 25$. We can obtain an unbiased estimate of the standard deviation from

$$\hat{\sigma} = \frac{s}{c_4} \qquad (4.18)$$

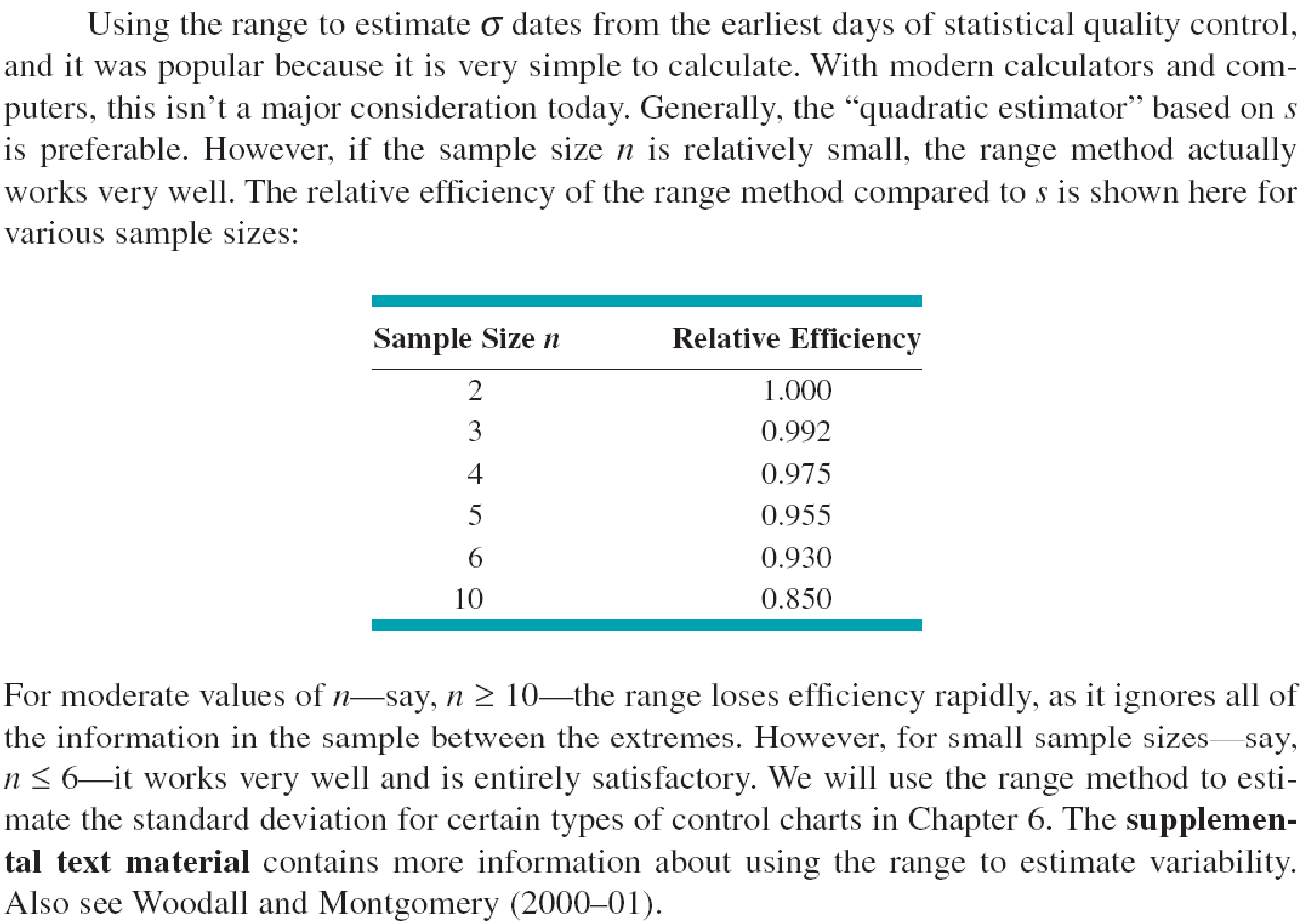
In many applications of statistics to quality-engineering problems, it is convenient to estimate the standard deviation by the **range method.** Let $x_1, x_2, \ldots, x_n$ be a random sample of n observations from a normal distribution with mean μ and variance σ^2. The **range** of the sample is

$$\begin{aligned} R &= \max(x_i) - \min(x_i) \\ &= x_{\max} - x_{\min} \end{aligned} \qquad (4.19)$$

That is, the range R is simply the difference between the largest and smallest sample observations. The random variable $W = R/\sigma$ is called the **relative range.** The distribution of W has been well studied. The mean of W is a constant d_2 that depends on the size of the sample. That is, $E(W) = d_2$. Therefore, an unbiased estimator of the standard deviation σ of a normal distribution is

$$\hat{\sigma} = \frac{R}{d_2} \qquad (4.20)$$

Values of d_2 for sample sizes $2 \leq n \leq 25$ are given in Appendix Table VI.



Using the range to estimate σ dates from the earliest days of statistical quality control, and it was popular because it is very simple to calculate. With modern calculators and computers, this isn't a major consideration today. Generally, the "quadratic estimator" based on s is preferable. However, if the sample size n is relatively small, the range method actually works very well. The relative efficiency of the range method compared to s is shown here for various sample sizes:

Sample Size n	**Relative Efficiency**
2	1.000
3	0.992
4	0.975
5	0.955
6	0.930
10	0.850

For moderate values of n—say, $n \geq 10$—the range loses efficiency rapidly, as it ignores all of the information in the sample between the extremes. However, for small sample sizes—say, $n \leq 6$—it works very well and is entirely satisfactory. We will use the range method to estimate the standard deviation for certain types of control charts in Chapter 6. The **supplemental text material** contains more information about using the range to estimate variability. Also see Woodall and Montgomery (2000–01).

4.3 Statistical Inference for a Single Sample

- Statistical inference = decision making
- Hypothesis testing
  - Null hypothesis, H_0
  - Alternative hypothesis, H_1
- Confidence intervals
- These two techniques are closely related

To test a hypothesis, we take a random sample from the population under study, compute an appropriate **test statistic,** and then either reject or fail to reject the null hypothesis H_0. The set of values of the test statistic leading to rejection of H_0 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, then a **type I error** has occurred. If the null hypothesis is not rejected when it is false, then a **type II error** has been made. The probabilities of these two types of errors are denoted as

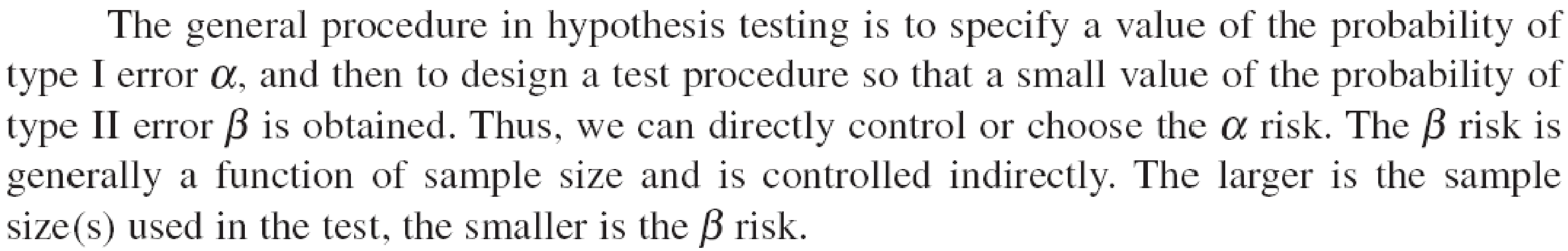
$$\alpha = \text{P}\{\text{type I error}\} = \text{P}\{\text{reject } H_0 \,|\, H_0 \text{ is true}\}$$

$$\beta = \text{P}\{\text{type II error}\} = \text{P}\{\text{fail to reject } H_0 \,|\, H_0 \text{ is false}\}$$

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

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

Thus, the power is the probability of *correctly* rejecting H_0. In quality control work, α is sometimes called the **producer's risk,** because it denotes the probability that a good lot will be rejected, or the probability that a process producing acceptable values of a particular quality characteristic will be rejected as performing unsatisfactorily. In addition, β is sometimes called the **consumer's risk,** because it denotes the probability of accepting a lot of poor quality, or allowing a process that is operating in an unsatisfactory manner relative to some quality characteristic to continue in operation.



The general procedure in hypothesis testing is to specify a value of the probability of type I error α, and then to design a test procedure so that a small value of the probability of type II error β is obtained. Thus, we can directly control or choose the α risk. The β risk is generally a function of sample size and is controlled indirectly. The larger is the sample size(s) used in the test, the smaller is the β risk.

4.3.1 Inference on the Mean of a Population, Variance Known

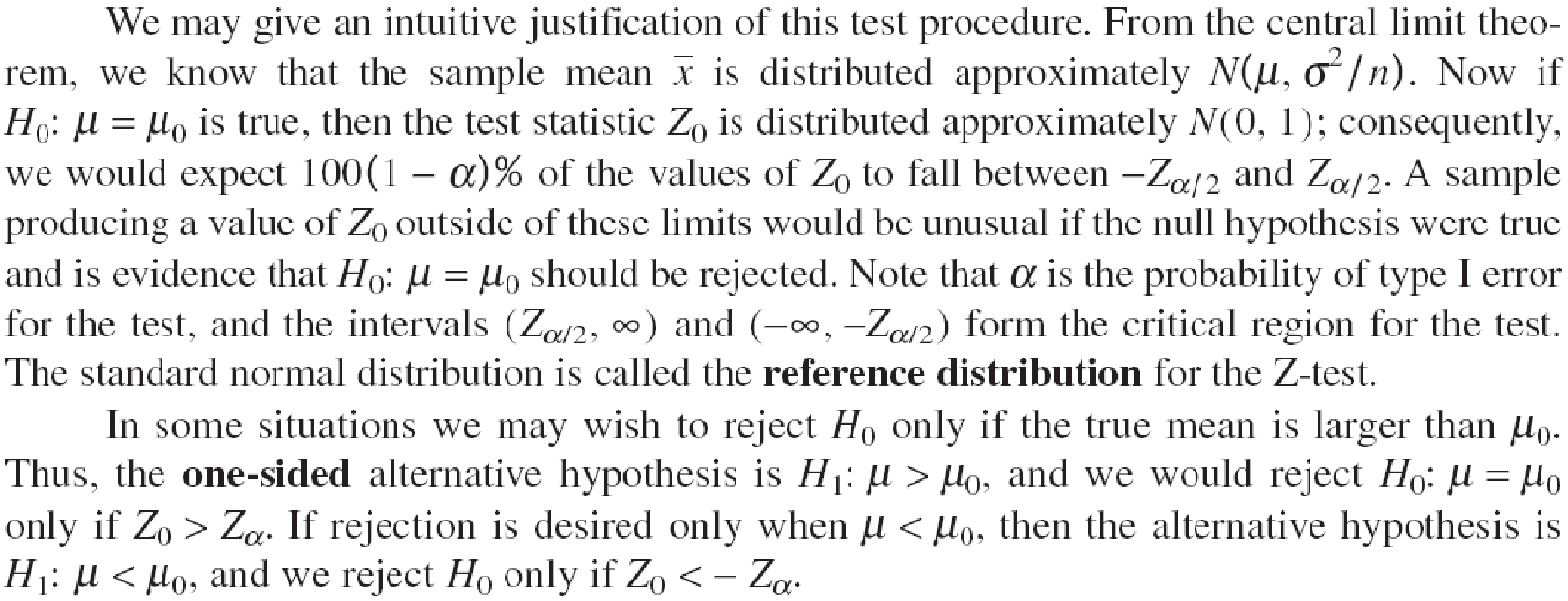
Hypothesis Testing. Suppose that x is a random variable with unknown mean μ and known variance σ^2. We wish to test the hypothesis that the mean is equal to a standard value—say, μ_0. The hypothesis may be formally stated as

$$\begin{aligned} H_0&: \quad \mu = \mu_0 \\ H_1&: \quad \mu \neq \mu_0 \end{aligned} \tag{4.22}$$

The procedure for testing this hypothesis is to take a random sample of n observations on the random variable x, compute the test statistic

$$\boxed{Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \qquad (4.23)}$$

and 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 one-sample *Z-test*.



We may give an intuitive justification of this test procedure. From the central limit theorem, we know that the sample mean $\bar{x}$ is distributed approximately $N(\mu, \sigma^2/n)$. Now if H_0: $\mu = \mu_0$ is true, then the test statistic Z_0 is distributed approximately $N(0, 1)$; consequently, we would expect $100(1 - \alpha)\%$ of the values of Z_0 to fall between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$. A sample producing a value of Z_0 outside of these limits would be unusual if the null hypothesis were true and is evidence that H_0: $\mu = \mu_0$ should be rejected. Note that α is the probability of type I error for the test, and the intervals $(Z_{\alpha/2}, \infty)$ and $(-\infty, -Z_{\alpha/2})$ form the critical region for the test. The standard normal distribution is called the **reference distribution** for the Z-test.

In some situations we may wish to reject H_0 only if the true mean is larger than μ_0. Thus, the **one-sided** alternative hypothesis is H_1: $\mu > \mu_0$, and we would reject H_0: $\mu = \mu_0$ only if $Z_0 > Z_\alpha$. If rejection is desired only when $\mu < \mu_0$, then the alternative hypothesis is H_1: $\mu < \mu_0$, and we reject H_0 only if $Z_0 < - Z_\alpha$.

EXAMPLE 4.1 Computer Response Time

The response time of a distributed computer system is an important quality characteristic. The system manager wants to know whether the mean response time to a specific type of command exceeds 75 millisec. From previous experience, he knows that the standard deviation of response time is 8 millisec. Use a type I error of $\alpha = 0.05$.

SOLUTION

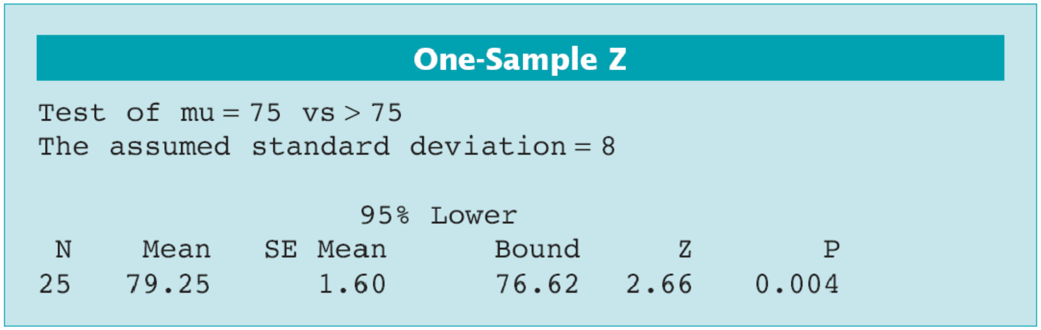
The appropriate hypotheses are

$$H_0: \quad \mu = 75$$
$$H_1: \quad \mu > 75$$

The command is executed 25 times and the response time for each trial is recorded. We assume that these observations can be considered as a random sample of the response times. The sample average response time is $\bar{x} = 79.25$ millisec. The value of the test statistic is

$$Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{79.25 - 75}{8/\sqrt{25}} = 2.66$$

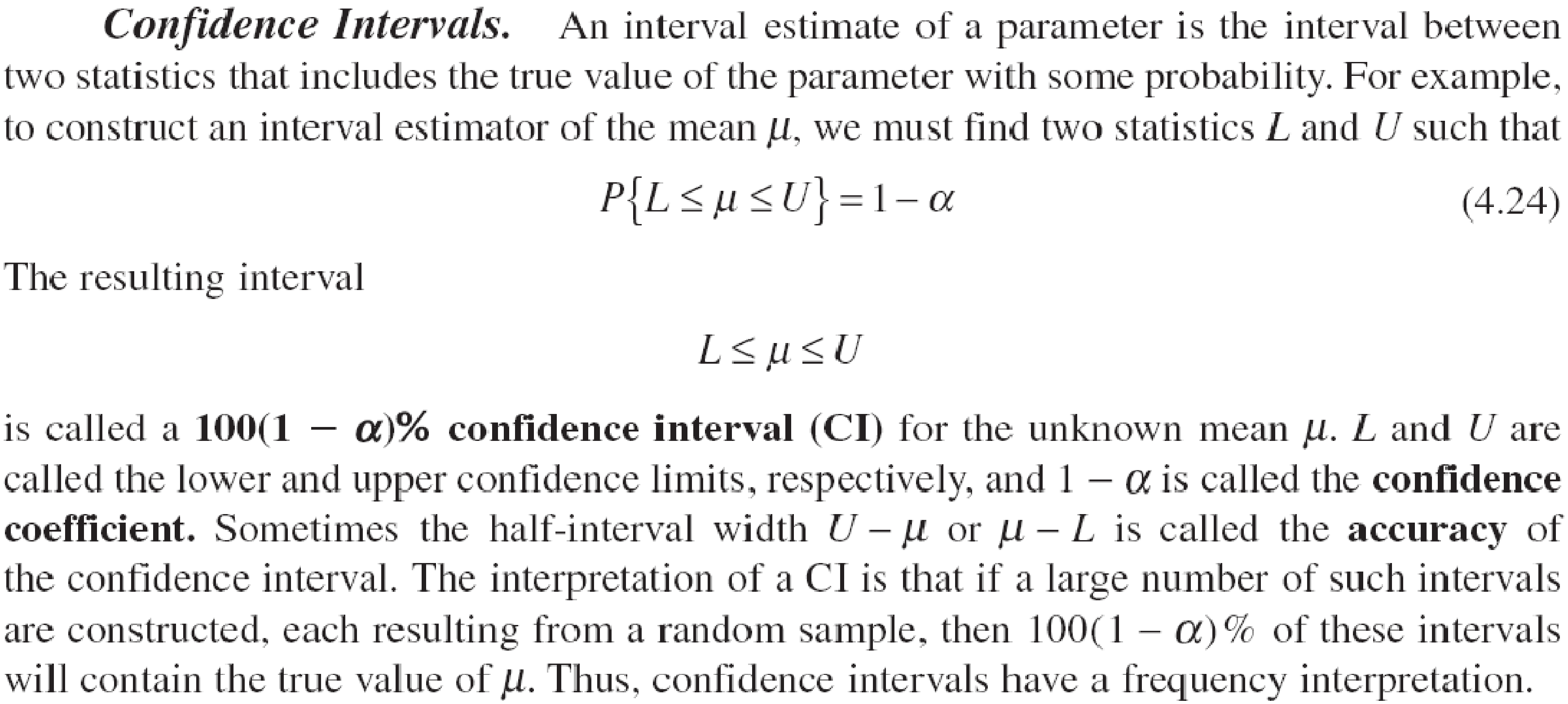
Because we specified a type I error of $\alpha = 0.05$ and the test is one-sided, then from Appendix Table II we find $Z_\alpha = Z_{0.05} = 1.645$. Therefore, we reject H_0: $\mu = 75$ and conclude that the mean response time exceeds 75 millisec.



One-Sample Z

```
Test of mu = 75 vs > 75
The assumed standard deviation = 8

                        95% Lower
 N    Mean   SE Mean        Bound      Z        P
25   79.25      1.60        76.62   2.66    0.004
```

Minitab output



Confidence Intervals. An interval estimate of a parameter is the interval between two statistics that includes the true value of the parameter with some probability. For example, to construct an interval estimator of the mean μ, we must find two statistics L and U such that

$$P\{L \le \mu \le U\} = 1 - \alpha \tag{4.24}$$

The resulting interval

$$L \le \mu \le U$$

is called a **100(1 − α)% confidence interval (CI)** for the unknown mean μ. L and U are called the lower and upper confidence limits, respectively, and $1 - \alpha$ is called the **confidence coefficient.** Sometimes the half-interval width $U - \mu$ or $\mu - L$ is called the **accuracy** of the confidence interval. The interpretation of a CI is that if a large number of such intervals are constructed, each resulting from a random sample, then $100(1 - \alpha)\%$ of these intervals will contain the true value of μ. Thus, confidence intervals have a frequency interpretation.

The CI (4.24) might be more properly called a **two-sided** confidence interval, as it specifies both a lower and an upper bound on μ. Sometimes in quality-control applications, a **one-sided** confidence bound might be more appropriate. A one-sided lower $100(1-\alpha)\%$ confidence bound on μ would be given by

$$L \le \mu \tag{4.25}$$

where L, the lower confidence bound, is chosen so that

$$P\{L \le \mu\} = 1-\alpha \tag{4.26}$$

A one-sided upper $100(1-\alpha)\%$ confidence bound on μ would be

$$\mu \le U \tag{4.27}$$

where U, the upper confidence bound, is chosen so that

$$P\{\mu \le U\} = 1-\alpha \tag{4.28}$$

Confidence Interval on the Mean with Variance Known. Consider the random variable x, with unknown mean μ and known variance σ^2. Suppose a random sample of n observations is taken—say, $x_1, x_2, \ldots, x_n$—and $\bar{x}$ is computed. Then the $100(1-\alpha)\%$ two-sided CI on μ is

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{4.29}$$

where $Z_{\alpha/2}$ is the percentage point of the $N(0, 1)$ distribution such that $P\{z \geq Z_{\alpha/2}\} = \alpha/2$.

$$\mu \leq \bar{x} + Z_{\alpha}\frac{\sigma}{\sqrt{n}} \tag{4.30}$$

$$\bar{x} - Z_{\alpha}\frac{\sigma}{\sqrt{n}} \leq \mu \tag{4.31}$$

Upper and lower confidence bounds

EXAMPLE 4.2 Computer Response Time

Reconsider the computer response time scenario from Example 4.1. Since $\bar{x} = 79.25$ millisec, we know that a reasonable point estimate of the mean response time is $\hat{\mu} = \bar{x} = 79.25$ millisec. Find a 95% two-sided confidence interval.

SOLUTION

From equation 4.29 we can compute

$$\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$79.25 - 1.96\frac{8}{\sqrt{25}} \le \mu \le 79.25 + 1.96\frac{8}{\sqrt{25}}$$

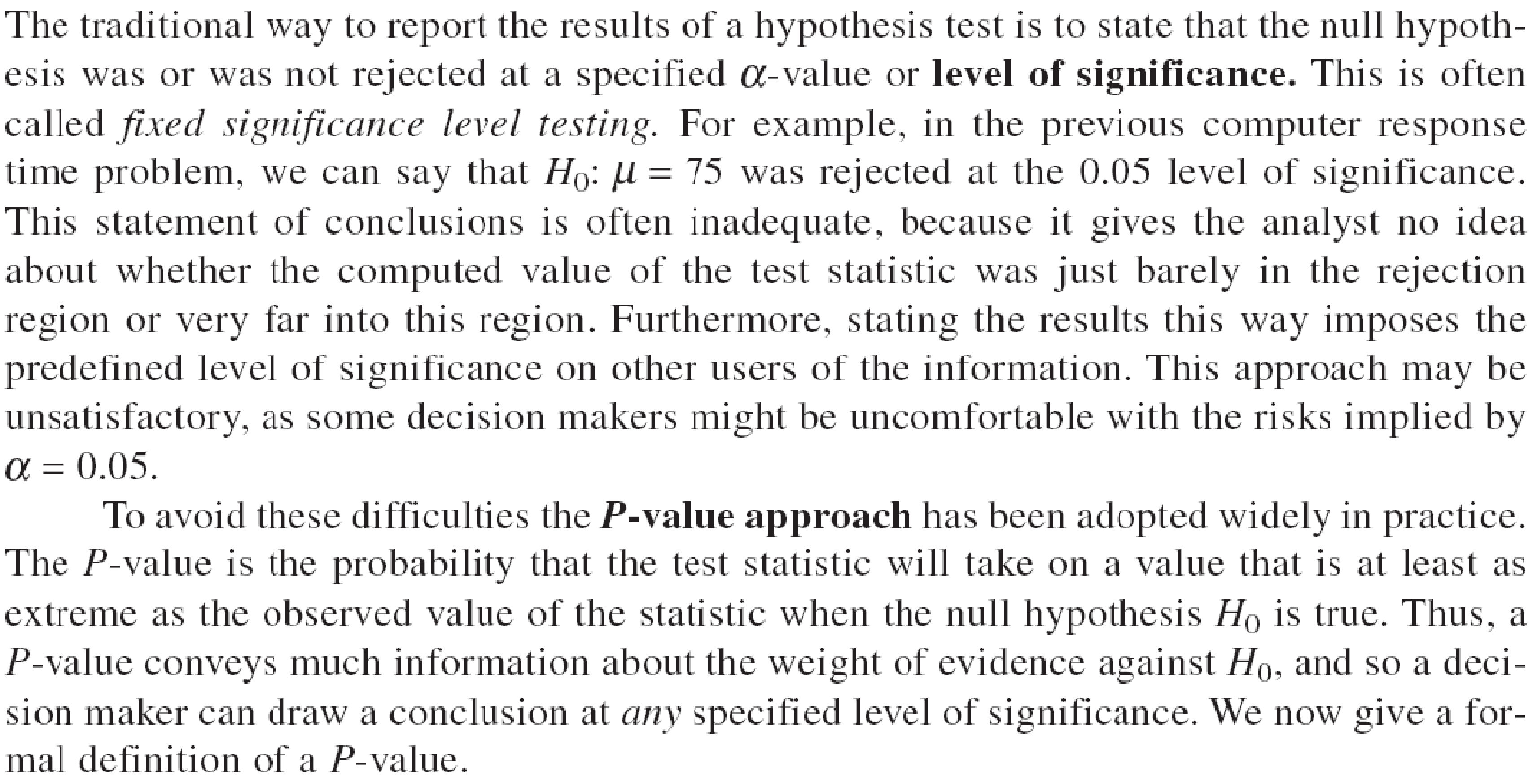
$$76.114 \le \mu \le 82.386$$

Another way to express this result is that our estimate of mean response time is 79.25 millisec ± 3.136 millisec with 95% confidence.

In the original Example 4.1, the alternative hypothesis was one-sided. In these situations, some analysts prefer to calculate a one-sided confidence bound. The Minitab output for Example 4.1 on p. 114 provider a 95% *lower confidence bound* on μ, which is computed from equation 4.31 as 76.62.

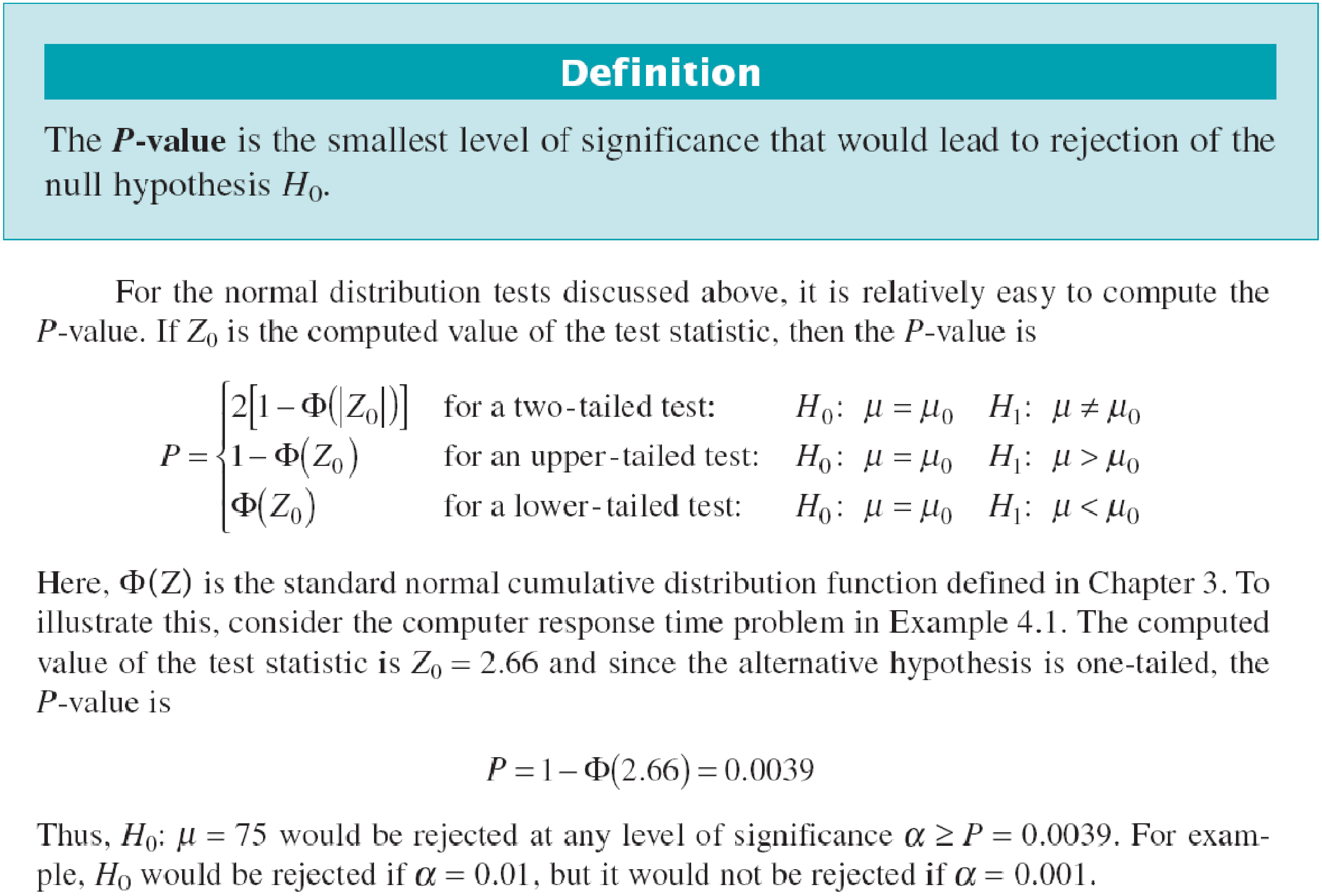
Notice that the CI from Minitab does *not* include the value $\mu = 75$. Furthermore, in Example 4.1 the hypothesis H_0: $\mu = 75$ was rejected at $\alpha = 0.05$. This is not a coincidence. In general, the test of significance for a parameter at level of significance α will lead to rejection of H_0 if, and only if, the parameter value specific in H_0 is not included in the $100(1 - \alpha)\%$ confidence interval.

4.3.2 *P*-Values



The traditional 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 previous computer response time problem, we can say that H_0: $\mu = 75$ was rejected at the 0.05 level of significance. This statement of conclusions is often inadequate, because it gives the analyst no idea about whether the computed value of the test statistic was just barely in the rejection region or 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, as 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. We now give a formal definition of a P-value.



Definition

The ***P*-value** is the smallest level of significance that would lead to rejection of the null hypothesis H_0.

For the normal distribution tests discussed above, it is relatively easy to compute the P-value. If Z_0 is the computed value of the test statistic, then the P-value is

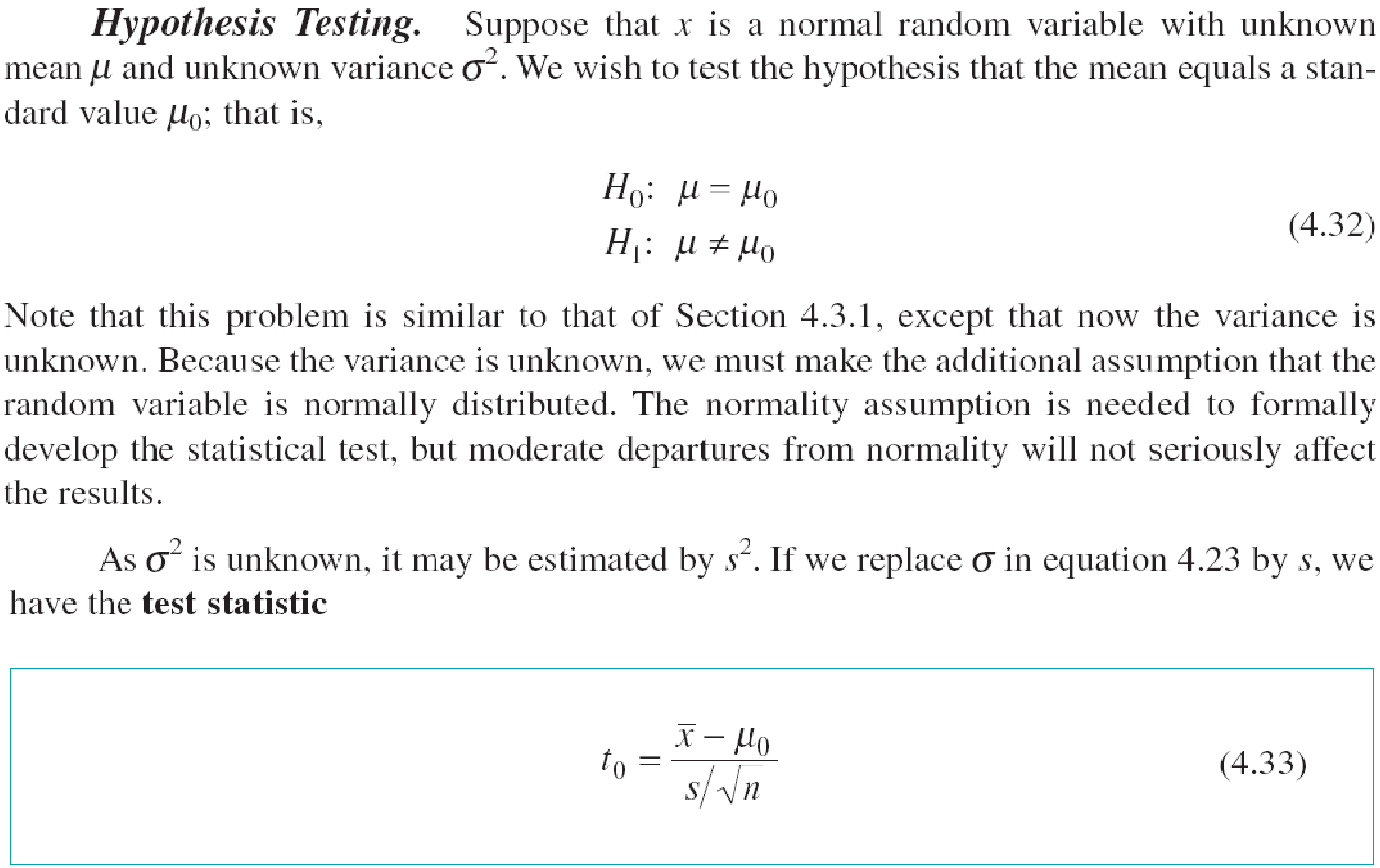
$$P = \begin{cases} 2\left[1-\Phi\left(\left|Z_0\right|\right)\right] & \text{for a two-tailed test:} \quad H_0\text{: } \mu = \mu_0 \quad H_1\text{: } \mu \neq \mu_0 \\ 1-\Phi\left(Z_0\right) & \text{for an upper-tailed test:} \quad H_0\text{: } \mu = \mu_0 \quad H_1\text{: } \mu > \mu_0 \\ \Phi\left(Z_0\right) & \text{for a lower-tailed test:} \quad H_0\text{: } \mu = \mu_0 \quad H_1\text{: } \mu < \mu_0 \end{cases}$$

Here, $\Phi(Z)$ is the standard normal cumulative distribution function defined in Chapter 3. To illustrate this, consider the computer response time problem in Example 4.1. The computed value of the test statistic is $Z_0 = 2.66$ and since the alternative hypothesis is one-tailed, the P-value is

$$P = 1-\Phi(2.66) = 0.0039$$

Thus, H_0: $\mu = 75$ would be rejected at any level of significance $\alpha \geq P = 0.0039$. For example, H_0 would be rejected if $\alpha = 0.01$, but it would not be rejected if $\alpha = 0.001$.

4.3.3 Inference on the Mean of a Normal Distribution, Variance Unknown



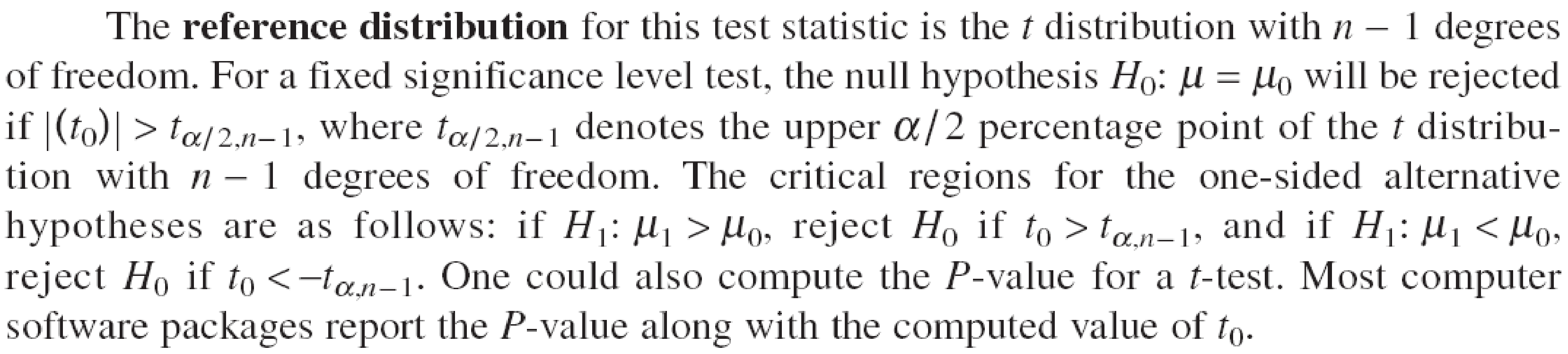
Hypothesis Testing. Suppose that x is a normal random variable with unknown mean μ and unknown variance σ^2. We wish to test the hypothesis that the mean equals a standard value μ_0; that is,

$$\begin{aligned} H_0&: \ \mu = \mu_0 \\ H_1&: \ \mu \neq \mu_0 \end{aligned} \qquad (4.32)$$

Note that this problem is similar to that of Section 4.3.1, except that now the variance is unknown. Because the variance is unknown, we must make the additional assumption that the random variable is normally distributed. The normality assumption is needed to formally develop the statistical test, but moderate departures from normality will not seriously affect the results.

As σ^2 is unknown, it may be estimated by s^2. If we replace σ in equation 4.23 by s, we have the **test statistic**

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad (4.33)$$



The **reference distribution** for this test statistic is the t distribution with $n - 1$ degrees of freedom. For a fixed significance level test, the null hypothesis H_0: $\mu = \mu_0$ will 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. The critical regions for the one-sided alternative hypotheses are as follows: if H_1: $\mu_1 > \mu_0$, reject H_0 if $t_0 > t_{\alpha,n-1}$, and if H_1: $\mu_1 < \mu_0$, reject H_0 if $t_0 < -t_{\alpha,n-1}$. One could also compute the P-value for a t-test. Most computer software packages report the P-value along with the computed value of t_0.

Checking the normality assumption – we will see how to do this in Example 4.3

EXAMPLE 4.3 Rubberized Asphalt

Rubber can be added to asphalt to reduce road noise when the material is used as pavement. Table 4.1 shows the stabilized viscosity (cP) of 15 specimens of asphalt paving material. To be suitable for the intended pavement application, the mean stabilized viscosity should be equal to 3200. Test this hypothesis using $\alpha = 0.05$. Based on experience we are willing to initially assume that stabilized viscosity is normally distributed.

■ TABLE 4.1
Stabilized Viscosity of Rubberized Asphalt

Specimen	Stabilized Viscosity
1	3193
2	3124
3	3153
4	3145
5	3093
6	3466
7	3355
8	2979
9	3182
10	3227
11	3256
12	3332
13	3204
14	3282
15	3170

SOLUTION

The appropriate hypotheses are

$$H_0\colon\ \mu = 3200$$
$$H_1\colon\ \mu \neq 3200$$

The sample mean and sample standard deviation are

$$\bar{x} = \frac{1}{15}\sum_{i=1}^{15} x_i = \frac{48,161}{15} = 3210.73$$

$$s = \sqrt{\frac{\sum_{i=1}^{15} x_i^2 - \frac{\left(\sum_{i=1}^{15} x_i\right)^2}{15}}{15-1}}$$

$$= \sqrt{\frac{154,825,783 - \frac{(48,161)^2}{15}}{14}} = 117.61$$

and the test statistic is

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3210.73 - 3200}{117.61/\sqrt{15}} = 0.35$$

Since the calculated value of the test statistic does not exceed $t_{0.025,\,14} = 2.145$ or $-t_{0.025,\,14} = -2.145$, we cannot reject the null hypothesis. Therefore, there is no strong evidence to conclude that the mean stabilized viscosity is different from 3200 cP.

The assumption of normality for the *t*-test can be checked by constructing a normal probability plot of the stabilized viscosity data. Figure 4.5 shows the normal probability plot. Because the observations lie along the straight line, there is no problem with the normality assumption.

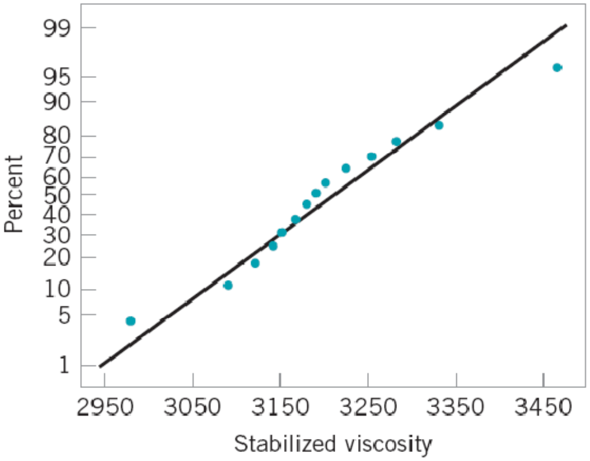
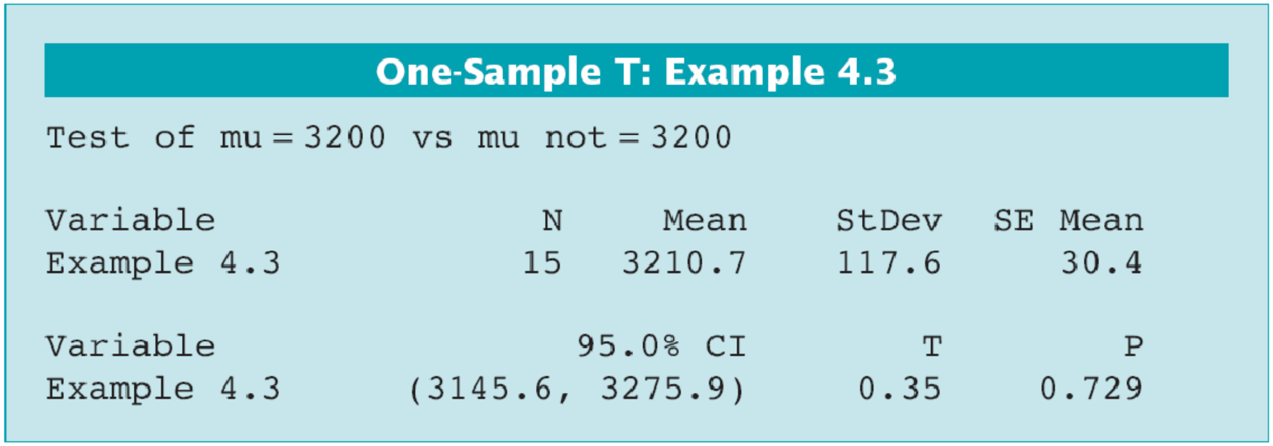


FIGURE 4.5 Normal probability plot of the stabilized viscosity data.



One-Sample T: Example 4.3

```
Test of mu = 3200 vs mu not = 3200

Variable                N     Mean    StDev  SE Mean
Example 4.3            15   3210.7    117.6     30.4

Variable                   95.0% CI        T        P
Example 4.3        (3145.6, 3275.9)     0.35    0.729
```

Confidence Interval on the Mean of a Normal Distribution with Variance Unknown. Suppose that x is a normal random variable with unknown mean μ and unknown variance σ^2. From a random sample of n observations the sample mean $\bar{x}$ and sample variance s^2 are computed. Then a $100(1-\alpha)\%$ two-sided CI on the true mean is

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \tag{4.34}$$

where $t_{\alpha/2,n-1}$ denotes the percentage point of the t distribution with $n-1$ degrees of freedom such that $P\{t_{n-1} \ge t_{\alpha/2,\,n-1}\} = \alpha/2$. The corresponding upper and lower $100(1-\alpha)\%$ confidence bounds are

$$\mu \le \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}} \tag{4.35}$$

and

$$\bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}} \le \mu \tag{4.36}$$

respectively.

EXAMPLE 4.4 Rubberized Viscosity

Reconsider the stabilized viscosity data from Example 4.3. Find a 95% confidence interval on the mean stabilized viscosity.

SOLUTION

Using equation 4.34, we can find the 95% CI on the mean stabilized viscosity as follows:

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

$$3210.73 - 2.145\frac{117.61}{\sqrt{15}} \le \mu \le 3210.73 + 2.145\frac{117.61}{\sqrt{15}}$$

$$3145.59 \le \mu \le 3275.87$$

Another way to express this result is that our estimate of the mean stabilized viscosity is 3210.73 ± 65.14 cP with 95% confidence. This confidence interval was reported by Minitab in the box on page 118.

The manufacturer may only be concerned about stabilized viscosity values that are too low and consequently may be interested in a one-sided confidence bound. The 95% lower confidence bound on mean stabilized viscosity is found from equation 4.36, using $t_{0.05,\,14} = 1.761$ as

$$3210.73 - 1.761\frac{117.61}{\sqrt{15}} \le \mu$$

or

$$3157.25 \le \mu$$

4.3.4 Inference on the Variance of a Normal Distribution

Hypothesis Testing. We now review hypothesis testing on the variance of a normal distribution. Whereas tests on means are relatively insensitive to the normality assumption, test procedures for variances are not.

Suppose we wish to test the hypothesis that the variance of a normal distribution equals a constant—say, σ_0^2. The hypotheses are

$$\begin{aligned} H_0&: \quad \sigma^2 = \sigma_0^2 \\ H_1&: \quad \sigma^2 \neq \sigma_0^2 \end{aligned} \tag{4.37}$$

The test statistic for this hypothesis is

$$\boxed{\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} \qquad (4.38)}$$

where s^2 is the sample variance computed from a random sample of n observations. For a fixed significance level test, 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. If a one-sided alternative is specified—say, H_1: $\sigma^2 < \sigma_0^2$, then we would reject if $\chi_0^2 < \chi_{1-\alpha,n-1}^2$. For the other one-sided alternative H_1: $\sigma^2 > \sigma_0^2$, reject if $\chi_0^2 > \chi_{\alpha,n-1}^2$.

Confidence Interval on the Variance of a Normal Distribution. Suppose that x is a normal random variable with unknown mean μ and unknown variance σ^2. Let the sample variance s^2 be computed from a random sample of n observations. Then a $100(1 - \alpha)\%$ two-sided CI on the variance is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \qquad (4.39)$$

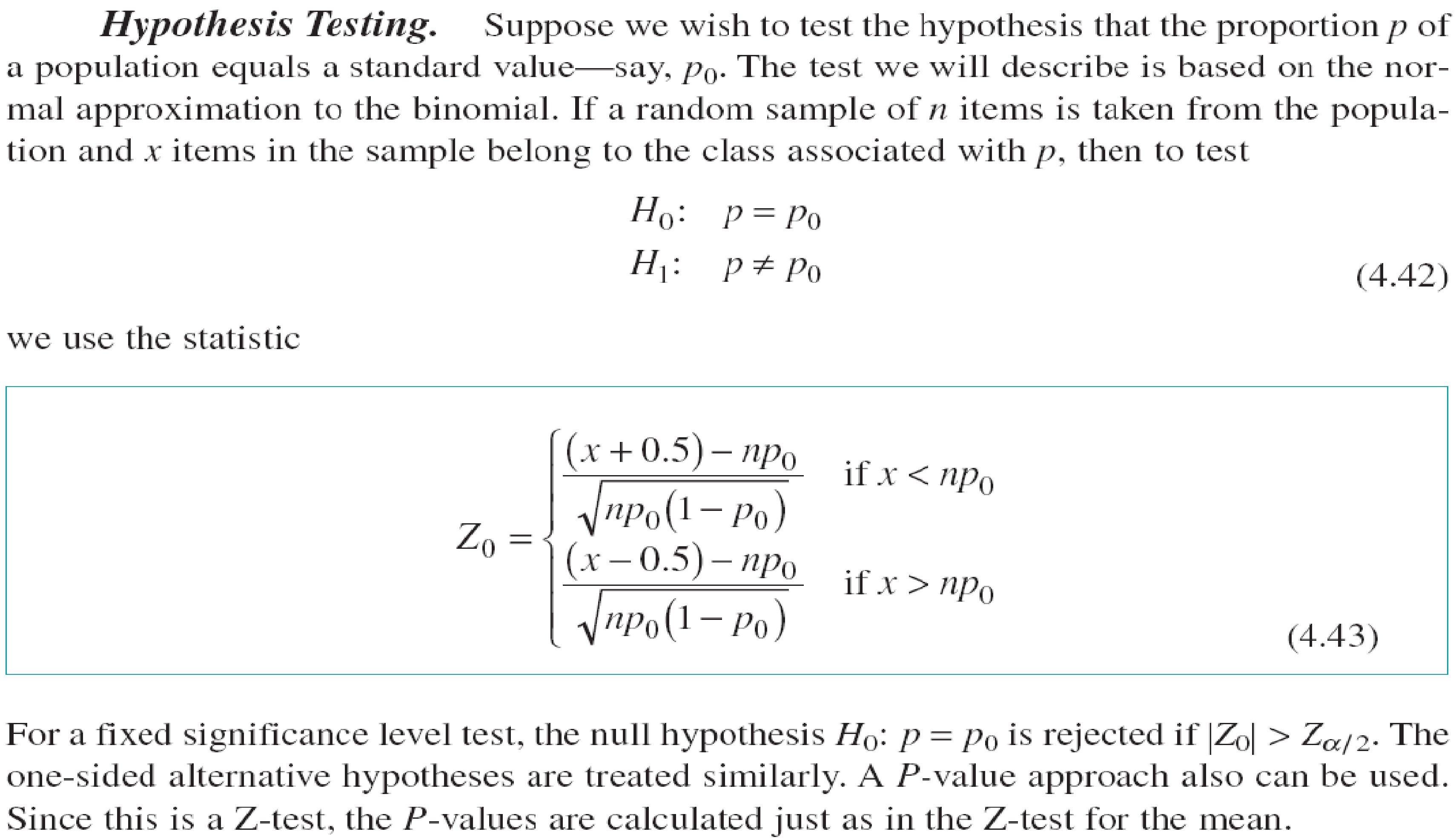
where $\chi^2_{\alpha/2,n-1}$ denotes the percentage point of the chi-square distribution such that $P\{\chi^2_{n-1} \ge \chi^2_{\alpha/2,n-1}\} = \alpha/2$. A CI or the standard deviation can be found by taking the square root throughout in equation (4.39).

$$\sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}} \qquad (4.40)$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha,n-1}} \le \sigma^2 \qquad (4.41)$$

Upper and lower confidence bounds

4.3.5 Inference on a Population Proportion



Hypothesis Testing. Suppose we wish to test the hypothesis that the proportion p of a population equals a standard value—say, p_0. The test we will describe is based on the normal approximation to the binomial. If a random sample of n items is taken from the population and x items in the sample belong to the class associated with p, then to test

$$\begin{aligned} H_0&: \quad p = p_0 \\ H_1&: \quad p \neq p_0 \end{aligned} \tag{4.42}$$

we use the statistic

$$Z_0 = \begin{cases} \dfrac{(x+0.5)-np_0}{\sqrt{np_0(1-p_0)}} & \text{if } x < np_0 \\ \dfrac{(x-0.5)-np_0}{\sqrt{np_0(1-p_0)}} & \text{if } x > np_0 \end{cases} \tag{4.43}$$

For a fixed significance level test, the null hypothesis H_0: $p = p_0$ is rejected if $|Z_0| > Z_{\alpha/2}$. The one-sided alternative hypotheses are treated similarly. A P-value approach also can be used. Since this is a Z-test, the P-values are calculated just as in the Z-test for the mean.

EXAMPLE 4.5 A Forging Process

A foundry produces steel forgings used in automobile manufacturing. We wish to test the hypothesis that the fraction conforming or fallout from this process is 10%. In a random sample of 250 forgings, 41 were found to be nonconforming. What are your conclusions using $\alpha = 0.05$?

SOLUTION

To test

$$H_0: \quad p = 0.1$$
$$H_1: \quad p \neq 0.1$$

we calculate the test statistic

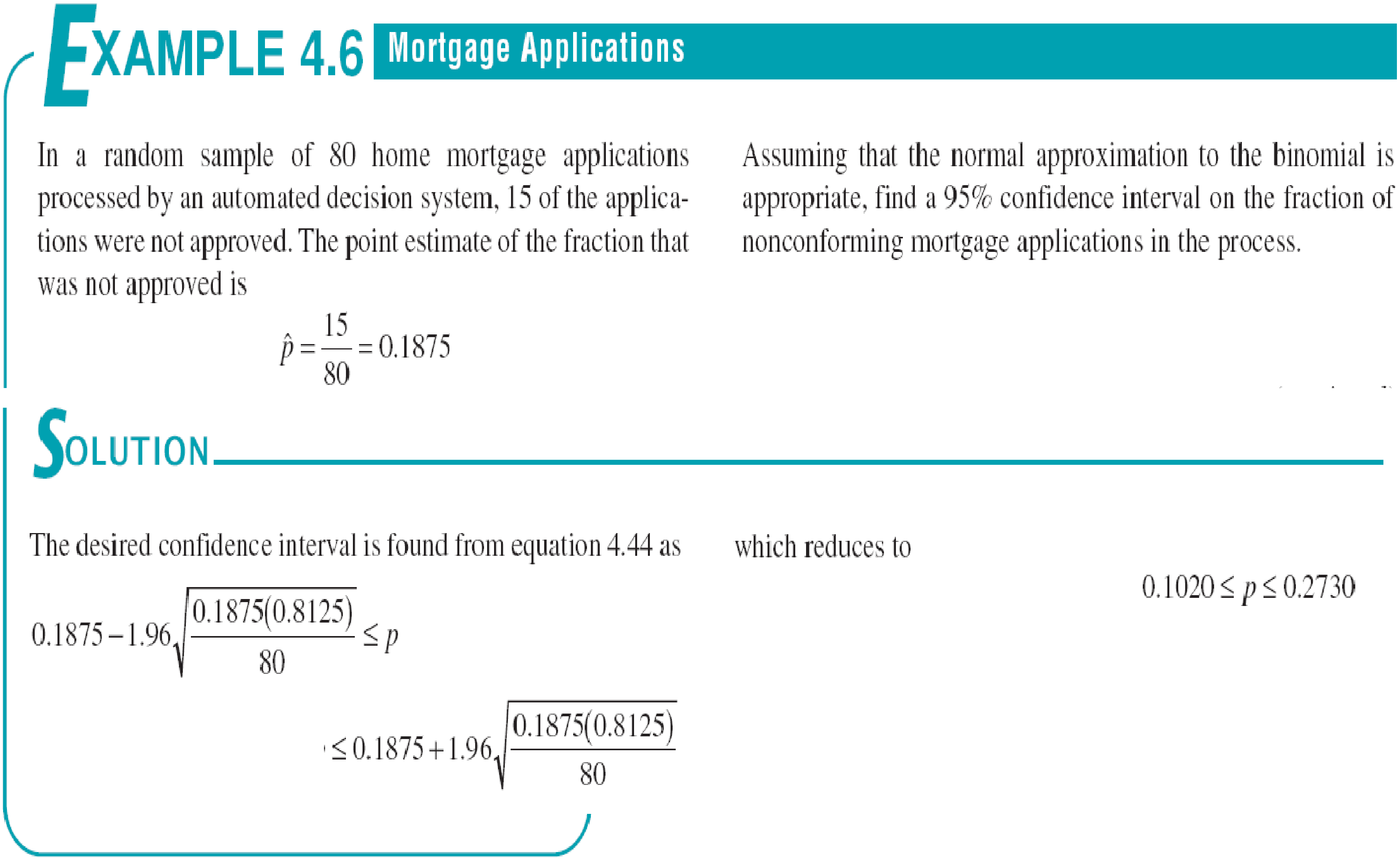
$$Z_0 = \frac{(x - 0.5) - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{(41 - 0.5) - (250)(0.1)}{\sqrt{250(0.1)(1 - 0.1)}} = 3.27$$

Using $\alpha = 0.05$ we find $Z_{0.025} = 1.96$, and therefore H_0: $p = 0.1$ is rejected (the P-value here is $P = 0.00108$). That is, the process fraction nonconforming or fallout is not equal to 10%.

Confidence Intervals on a Population Proportion. It is frequently necessary to construct 100(1 – α)% CIs on a population proportion p. This parameter frequently corresponds to a lot or process fraction nonconforming. Now p is only one of the parameters of a binomial distribution, and we usually assume that the other binomial parameter n is known. If a random sample of n observations from the population has been taken, and x "nonconforming" observations have been found in this sample, then the unbiased point estimator of p is $\hat{p} = x/n$.

There are several approaches to constructing the CI on p. If n is large and $p \geq 0.1$ (say), then the normal approximation to the binomial can be used, resulting in the 100(1 – α)% confidence interval:

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (4.44)$$



EXAMPLE 4.6 Mortgage Applications

In a random sample of 80 home mortgage applications processed by an automated decision system, 15 of the applications were not approved. The point estimate of the fraction that was not approved is

$$\hat{p} = \frac{15}{80} = 0.1875$$

Assuming that the normal approximation to the binomial is appropriate, find a 95% confidence interval on the fraction of nonconforming mortgage applications in the process.

SOLUTION

The desired confidence interval is found from equation 4.44 as

$$0.1875 - 1.96\sqrt{\frac{0.1875(0.8125)}{80}} \leq p$$

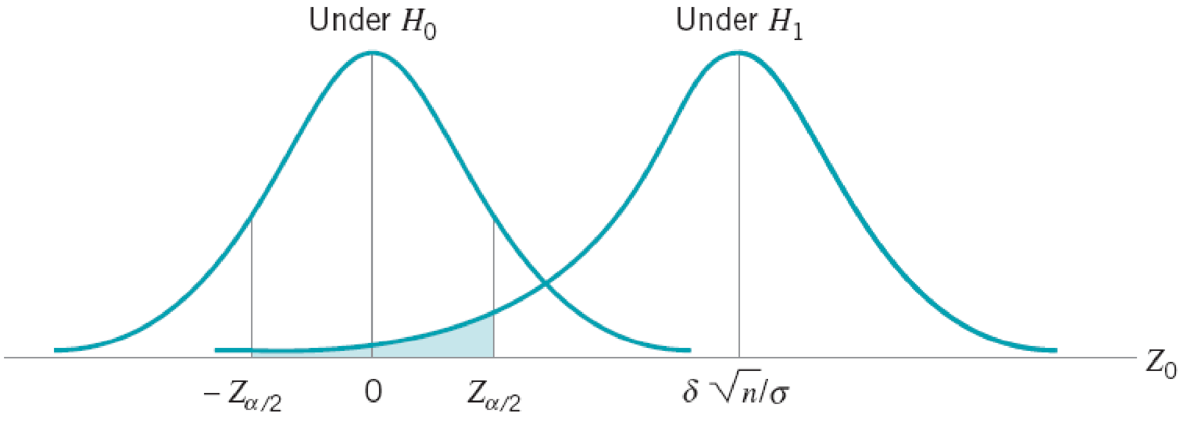
$$\leq 0.1875 + 1.96\sqrt{\frac{0.1875(0.8125)}{80}}$$

which reduces to

$$0.1020 \leq p \leq 0.2730$$

4.3.6 Type II Error and Sample Size

$$\beta = \Phi\left(Z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma} \right) - \Phi\left(-Z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma} \right) \qquad (4.46)$$



■ FIGURE 4.6 The distribution of Z_0 under H_0 and H_1.

EXAMPLE 4.7 Finding the Power of a Test

The mean contents of coffee cans filled on a particular production line are being studied. Standards specify that the mean contents must be 16.0 oz, and from past experience it is known that the standard deviation of the can contents is 0.1 oz. The hypotheses are

$$H_0: \quad \mu = 16.0$$
$$H_1: \quad \mu \neq 16.0$$

A random sample of nine cans is to be used, and the type I error probability is specified as $\alpha = 0.05$. Therefore, the test statistic is

$$Z_0 = \frac{\bar{x} - 16.0}{0.1\sqrt{9}}$$

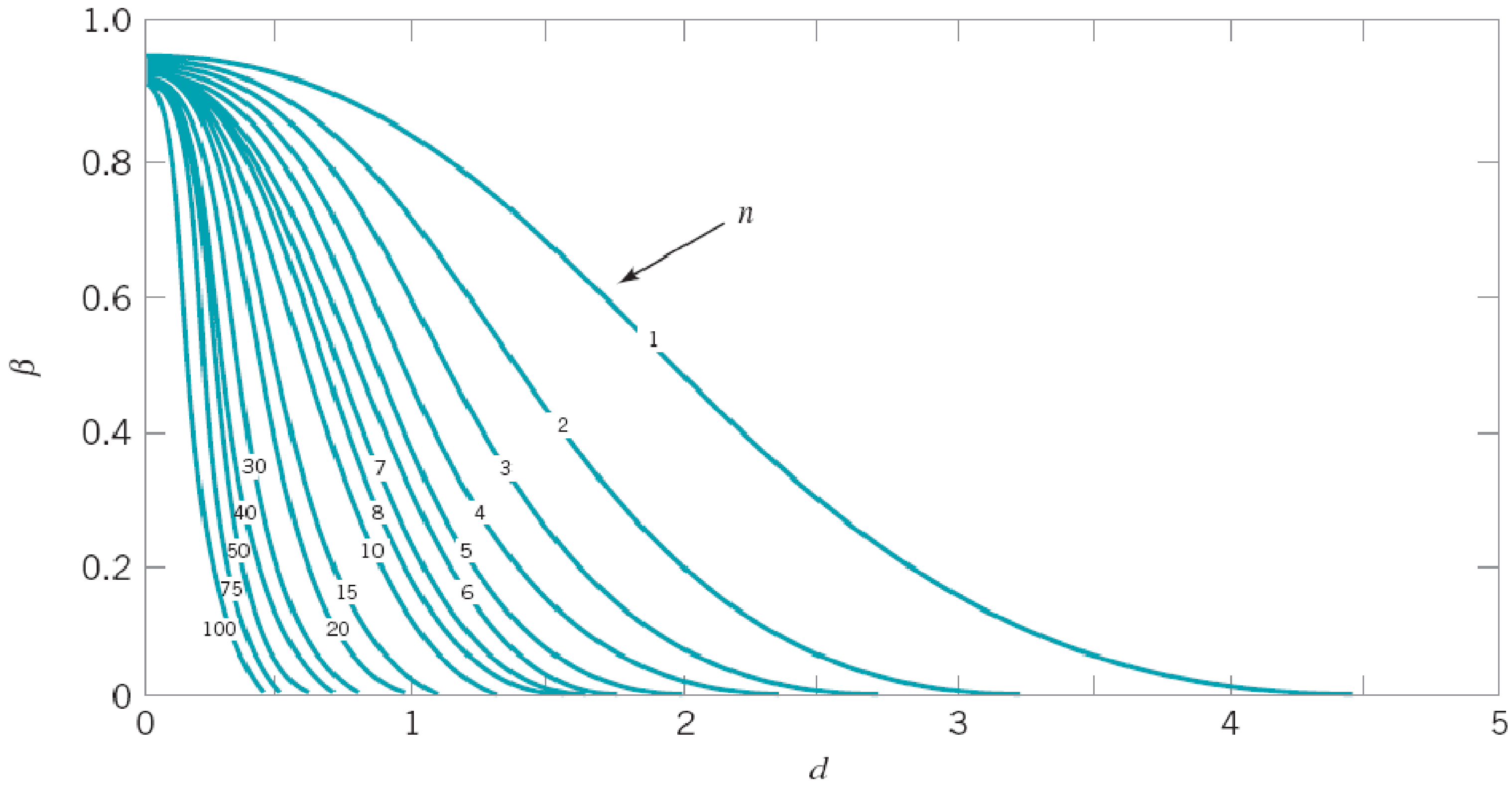
and H_0 is rejected if $|Z_0| > Z_{0.025} = 1.96$. Find the probability of type II error and the power of the test, if the true mean contents are $\mu_1 = 16.1$ oz.

SOLUTION

Since we are given that $\delta = \mu_1 - \mu_0 = 16.1 - 16.0 = 0.1$, we have

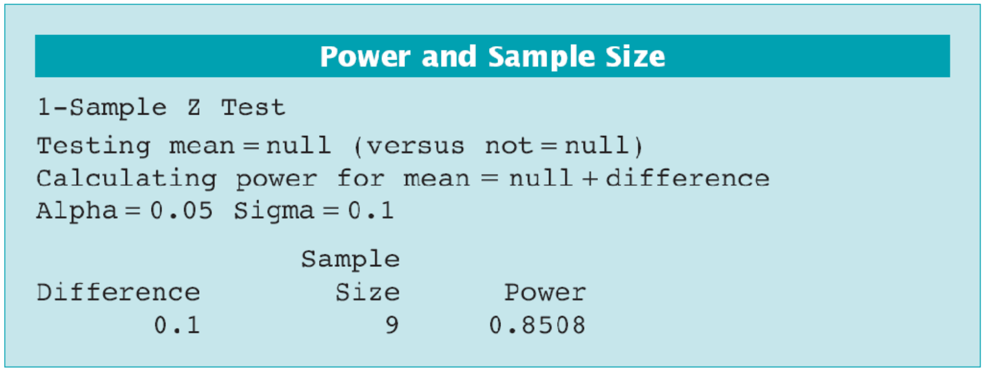
$$\begin{aligned}\beta &= \Phi\left(Z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-Z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right)\\ &= \Phi\left(1.96 - \frac{(0.1)(3)}{0.1}\right) - \Phi\left(-1.96 - \frac{(0.1)(3)}{0.1}\right)\\ &= \Phi(-1.04) - \Phi(-4.96)\\ &= 0.1492\end{aligned}$$

That is, the probability that we will incorrectly fail to reject H_0 if the true mean contents are 16.1 oz is 0.1492. Equivalently, we can say that the power of the test is $1 - \beta = 1 - 0.1492 = 0.8508$.



■ **FIGURE 4.7** Operating-characteristic curves for the two-sided normal test with $\alpha = 0.05$. (Reproduced with permission from C. L. Ferris, F. E. Grubbs, and C. L. Weaver, "Operating Characteristic Curves for the Common Statistical Tests of Significance," *Annals of Mathematical Statistics*, June 1946.)

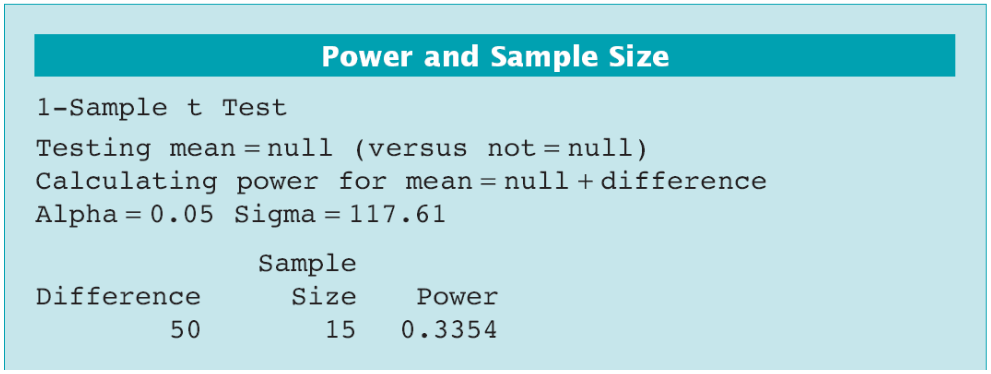
Minitab can perform sample size and power calculations. From Example 4.7:



Power and Sample Size

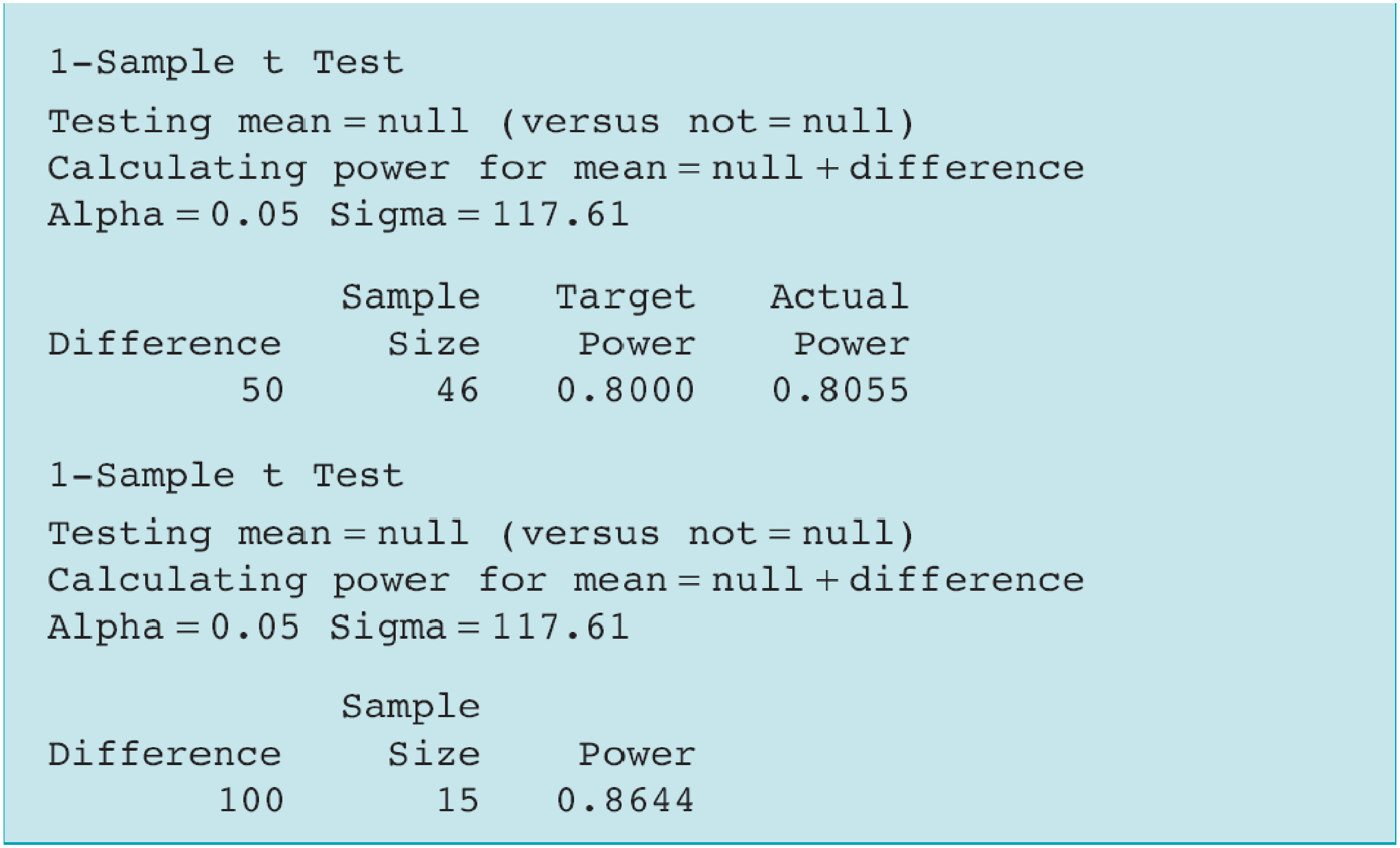
```
1-Sample Z Test
Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0.05  Sigma = 0.1

                 Sample
Difference         Size      Power
       0.1            9     0.8508
```

From Example 4.3:



Power and Sample Size

```
1-Sample t Test
Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0.05  Sigma = 117.61

              Sample
Difference      Size    Power
        50        15   0.3354
```



```
1-Sample t Test

Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0.05  Sigma = 117.61

                Sample    Target    Actual
Difference        Size     Power     Power
        50          46    0.8000    0.8055

1-Sample t Test

Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0.05  Sigma = 117.61

                Sample
Difference        Size     Power
       100          15    0.8644
```

4.4 Statistical Inference for Two Samples

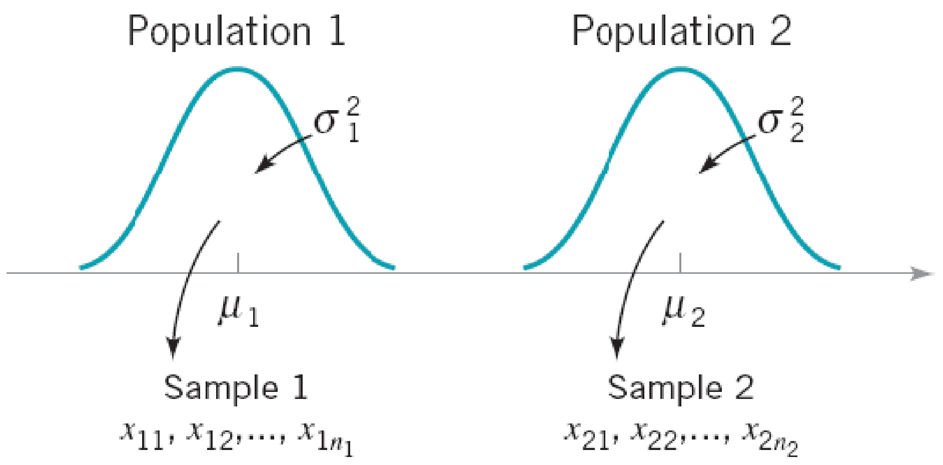
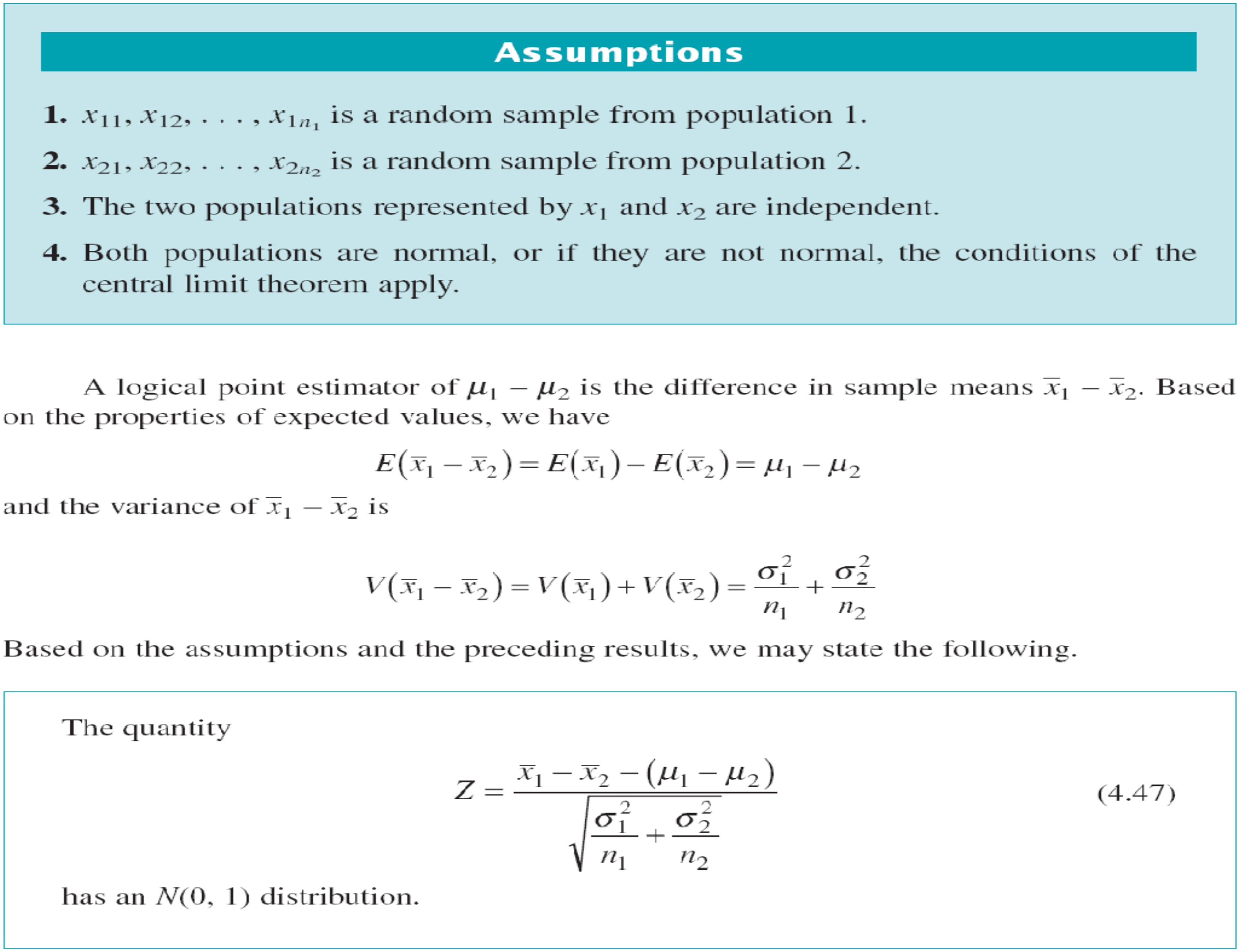


FIGURE 4.8 Two independent populations

Comparing means, comparing variances, comparing proportions

4.4.1 Inference on the Difference in Means, Variances Known



Assumptions

1. $x_{11}, x_{12}, \ldots, x_{1n_1}$ is a random sample from population 1.
2. $x_{21}, x_{22}, \ldots, x_{2n_2}$ is a random sample from population 2.
3. The two populations represented by x_1 and x_2 are independent.
4. Both populations are normal, or if they are not normal, the conditions of the central limit theorem apply.

A logical point estimator of $\mu_1 - \mu_2$ is the difference in sample means $\overline{x}_1 - \overline{x}_2$. Based on the properties of expected values, we have

$$E(\overline{x}_1 - \overline{x}_2) = E(\overline{x}_1) - E(\overline{x}_2) = \mu_1 - \mu_2$$

and the variance of $\overline{x}_1 - \overline{x}_2$ is

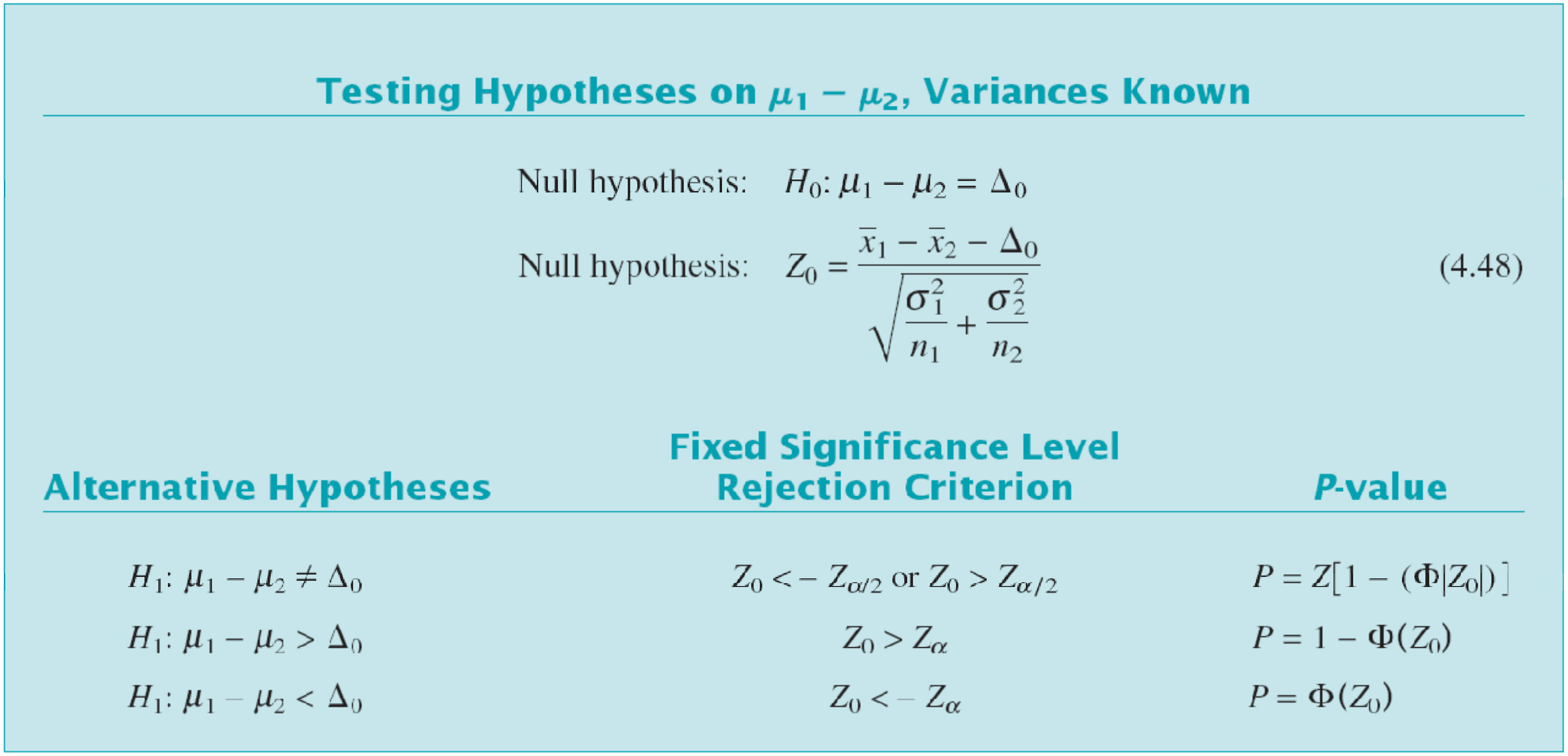
$$V(\overline{x}_1 - \overline{x}_2) = V(\overline{x}_1) + V(\overline{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Based on the assumptions and the preceding results, we may state the following.

The quantity

$$Z = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad (4.47)$$

has an $N(0,\ 1)$ distribution.

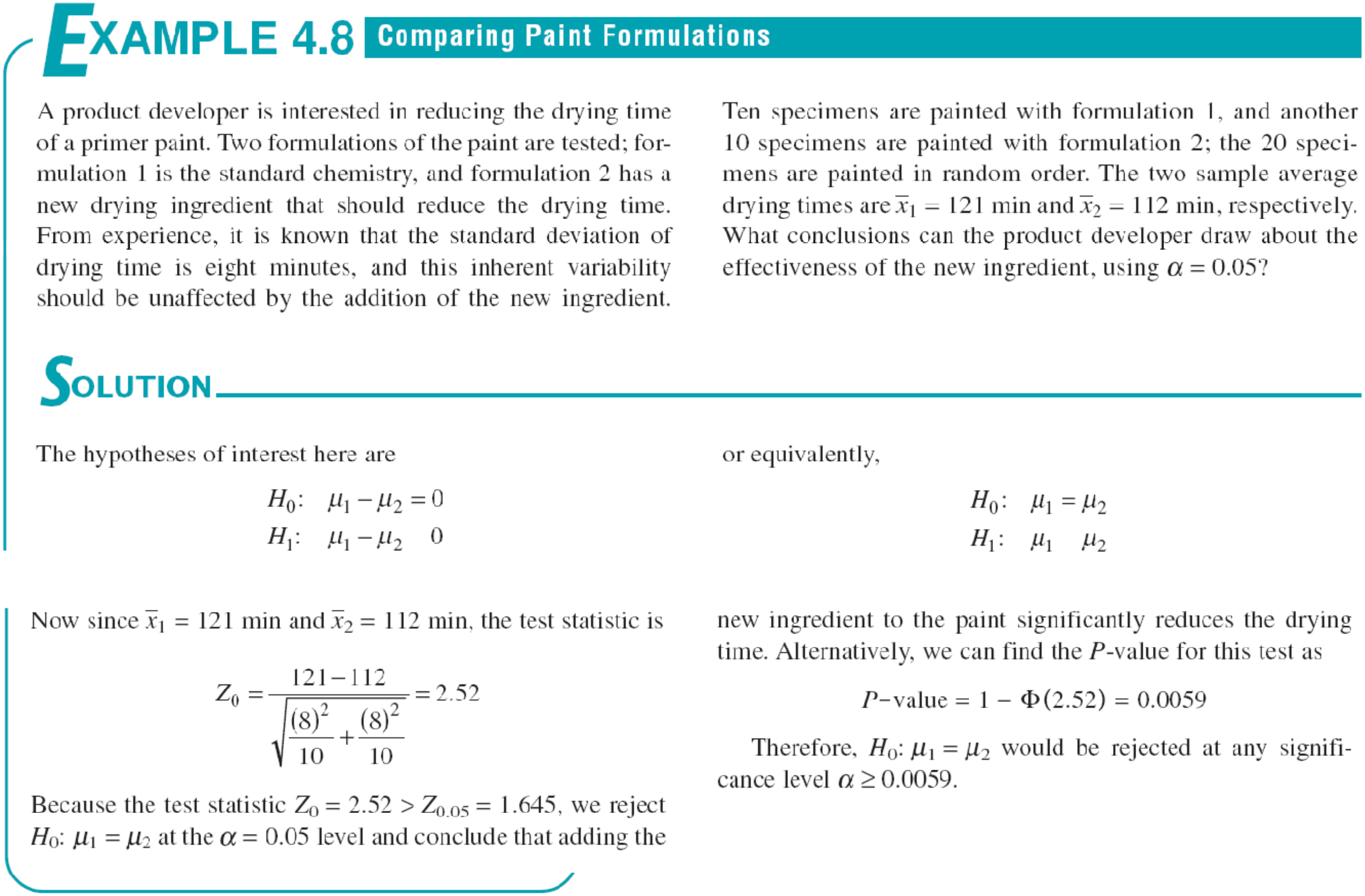


Testing Hypotheses on $\mu_1 - \mu_2$, Variances Known

Null hypothesis: $H_0\colon \mu_1 - \mu_2 = \Delta_0$

Null hypothesis: $$Z_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad (4.48)$$

Alternative Hypotheses	Fixed Significance Level Rejection Criterion	*P*-value
$H_1\colon \mu_1 - \mu_2 \neq \Delta_0$	$Z_0 < -Z_{\alpha/2}$ or $Z_0 > Z_{\alpha/2}$	$P = Z[1 - (\Phi\|Z_0\|)]$
$H_1\colon \mu_1 - \mu_2 > \Delta_0$	$Z_0 > Z_\alpha$	$P = 1 - \Phi(Z_0)$
$H_1\colon \mu_1 - \mu_2 < \Delta_0$	$Z_0 < -Z_\alpha$	$P = \Phi(Z_0)$



EXAMPLE 4.8 Comparing Paint Formulations

A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is eight minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order. The two sample average drying times are $\bar{x}_1 = 121$ min and $\bar{x}_2 = 112$ min, respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

SOLUTION

The hypotheses of interest here are

$$H_0\colon \quad \mu_1 - \mu_2 = 0$$
$$H_1\colon \quad \mu_1 - \mu_2 \quad 0$$

or equivalently,

$$H_0\colon \quad \mu_1 = \mu_2$$
$$H_1\colon \quad \mu_1 \quad \mu_2$$

Now since $\bar{x}_1 = 121$ min and $\bar{x}_2 = 112$ min, the test statistic is

$$Z_0 = \frac{121 - 112}{\sqrt{\frac{(8)^2}{10} + \frac{(8)^2}{10}}} = 2.52$$

Because the test statistic $Z_0 = 2.52 > Z_{0.05} = 1.645$, we reject $H_0\colon \mu_1 = \mu_2$ at the $\alpha = 0.05$ level and conclude that adding the new ingredient to the paint significantly reduces the drying time. Alternatively, we can find the P-value for this test as

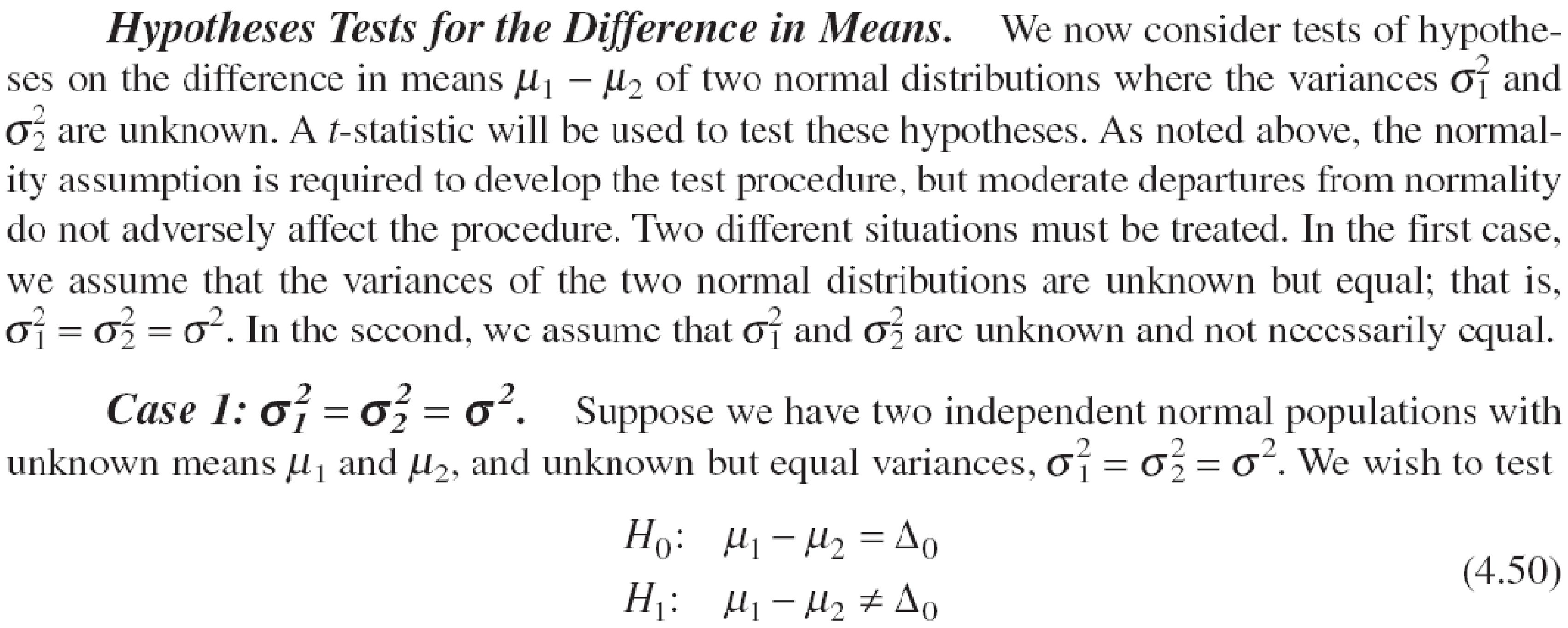
$$P\text{-value} = 1 - \Phi(2.52) = 0.0059$$

Therefore, $H_0\colon \mu_1 = \mu_2$ would be rejected at any significance level $\alpha \geq 0.0059$.

$$\bar{x}_1 - \bar{x}_2 - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (4.49)$$

This is a two-sided CI. The one-sided confidence bounds would be found by using only one of the limits in Equation (4.49) with α/2 replaced by α.

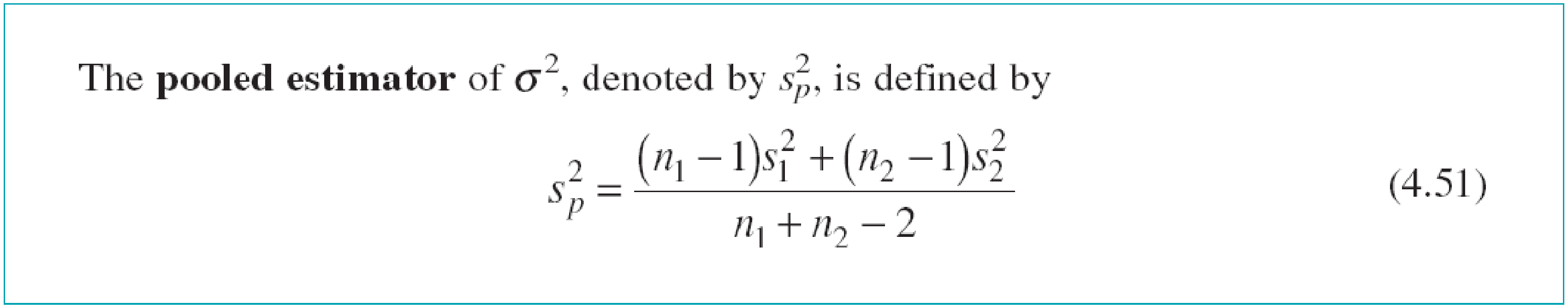
4.4.2 Inference on the Difference in Means of Two Normal Distributions, Variances Unknown



Hypotheses Tests for the Difference in Means. We now consider tests of hypotheses on the difference in means $\mu_1 - \mu_2$ of two normal distributions where the variances σ_1^2 and σ_2^2 are unknown. A t-statistic will be used to test these hypotheses. As noted above, the normality assumption is required to develop the test procedure, but moderate departures from normality do not adversely affect the procedure. Two different situations must be treated. In the first case, we assume that the variances of the two normal distributions are unknown but equal; that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. In the second, we assume that σ_1^2 and σ_2^2 are unknown and not necessarily equal.

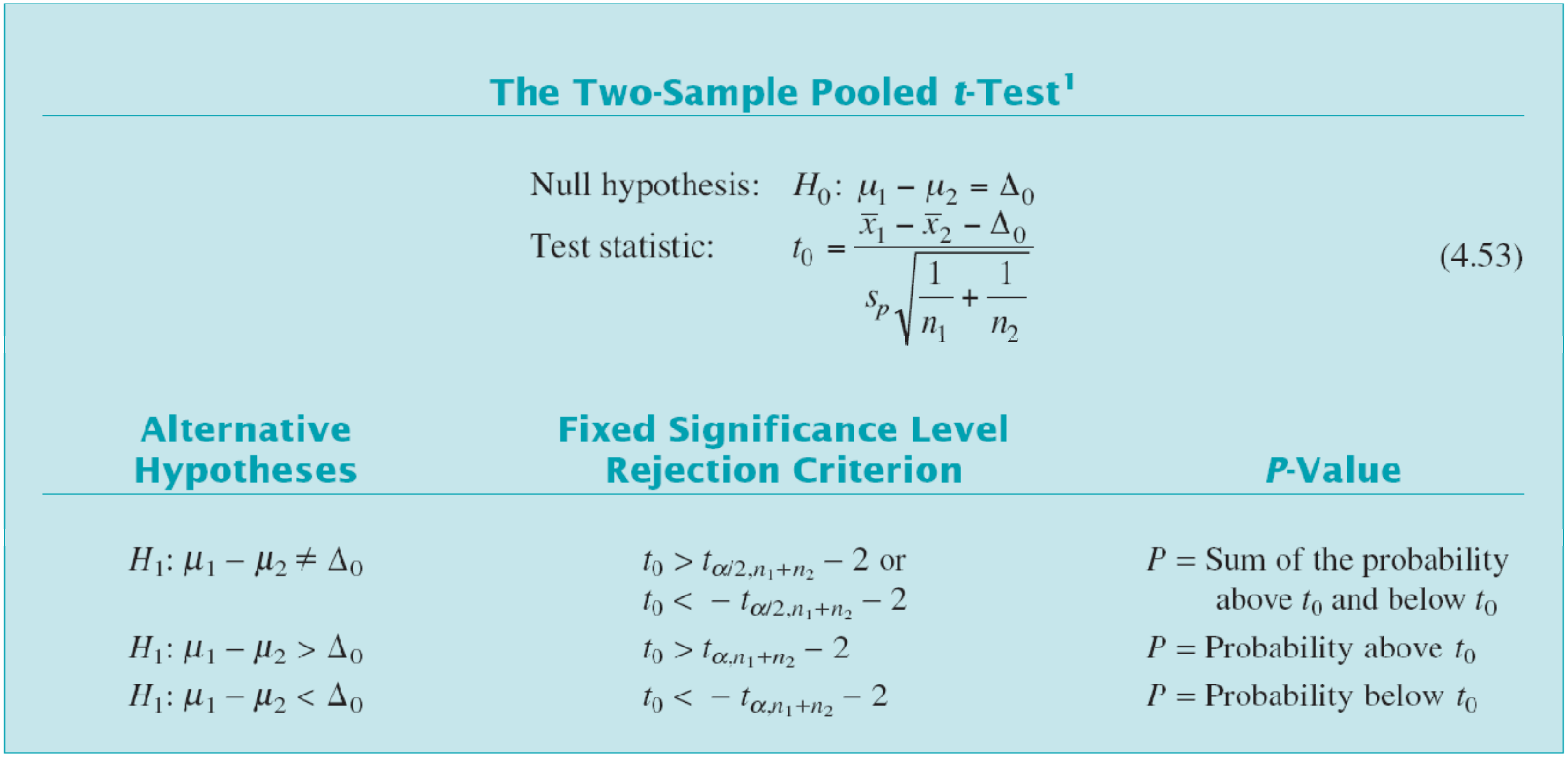
Case 1: $\boldsymbol{\sigma_1^2 = \sigma_2^2 = \sigma^2}$**.** Suppose we have two independent normal populations with unknown means μ_1 and μ_2, and unknown but equal variances, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. We wish to test

$$\begin{aligned} H_0&: \quad \mu_1 - \mu_2 = \Delta_0 \\ H_1&: \quad \mu_1 - \mu_2 \neq \Delta_0 \end{aligned} \tag{4.50}$$



The **pooled estimator** of σ^2, denoted by s_p^2, is defined by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad (4.51)$$

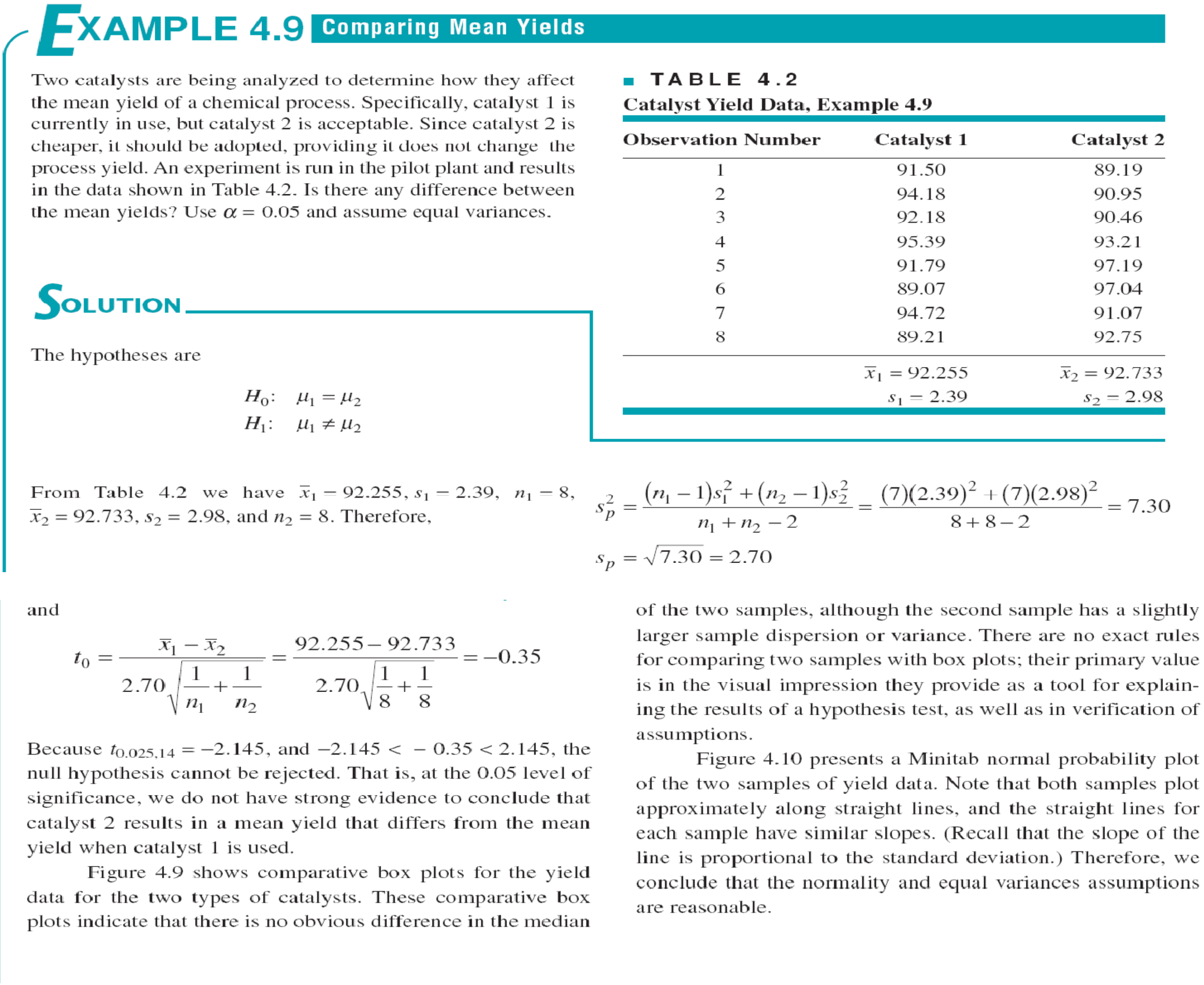


The Two-Sample Pooled *t*-Test[1]

Null hypothesis: $H_0\colon \mu_1 - \mu_2 = \Delta_0$

Test statistic: $$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \qquad (4.53)$$

Alternative Hypotheses	Fixed Significance Level Rejection Criterion	*P*-Value
$H_1\colon \mu_1 - \mu_2 \neq \Delta_0$	$t_0 > t_{\alpha/2,n_1+n_2} - 2$ or $t_0 < -t_{\alpha/2,n_1+n_2} - 2$	P = Sum of the probability above t_0 and below t_0
$H_1\colon \mu_1 - \mu_2 > \Delta_0$	$t_0 > t_{\alpha,n_1+n_2} - 2$	P = Probability above t_0
$H_1\colon \mu_1 - \mu_2 < \Delta_0$	$t_0 < -t_{\alpha,n_1+n_2} - 2$	P = Probability below t_0



EXAMPLE 4.9 Comparing Mean Yields

Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable. Since catalyst 2 is cheaper, it should be adopted, providing it does not change the process yield. An experiment is run in the pilot plant and results in the data shown in Table 4.2. Is there any difference between the mean yields? Use $\alpha = 0.05$ and assume equal variances.

■ TABLE 4.2
Catalyst Yield Data, Example 4.9

Observation Number	Catalyst 1	Catalyst 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75
	$\bar{x}_1 = 92.255$	$\bar{x}_2 = 92.733$
	$s_1 = 2.39$	$s_2 = 2.98$

SOLUTION

The hypotheses are

$$H_0\colon \quad \mu_1 = \mu_2$$
$$H_1\colon \quad \mu_1 \neq \mu_2$$

From Table 4.2 we have $\bar{x}_1 = 92.255$, $s_1 = 2.39$, $n_1 = 8$, $\bar{x}_2 = 92.733$, $s_2 = 2.98$, and $n_2 = 8$. Therefore,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(7)(2.39)^2 + (7)(2.98)^2}{8 + 8 - 2} = 7.30$$

$$s_p = \sqrt{7.30} = 2.70$$

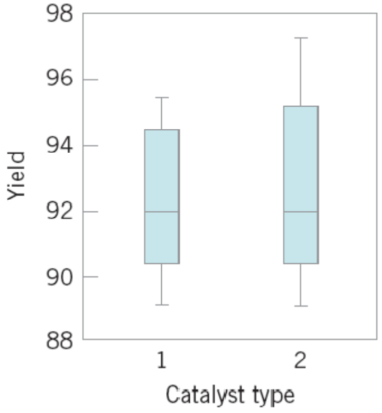
and

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{2.70\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{92.255 - 92.733}{2.70\sqrt{\frac{1}{8} + \frac{1}{8}}} = -0.35$$

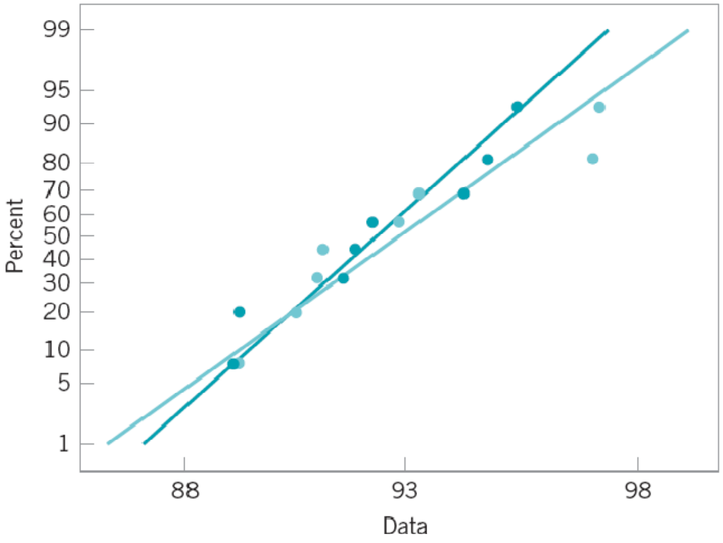
Because $t_{0.025,14} = -2.145$, and $-2.145 < -0.35 < 2.145$, the null hypothesis cannot be rejected. That is, at the 0.05 level of significance, we do not have strong evidence to conclude that catalyst 2 results in a mean yield that differs from the mean yield when catalyst 1 is used.

Figure 4.9 shows comparative box plots for the yield data for the two types of catalysts. These comparative box plots indicate that there is no obvious difference in the median of the two samples, although the second sample has a slightly larger sample dispersion or variance. There are no exact rules for comparing two samples with box plots; their primary value is in the visual impression they provide as a tool for explaining the results of a hypothesis test, as well as in verification of assumptions.

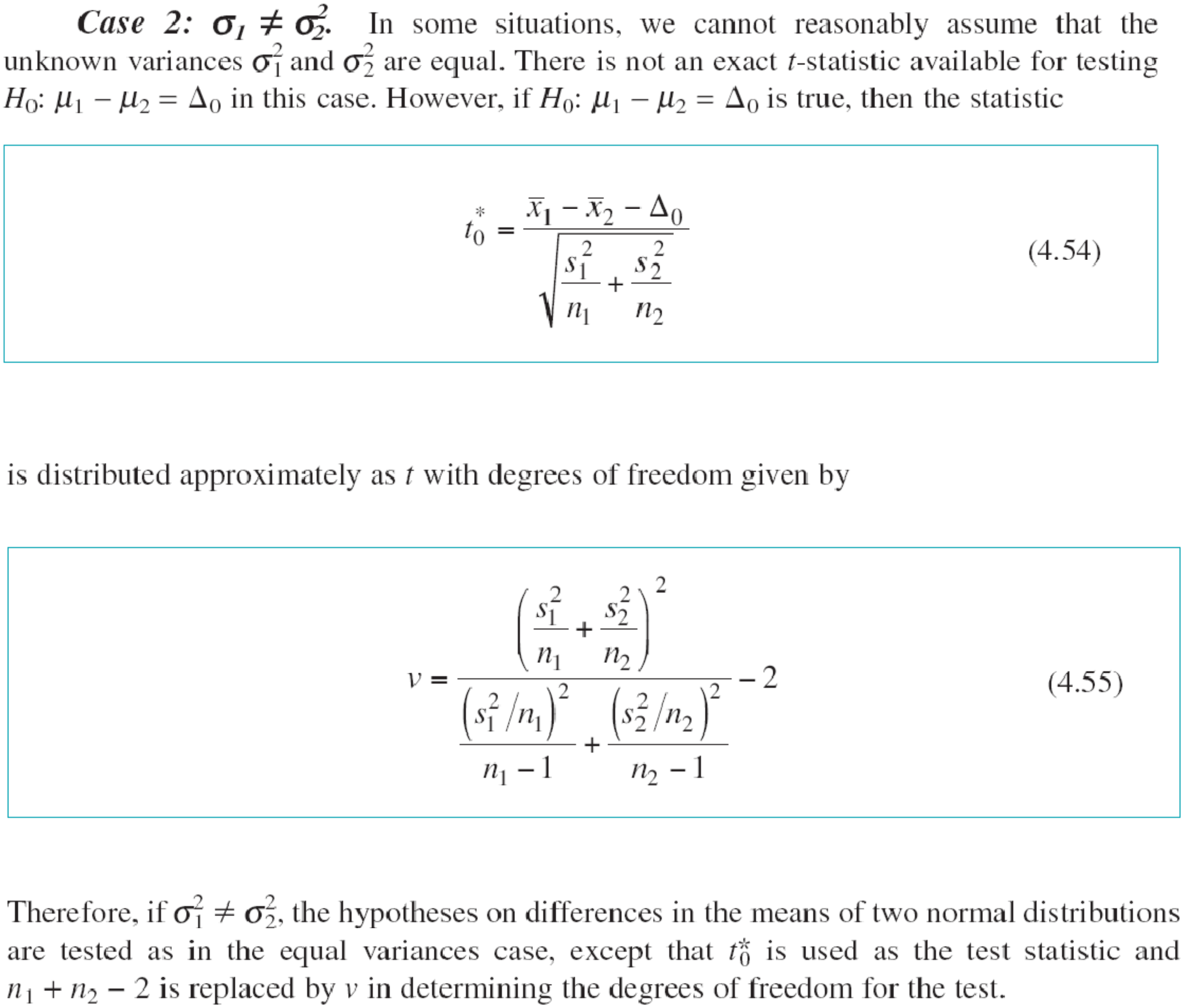
Figure 4.10 presents a Minitab normal probability plot of the two samples of yield data. Note that both samples plot approximately along straight lines, and the straight lines for each sample have similar slopes. (Recall that the slope of the line is proportional to the standard deviation.) Therefore, we conclude that the normality and equal variances assumptions are reasonable.



■ **FIGURE 4.9** Comparative box plots for the catalyst yield data.



■ **FIGURE 4.10** Minitab normal probability plot of the catalyst yield-data.



***Case 2:* $\sigma_1 \neq \sigma_2^2$.** In some situations, we cannot reasonably assume that the unknown variances σ_1^2 and σ_2^2 are equal. There is not an exact t-statistic available for testing H_0: $\mu_1 - \mu_2 = \Delta_0$ in this case. However, if H_0: $\mu_1 - \mu_2 = \Delta_0$ is true, then the statistic

$$t_0^* = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \qquad (4.54)$$

is distributed approximately as t with degrees of freedom given by

$$v = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(s_2^2/n_2\right)^2}{n_2 - 1}} - 2 \qquad (4.55)$$

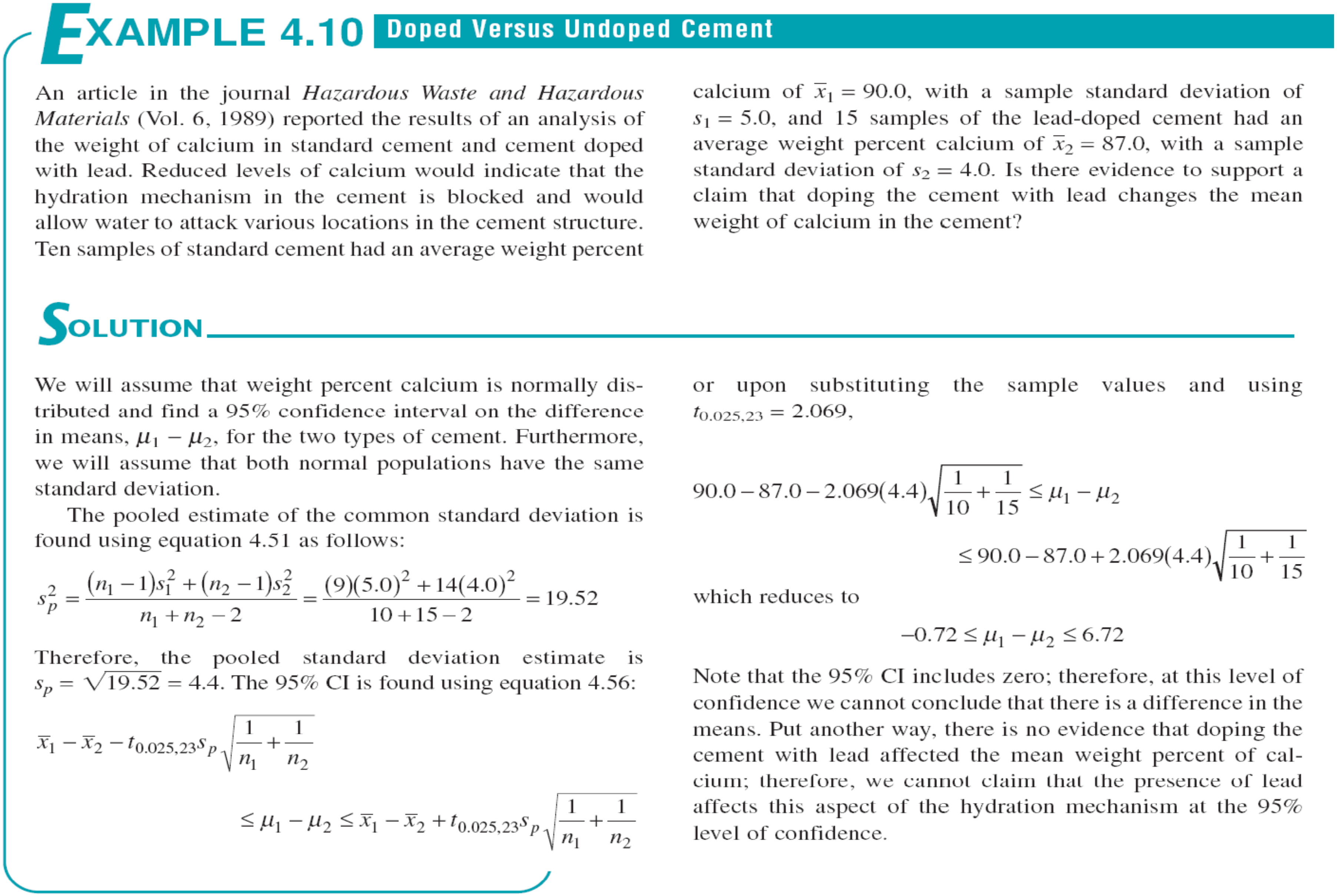
Therefore, if $\sigma_1^2 \neq \sigma_2^2$, the hypotheses on differences in the means of two normal distributions are tested as in the equal variances case, except that t_0^* is used as the test statistic and $n_1 + n_2 - 2$ is replaced by v in determining the degrees of freedom for the test.

Confidence Intervals – Case 1:

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,\, n_1+n_2-2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,\, n_1+n_2-2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (4.56)$$

Case 2:

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,\, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha/2, v} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (4.57)$$



EXAMPLE 4.10 Doped Versus Undoped Cement

An article in the journal *Hazardous Waste and Hazardous Materials* (Vol. 6, 1989) reported the results of an analysis of the weight of calcium in standard cement and cement doped with lead. Reduced levels of calcium would indicate that the hydration mechanism in the cement is blocked and would allow water to attack various locations in the cement structure. Ten samples of standard cement had an average weight percent calcium of $\bar{x}_1 = 90.0$, with a sample standard deviation of $s_1 = 5.0$, and 15 samples of the lead-doped cement had an average weight percent calcium of $\bar{x}_2 = 87.0$, with a sample standard deviation of $s_2 = 4.0$. Is there evidence to support a claim that doping the cement with lead changes the mean weight of calcium in the cement?

SOLUTION

We will assume that weight percent calcium is normally distributed and find a 95% confidence interval on the difference in means, $\mu_1 - \mu_2$, for the two types of cement. Furthermore, we will assume that both normal populations have the same standard deviation.

The pooled estimate of the common standard deviation is found using equation 4.51 as follows:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(9)(5.0)^2 + 14(4.0)^2}{10 + 15 - 2} = 19.52$$

Therefore, the pooled standard deviation estimate is $s_p = \sqrt{19.52} = 4.4$. The 95% CI is found using equation 4.56:

$$\bar{x}_1 - \bar{x}_2 - t_{0.025,23}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{0.025,23}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

or upon substituting the sample values and using $t_{0.025,23} = 2.069$,

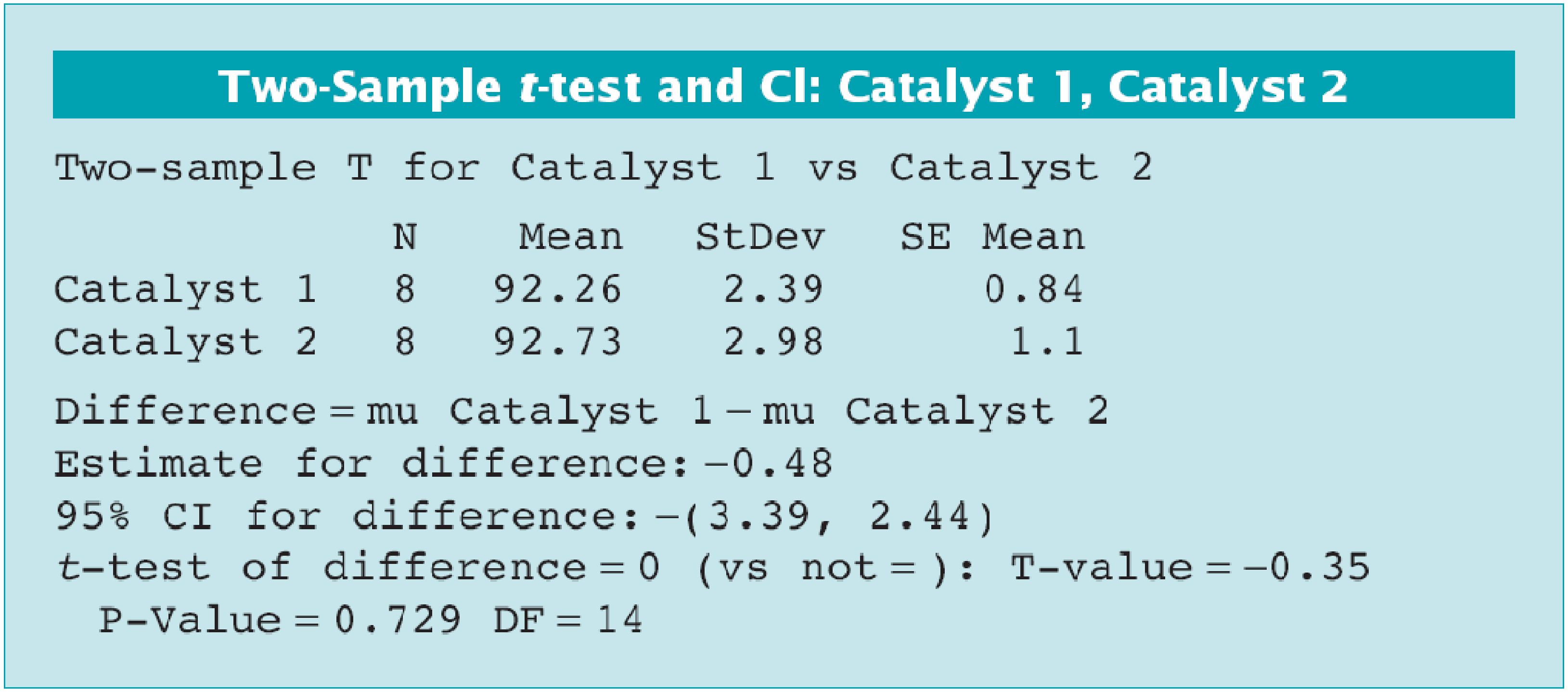
$$90.0 - 87.0 - 2.069(4.4)\sqrt{\frac{1}{10} + \frac{1}{15}} \le \mu_1 - \mu_2$$

$$\le 90.0 - 87.0 + 2.069(4.4)\sqrt{\frac{1}{10} + \frac{1}{15}}$$

which reduces to

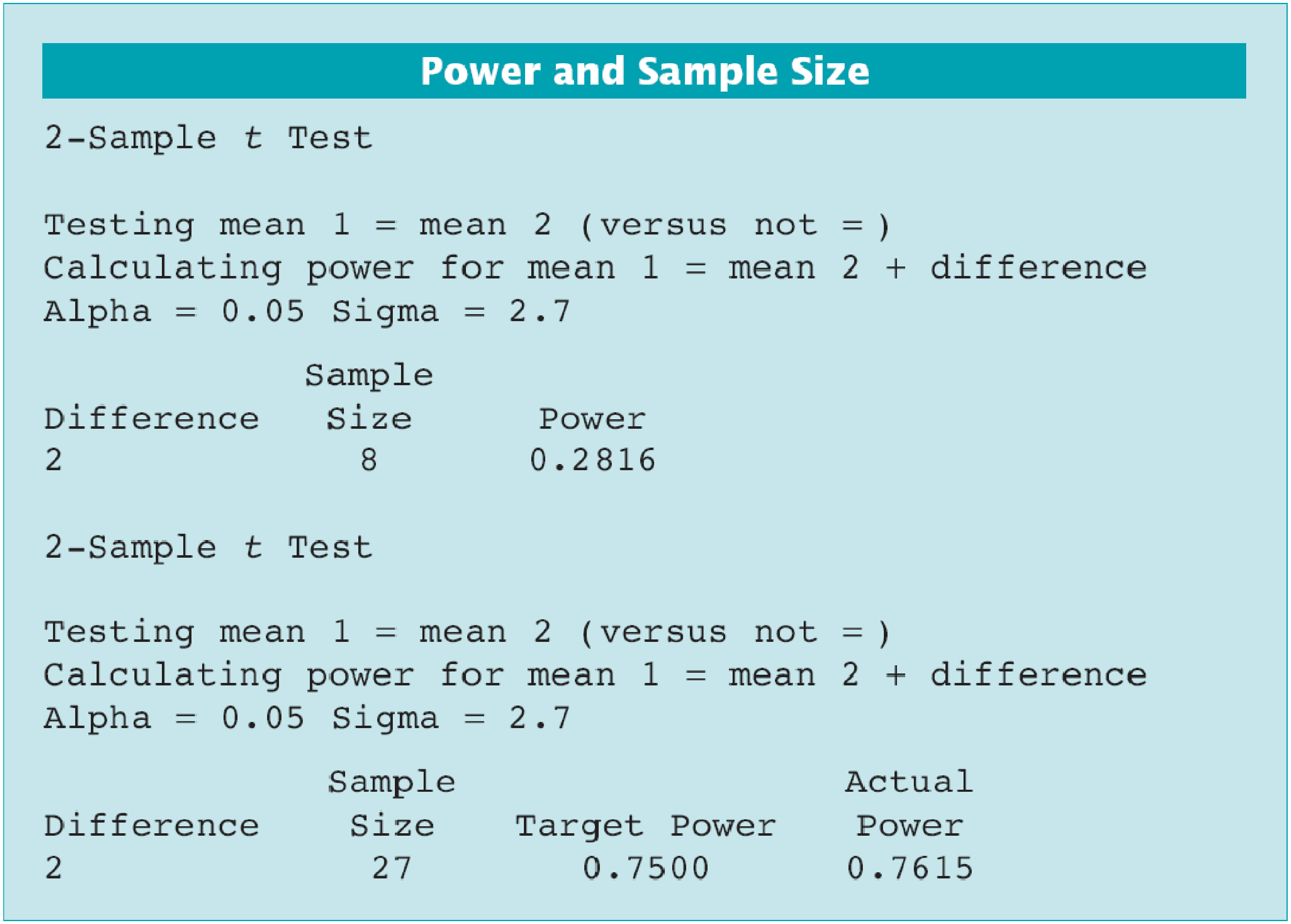
$$-0.72 \le \mu_1 - \mu_2 \le 6.72$$

Note that the 95% CI includes zero; therefore, at this level of confidence we cannot conclude that there is a difference in the means. Put another way, there is no evidence that doping the cement with lead affected the mean weight percent of calcium; therefore, we cannot claim that the presence of lead affects this aspect of the hydration mechanism at the 95% level of confidence.



Two-Sample *t*-test and CI: Catalyst 1, Catalyst 2

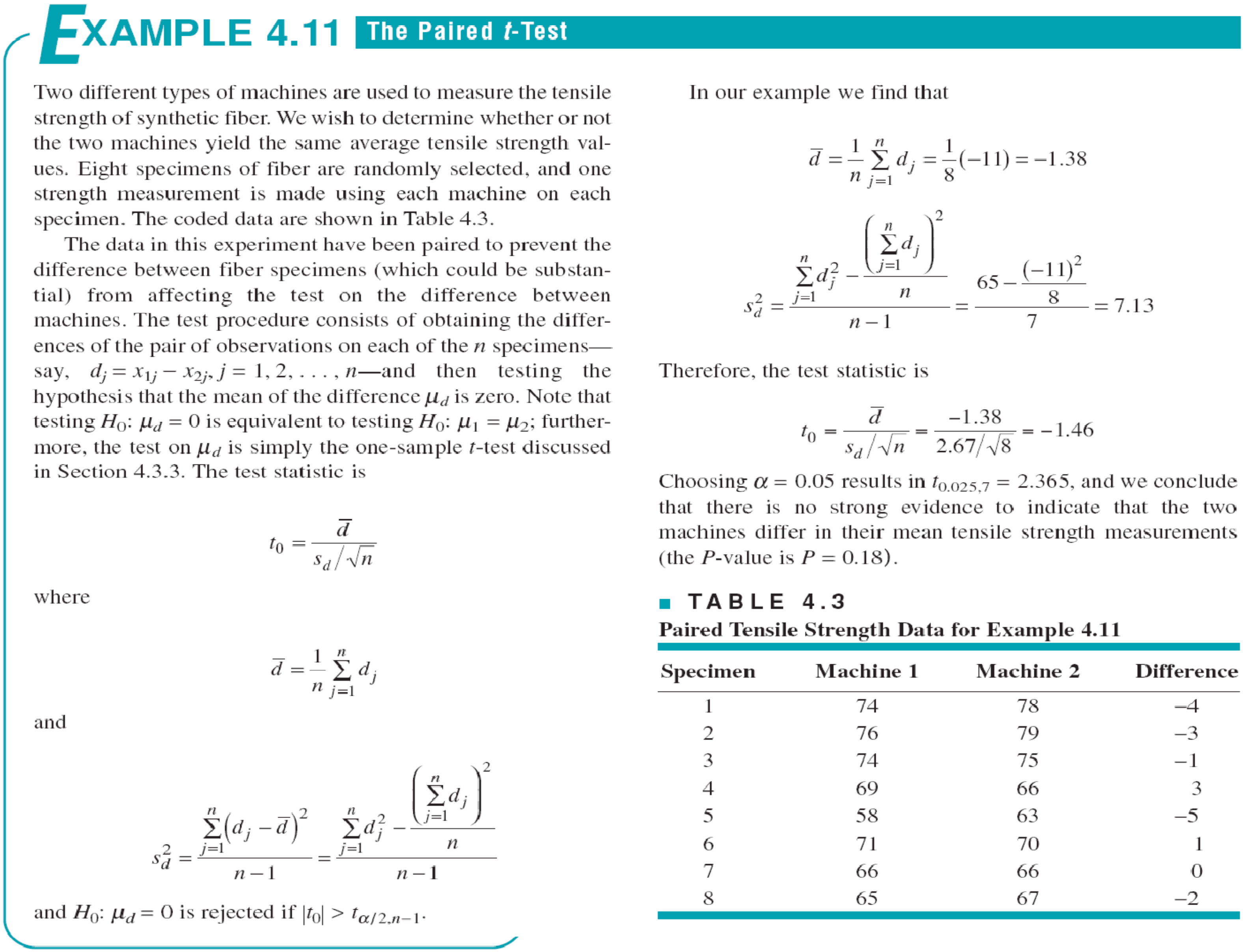
```
Two-sample T for Catalyst 1 vs Catalyst 2

             N    Mean   StDev   SE Mean
Catalyst 1   8   92.26    2.39      0.84
Catalyst 2   8   92.73    2.98       1.1

Difference = mu Catalyst 1 – mu Catalyst 2
Estimate for difference: –0.48
95% CI for difference: –(3.39, 2.44)
t-test of difference = 0 (vs not = ): T-value = –0.35
  P-Value = 0.729 DF = 14
```



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 = 2.7

                 Sample
Difference        Size        Power
2                  8         0.2816

2-Sample t Test

Testing mean 1 = mean 2 (versus not =)
Calculating power for mean 1 = mean 2 + difference
Alpha = 0.05 Sigma = 2.7

                   Sample                     Actual
Difference          Size    Target Power      Power
2                    27        0.7500        0.7615
```

Paired Data:



EXAMPLE 4.11 The Paired *t*-Test

Two different types of machines are used to measure the tensile strength of synthetic fiber. We wish to determine whether or not the two machines yield the same average tensile strength values. Eight specimens of fiber are randomly selected, and one strength measurement is made using each machine on each specimen. The coded data are shown in Table 4.3.

The data in this experiment have been paired to prevent the difference between fiber specimens (which could be substantial) from affecting the test on the difference between machines. The test procedure consists of obtaining the differences of the pair of observations on each of the n specimens—say, $d_j = x_{1j} - x_{2j}$, $j = 1, 2, \ldots, n$—and then testing the hypothesis that the mean of the difference μ_d is zero. Note that testing H_0: $\mu_d = 0$ is equivalent to testing H_0: $\mu_1 = \mu_2$; furthermore, the test on μ_d is simply the one-sample t-test discussed in Section 4.3.3. The test statistic is

$$t_0 = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where

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

and

$$s_d^2 = \frac{\sum_{j=1}^{n}\left(d_j - \bar{d}\right)^2}{n-1} = \frac{\sum_{j=1}^{n} d_j^2 - \frac{\left(\sum_{j=1}^{n} d_j\right)^2}{n}}{n-1}$$

and H_0: $\mu_d = 0$ is rejected if $|t_0| > t_{\alpha/2,n-1}$.

In our example we find that

$$\bar{d} = \frac{1}{n}\sum_{j=1}^{n} d_j = \frac{1}{8}(-11) = -1.38$$

$$s_d^2 = \frac{\sum_{j=1}^{n} d_j^2 - \frac{\left(\sum_{j=1}^{n} d_j\right)^2}{n}}{n-1} = \frac{65 - \frac{(-11)^2}{8}}{7} = 7.13$$

Therefore, the test statistic is

$$t_0 = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{-1.38}{2.67/\sqrt{8}} = -1.46$$

Choosing $\alpha = 0.05$ results in $t_{0.025,7} = 2.365$, and we conclude that there is no strong evidence to indicate that the two machines differ in their mean tensile strength measurements (the P-value is $P = 0.18$).

TABLE 4.3

Paired Tensile Strength Data for Example 4.11

Specimen	Machine 1	Machine 2	Difference
1	74	78	–4
2	76	79	–3
3	74	75	–1
4	69	66	3
5	58	63	–5
6	71	70	1
7	66	66	0
8	65	67	–2

4.4.3 Inference on the Variances of Two Normal Distributions

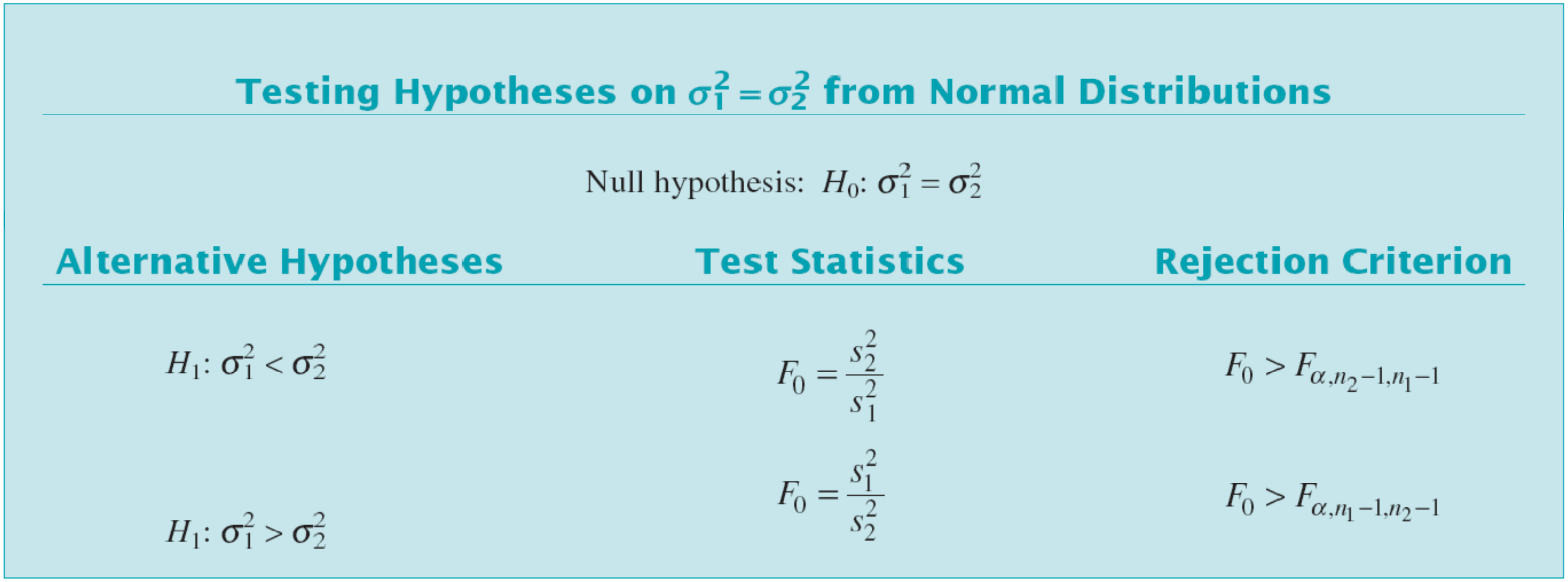
Hypothesis Testing. Consider testing the hypothesis that the variances of two independent normal distributions are equal. If random samples of sizes n_1 and n_2 are taken from populations 1 and 2, respectively, then the test statistic for

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

is simply the ratio of the two sample variances,

$$F_0 = \frac{s_1^2}{s_2^2} \qquad (4.58)$$

We would reject H_0 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, respectively. The following display summarizes the test procedures for the one-sided alternative hypotheses.



Testing Hypotheses on $\sigma_1^2 = \sigma_2^2$ from Normal Distributions

Null hypothesis: $H_0\colon \sigma_1^2 = \sigma_2^2$

Alternative Hypotheses	Test Statistics	Rejection Criterion
$H_1\colon \sigma_1^2 < \sigma_2^2$	$F_0 = \dfrac{s_2^2}{s_1^2}$	$F_0 > F_{\alpha, n_2-1, n_1-1}$
$H_1\colon \sigma_1^2 > \sigma_2^2$	$F_0 = \dfrac{s_1^2}{s_2^2}$	$F_0 > F_{\alpha, n_1-1, n_2-1}$

The two-sided CI is:

$$\frac{s_1^2}{s_2^2} F_{1-\alpha/2, n_2-1, n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} F_{\alpha/2, n_2-1, n_1-1} \qquad (4.59)$$

where $F_{\alpha/2,u,v}$ is the percentage point of the F distribution with u and v degrees of freedom such that $P\{F_{u,v} \ge F_{\alpha/2,u,v}\} = \alpha/2$. The corresponding upper and lower confidence bounds are

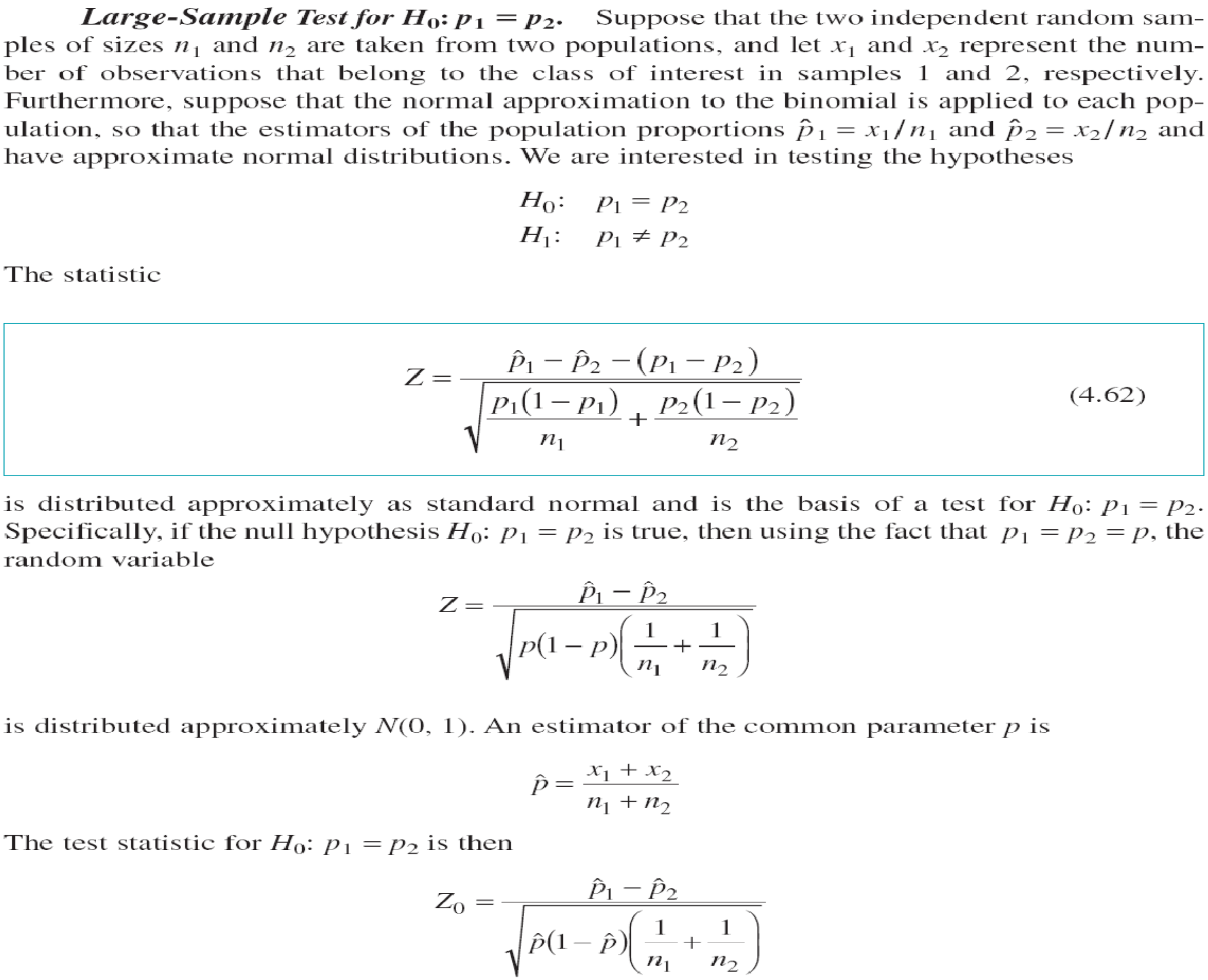
$$\frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} F_{\alpha, n_2-1, n_1-1} \qquad (4.60)$$

and

$$\frac{s_1^2}{s_2^2} F_{1-\alpha, n_2-1, n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \qquad (4.61)$$

respectively.[2]

4.4.4 Inference on Two Proportions



Large-Sample Test for H_0: $p_1 = p_2$. Suppose that the two independent random samples of sizes n_1 and n_2 are taken from two populations, and let x_1 and x_2 represent the number of observations that belong to the class of interest in samples 1 and 2, respectively. Furthermore, suppose that the normal approximation to the binomial is applied to each population, so that the estimators of the population proportions $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$ and have approximate normal distributions. We are interested in testing the hypotheses

$$\begin{aligned} H_0&: \quad p_1 = p_2 \\ H_1&: \quad p_1 \neq p_2 \end{aligned}$$

The statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \qquad (4.62)$$

is distributed approximately as standard normal and is the basis of a test for H_0: $p_1 = p_2$. Specifically, if the null hypothesis H_0: $p_1 = p_2$ is true, then using the fact that $p_1 = p_2 = p$, the random variable

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

is distributed approximately $N(0, 1)$. An estimator of the common parameter p is

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

The test statistic for H_0: $p_1 = p_2$ is then

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

Testing Hypothesis on Two Population Proportions

Null hypothesis: $H_0\colon p_1 = p_2$

Test statistic:

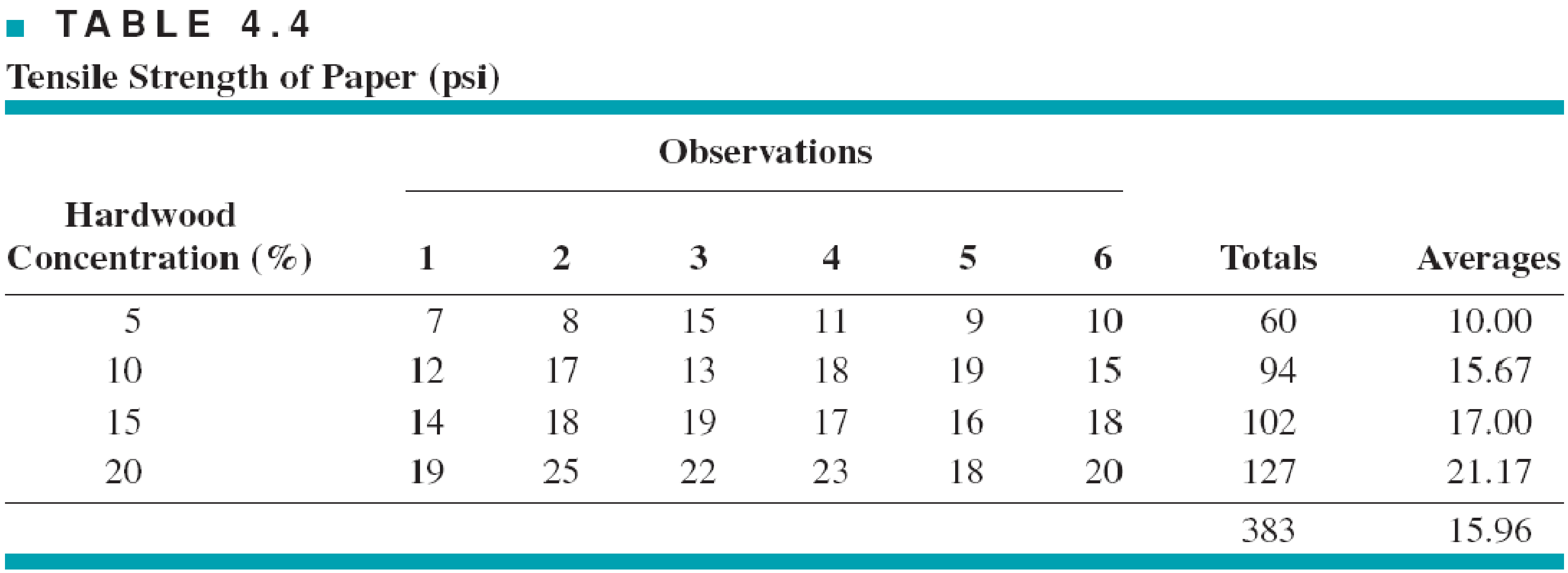
$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{4.63}$$

Alternative Hypotheses	Fixed Significance Level Rejection Criterion	*P*-value
$H_1\colon p_1 \neq p_2$	$Z_0 > Z_{\alpha/2}$ or $Z_0 < -Z_{\alpha/2}$	$P = 2[1 - \Phi(\lvert Z_0\rvert)]$
$H_1\colon p_1 > p_2$	$Z_0 > Z_\alpha$	$P = 1 - \Phi(Z_0)$
$H_1\colon p_1 < p_2$	$Z_0 < -Z_\alpha$	$P = \Phi(Z_0)$

$$\hat{p}_1 - \hat{p}_2 - Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{4.64}$$

4.5 What if There Are More Than Two Populations? The Analysis of Variance

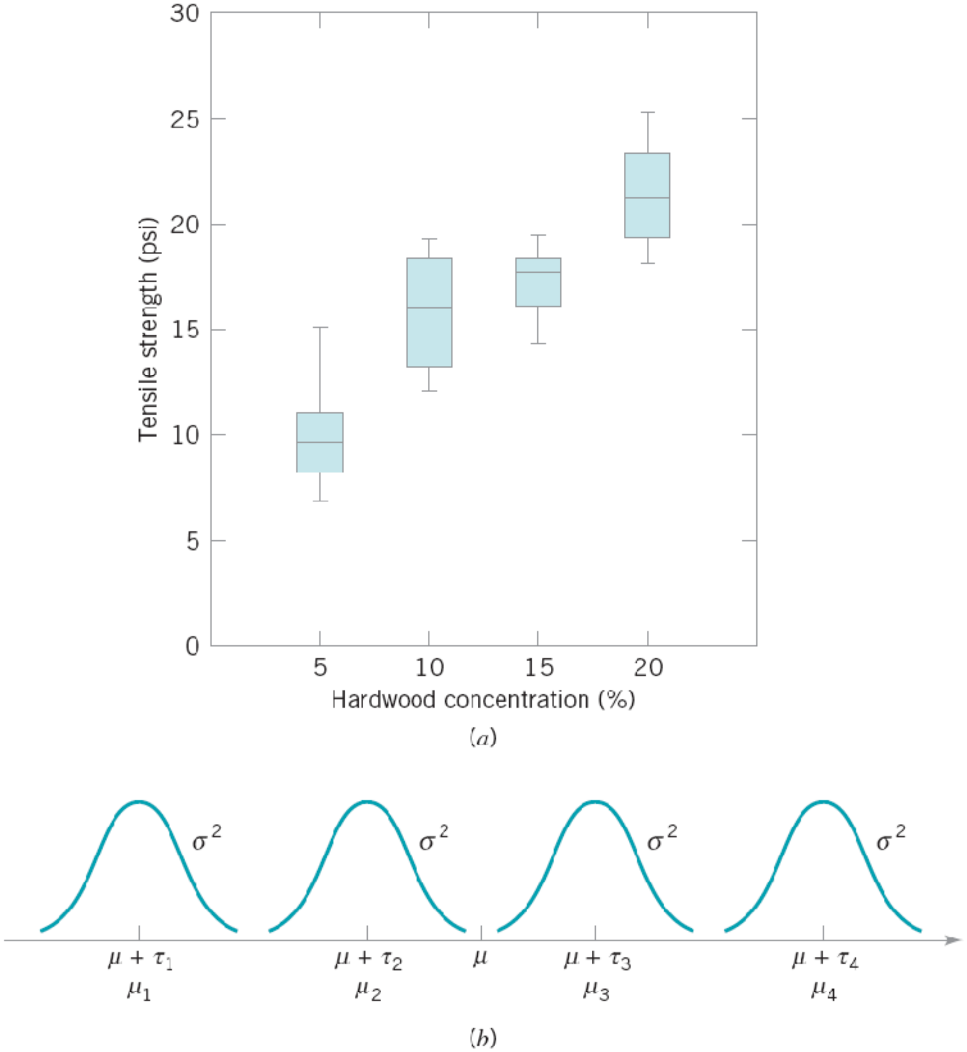
Example: Does changing the hardwood concentration in the pulp affect the mean tensile strength of paper?



■ **TABLE 4.4**

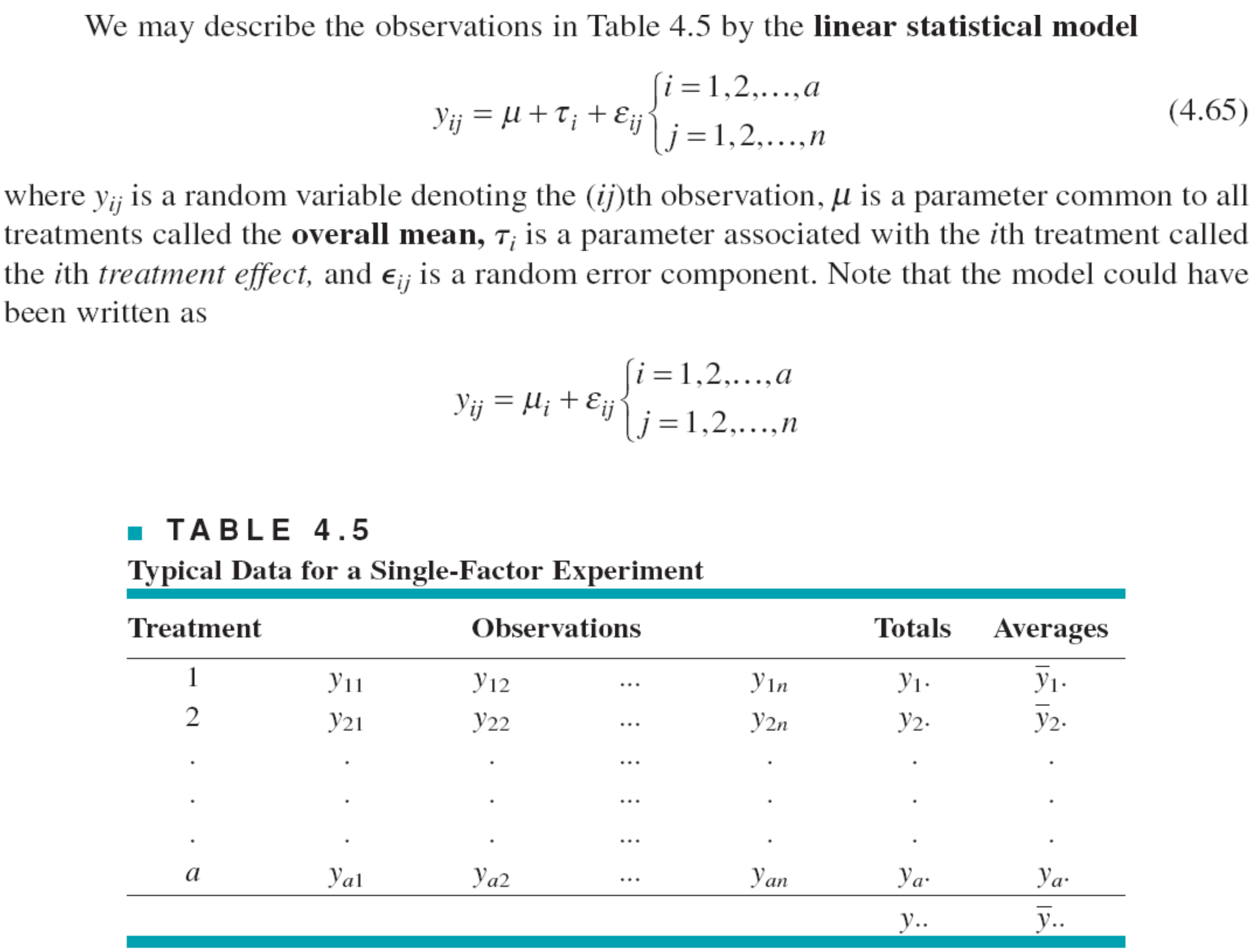
Tensile Strength of Paper (psi)

	Observations							
Hardwood Concentration (%)	**1**	**2**	**3**	**4**	**5**	**6**	**Totals**	**Averages**
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96



■ **FIGURE 4.11** (*a*) Box plots of hardwood concentration data. (*b*) Display of the model in equation 4.65 for the completely randomized single-factor experiment.

The Analysis of Variance (ANOVA)



We may describe the observations in Table 4.5 by the **linear statistical model**

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,n \end{cases} \tag{4.65}$$

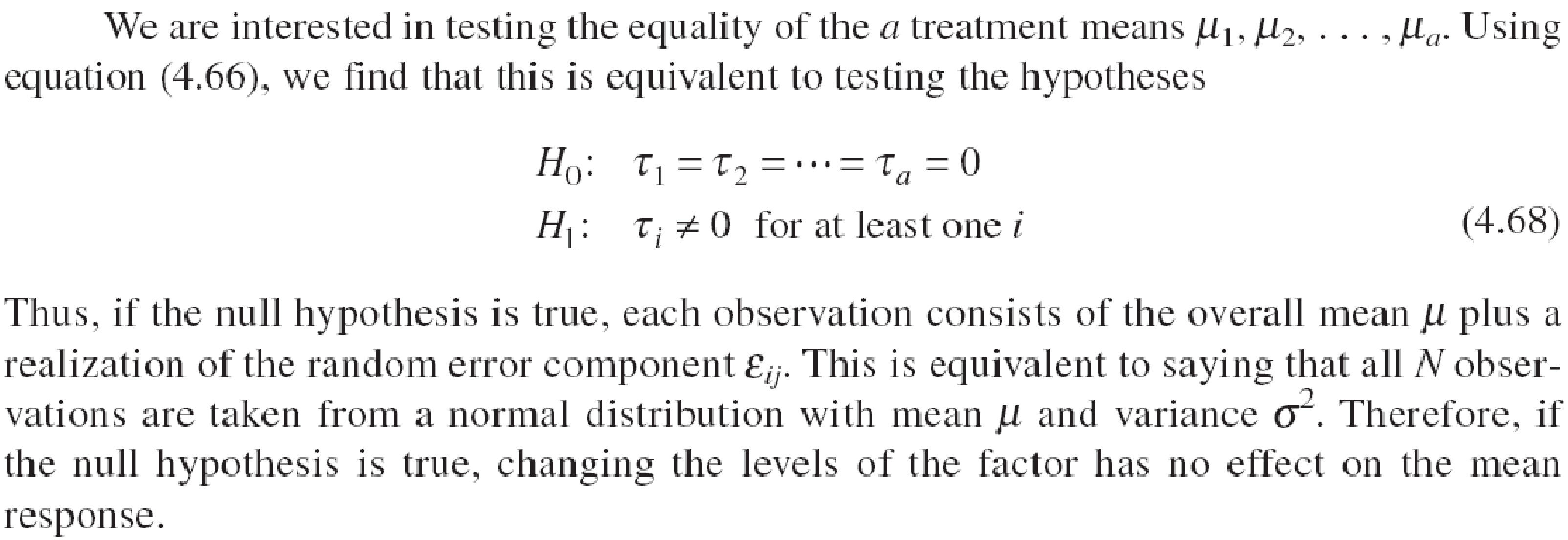
where y_{ij} is a random variable denoting the (ij)th observation, μ is a parameter common to all treatments called the **overall mean,** τ_i is a parameter associated with the ith treatment called the ith *treatment effect,* and ϵ_{ij} is a random error component. Note that the model could have been written as

$$y_{ij} = \mu_i + \varepsilon_{ij} \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,n \end{cases}$$

■ TABLE 4.5

Typical Data for a Single-Factor Experiment

Treatment	Observations				Totals	Averages
1	y_{11}	y_{12}	...	y_{1n}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	...	y_{2n}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
.	.	.	...	.	.	.
.	.	.	...	.	.	.
.	.	.	...	.	.	.
a	y_{a1}	y_{a2}	...	y_{an}	$y_{a\cdot}$	$y_{a\cdot}$
					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

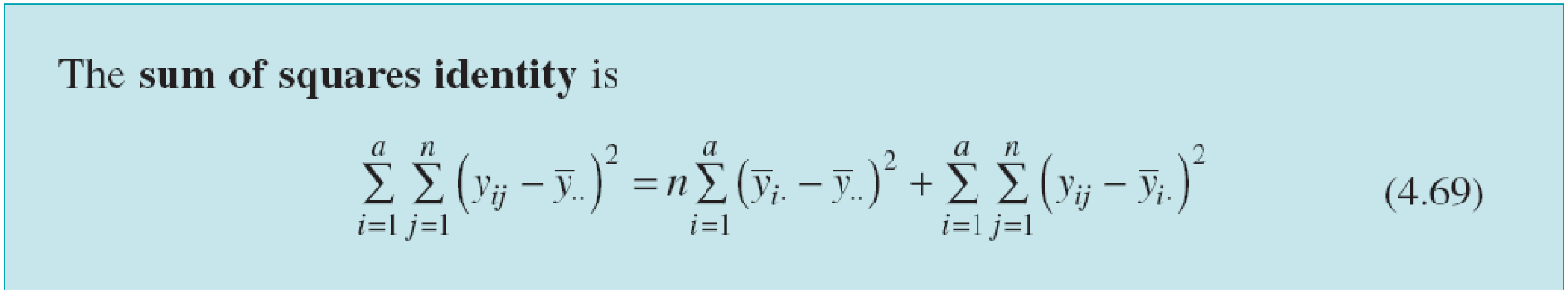


We are interested in testing the equality of the a treatment means $\mu_1, \mu_2, \ldots, \mu_a$. Using equation (4.66), we find that this is equivalent to testing the hypotheses

$$\begin{aligned} H_0&: \quad \tau_1 = \tau_2 = \cdots = \tau_a = 0 \\ H_1&: \quad \tau_i \neq 0 \ \text{ for at least one } i \end{aligned} \tag{4.68}$$

Thus, if the null hypothesis is true, each observation consists of the overall mean μ plus a realization of the random error component ε_{ij}. This is equivalent to saying that all N observations are taken from a normal distribution with mean μ and variance σ^2. Therefore, if the null hypothesis is true, changing the levels of the factor has no effect on the mean response.

The ANOVA is based on the following partitioning of the total sum of squares (which measures the total variability in the sample data):



The **sum of squares identity** is

$$\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2 = n\sum_{i=1}^{a}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2 \tag{4.69}$$

$$SS_T = SS_{\text{Treatments}} + SS_E \tag{4.71}$$

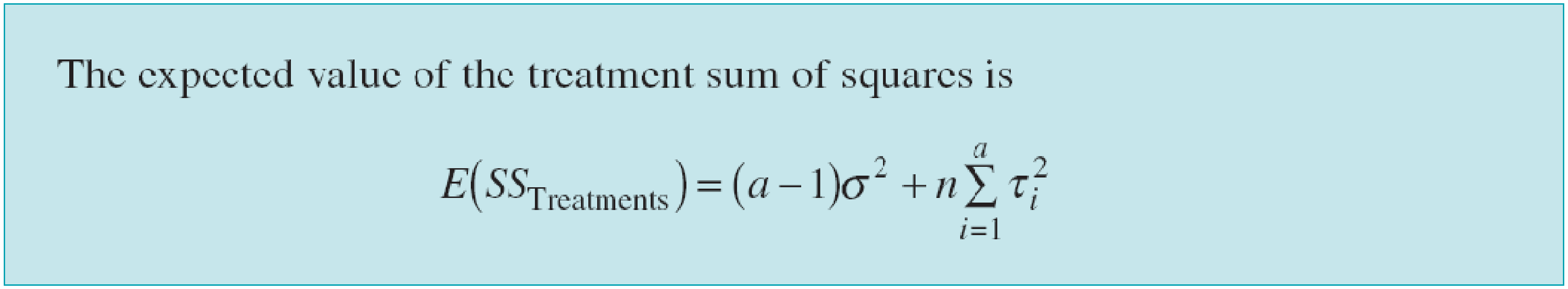
where

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_{..}\right)^2 = \text{total sum of squares}$$

$$SS_{\text{Treatments}} = n\sum_{i=1}^{a}\left(\overline{y}_{i.} - \overline{y}_{..}\right)^2 = \text{treatment sum of squares}$$

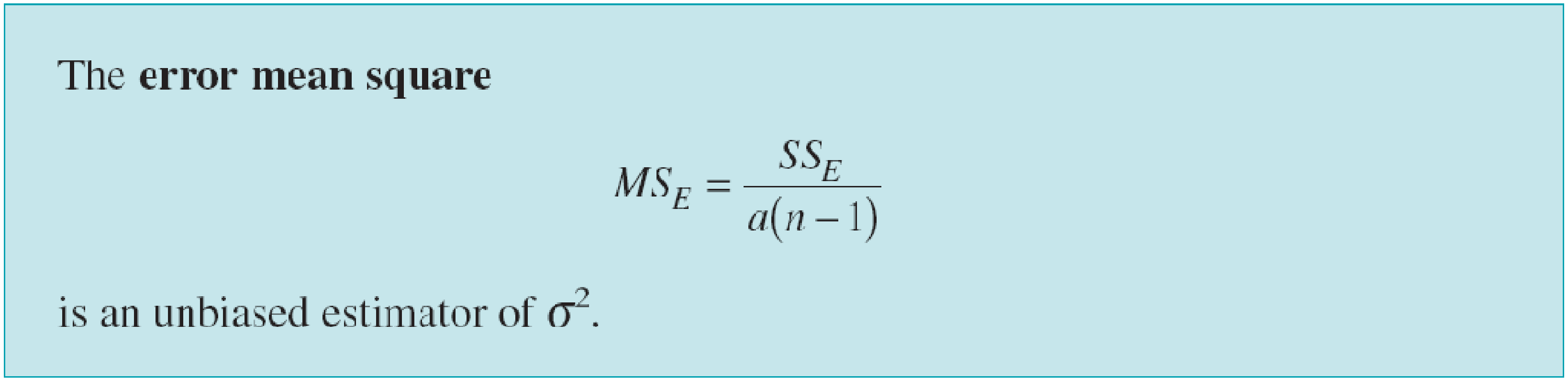
and

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_{j.}\right)^2 = \text{error sum of squares}$$



The expected value of the treatment sum of squares is

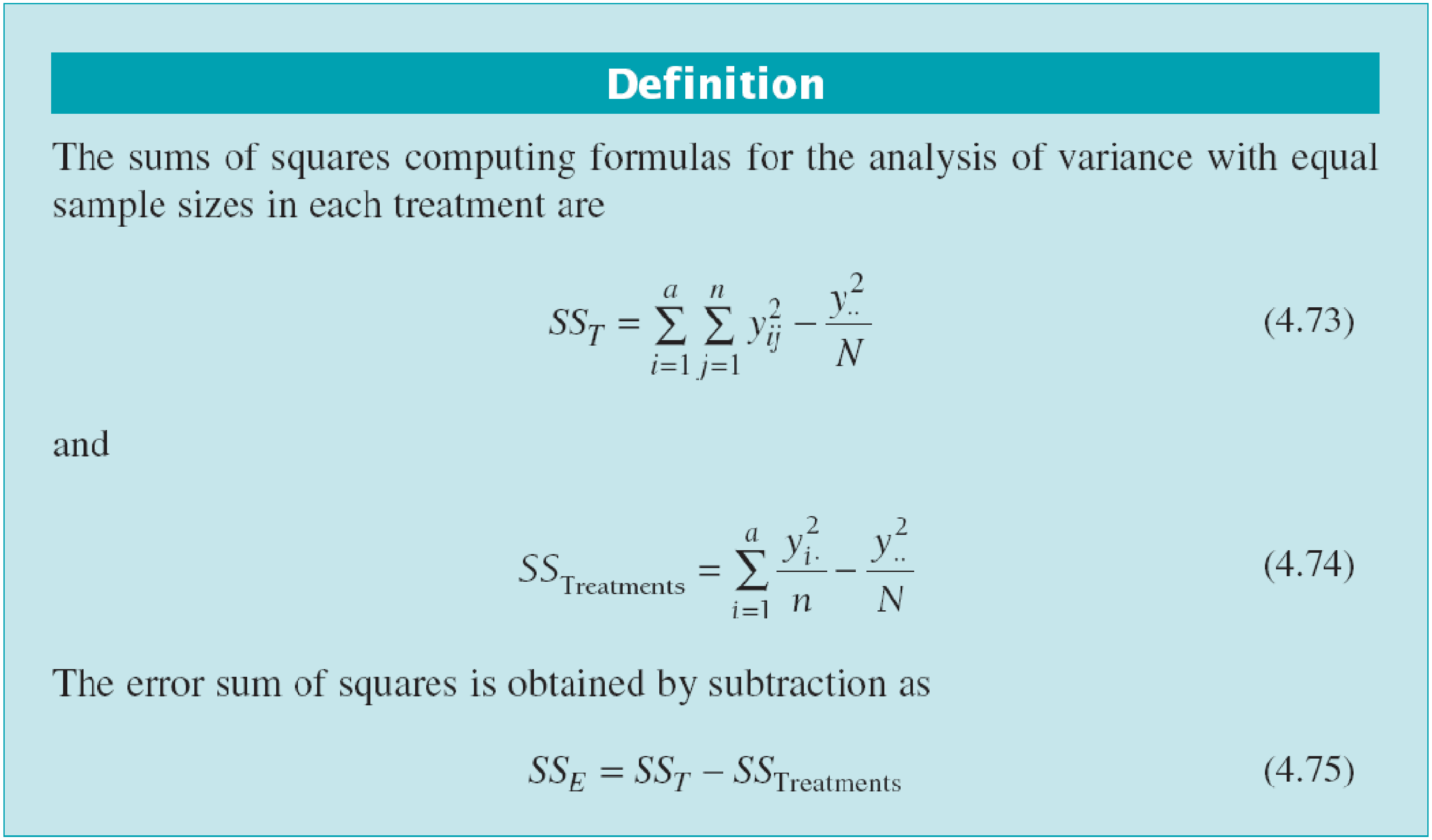
$$E\left(SS_{\text{Treatments}}\right) = (a-1)\sigma^2 + n\sum_{i=1}^{a}\tau_i^2$$



The **error mean square**

$$MS_E = \frac{SS_E}{a(n-1)}$$

is an unbiased estimator of σ^2.



Definition

The sums of squares computing formulas for the analysis of variance with equal sample sizes in each treatment are

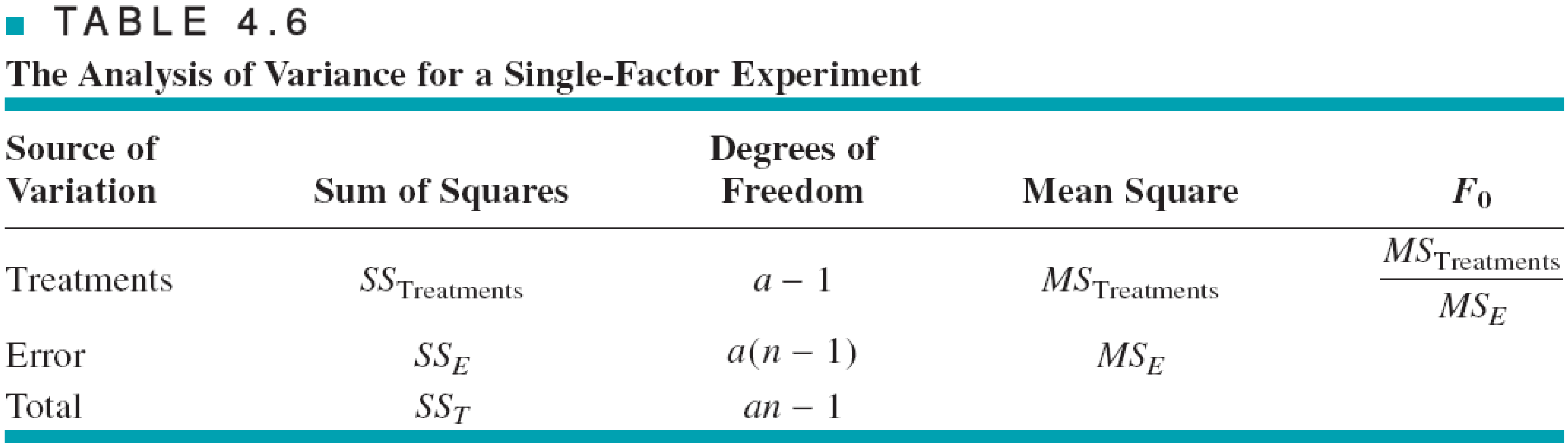
$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{N} \qquad (4.73)$$

and

$$SS_{\text{Treatments}} = \sum_{i=1}^{a} \frac{y_{i.}^2}{n} - \frac{y_{..}^2}{N} \qquad (4.74)$$

The error sum of squares is obtained by subtraction as

$$SS_E = SS_T - SS_{\text{Treatments}} \qquad (4.75)$$



■ TABLE 4.6

The Analysis of Variance for a Single-Factor Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}}$	$a-1$	$MS_{\text{Treatments}}$	$\dfrac{MS_{\text{Treatments}}}{MS_E}$
Error	SS_E	$a(n-1)$	MS_E	
Total	SS_T	$an-1$		

The ANOVA test statistic is:

$$F_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/[a(n-1)]} = \frac{MS_{\text{Treatments}}}{MS_E}$$

If F_0 is greater than the critical value $F_{\alpha, a\text{-}1,\ a(n\text{-}1)}$ then the null hypothesis of equal treatment means is rejected. A *P*-value approach can also be used. The *P*-value would be the probability above F_0 in the $F_{a\text{-}1,\ a(n\text{-}1)}$ distribution.

EXAMPLE 4.12 The Paper Tensile Strength Experiment

Consider the paper tensile strength experiment described in Section 4.5.1. Use the analysis of variance to test the hypothesis that different hardwood concentrations do not affect the mean tensile strength of the paper.

SOLUTION

The hypotheses are

$$H_0: \quad \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$$
$$H_1: \quad \tau_i \neq 0 \text{ for at least one } i$$

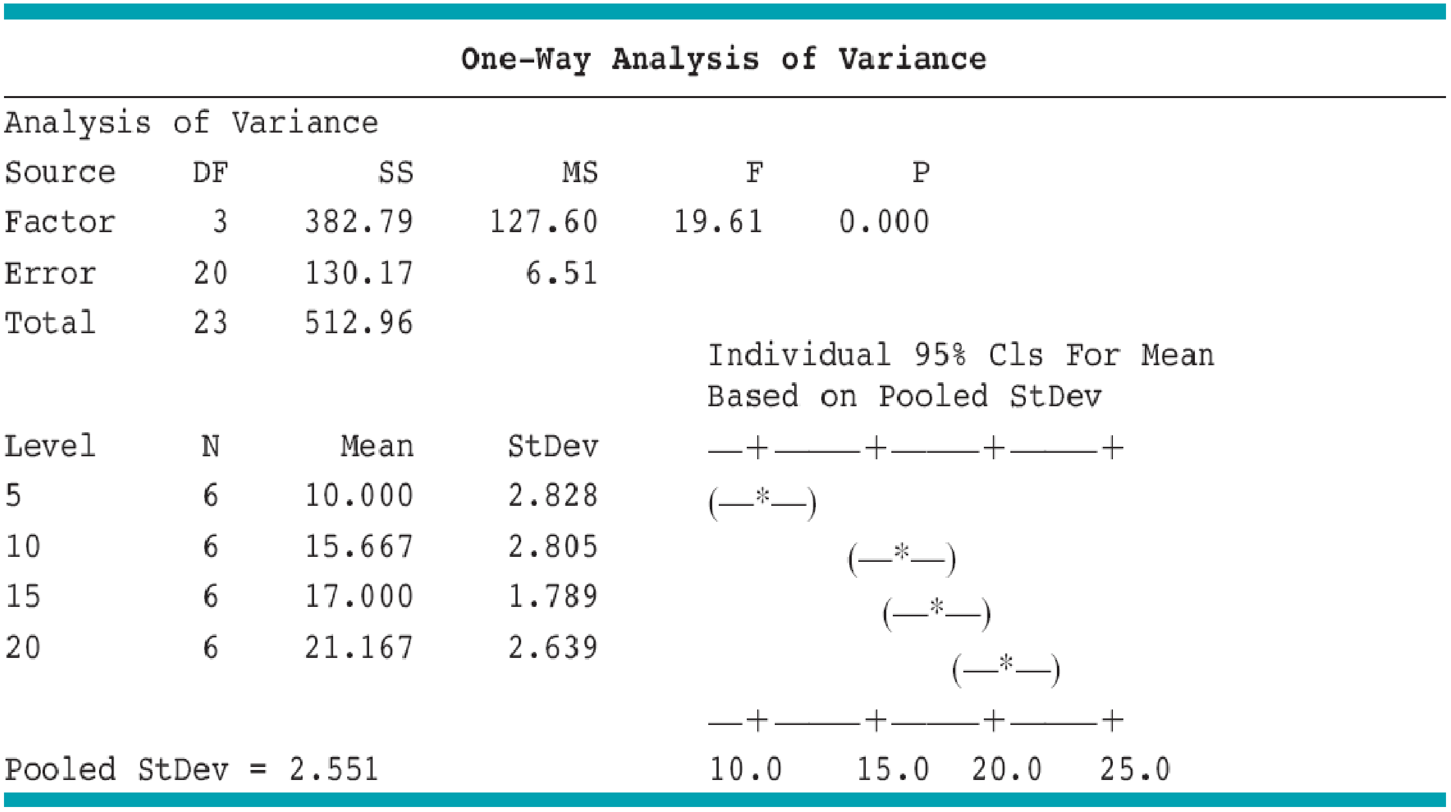
We will use $\alpha = 0.01$. The sums of squares for the ANOVA are computed from equations 4.73, 4.74, and 4.75 as follows:

$$\begin{aligned} SS_T &= \sum_{i=1}^{4}\sum_{j=1}^{6} y_{ij}^2 - \frac{y_{..}^2}{N} \\ &= (7)^2 + (8)^2 + \cdots + (20)^2 - \frac{(383)^2}{24} = 512.96 \\ SS_{\text{Treatments}} &= \sum_{i=1}^{4} \frac{y_{i.}^2}{n} - \frac{y_{..}^2}{N} \\ &= \frac{(60)^2 + (94)^2 + (102)^2 + (127)^2}{6} - \frac{(383)^2}{24} = 382.79 \\ SS_E &= SS_T - SS_{\text{Treatments}} \\ &= 512.96 - 382.79 = 130.17 \end{aligned}$$

We usually do not perform these calculations by hand. The ANOVA from Minitab is presented in Table 4.7. Since $F_{0.01,3,20} = 4.94$, we reject H_0 and conclude that hardwood concentration in the pulp significantly affects the strength of the paper. Note that the computer output reports a P-value for the test statistic $F = 19.61$ in Table 4.7 of zero. This is a truncated value; the actual P-value is $P = 3.59 \times 10^{-6}$. However, since the P-value is considerably smaller than $\alpha = 0.01$, we have strong evidence to conclude that H_0 is not true. Note that Minitab also provides some summary information about each level of hardwood concentration, including the confidence interval on each mean.

TABLE 4.7

Minitab Analysis of Variance Output for the Paper Tensile Strength Experiment



```
                      One-Way Analysis of Variance

Analysis of Variance
Source    DF        SS        MS        F        P
Factor     3    382.79    127.60    19.61    0.000
Error     20    130.17      6.51
Total     23    512.96
                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level      N      Mean     StDev    --+-------+-------+-------+
5          6    10.000     2.828    (--*--)
10         6    15.667     2.805           (--*--)
15         6    17.000     1.789             (--*--)
20         6    21.167     2.639                 (--*--)
                                    --+-------+-------+-------+
Pooled StDev = 2.551                10.0    15.0    20.0    25.0
```

Graphical comparison of individual means

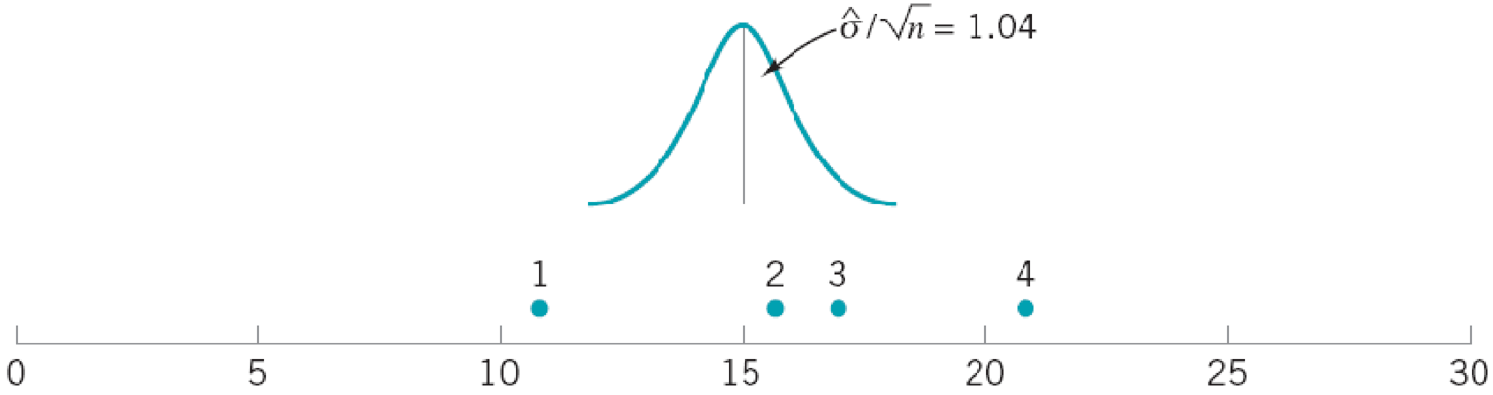


FIGURE 4.12 Tensile strength averages from the hardwood concentration experiment in relation to a normal distribution with standard deviation $\sqrt{MS_E/n} = \sqrt{6.51/6} = 1.04$

20% hardwood produces higher mean strength than the others; 5% hardwood produces lower strength; 10% and 15% hardwood don’t differ but give lower strength than 20%.

4.5.3 Checking Assumptions: Residual Analysis

■ **TABLE 4.8**

Residuals for the Hardwood Experiment

Hardwood Concentration	Residuals					
5%	−3.00	−2.00	5.00	1.00	−1.00	0.00
10%	−3.67	1.33	−2.67	2.33	+3.33	−0.67
15%	−3.00	1.00	2.00	0.00	−1.00	1.00
20%	−2.17	3.83	0.83	1.83	−3.17	−1.17

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i.}$$

Residual Plots:

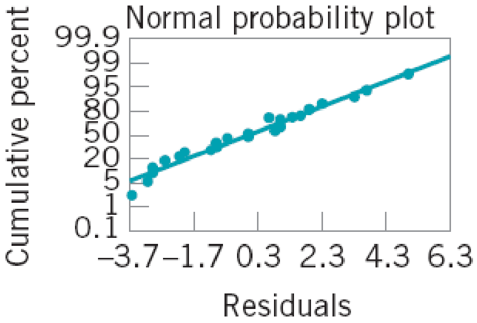


FIGURE 4.13 Normal probability plot of residuals from the hardwood concentration experiment.

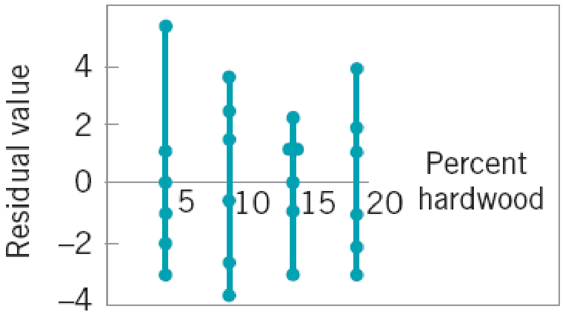


FIGURE 4.14 Plot of residuals versus factor levels.

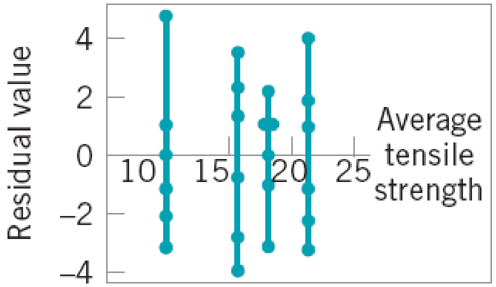
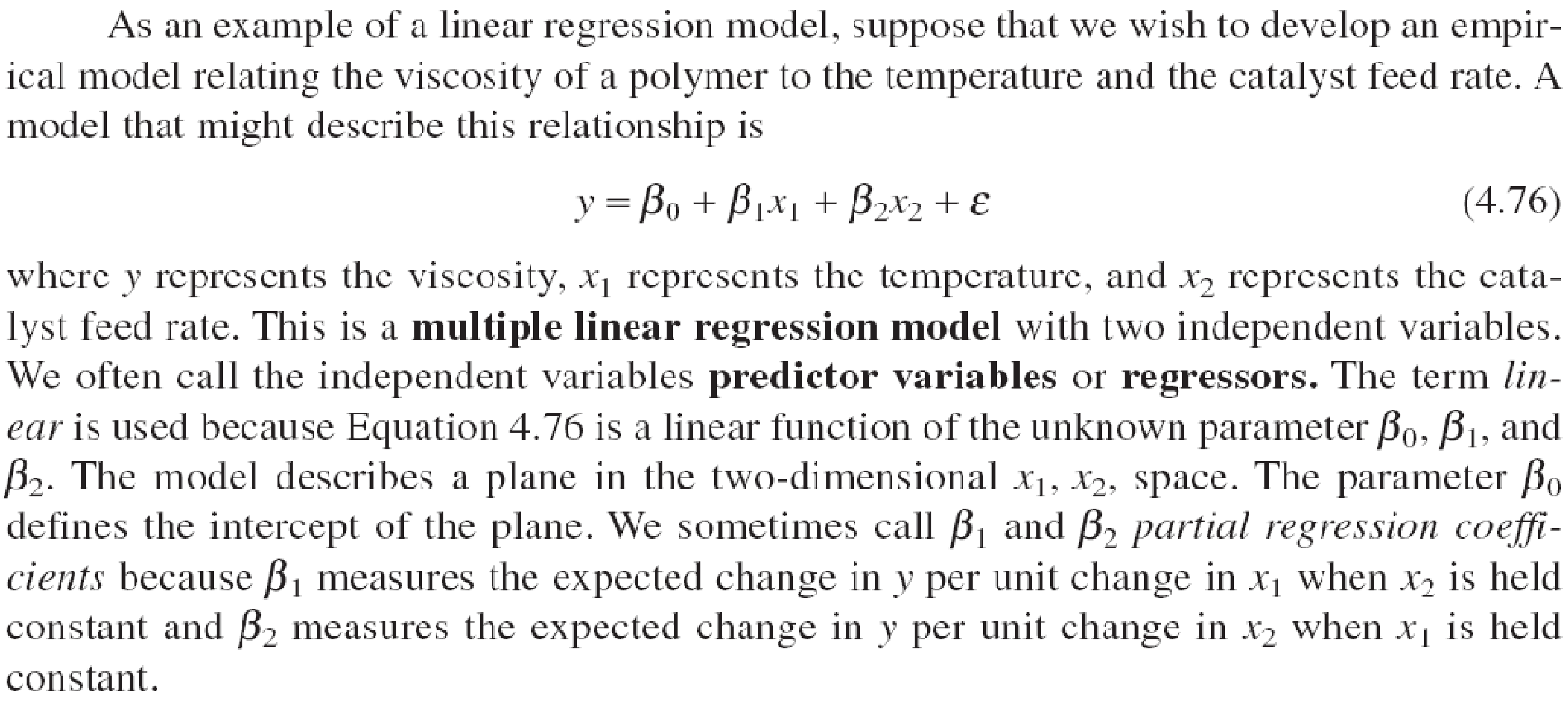


FIGURE 4.15 Plot of residuals verus $\overline{y}_{i.}$

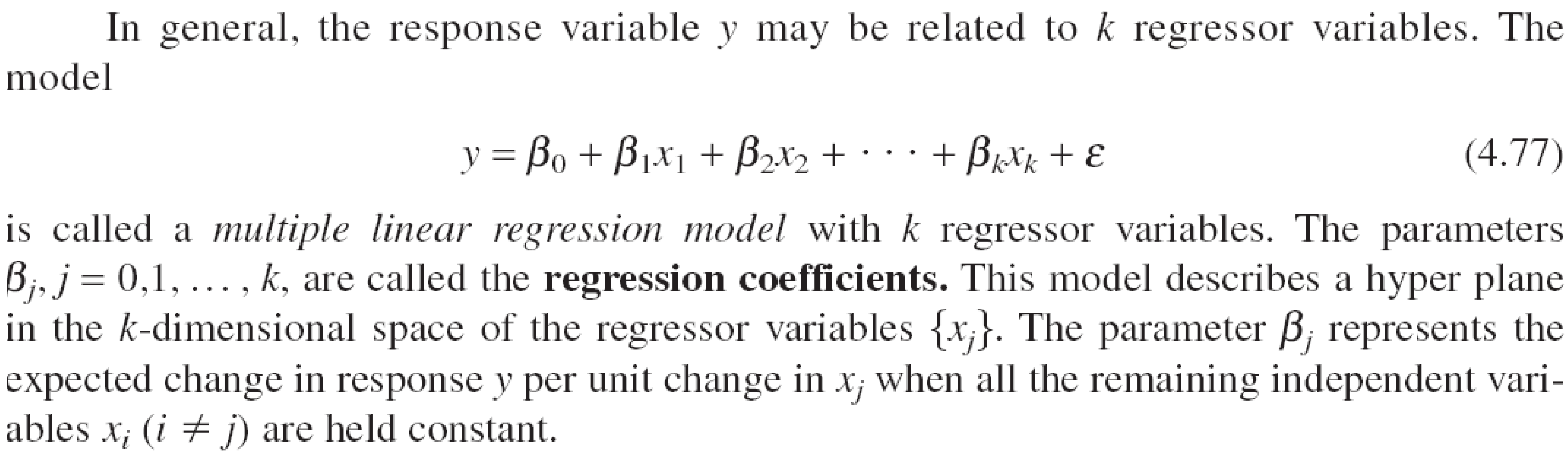
4.6 Linear Regression Models



As an example of a linear regression model, suppose that we wish to develop an empirical model relating the viscosity of a polymer to the temperature and the catalyst feed rate. A model that might describe this relationship is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{4.76}$$

where y represents the viscosity, x_1 represents the temperature, and x_2 represents the catalyst feed rate. This is a **multiple linear regression model** with two independent variables. We often call the independent variables **predictor variables** or **regressors.** The term *linear* is used because Equation 4.76 is a linear function of the unknown parameter β_0, β_1, and β_2. The model describes a plane in the two-dimensional x_1, x_2, space. The parameter β_0 defines the intercept of the plane. We sometimes call β_1 and β_2 *partial regression coefficients* because β_1 measures the expected change in y per unit change in x_1 when x_2 is held constant and β_2 measures the expected change in y per unit change in x_2 when x_1 is held constant.



In general, the response variable y may be related to k regressor variables. The model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \qquad (4.77)$$

is called a *multiple linear regression model* with k regressor variables. The parameters β_j, $j = 0,1, \ldots , k$, are called the **regression coefficients.** This model describes a hyper plane in the k-dimensional space of the regressor variables $\{x_j\}$. The parameter β_j represents the expected change in response y per unit change in x_j when all the remaining independent variables x_i $(i \neq j)$ are held constant.

Models that are not linear in the regressors can still be fit using linear regression techniques, so long as they are linear in the parameters.

Important cases include models with interaction terms and polynomials.

We may write the model equation [equation (4.77)] in terms of the observations in Table 4.9 as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \quad i = 1, 2, \ldots, n \end{aligned} \tag{4.82}$$

■ TABLE 4.9

Data for Multiple Linear Regression

y	x_1	x_2	...	x_k
y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

The method of least squares:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{4.83}$$

The function L is to be minimized with respect to $\beta_0, \beta_1, \ldots, \beta_k$. The least squares estimators, say $\beta_0, \beta_1, \ldots, \beta_k$, must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0 \tag{4.84a}$$

and

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \ldots, k \tag{4.84b}$$

Simplifying Equation 4.84 we obtain

$$\begin{array}{lllll}
n\hat{\beta}_0 + \hat{\beta}_i \sum_{i=1}^{n} x_{i1} & + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} & + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} & = & \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 \, \square & + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} & + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} & = & \sum_{i=1}^{n} x_{i1}y_i \\
\vdots \qquad \vdots & \vdots & \vdots & & \vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} & + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} & + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 & = & \sum_{i=1}^{n} x_{ik}y_i
\end{array} \tag{4.85}$$

Least squares normal equations

It is simpler to solve the normal equations if they are expressed in matrix notation. We now give a matrix development of the normal equations that parallels the development of equation (4.85). The model in terms of the observations, equation (4.82) may be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

In general, **y** is an $(n \times 1)$ vector of the observations, **X** is an $(n \times p)$ matrix of the levels of the independent variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of the regression coefficients, and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of random errors.

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{4.87}$$

Equation (4.87) is the matrix form of the least squares normal equations. It is identical to equation (4.85). To solve the normal equations, multiply both sides of equation (4.87) by the inverse of $\mathbf{X}'\mathbf{X}$**.** Thus, the least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{4.88}$$

The fitted regression model is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{4.89}$$

The difference between the actual observation y_i and the corresponding fitted value $\hat{y}_i$ is the **residual,** say $e_i = y_i - \hat{y}_i$. The $(n \times 1)$ vector of residuals is denoted by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{4.90}$$

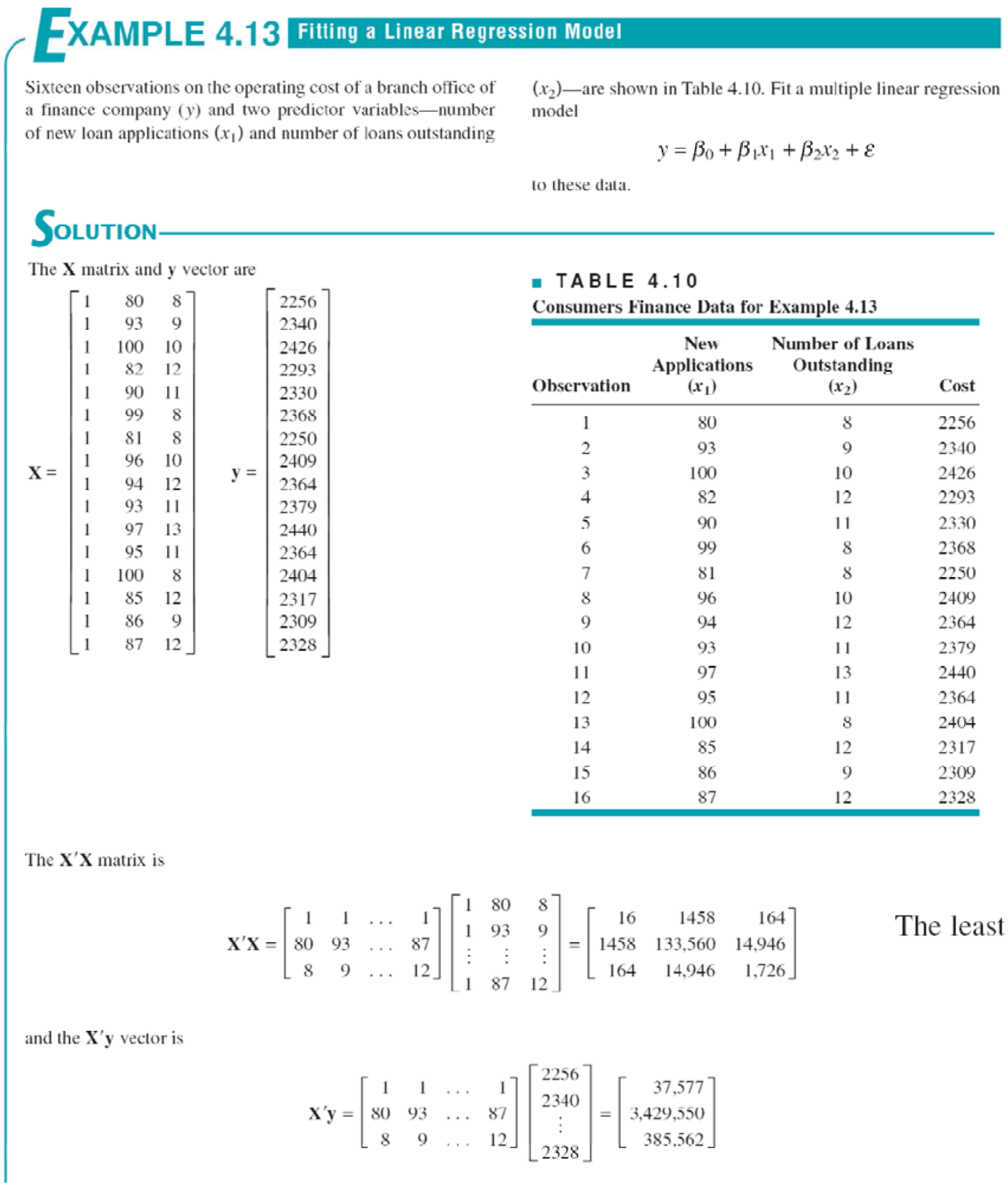
$$SS_E = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \tag{4.91}$$

Equation (4.91) is called the **error** or **residual sum of squares,** and it has $n - p$ degrees of freedom associated with it. It can be shown that

$$E(SS_E) = \sigma^2(n - p)$$

so an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{SS_E}{n - p} \tag{4.92}$$



EXAMPLE 4.13 Fitting a Linear Regression Model

Sixteen observations on the operating cost of a branch office of a finance company (y) and two predictor variables—number of new loan applications (x_1) and number of loans outstanding (x_2)—are shown in Table 4.10. Fit a multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

to these data.

SOLUTION

The **X** matrix and **y** vector are

$$\mathbf{X} = \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ 1 & 100 & 10 \\ 1 & 82 & 12 \\ 1 & 90 & 11 \\ 1 & 99 & 8 \\ 1 & 81 & 8 \\ 1 & 96 & 10 \\ 1 & 94 & 12 \\ 1 & 93 & 11 \\ 1 & 97 & 13 \\ 1 & 95 & 11 \\ 1 & 100 & 8 \\ 1 & 85 & 12 \\ 1 & 86 & 9 \\ 1 & 87 & 12 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 2256 \\ 2340 \\ 2426 \\ 2293 \\ 2330 \\ 2368 \\ 2250 \\ 2409 \\ 2364 \\ 2379 \\ 2440 \\ 2364 \\ 2404 \\ 2317 \\ 2309 \\ 2328 \end{bmatrix}$$

■ **TABLE 4.10**
Consumers Finance Data for Example 4.13

Observation	New Applications (x_1)	Number of Loans Outstanding (x_2)	Cost
1	80	8	2256
2	93	9	2340
3	100	10	2426
4	82	12	2293
5	90	11	2330
6	99	8	2368
7	81	8	2250
8	96	10	2409
9	94	12	2364
10	93	11	2379
11	97	13	2440
12	95	11	2364
13	100	8	2404
14	85	12	2317
15	86	9	2309
16	87	12	2328

The $\mathbf{X}'\mathbf{X}$ matrix is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 80 & 93 & \cdots & 87 \\ 8 & 9 & \cdots & 12 \end{bmatrix} \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ \vdots & \vdots & \vdots \\ 1 & 87 & 12 \end{bmatrix} = \begin{bmatrix} 16 & 1458 & 164 \\ 1458 & 133{,}560 & 14{,}946 \\ 164 & 14{,}946 & 1{,}726 \end{bmatrix}$$

and the $\mathbf{X}'\mathbf{y}$ vector is

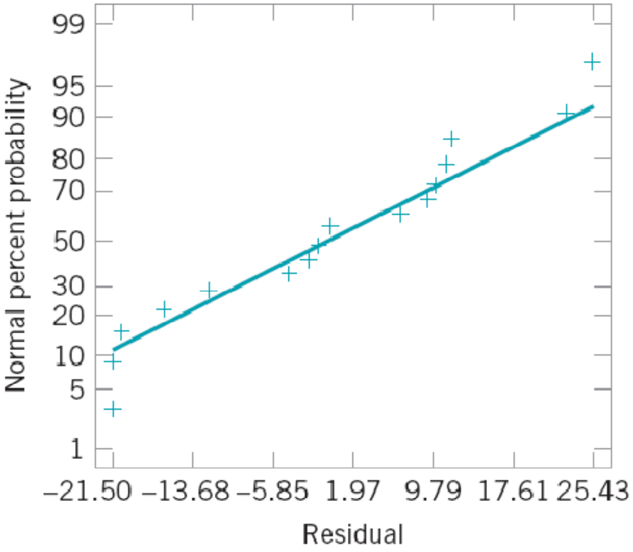
$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 80 & 93 & \cdots & 87 \\ 8 & 9 & \cdots & 12 \end{bmatrix} \begin{bmatrix} 2256 \\ 2340 \\ \vdots \\ 2328 \end{bmatrix} = \begin{bmatrix} 37{,}577 \\ 3{,}429{,}550 \\ 385{,}562 \end{bmatrix}$$

The least squares estimate of $\boldsymbol{\beta}$ is

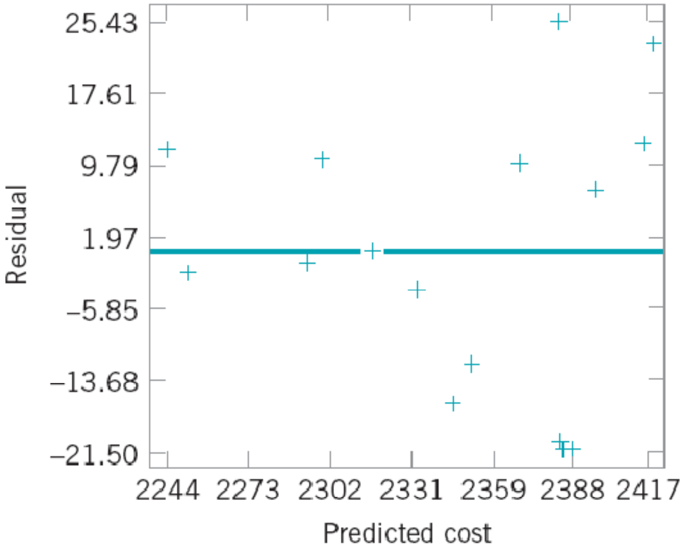
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\hat{\beta} = \begin{bmatrix} 14.176004 & -0.129746 & -0.223453 \\ -0.129746 & 1.429184 \times 10^{-3} & -4.763947 \times 10^{-5} \\ -0.223453 & -4{,}763947 \times 10^{-5} & 2.222381 \times 10^{-2} \end{bmatrix} \begin{bmatrix} 37{,}577 \\ 3{,}429{,}550 \\ 385{,}562 \end{bmatrix}$$

$$= \begin{bmatrix} 1566.07777 \\ 7.62129 \\ 8.58485 \end{bmatrix}$$



■ FIGURE 4.16 Normal probability plot of residuals, Example 4.13.



■ FIGURE 4.17 Plot of residuals versus predicted cost, Example 4.13.

The least squares fit, with the regression coefficients reported to two decimal places, is

$$\hat{y} = 1566.08 + 7.62x_1 + 8.58x_2$$

The first three columns of Table 4.11 present the actual observations y_i, the predicted or fitted values $\hat{y}_i$, and the residuals. Figure 4.16 is a normal probability plot of the residuals. Plots of the residuals versus the predicted values $\hat{y}_i$ and versus the two variables x_1 and x_2 are shown in Figures 4.17, 4.18, and 4.19, respectively. Just as in ANOVA, residual plotting is an integral part of regression model building. These plots indicate that variance of the observed cost tends to increase with the magnitude of cost. Figure 4.18 suggests that the variability in cost may be increasing as the number of new applications increases.

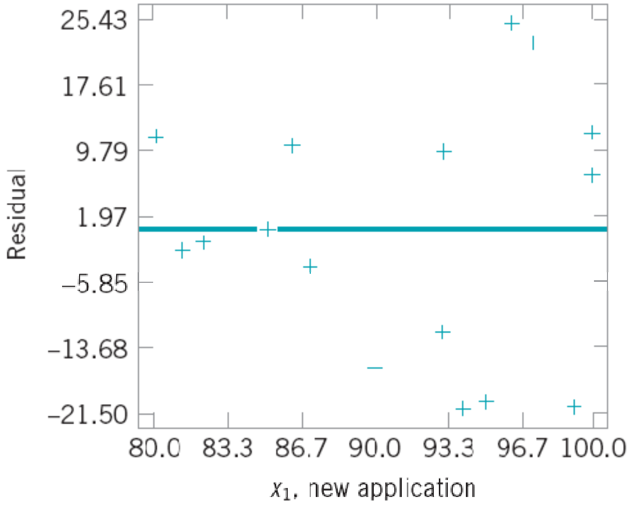


FIGURE 4.18 Plot of residuals versus x_1 (new applications), Example 4.13.

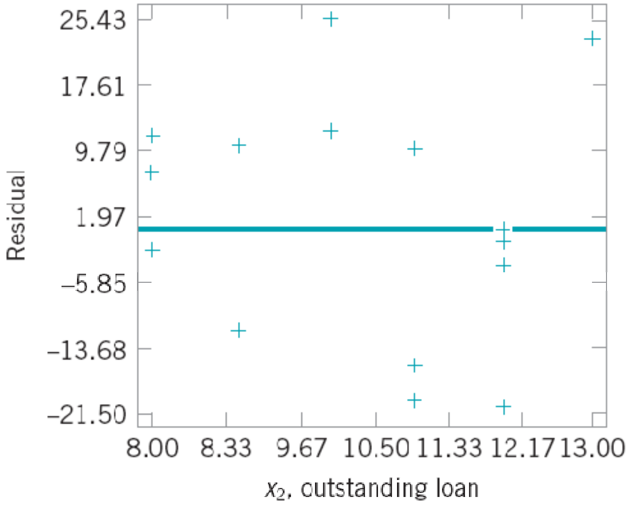
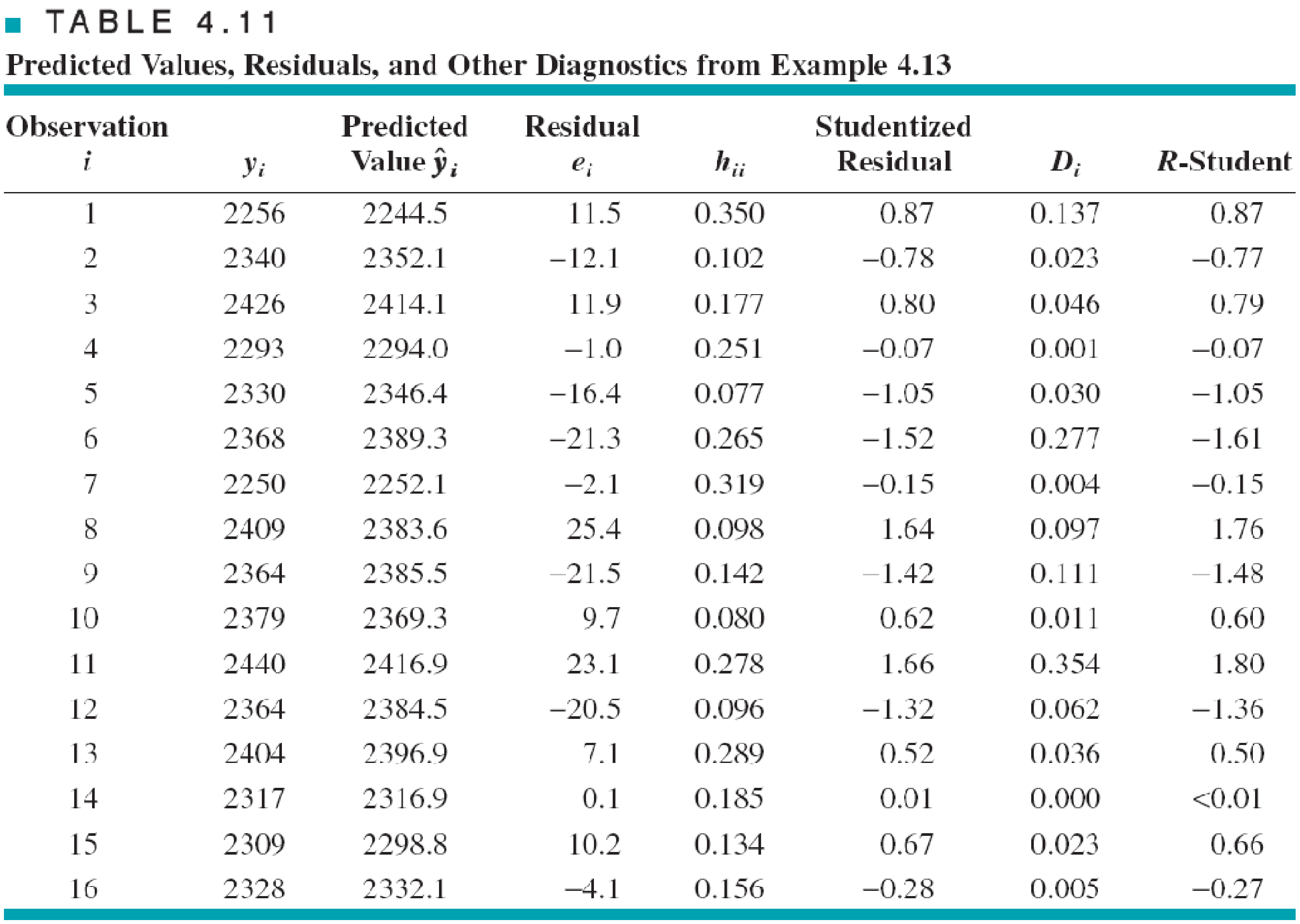


FIGURE 4.19 Plot of residuals versus x_2 (outstanding loan), Example 4.13.



■ TABLE 4.11

Predicted Values, Residuals, and Other Diagnostics from Example 4.13

Observation i	y_i	Predicted Value $\hat{y}_i$	Residual e_i	h_{ii}	Studentized Residual	D_i	R-Student
1	2256	2244.5	11.5	0.350	0.87	0.137	0.87
2	2340	2352.1	−12.1	0.102	−0.78	0.023	−0.77
3	2426	2414.1	11.9	0.177	0.80	0.046	0.79
4	2293	2294.0	−1.0	0.251	−0.07	0.001	−0.07
5	2330	2346.4	−16.4	0.077	−1.05	0.030	−1.05
6	2368	2389.3	−21.3	0.265	−1.52	0.277	−1.61
7	2250	2252.1	−2.1	0.319	−0.15	0.004	−0.15
8	2409	2383.6	25.4	0.098	1.64	0.097	1.76
9	2364	2385.5	−21.5	0.142	−1.42	0.111	−1.48
10	2379	2369.3	9.7	0.080	0.62	0.011	0.60
11	2440	2416.9	23.1	0.278	1.66	0.354	1.80
12	2364	2384.5	−20.5	0.096	−1.32	0.062	−1.36
13	2404	2396.9	7.1	0.289	0.52	0.036	0.50
14	2317	2316.9	0.1	0.185	0.01	0.000	<0.01
15	2309	2298.8	10.2	0.134	0.67	0.023	0.66
16	2328	2332.1	−4.1	0.156	−0.28	0.005	−0.27

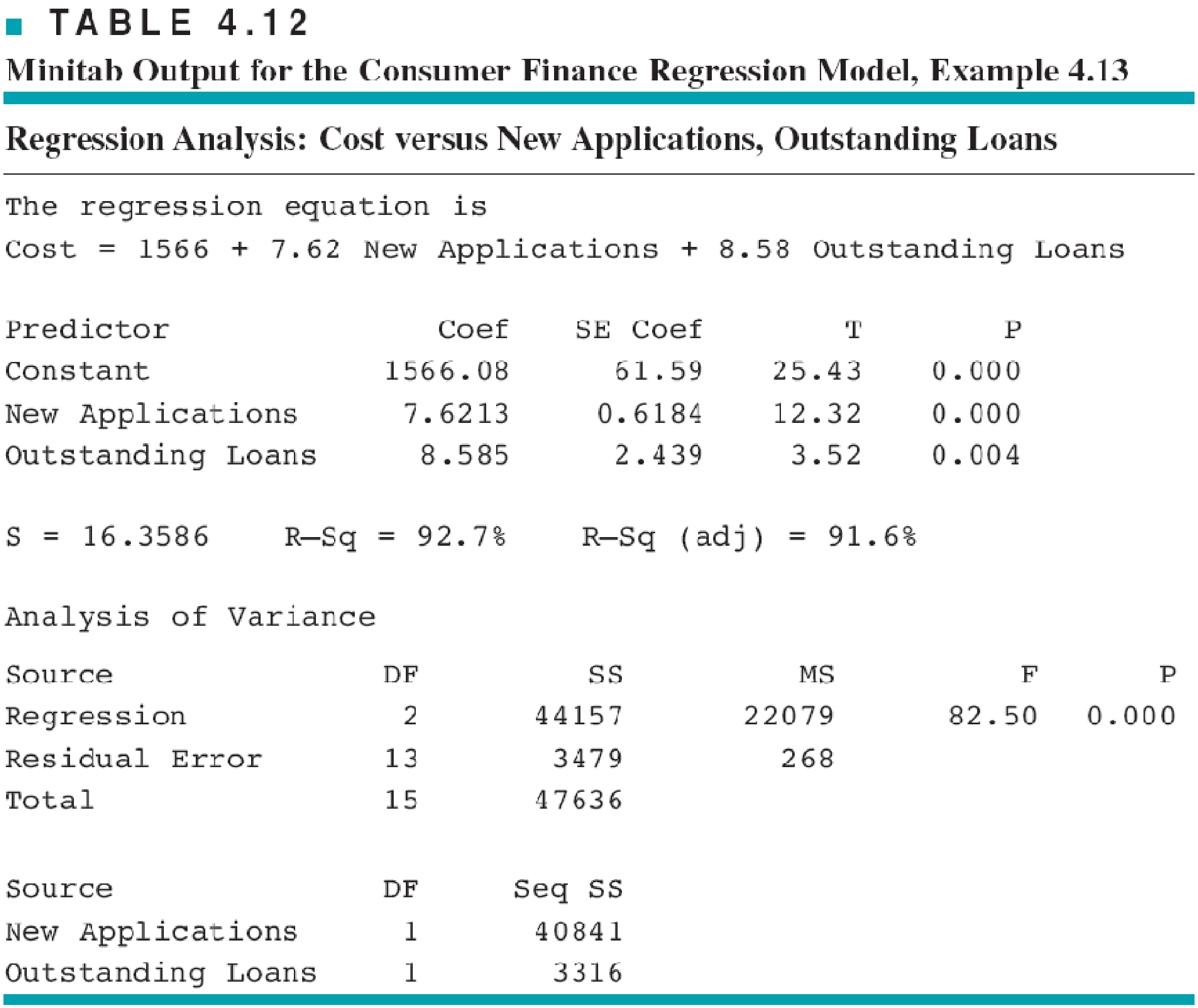
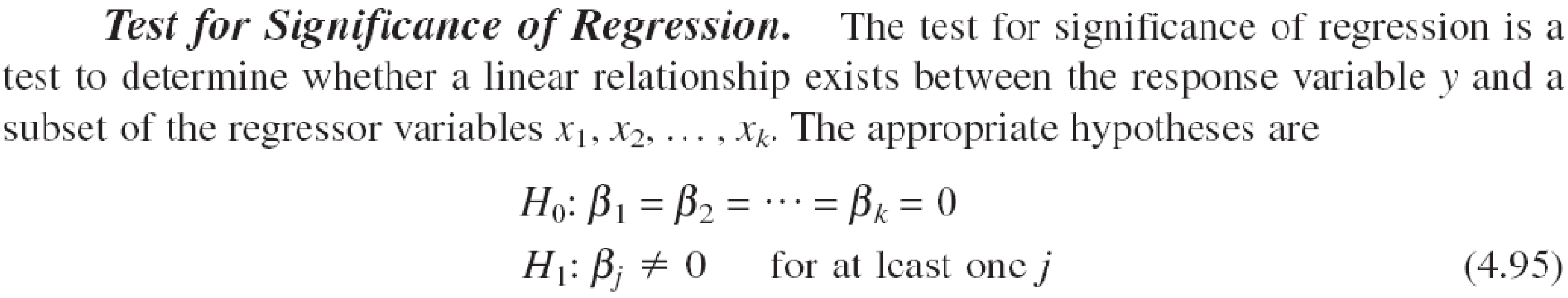


TABLE 4.12

Minitab Output for the Consumer Finance Regression Model, Example 4.13

Regression Analysis: Cost versus New Applications, Outstanding Loans

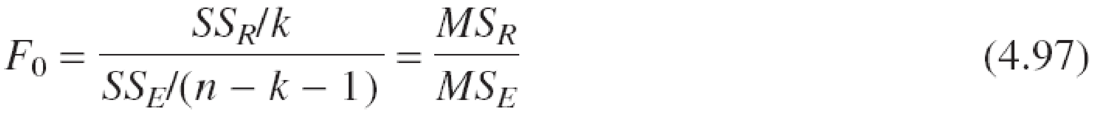
```
The regression equation is
Cost = 1566 + 7.62 New Applications + 8.58 Outstanding Loans

Predictor               Coef    SE Coef       T        P
Constant             1566.08      61.59   25.43    0.000
New Applications      7.6213     0.6184   12.32    0.000
Outstanding Loans      8.585      2.439    3.52    0.004

S = 16.3586    R–Sq = 92.7%    R–Sq (adj) = 91.6%

Analysis of Variance

Source            DF        SS        MS         F       P
Regression         2     44157     22079     82.50   0.000
Residual Error    13      3479       268
Total             15     47636

Source            DF    Seq SS
New Applications   1     40841
Outstanding Loans  1      3316
```



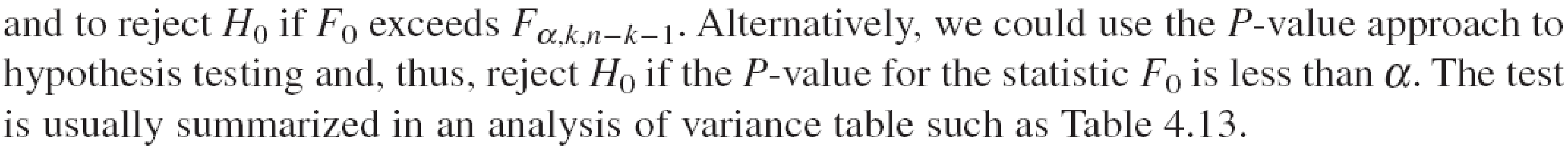
Test for Significance of Regression. The test for significance of regression is a test to determine whether a linear relationship exists between the response variable y and a subset of the regressor variables $x_1, x_2, \ldots, x_k$. The appropriate hypotheses are

$$\begin{aligned} &H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ &H_1\colon \beta_j \neq 0 \quad \text{for at lcast onc } j \end{aligned} \tag{4.95}$$

The test procedure is to calculate the test statistic

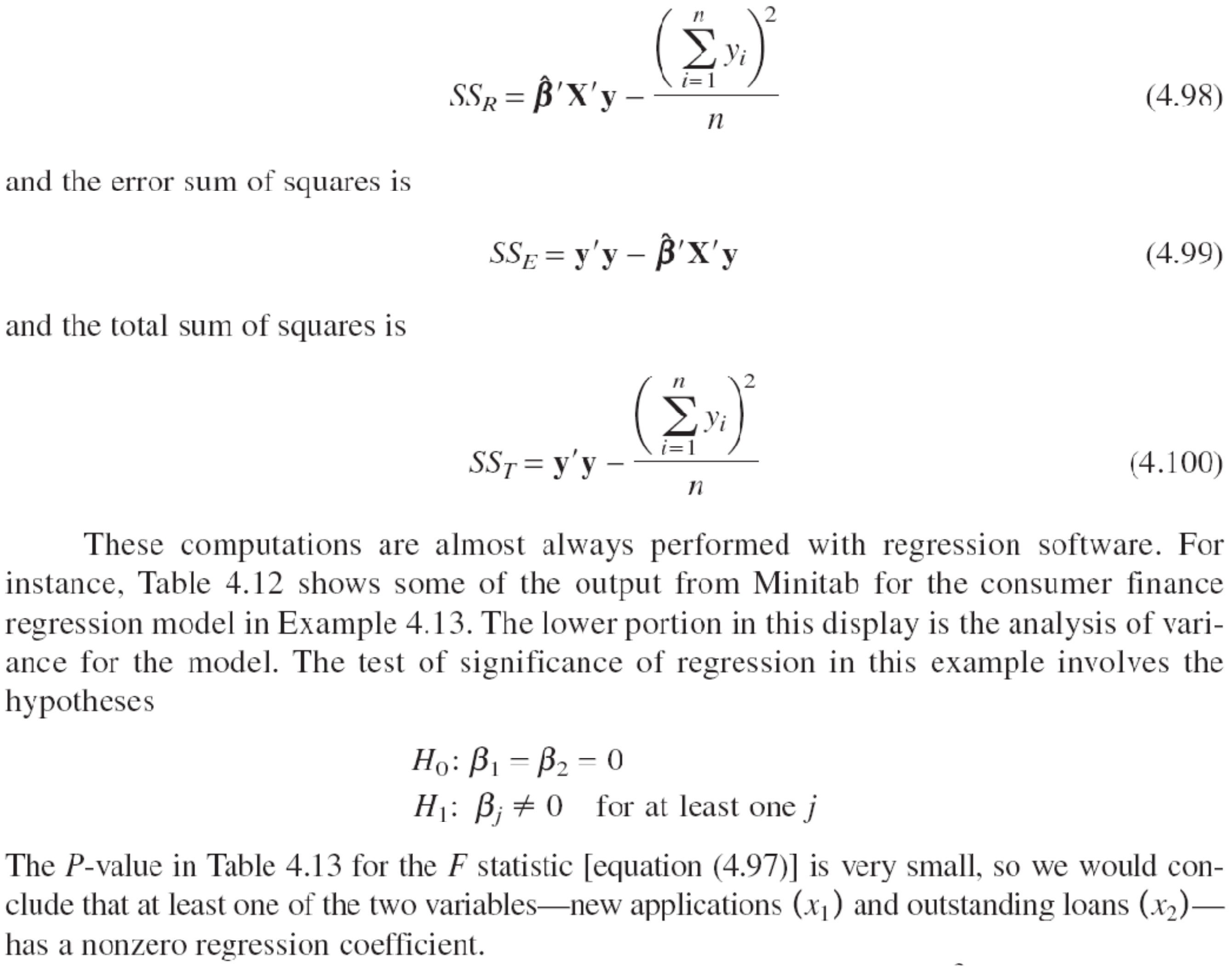


$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E} \tag{4.97}$$



and to reject H_0 if F_0 exceeds $F_{\alpha,k,n-k-1}$. Alternatively, we could use the P-value approach to hypothesis testing and, thus, reject H_0 if the P-value for the statistic F_0 is less than α. The test is usually summarized in an analysis of variance table such as Table 4.13.

The regression sum of squares is



$$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \qquad (4.98)$$

and the error sum of squares is

$$SS_E = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \qquad (4.99)$$

and the total sum of squares is

$$SS_T = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \qquad (4.100)$$

These computations are almost always performed with regression software. For instance, Table 4.12 shows some of the output from Minitab for the consumer finance regression model in Example 4.13. The lower portion in this display is the analysis of variance for the model. The test of significance of regression in this example involves the hypotheses

$$H_0\colon \beta_1 = \beta_2 = 0$$
$$H_1\colon \beta_j \neq 0 \quad \text{for at least one } j$$

The P-value in Table 4.13 for the F statistic [equation (4.97)] is very small, so we would conclude that at least one of the two variables—new applications (x_1) and outstanding loans (x_2)—has a nonzero regression coefficient.

Table 4.13 also reports the coefficient to multiple determination R^2, where

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{4.101}$$

The statistic R^2 is a measure of the amount of reduction in the variability of y obtained by using the regressor variables $x_1, x_2, \dots, x_k$ in the model. However, a large value of R^2 does not necessarily imply that the regression model is a good one. Adding a variable to the model will always increase R^2, regardless of whether the additional variable is statistically significant or not. Thus, it is possible for models that have large values of R^2 to yield poor predictions of new observations or estimates of the mean response.

Because R^2 always increases as we add terms to the model, some regression model builders prefer to use an **adjusted $\mathbf{R^2}$ statistic** defined as

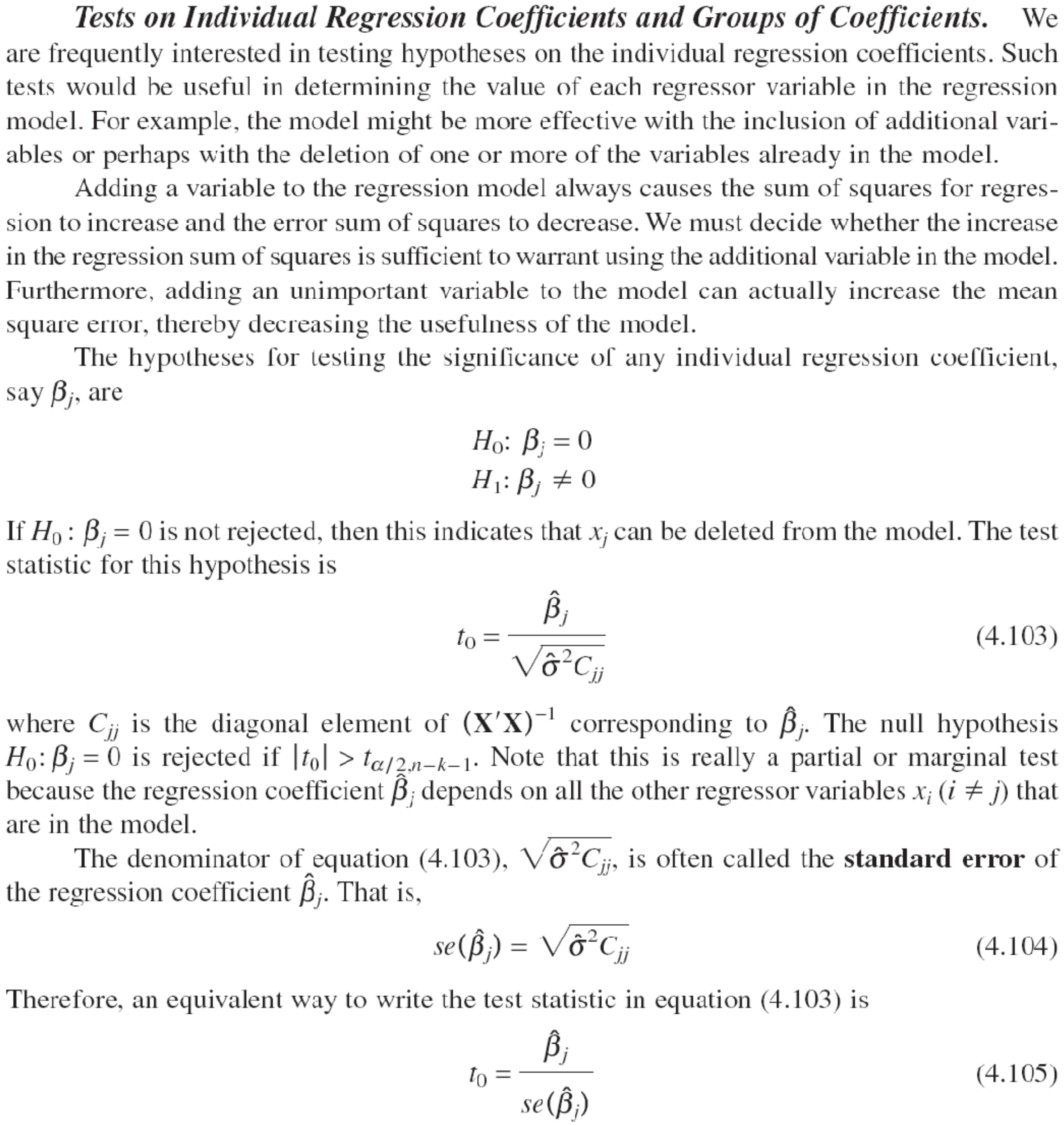
$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2) \tag{4.102}$$

In general, the adjusted R^2 statistic will not always increase as variables are added to the model. In fact, if unnecessary terms are added, the value of R^2_{adj} will often decrease.

For example, consider the consumer finance regression model. The adjusted R^2 for the model is shown in Table 4.12. It is computed as

$$\begin{aligned} R^2_{\text{adj}} &= 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2) \\ &= 1 - \left(\frac{15}{13}\right)(1 - 0.92697) = 0.915735 \end{aligned}$$

which is very close to the ordinary R^2. When R^2 and R^2_{adj} differ dramatically, there is a good chance that nonsignificant terms have been included in the model.



Tests on Individual Regression Coefficients and Groups of Coefficients. We are frequently interested in testing hypotheses on the individual regression coefficients. Such tests would be useful in determining the value of each regressor variable in the regression model. For example, the model might be more effective with the inclusion of additional variables or perhaps with the deletion of one or more of the variables already in the model.

Adding a variable to the regression model always causes the sum of squares for regression to increase and the error sum of squares to decrease. We must decide whether the increase in the regression sum of squares is sufficient to warrant using the additional variable in the model. Furthermore, adding an unimportant variable to the model can actually increase the mean square error, thereby decreasing the usefulness of the model.

The hypotheses for testing the significance of any individual regression coefficient, say β_j, are

$$\begin{aligned} &H_0\colon \beta_j = 0 \\ &H_1\colon \beta_j \neq 0 \end{aligned}$$

If $H_0: \beta_j = 0$ is not rejected, then this indicates that x_j can be deleted from the model. The test statistic for this hypothesis is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \tag{4.103}$$

where C_{jj} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_j$. The null hypothesis $H_0\colon \beta_j = 0$ is rejected if $|t_0| > t_{\alpha/2,n-k-1}$. Note that this is really a partial or marginal test because the regression coefficient $\hat{\beta}_j$ depends on all the other regressor variables x_i $(i \neq j)$ that are in the model.

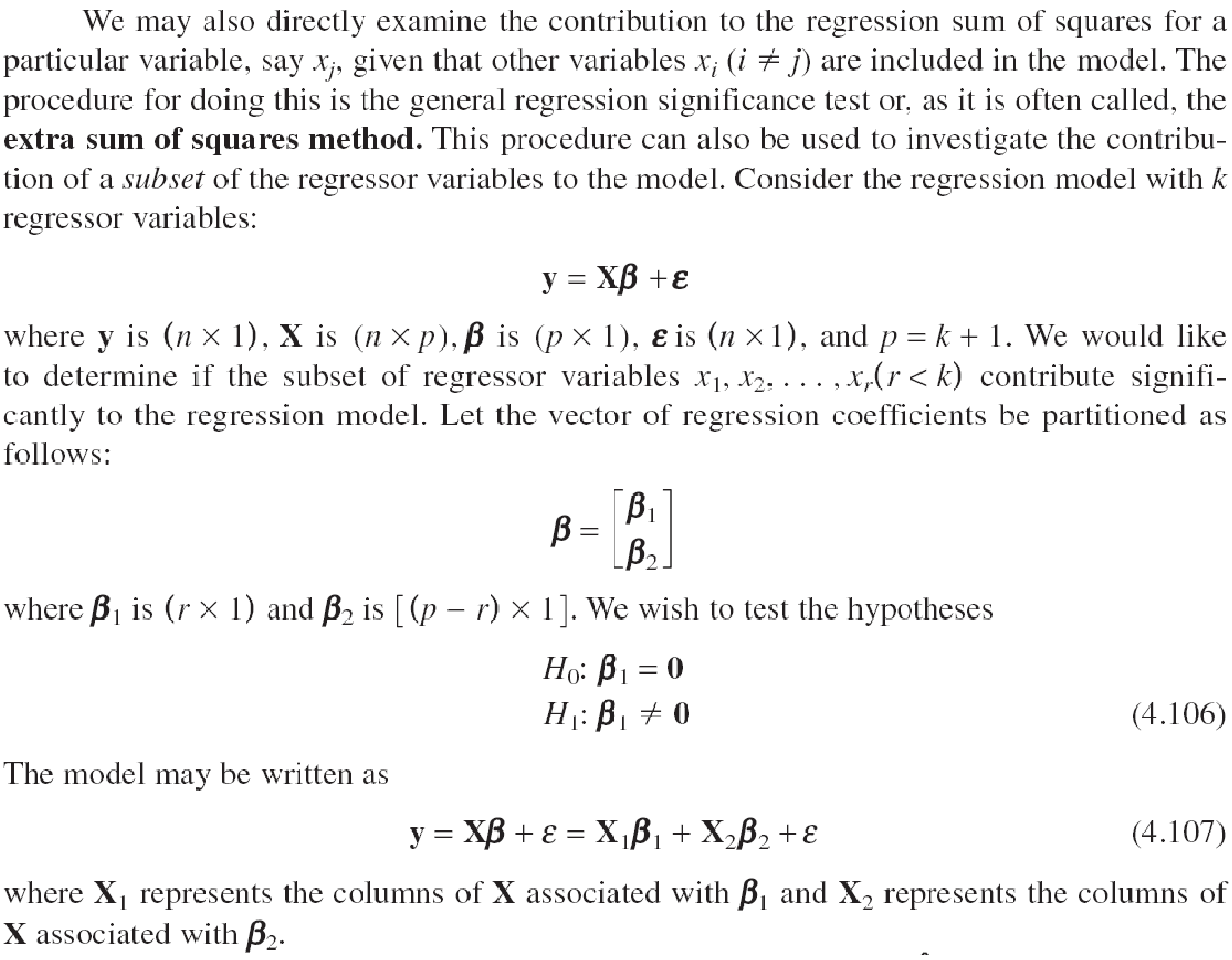
The denominator of equation (4.103), $\sqrt{\hat{\sigma}^2 C_{jj}}$, is often called the **standard error** of the regression coefficient $\hat{\beta}_j$. That is,

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}} \tag{4.104}$$

Therefore, an equivalent way to write the test statistic in equation (4.103) is

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{4.105}$$

See Table 4.12 (slide 91) for the *t*-tests on the individual regressors in the consumer finance model – both variables are significant



We may also directly examine the contribution to the regression sum of squares for a particular variable, say x_j, given that other variables x_i $(i \neq j)$ are included in the model. The procedure for doing this is the general regression significance test or, as it is often called, the **extra sum of squares method.** This procedure can also be used to investigate the contribution of a *subset* of the regressor variables to the model. Consider the regression model with k regressor variables:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ is $(n \times 1)$, $\mathbf{X}$ is $(n \times p)$, $\boldsymbol{\beta}$ is $(p \times 1)$, $\boldsymbol{\varepsilon}$ is $(n \times 1)$, and $p = k + 1$. We would like to determine if the subset of regressor variables $x_1, x_2, \ldots, x_r (r < k)$ contribute significantly to the regression model. Let the vector of regression coefficients be partitioned as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

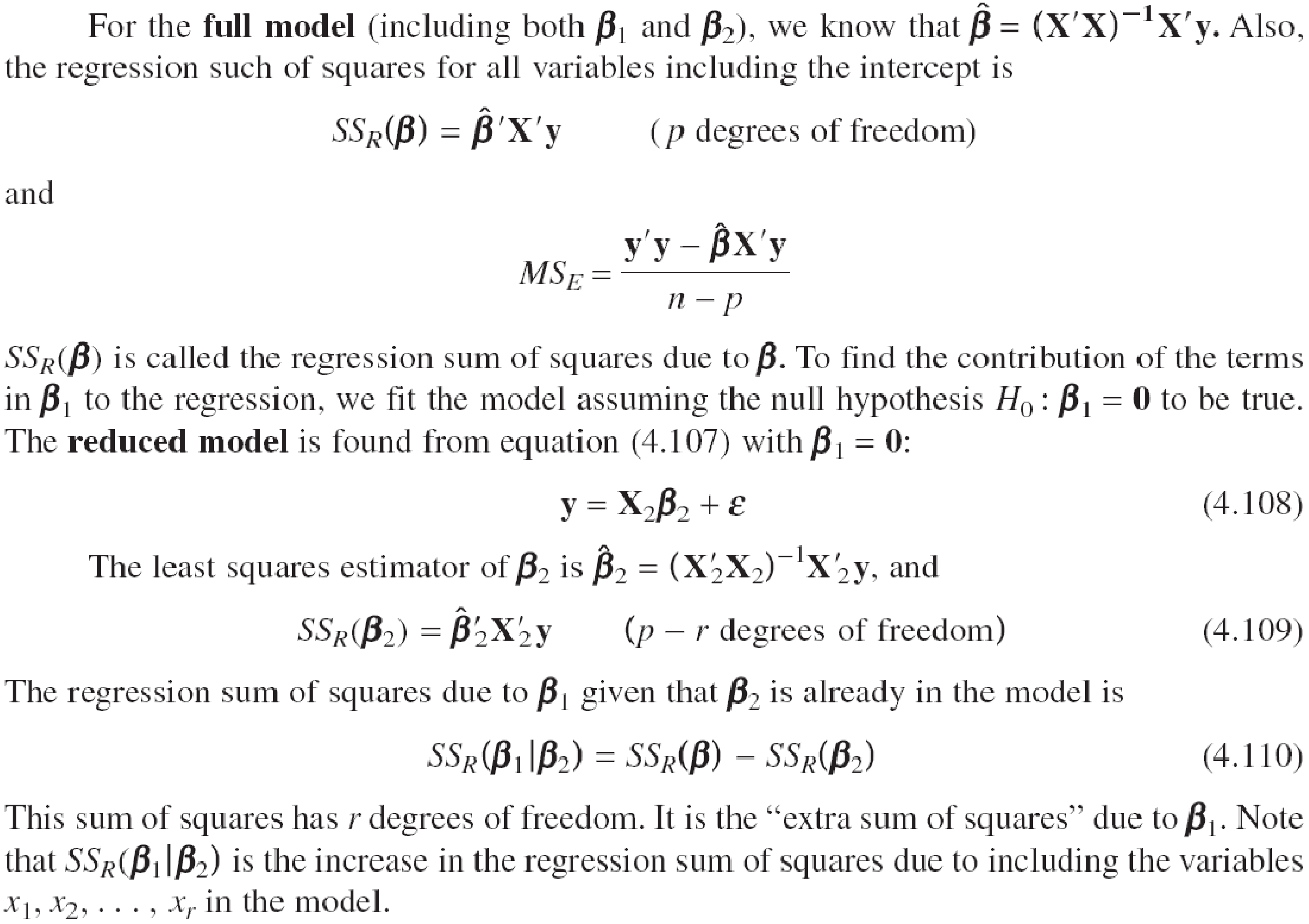
where $\boldsymbol{\beta}_1$ is $(r \times 1)$ and $\boldsymbol{\beta}_2$ is $[(p - r) \times 1]$. We wish to test the hypotheses

$$\begin{aligned} &H_0\colon \boldsymbol{\beta}_1 = \mathbf{0} \\ &H_1\colon \boldsymbol{\beta}_1 \neq \mathbf{0} \end{aligned} \tag{4.106}$$

The model may be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \varepsilon \tag{4.107}$$

where $\mathbf{X}_1$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_1$ and $\mathbf{X}_2$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_2$.



For the **full model** (including both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$), we know that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Also, the regression such of squares for all variables including the intercept is

$$SS_R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \qquad (p \text{ degrees of freedom})$$

and

$$MS_E = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y}}{n-p}$$

$SS_R(\boldsymbol{\beta})$ is called the regression sum of squares due to $\boldsymbol{\beta}$. To find the contribution of the terms in $\boldsymbol{\beta}_1$ to the regression, we fit the model assuming the null hypothesis $H_0: \boldsymbol{\beta}_1 = \mathbf{0}$ to be true. The **reduced model** is found from equation (4.107) with $\boldsymbol{\beta}_1 = \mathbf{0}$:

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{4.108}$$

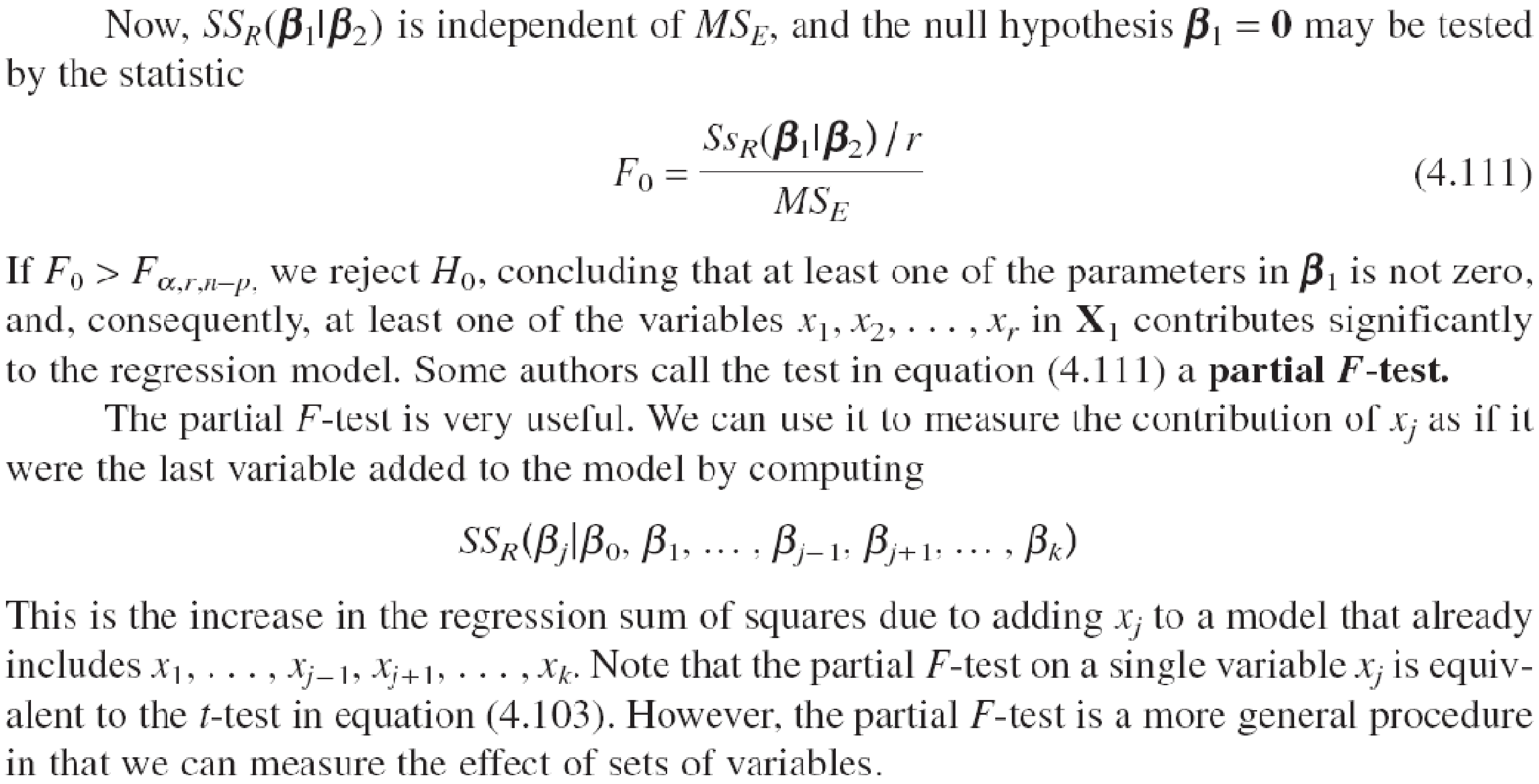
The least squares estimator of $\boldsymbol{\beta}_2$ is $\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{y}$, and

$$SS_R(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}'_2\mathbf{X}'_2\mathbf{y} \qquad (p-r \text{ degrees of freedom}) \tag{4.109}$$

The regression sum of squares due to $\boldsymbol{\beta}_1$ given that $\boldsymbol{\beta}_2$ is already in the model is

$$SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_2) \tag{4.110}$$

This sum of squares has r degrees of freedom. It is the "extra sum of squares" due to $\boldsymbol{\beta}_1$. Note that $SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$ is the increase in the regression sum of squares due to including the variables $x_1, x_2, \ldots, x_r$ in the model.



Now, $SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$ is independent of MS_E, and the null hypothesis $\boldsymbol{\beta}_1 = \mathbf{0}$ may be tested by the statistic

$$F_0 = \frac{Ss_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)\,/\,r}{MS_E} \qquad (4.111)$$

If $F_0 > F_{\alpha,r,n-p,}$ we reject H_0, concluding that at least one of the parameters in $\boldsymbol{\beta}_1$ is not zero, and, consequently, at least one of the variables $x_1, x_2, \ldots, x_r$ in $\mathbf{X}_1$ contributes significantly to the regression model. Some authors call the test in equation (4.111) a **partial *F*-test.**

The partial F-test is very useful. We can use it to measure the contribution of x_j as if it were the last variable added to the model by computing

$$SS_R(\beta_j|\beta_0, \beta_1, \ldots ; \beta_{j-1}; \beta_{j+1}; \ldots , \beta_k)$$

This is the increase in the regression sum of squares due to adding x_j to a model that already includes $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$. Note that the partial F-test on a single variable x_j is equivalent to the t-test in equation (4.103). However, the partial F-test is a more general procedure in that we can measure the effect of sets of variables.

EXAMPLE 4.14 The Extra Sum of Squares Method

Consider the consumer finance data in Example 4.13. Evaluate the contribution of x_2 (outstanding loans) to the model.

SOLUTION

The hypotheses we wish to test are

$$H_0\colon \beta_2 = 0$$
$$H_1\colon \beta_2 \neq 0$$

This will require the extra sum of squares due to β_2, or

$$\begin{aligned} SS_R(\beta_2|\beta_1, \beta_0) &= SS_R(\beta_0, \beta_1, \beta_2) - SS_R(\beta_0, \beta_1) \\ &= Ss_R(\beta_1, \beta_2|\beta_0) - SS_R(\beta_2|\beta_0) \end{aligned}$$

Now from Table 4.12, where we tested for significance of regression, we have

$$SS_R(\beta_1{,}\beta_2|\beta_0) = 44{,}157.1$$

which was called the model sum of squares in the table. This sum of squares has two degrees of freedom.

The reduced model is

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

The least squares fit for this model is

$$\hat{y} = 1652.3955 + 7.6397x_1$$

and the regression sum of squares for this model (with one degree of freedom) is

$$SS_R(\beta_1|\beta_0) = 40{,}840.8$$

Note that $SS_R(\beta_1|\beta_0)$ is shown at the bottom of the Minitab output in Table 4.12 under the heading "Seq SS." Therefore,

$$\begin{aligned} SS_R(\beta_2|\beta_0, \beta_1) &= 44{,}157.1 - 40{,}840.8 \\ &= 3316.3 \end{aligned}$$

with $2 - 1 = 1$ degree of freedom. This is the increase in the regression sum of squares that results from adding x_2 to model already containing x_1, and it is shown at the bottom of the Minitab output in Table 4.12. To test $H_0\colon \beta_2 = 0$, from the test statistic we obtain

$$F_0 = \frac{Ss_R(\beta_2|\beta_0, \beta_1)/1}{MS_E} = \frac{3316.3/1}{267.604} = 12.3926$$

Note that MS_E from the full model (Table 4.12) is used in the denominator of F_0. Now, because $F_{0.05,1,13} = 4.67$, we would reject $H_0\colon \beta_2 = 0$ and conclude that x_2 (outstanding loans) contributes significantly to the model.

Because this partial F-test involves only a single regressor, it is equivalent to the t-test because the square of a t random variable with v degrees of freedom is an F random variable with 1 and v degrees of freedom. To see this, note from Table 4.12 that the t-statistic for $H_0 : \beta_2 = 0$ resulted in $t_0 = 3.5203$ and that $t_0^2 = (3.5203)^2 = 12.3925 \simeq F_0$.

Confidence Intervals on the Individual Regression Coefficients. Because the least squares estimator $\hat{\boldsymbol{\beta}}$ is a linear combination of the observations, it follows that $\hat{\boldsymbol{\beta}}$ is normally distributed with mean vector $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Then each of the statistics

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \qquad j = 0, 1, \ldots, k \tag{4.112}$$

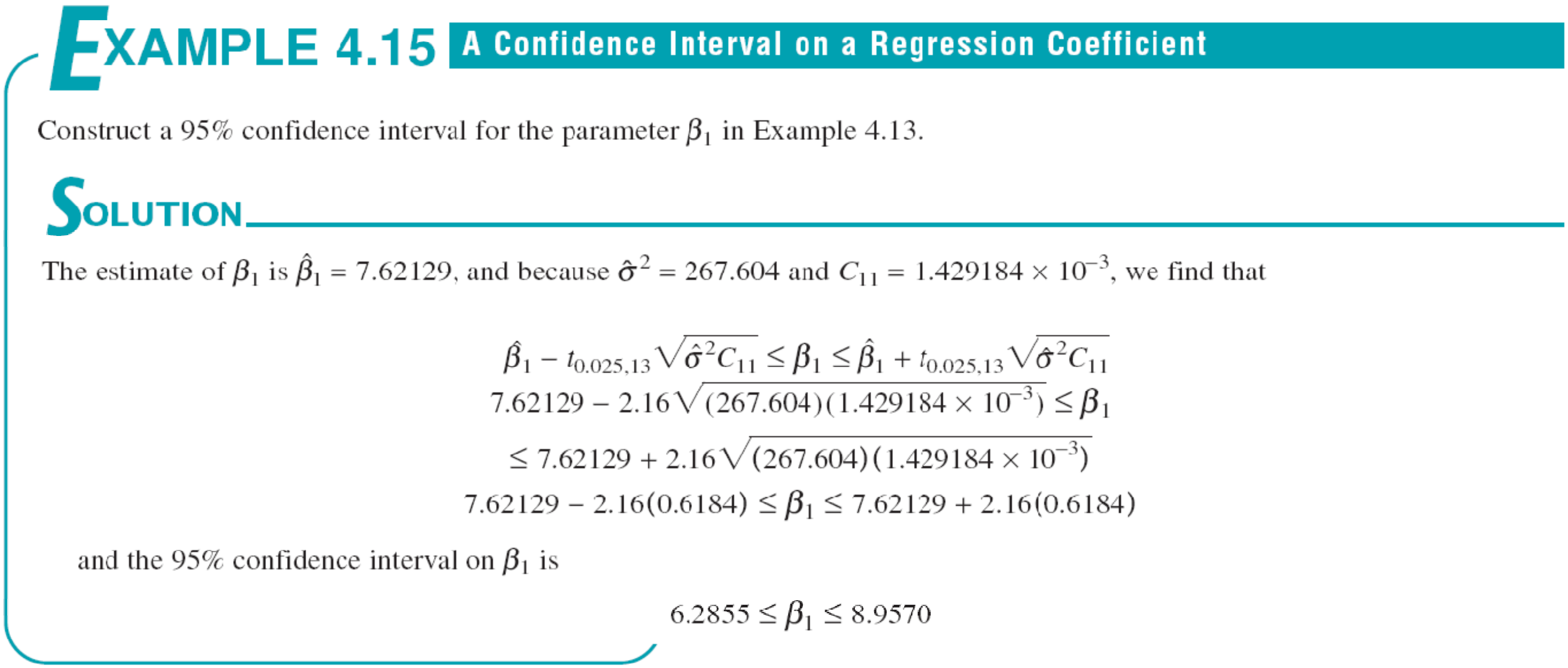
is distributed as t with $n - p$ degrees of freedom, where C_{jj} is the (jj)th element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix, and $\hat{\sigma}^2$ is the estimate of the error variance, obtained from equation (4.92). Therefore, a $100(1 - \alpha)$ percent CI for the regression coefficient β_j, $j = 0, 1, \ldots, k$, is

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \tag{4.113}$$

Note that this CI could also be written as

$$\hat{\beta}_j - t_{\alpha/2,n-p}se(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}se(\hat{\beta}_j)$$

because $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$.



EXAMPLE 4.15 A Confidence Interval on a Regression Coefficient

Construct a 95% confidence interval for the parameter β_1 in Example 4.13.

SOLUTION

The estimate of β_1 is $\hat{\beta}_1 = 7.62129$, and because $\hat{\sigma}^2 = 267.604$ and $C_{11} = 1.429184 \times 10^{-3}$, we find that

$$\hat{\beta}_1 - t_{0.025,13}\sqrt{\hat{\sigma}^2 C_{11}} \le \beta_1 \le \hat{\beta}_1 + t_{0.025,13}\sqrt{\hat{\sigma}^2 C_{11}}$$

$$7.62129 - 2.16\sqrt{(267.604)(1.429184 \times 10^{-3})} \le \beta_1$$

$$\le 7.62129 + 2.16\sqrt{(267.604)(1.429184 \times 10^{-3})}$$

$$7.62129 - 2.16(0.6184) \le \beta_1 \le 7.62129 + 2.16(0.6184)$$

and the 95% confidence interval on β_1 is

$$6.2855 \le \beta_1 \le 8.9570$$

Confidence Interval on the Mean Response. We may also obtain a confidence interval on the mean response at a particular point, say, $x_{01}, x_{02}, \ldots, x_{0k}$. We first define the vector

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

The mean response at this point is

$$\mu_{y|\mathbf{x}_0} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_k x_{0k} = \mathbf{x}_0' \boldsymbol{\beta}$$

The estimated mean response at this point is

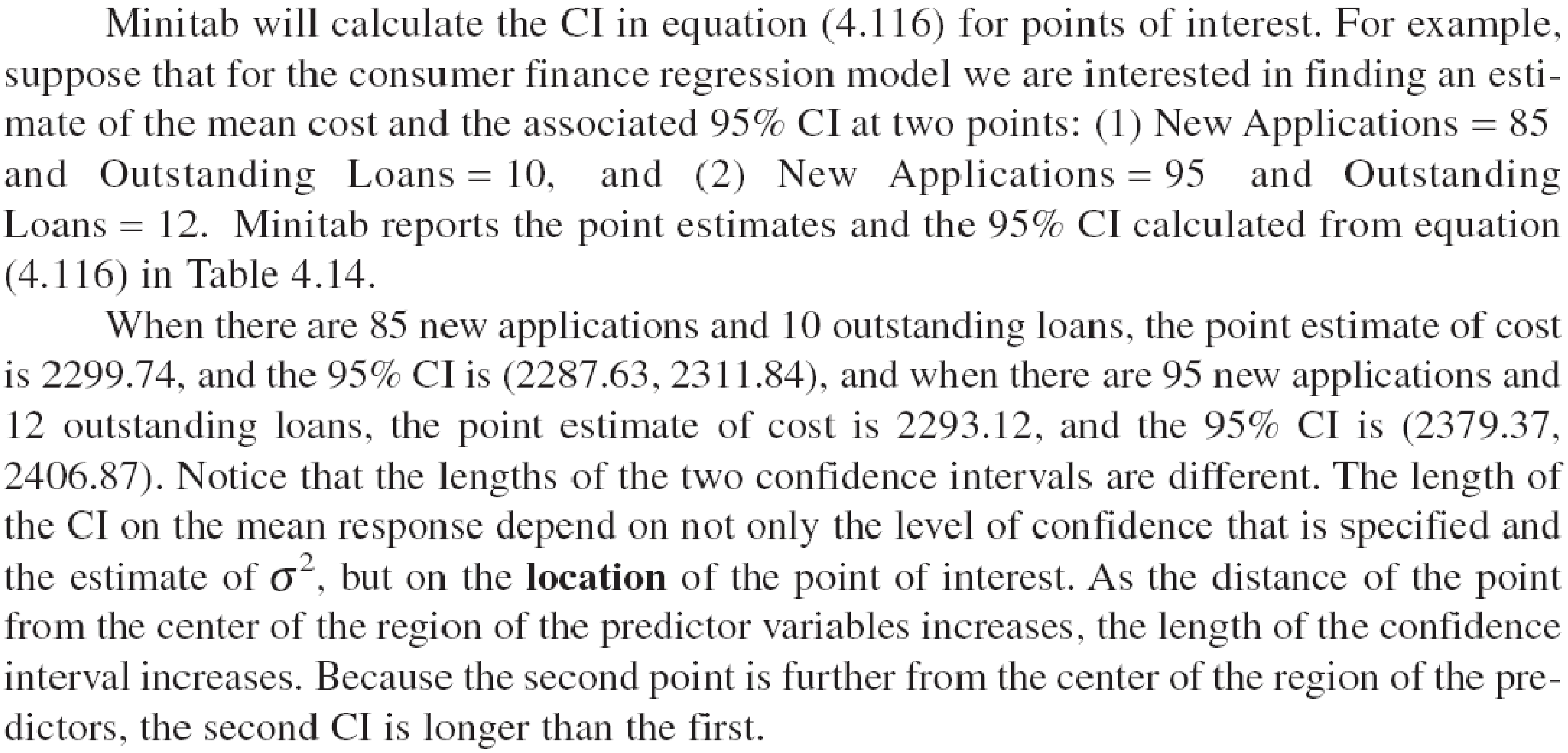
$$\hat{y}(\mathbf{x}_0) = \mathbf{x}_0' \hat{\boldsymbol{\beta}} \tag{4.114}$$

This estimator is unbiased because $E[\hat{y}(\mathbf{x}_0)] = E(\mathbf{x}_0' \hat{\boldsymbol{\beta}}) = \mathbf{x}_0' \boldsymbol{\beta} = \mu_{y|\mathrm{X}_0}$, and the variance of $\hat{y}(\mathbf{x}_0)$ is

$$V[\hat{y}(\mathbf{x}_0)] = \sigma^2 \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 \tag{4.115}$$

Therefore, a $100(1-\alpha)$ percent CI on the mean response at the point $x_{01}, x_{02}, \ldots, x_{0k}$ is

$$\hat{y}(\mathbf{x}_0) - t_{\alpha/2,n-p} \sqrt{\hat{\sigma}^2 \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0} \le \mu_{y|x_0} \le y(\mathbf{x}_0) + t_{\alpha/2,n-p} \sqrt{\hat{\sigma}^2 \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0} \tag{4.116}$$

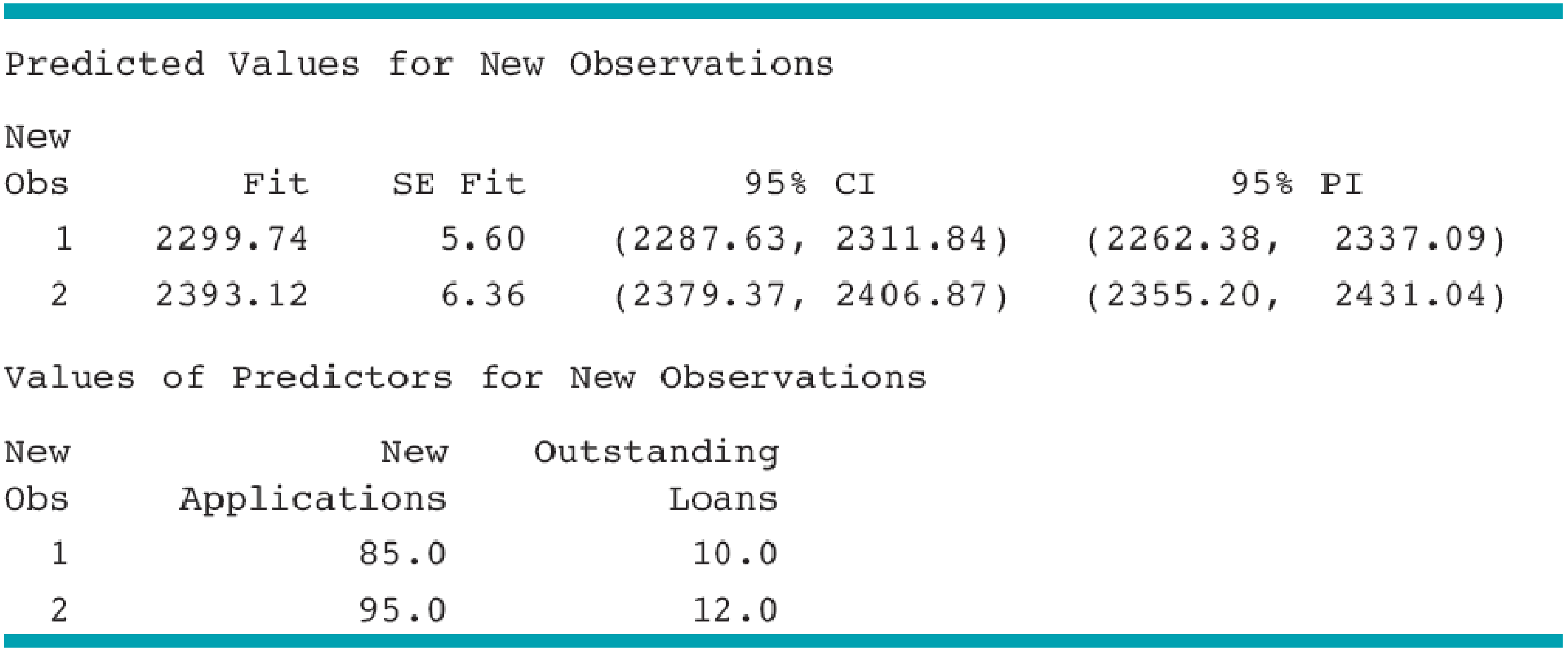


Minitab will calculate the CI in equation (4.116) for points of interest. For example, suppose that for the consumer finance regression model we are interested in finding an estimate of the mean cost and the associated 95% CI at two points: (1) New Applications = 85 and Outstanding Loans = 10, and (2) New Applications = 95 and Outstanding Loans = 12. Minitab reports the point estimates and the 95% CI calculated from equation (4.116) in Table 4.14.

When there are 85 new applications and 10 outstanding loans, the point estimate of cost is 2299.74, and the 95% CI is (2287.63, 2311.84), and when there are 95 new applications and 12 outstanding loans, the point estimate of cost is 2293.12, and the 95% CI is (2379.37, 2406.87). Notice that the lengths of the two confidence intervals are different. The length of the CI on the mean response depend on not only the level of confidence that is specified and the estimate of σ^2, but on the **location** of the point of interest. As the distance of the point from the center of the region of the predictor variables increases, the length of the confidence interval increases. Because the second point is further from the center of the region of the predictors, the second CI is longer than the first.

■ TABLE 4.14

Minitab Output



```
Predicted Values for New Observations

New
Obs       Fit    SE Fit         95% CI                 95% PI
  1   2299.74      5.60   (2287.63, 2311.84)   (2262.38,  2337.09)
  2   2393.12      6.36   (2379.37, 2406.87)   (2355.20,  2431.04)

Values of Predictors for New Observations

New              New   Outstanding
Obs     Applications         Loans
  1             85.0          10.0
  2             95.0          12.0
```

A regression model can be used to predict future observations on the response y corresponding to particular values of the regressor variables, say $x_{01}, x_{02}, \ldots, x_{0k}$. If $\mathbf{x}_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$, then a point estimate for the future observation y_0 at the point $x_{01}, x_{02}, \ldots, x_{0k}$ is computed from equation (4.114):

$$\hat{y}(\mathbf{x}_0) = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$$

A $100(1-\alpha)$ percent **prediction interval** (PI) for this future observation is

$$\hat{y}(\mathbf{x}_0) = t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \le y_0$$
$$\le \hat{y}(\mathbf{x}_0) + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \quad (4.117)$$

In predicting new observations and in estimating the mean response at a given point $x_{01}, x_{02}, \ldots, x_{0k}$, we must be careful about extrapolating beyond the region containing the original observations. It is very possible that a model that fits well in the region of the original data will no longer fit well outside of that region.

The Minitab output in Table 4.14 shows the 95% prediction intervals on cost for the consumer finance regression model at the two points considered previously: (1) New Applications = 85 and Outstanding Loans = 10, and (2) New Applications = 95 and Outstanding Loans = 12. The predicted value of the future observation is exactly equal to the estimate of the mean at the point of interest. Notice that the prediction intervals are longer that the corresponding confidence intervals. You should be able to see why this happens from examining equations (4.116) and (4.117). The prediction intervals also get longer as the point where the prediction is made moves further away from the center of the predictor variable region.

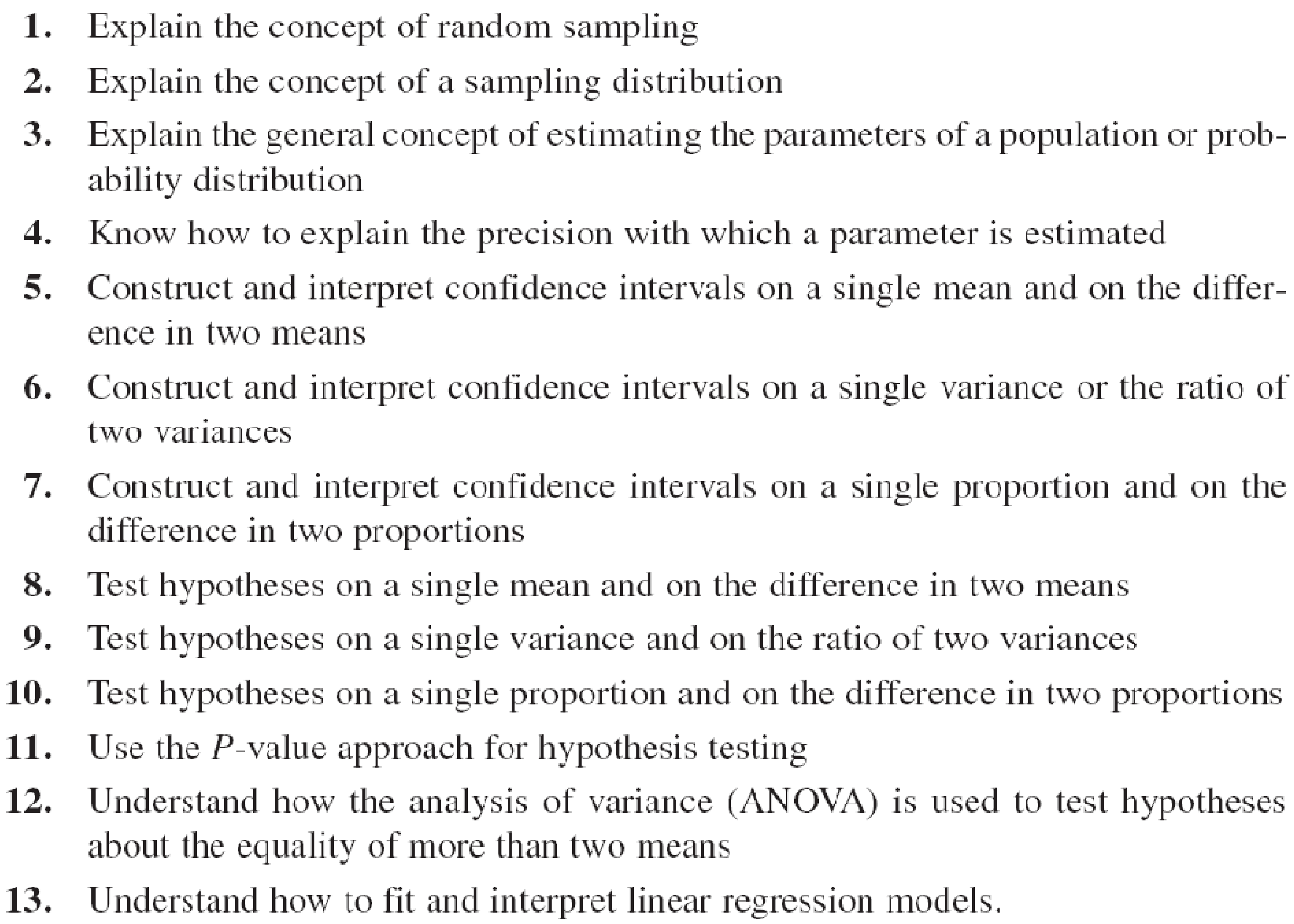
Other Diagnostic Tools

- Standardized and Studentized residuals
- *R*-student – an outlier diagnostic
- The PRESS statistic
- R^2 for prediction based on PRESS – a measure of how well the model will predict new data
- Measure of leverage – hat diagonals
- Cook's distance – a measure of influence

Important Terms and Concepts

Alternative hypothesis
Analysis of variance (ANOVA)
Binomial distribution
Checking assumptions for statistical inference procedures
Chi-square distribution
Confidence interval
Confidence intervals on means, known variance(s)
Confidence intervals on means, unknown variance(s)
Confidence intervals on proportions
Confidence intervals on the variance of a normal distribution
Confidence intervals on the variances of two normal distributions
Critical region for a test statistic
F-distribution
Hypothesis testing
Least square estimator
Linear statistical model
Minimum variance estimator
Null hypothesis
P-value
P-value approach
Parameters of a distribution
Point estimator
Poisson distribution
Pooled estimator
Power of a statistical test
Random sample
Regression model
Residual analysis
Sampling distribution
Scaled residuals
Statistic
t-distribution
Test statistic
Tests of hypotheses on means, known variance(s)
Tests of hypotheses on means, unknown variance(s)
Tests of hypotheses on proportions
Tests of hypotheses on the variance of a normal distribution
Tests of hypotheses on the variances of two normal distributions
Type I error
Type II error
Unbiased estimator

Learning Objectives



1. Explain the concept of random sampling
2. Explain the concept of a sampling distribution
3. Explain the general concept of estimating the parameters of a population or probability distribution
4. Know how to explain the precision with which a parameter is estimated
5. Construct and interpret confidence intervals on a single mean and on the difference in two means
6. Construct and interpret confidence intervals on a single variance or the ratio of two variances
7. Construct and interpret confidence intervals on a single proportion and on the difference in two proportions
8. Test hypotheses on a single mean and on the difference in two means
9. Test hypotheses on a single variance and on the ratio of two variances
10. Test hypotheses on a single proportion and on the difference in two proportions
11. Use the *P*-value approach for hypothesis testing
12. Understand how the analysis of variance (ANOVA) is used to test hypotheses about the equality of more than two means
13. Understand how to fit and interpret linear regression models.