10

Statistical Inference for Two Samples

CHAPTER OUTLINE

LEARNING OBJECTIVES

After careful study of this chapter, you should be able to do the following:

1. Structure comparative experiments involving two samples as hypothesis tests
2. Test hypotheses and construct confidence intervals on the difference in means of two normal distributions

3. Test hypotheses and construct confidence intervals on the ratio of the variances or standard deviations of two normal distributions

4. Test hypotheses and construct confidence intervals on the difference in two population proportions

5. Use the *P*-value approach for making decisions in hypotheses tests

6. Compute power, type II error probability, and make sample size decisions for two-sample tests on means, variances, and proportions

7. Explain and use the relationship between confidence intervals and hypothesis tests

CD MATERIAL

8. Use the Fisher-Irwin test to compare two population proportions when the normal approximation to the binomial distribution does not apply

---

Answers for many odd numbered exercises are at the end of the book. Answers to exercises whose numbers are surrounded by a box can be accessed in the e-Text by clicking on the box. Complete worked solutions to certain exercises are also available in the e-Text. These are indicated in the Answers to Selected Exercises section by a box around the exercise number. Exercises are also available for some of the text sections that appear on CD only. These exercises may be found within the e-Text immediately following the section they accompany.

10-1 INTRODUCTION

The previous chapter presented hypothesis tests and confidence intervals for a single population parameter (the mean μ, the variance σ^2, or a proportion p). This chapter extends those results to the case of two independent populations.

The general situation is shown in Fig. 10-1. Population 1 has mean μ_1 and variance σ_1^2, while population 2 has mean μ_2 and variance σ_2^2. Inferences will be based on two random samples of sizes n1 and n2, respectively. That is, X11, X12, p , X_{1n_1} is a random sample of n1 observations from population 1, and X21, X22, p , X_{2n_2} is a random sample of n2 observations from population 2. Most of the practical applications of the procedures in this chapter arise in the context of **simple comparative experiments** in which the objective is to study the difference in the parameters of the two populations.

10-2 INFERENCE FOR A DIFFERENCE IN MEANS OF TWO NORMAL DISTRIBUTIONS, VARIANCES KNOWN

In this section we consider statistical inferences on the difference in means $\mu_1 - \mu_2$ of two normal distributions, where the variances σ_1^2 and σ_2^2 are known. The assumptions for this section are summarized as follows.

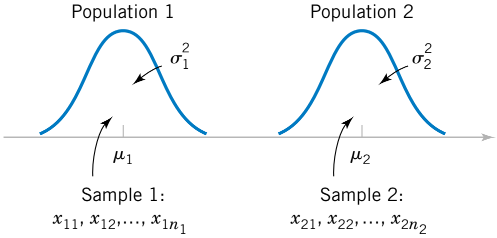

Figure 10-1 Two independent populations.

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.

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

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

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

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{10-1}$$

has a $N(0, 1)$ distribution.

This result will be used to form tests of hypotheses and confidence intervals on $\mu_1 - \mu_2$. Essentially, we may think of $\mu_1 - \mu_2$ as a parameter θ, and its estimator is $\hat{\Theta} = \overline{X}_1 - \overline{X}_2$ with variance $\sigma_{\hat{\Theta}}^2 = \sigma_1^2/n_1 + \sigma_2^2/n_2$. If θ_0 is the null hypothesis value specified for θ, the test statistic will be $(\hat{\Theta} - \theta_0)/\sigma_{\hat{\Theta}}$. Notice how similar this is to the test statistic for a single mean used in Equation 9-8 of Chapter 9.

10-2.1 Hypothesis Tests for a Difference in Means, Variances Known

We now consider hypothesis testing on the difference in the means $\mu_1 - \mu_2$ of two normal populations. Suppose that we are interested in testing that the difference in means $\mu_1 - \mu_2$ is equal to a specified value Δ_0. Thus, the null hypothesis will be stated as H_0: $\mu_1 - \mu_2 = \Delta_0$. Obviously, in many cases, we will specify $\Delta_0 = 0$ so that we are testing the equality of two means (i.e., H_0: $\mu_1 = \mu_2$). The appropriate test statistic would be found by replacing $\mu_1 - \mu_2$ in Equation 10-1 by Δ_0, and this test statistic would have a standard normal distribution under H_0. That is, the standard normal distribution is the **reference distribution** for the test statistic. Suppose that the alternative hypothesis is H_1: $\mu_1 - \mu_2 \neq \Delta_0$. Now, a sample value of $\bar{x}_1 - \bar{x}_2$ that is considerably different from Δ_0 is evidence that H_1 is true. Because Z_0 has the $N(0, 1)$

distribution when H_0 is true, we would take $-z_{\alpha/2}$ and $z_{\alpha/2}$ as the boundaries of the critical region just as we did in the single-sample hypothesis-testing problem of Section 9-2.1. This would give a test with level of significance α. Critical regions for the one-sided alternatives would be located similarly. Formally, we summarize these results below.

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

Test statistic:
$$Z_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{10-2}$$

Alternative Hypotheses	Rejection Criterion
$H_1\colon \mu_1 - \mu_2 \neq \Delta_0$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1\colon \mu_1 - \mu_2 > \Delta_0$	$z_0 > z_\alpha$
$H_1\colon \mu_1 - \mu_2 < \Delta_0$	$z_0 < -z_\alpha$

EXAMPLE 10-1

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 8 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 $\bar{x}_1 = 121$ minutes and $\bar{x}_2 = 112$ minutes, respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

We apply the eight-step procedure to this problem as follows:

1. The quantity of interest is the difference in mean drying times, $\mu_1 - \mu_2$, and $\Delta_0 = 0$.
2. $H_0\colon \mu_1 - \mu_2 = 0$, or $H_0\colon \mu_1 = \mu_2$.
3. $H_1\colon \mu_1 > \mu_2$. We want to reject H_0 if the new ingredient reduces mean drying time.
4. $\alpha = 0.05$
5. The test statistic is

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

where $\sigma_1^2 = \sigma_2^2 = (8)^2 = 64$ and $n_1 = n_2 = 10$.

6. Reject $H_0\colon \mu_1 = \mu_2$ if $z_0 > 1.645 = z_{0.05}$.
7. Computations: Since $\bar{x}_1 = 121$ minutes and $\bar{x}_2 = 112$ minutes, the test statistic is

$$z_0 = \frac{121 - 112}{\sqrt{\dfrac{(8)^2}{10} + \dfrac{(8)^2}{10}}} = 2.52$$

8. Conclusion: Since $z_0 = 2.52 > 1.645$, we reject H_0: $\mu_1 = \mu_2$ 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: $\mu_1 = \mu_2$ would be rejected at any significance level $\alpha \geq 0.0059$.

When the population variances are unknown, the sample variances s_1^2 and s_2^2 can be substituted into the test statistic Equation 10-2 to produce a **large-sample test** for the difference in means. This procedure will also work well when the populations are not necessarily normally distributed. However, both n_1 and n_2 should exceed 40 for this large-sample test to be valid.

10-2.2 Choice of Sample Size

Use of Operating Characteristic Curves

The operating characteristic curves in Appendix Charts VI*a*, VI*b*, VI*c*, and VI*d* may be used to evaluate the type II error probability for the hypotheses in the display (10-2). These curves are also useful in determining sample size. Curves are provided for $\alpha = 0.05$ and $\alpha = 0.01$. For the two-sided alternative hypothesis, the abscissa scale of the operating characteristic curve in charts VI*a* and VI*b* is d, where

$$d = \frac{|\mu_1 - \mu_2 - \Delta_0|}{\sqrt{\sigma_1^2 + \sigma_2^2}} = \frac{|\Delta - \Delta_0|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{10-3}$$

and one must choose equal sample sizes, say, $n = n_1 = n_2$. The one-sided alternative hypotheses require the use of Charts VI*c* and VI*d*. For the one-sided alternatives H_1: $\mu_1 - \mu_2 > \Delta_0$ or H_1: $\mu_1 - \mu_2 < \Delta_0$, the abscissa scale is also given by

$$d = \frac{|\mu_1 - \mu_2 - \Delta_0|}{\sqrt{\sigma_1^2 + \sigma_2^2}} = \frac{|\Delta - \Delta_0|}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

It is not unusual to encounter problems where the costs of collecting data differ substantially between the two populations, or where one population variance is much greater than the other. In those cases, we often use unequal sample sizes. If $n_1 \neq n_2$, the operating characteristic curves may be entered with an *equivalent* value of n computed from

$$n = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2/n_1 + \sigma_2^2/n_2} \tag{10-4}$$

If $n_1 \neq n_2$, and their values are fixed in advance, Equation 10-4 is used directly to calculate n, and the operating characteristic curves are entered with a specified d to obtain β. If we are given d and it is necessary to determine n_1 and n_2 to obtain a specified β, say, β^*, we guess at trial values of n_1 and n_2, calculate n in Equation 10-4, and enter the curves with the specified value of d to find β. If $\beta = \beta^*$, the trial values of n_1 and n_2 are satisfactory. If $\beta \neq \beta^*$, adjustments to n_1 and n_2 are made and the process is repeated.

EXAMPLE 10-2 Consider the paint drying time experiment from Example 10-1. If the true difference in mean drying times is as much as 10 minutes, find the sample sizes required to detect this difference with probability at least 0.90.

The appropriate value of the abscissa parameter is (since $\Delta_0 = 0$, and $\Delta = 10$)

$$d = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} = \frac{10}{\sqrt{8^2 + 8^2}} = 0.88$$

and since the detection probability or power of the test must be at least 0.9, with $\alpha = 0.05$, we find from Appendix Chart VI*c* that $n = n_1 = n_2 \simeq 11$.

Sample Size Formulas

It is also possible to obtain formulas for calculating the sample sizes directly. Suppose that the null hypothesis H_0: $\mu_1 - \mu_2 = \Delta_0$ is false and that the true difference in means is $\mu_1 - \mu_2 = \Delta$, where $\Delta > \Delta_0$. One may find formulas for the sample size required to obtain a specific value of the type II error probability β for a given difference in means Δ and level of significance α.

For the two-sided alternative hypothesis with significance level α, the sample size $n_1 = n_2 = n$ required to detect a true difference in means of Δ with power at least $1 - \beta$ is

$$n \simeq \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2} \qquad (10\text{-}5)$$

This approximation is valid when $\Phi(-z_{\alpha/2} - (\Delta - \Delta_0)\sqrt{n}/\sqrt{\sigma_1^2 + \sigma_2^2})$ is small compared to β.

For a one-sided alternative hypothesis with significance level α, the sample size $n_1 = n_2 = n$ required to detect a true difference in means of $\Delta(\neq \Delta_0)$ with power at least $1 - \beta$ is

$$n = \frac{(z_\alpha + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2} \qquad (10\text{-}6)$$

The derivation of Equations 10-5 and 10-6 closely follows the single-sample case in Section 9-2.3. For example, to obtain Equation 10-6, we first write the expression for the β-error for the two-sided alternate, which is

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\Delta - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \right) - \Phi\left(-z_{\alpha/2} - \frac{\Delta - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \right)$$

where Δ is the true difference in means of interest. Then by following a procedure similar to that used to obtain Equation 9-17, the expression for β can be obtained for the case where $n = n_1 = n_2$.

EXAMPLE 10-3 To illustrate the use of these sample size equations, consider the situation described in Example 10-1, and suppose that if the true difference in drying times is as much as 10 minutes, we want to detect this with probability at least 0.90. Under the null hypothesis, $\Delta_0 = 0$. We have a one-sided alternative hypothesis with $\Delta = 10$, $\alpha = 0.05$ (so $z_\alpha = z_{0.05} = 1.645$), and since the power is 0.9, $\beta = 0.10$ (so $z_\beta = z_{0.10} = 1.28$). Therefore we may find the required sample size from Equation 10-6 as follows:

$$n = \frac{(z_\alpha + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2} = \frac{(1.645 + 1.28)^2[(8)^2 + (8)^2]}{(10 - 0)^2} = 11$$

This is exactly the same as the result obtained from using the O.C. curves.

10-2.3 Identifying Cause and Effect

Engineers and scientists are often interested in comparing two different conditions to determine whether either condition produces a significant effect on the response that is observed. These conditions are sometimes called **treatments.** Example 10-1 illustrates such a situation; the two different treatments are the two paint formulations, and the response is the drying time. The purpose of the study is to determine whether the new formulation results in a significant effect—reducing drying time. In this situation, the product developer (the experimenter) randomly assigned 10 test specimens to one formulation and 10 test specimens to the other formulation. Then the paints were applied to the test specimens in random order until all 20 specimens were painted. This is an example of a **completely randomized experiment.**

When statistical significance is observed in a randomized experiment, the experimenter can be confident in the conclusion that it was the difference in treatments that resulted in the difference in response. That is, we can be confident that a cause-and-effect relationship has been found.

Sometimes the objects to be used in the comparison are not assigned at random to the treatments. For example, the September 1992 issue of *Circulation* (a medical journal published by the American Heart Association) reports a study linking high iron levels in the body with increased risk of heart attack. The study, done in Finland, tracked 1931 men for five years and showed a statistically significant effect of increasing iron levels on the incidence of heart attacks. In this study, the comparison was not performed by randomly selecting a sample of men and then assigning some to a "low iron level" treatment and the others to a "high iron level" treatment. The researchers just tracked the subjects over time. Recall from Chapter 1 that this type of study is called an **observational study.**

It is difficult to identify causality in observational studies, because the observed statistically significant difference in response between the two groups may be due to some other underlying factor (or group of factors) that was not equalized by randomization and not due to the treatments. For example, the difference in heart attack risk could be attributable to the difference in iron levels, or to other underlying factors that form a reasonable explanation for the observed results—such as cholesterol levels or hypertension.

The difficulty of establishing causality from observational studies is also seen in the smoking and health controversy. Numerous studies show that the incidence of lung cancer and other respiratory disorders is higher among smokers than nonsmokers. However, establishing cause and effect here has proven enormously difficult. Many individuals had decided to smoke long before the start of the research studies, and many factors other than smoking could have a role in contracting lung cancer.

10-2.4 Confidence Interval on a Difference in Means, Variances Known

The $100(1 - \alpha)\%$ confidence interval on the difference in two means $\mu_1 - \mu_2$ when the variances are known can be found directly from results given previously in this section. Recall that $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample of n_1 observations from the first population and $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample of n_2 observations from the second population. The difference in sample means $\overline{X}_1 - \overline{X}_2$ is a point estimator of $\mu_1 - \mu_2$, and

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

has a standard normal distribution if the two populations are normal or is approximately standard normal if the conditions of the central limit theorem apply, respectively. This implies that $P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$, or

$$P\left[-z_{\alpha/2} \le \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \le z_{\alpha/2}\right] = 1 - \alpha$$

This can be rearranged as

$$P\left(\overline{X}_1 - \overline{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

Therefore, the $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is defined as follows.

Definition

If $\bar{x}_1$ and $\bar{x}_2$ are the means of independent random samples of sizes n_1 and n_2 from two independent normal populations with known variances σ_1^2 and σ_2^2, respectively, a **$100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$** is

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (10\text{-}7)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

The confidence level $1 - \alpha$ is exact when the populations are normal. For nonnormal populations, the confidence level is approximately valid for large sample sizes.

EXAMPLE 10-4 Tensile strength tests were performed on two different grades of aluminum spars used in manufacturing the wing of a commercial transport aircraft. From past experience with the spar manufacturing process and the testing procedure, the standard deviations of tensile strengths are assumed to be known. The data obtained are as follows: $n_1 = 10$, $\bar{x}_1 = 87.6$, $\sigma_1 = 1$, $n_2 = 12$, $\bar{x}_2 = 74.5$, and $\sigma_2 = 1.5$. If μ_1 and μ_2 denote the true mean tensile strengths for the two grades of spars, we may find a 90% confidence interval on the difference in mean strength $\mu_1 - \mu_2$ as follows:

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$87.6 - 74.5 - 1.645\sqrt{\frac{(1)^2}{10} + \frac{(1.5)^3}{12}} \le \mu_1 - \mu_2 \le 87.6 - 74.5 + 1.645\sqrt{\frac{(1^2)}{10} + \frac{(1.5)^2}{12}}$$

Therefore, the 90% confidence interval on the difference in mean tensile strength (in kilograms per square millimeter) is

$$12.22 \le \mu_1 - \mu_2 \le 13.98 \text{ (in kilograms per square millimeter)}$$

Notice that the confidence interval does not include zero, implying that the mean strength of aluminum grade 1 (μ_1) exceeds the mean strength of aluminum grade 2 (μ_2). In fact, we can state that we are 90% confident that the mean tensile strength of aluminum grade 1 exceeds that of aluminum grade 2 by between 12.22 and 13.98 kilograms per square millimeter.

Choice of Sample Size

If the standard deviations σ_1 and σ_2 are known (at least approximately) and the two sample sizes n_1 and n_2 are equal ($n_1 = n_2 = n$, say), we can determine the sample size required so that the error in estimating $\mu_1 - \mu_2$ by $\bar{x}_1 - \bar{x}_2$ will be less than E at $100(1 - \alpha)\%$ confidence. The required sample size from each population is

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 (\sigma_1^2 + \sigma_2^2) \qquad (10\text{-}8)$$

Remember to round up if n is not an integer. This will ensure that the level of confidence does not drop below $100(1 - \alpha)\%$.

One-Sided Confidence Bounds

One-sided confidence bounds on $\mu_1 - \mu_2$ may also be obtained. A $100(1 - \alpha)\%$ upper-confidence bound on $\mu_1 - \mu_2$ is

$$\mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + z_\alpha\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (10\text{-}9)$$

and a $100(1-\alpha)\%$ lower-confidence bound is

$$\bar{x}_1 - \bar{x}_2 - z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \qquad (10\text{-}10)$$

EXERCISES FOR SECTION 10-2

10-1. Two machines are used for filling plastic bottles with a net volume of 16.0 ounces. The fill volume can be assumed normal, with standard deviation $\sigma_1 = 0.020$ and $\sigma_2 = 0.025$ ounces. A member of the quality engineering staff suspects that both machines fill to the same mean net volume, whether or not this volume is 16.0 ounces. A random sample of 10 bottles is taken from the output of each machine.

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

(a) Do you think the engineer is correct? Use $\alpha = 0.05$.
(b) What is the *P*-value for this test?
(c) What is the power of the test in part (a) for a true difference in means of 0.04?
(d) Find a 95% confidence interval on the difference in means. Provide a practical interpretation of this interval.
(e) Assuming equal sample sizes, what sample size should be used to assure that $\beta = 0.05$ if the true difference in means is 0.04? Assume that $\alpha = 0.05$.

10-2. Two types of plastic are suitable for use by an electronics component manufacturer. The breaking strength of this plastic is important. It is known that $\sigma_1 = \sigma_2 = 1.0$ psi. From a random sample of size $n_1 = 10$ and $n_2 = 12$, we obtain $\bar{x}_1 = 162.5$ and $\bar{x}_2 = 155.0$. The company will not adopt plastic 1 unless its mean breaking strength exceeds that of plastic 2 by at least 10 psi. Based on the sample information, should it use plastic 1? Use $\alpha = 0.05$ in reaching a decision.

10-3. Reconsider the situation in Exercise 10-2. Suppose that the true difference in means is really 12 psi. Find the power of the test assuming that $\alpha = 0.05$. If it is really important to detect this difference, are the sample sizes employed in Exercise 10-2 adequate, in your opinion?

10-4. The burning rates of two different solid-fuel propellants used in aircrew escape systems are being studied. It is known that both propellants have approximately the same standard deviation of burning rate; that is $\sigma_1 = \sigma_2 = 3$ centimeters per second. Two random samples of $n_1 = 20$ and $n_2 = 20$ specimens are tested; the sample mean burning rates are $\bar{x}_1 = 18$ centimeters per second and $\bar{x}_2 = 24$ centimeters per second.
(a) Test the hypothesis that both propellants have the same mean burning rate. Use $\alpha = 0.05$.
(b) What is the *P*-value of the test in part (a)?
(c) What is the β-error of the test in part (a) if the true difference in mean burning rate is 2.5 centimeters per second?
(d) Construct a 95% confidence interval on the difference in means $\mu_1 - \mu_2$. What is the practical meaning of this interval?

10-5. Two machines are used to fill plastic bottles with dishwashing detergent. The standard deviations of fill volume are known to be $\sigma_1 = 0.10$ fluid ounces and $\sigma_2 = 0.15$ fluid ounces for the two machines, respectively. Two random samples of $n_1 = 12$ bottles from machine 1 and $n_2 = 10$ bottles from machine 2 are selected, and the sample mean fill volumes are $\bar{x}_1 = 30.87$ fluid ounces and $\bar{x}_2 = 30.68$ fluid ounces. Assume normality.
(a) Construct a 90% two-sided confidence interval on the mean difference in fill volume. Interpret this interval.
(b) Construct a 95% two-sided confidence interval on the mean difference in fill volume. Compare and comment on the width of this interval to the width of the interval in part (a).
(c) Construct a 95% upper-confidence interval on the mean difference in fill volume. Interpret this interval.

10-6. Reconsider the situation described in Exercise 10-5.
(a) Test the hypothesis that both machines fill to the same mean volume. Use $\alpha = 0.05$.
(b) What is the *P*-value of the test in part (a)?
(c) If the β-error of the test when the true difference in fill volume is 0.2 fluid ounces should not exceed 0.1, what sample sizes must be used? Use $\alpha = 0.05$.

10-7. Two different formulations of an oxygenated motor fuel are being tested to study their road octane numbers. The variance of road octane number for formulation 1 is $\sigma_1^2 = 1.5$, and for formulation 2 it is $\sigma_2^2 = 1.2$. Two random samples of size $n_1 = 15$ and $n_2 = 20$ are tested, and the mean road octane numbers observed are $\bar{x}_1 = 89.6$ and $\bar{x}_2 = 92.5$. Assume normality.
(a) Construct a 95% two-sided confidence interval on the difference in mean road octane number.
(b) If formulation 2 produces a higher road octane number than formulation 1, the manufacturer would like to detect

it. Formulate and test an appropriate hypothesis, using $\alpha = 0.05$.

(c) What is the *P*-value for the test you conducted in part (b)?

10-8. Consider the situation described in Exercise 10-4. What sample size would be required in each population if we wanted the error in estimating the difference in mean burning rates to be less than 4 centimeters per second with 99% confidence?

10-9. Consider the road octane test situation described in Exercise 10-7. What sample size would be required in each population if we wanted to be 95% confident that the error in estimating the difference in mean road octane number is less than 1?

10-10. A polymer is manufactured in a batch chemical process. Viscosity measurements are normally made on each batch, and long experience with the process has indicated that the variability in the process is fairly stable with $\sigma = 20$. Fifteen batch viscosity measurements are given as follows: 724, 718, 776, 760, 745, 759, 795, 756, 742, 740, 761, 749, 739, 747, 742. A process change is made which involves switching the type of catalyst used in the process. Following the process change, eight batch viscosity measurements are taken: 735, 775, 729, 755, 783, 760, 738, 780. Assume that process variability is unaffected by the catalyst change. Find a 90% confidence interval on the difference in mean batch viscosity resulting from the process change.

10-11. The concentration of active ingredient in a liquid laundry detergent is thought to be affected by the type of catalyst used in the process. The standard deviation of active concentration is known to be 3 grams per liter, regardless of the catalyst type. Ten observations on concentration are taken with each catalyst, and the data follow:

Catalyst 1: 57.9, 66.2, 65.4, 65.4, 65.2, 62.6, 67.6, 63.7, 67.2, 71.0

Catalyst 2: 66.4, 71.7, 70.3, 69.3, 64.8, 69.6, 68.6, 69.4, 65.3, 68.8

(a) Find a 95% confidence interval on the difference in mean active concentrations for the two catalysts.

(b) Is there any evidence to indicate that the mean active concentrations depend on the choice of catalyst? Base your answer on the results of part (a).

10-12. Consider the polymer batch viscosity data in Exercise 10-10. If the difference in mean batch viscosity is 10 or less, the manufacturer would like to detect it with a high probability.

(a) Formulate and test an appropriate hypothesis using $\alpha = 0.10$. What are your conclusions?

(b) Calculate the *P*-value for this test.

(c) Compare the results of parts (a) and (b) to the length of the 90% confidence interval obtained in Exercise 10-10 and discuss your findings.

10-13. For the laundry detergent problem in Exercise 10-11, test the hypothesis that the mean active concentrations are the same for both types of catalyst. Use $\alpha = 0.05$. What is the *P*-value for this test? Compare your answer to that found in part (b) of Exercise 10-11, and comment on why they are the same or different.

10-14. Reconsider the laundry detergent problem in Exercise 10-11. Suppose that the true mean difference in active concentration is 5 grams per liter. What is the power of the test to detect this difference if $\alpha = 0.05$? If this difference is really important, do you consider the sample sizes used by the experimenter to be adequate?

10-15. Consider the polymer viscosity data in Exercise 10-10. Does the assumption of normality seem reasonable for both samples?

10-16. Consider the concentration data in Exercise 10-11. Does the assumption of normality seem reasonable?

10-3 INFERENCE FOR THE DIFFERENCE IN MEANS OF TWO NORMAL DISTRIBUTIONS, VARIANCES UNKNOWN

We now extend the results of the previous section to the difference in means of the two distributions in Fig. 10-1 when the variances of both distributions σ_1^2 and σ_2^2 are unknown. If the sample sizes n_1 and n_2 exceed 40, the normal distribution procedures in Section 10-2 could be used. However, when small samples are taken, we will assume that the populations are normally distributed and base our hypotheses tests and confidence intervals on the *t* distribution. This nicely parallels the case of inference on the mean of a single sample with unknown variance.

10-3.1 Hypotheses Tests for a Difference in Means, Variances Unknown

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 and in Section 9-3, 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: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

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

$$H_0\colon \mu_1 - \mu_2 = \Delta_0$$
$$H_1\colon \mu_1 - \mu_2 \neq \Delta_0 \qquad (10\text{-}11)$$

Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be a random sample of n_1 observations from the first population and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be a random sample of n_2 observations from the second population. Let $\overline{X}_1$, $\overline{X}_2$, S_1^2, and S_2^2 be the sample means and sample variances, respectively. Now the expected value of the difference in sample means $\overline{X}_1 - \overline{X}_2$ is $E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$, so $\overline{X}_1 - \overline{X}_2$ is an unbiased estimator of the difference in means. The variance of $\overline{X}_1 - \overline{X}_2$ is

$$V(\overline{X}_1 - \overline{X}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

It seems reasonable to combine the two sample variances S_1^2 and S_2^2 to form an estimator of σ^2. The **pooled estimator** of σ^2 is defined as follows.

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

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \qquad (10\text{-}12)$$

It is easy to see that the pooled estimator S_p^2 can be written as

$$S_p^2 = \frac{n_1 - 1}{n_1 + n_2 - 2}S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2}S_2^2 = wS_1^2 + (1 - w)S_2^2$$

where $0 < w \leq 1$. Thus S_p^2 is a **weighted average** of the two sample variances S_1^2 and S_2^2, where the weights w and $1 - w$ depend on the two sample sizes n_1 and n_2. Obviously, if $n_1 = n_2 = n$, $w = 0.5$ and S_p^2 is just the arithmetic average of S_1^2 and S_2^2. If $n_1 = 10$ and $n_2 = 20$ (say), $w = 0.32$ and $1 - w = 0.68$. The first sample contributes $n_1 - 1$ degrees of freedom to S_p^2 and the second sample contributes $n_2 - 1$ degrees of freedom. Therefore, S_p^2 has $n_1 + n_2 - 2$ degrees of freedom.

Now we know that

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

has a $N(0, 1)$ distribution. Replacing σ by S_p gives the following.

Given the assumptions of this section, the quantity

$$T = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{10-13}$$

has a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

The use of this information to test the hypotheses in Equation 10-11 is now straightforward: simply replace $\mu_1 - \mu_2$ by Δ_0, and the resulting **test statistic** has a t distribution with $n_1 + n_2 - 2$ degrees of freedom under H_0: $\mu_1 - \mu_2 = \Delta_0$. Therefore, the reference distribution for the test statistic is the t distribution with $n_1 + n_2 - 2$ degrees of freedom. The location of the critical region for both two- and one-sided alternatives parallels those in the one-sample case. Because a pooled estimate of variance is used, the procedure is often called the **pooled *t*-test.**

Definition: The Two-Sample or Pooled *t*-Test*

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

Test statistic:
$$T_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{10-14}$$

Alternative Hypothesis	Rejection Criterion
H_1: $\mu_1 - \mu_2 \neq \Delta_0$	$t_0 > t_{\alpha/2, n_1+n_2-2}$ or $t_0 < -t_{\alpha/2, n_1+n_2-2}$
H_1: $\mu_1 - \mu_2 > \Delta_0$	$t_0 > t_{\alpha, n_1+n_2-2}$
H_1: $\mu_1 - \mu_2 < \Delta_0$	$t_0 < -t_{\alpha, n_1+n_2-2}$

EXAMPLE 10-5 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. A test is run in the pilot plant and results in the data shown in Table 10-1. Is there any difference between the mean yields? Use $\alpha = 0.05$, and assume equal variances.

The solution using the eight-step hypothesis-testing procedure is as follows:

1. The parameters of interest are μ_1 and μ_2, the mean process yield using catalysts 1 and 2, respectively, and we want to know if $\mu_1 - \mu_2 = 0$.
2. H_0: $\mu_1 - \mu_2 = 0$, or H_0: $\mu_1 = \mu_2$

*While we have given the development of this procedure for the case where the sample sizes could be different, there is an advantage to using equal sample sizes $n_1 = n_2 = n$. When the sample sizes are the same from both populations, the t-test is more robust to the assumption of equal variances. Please see Section 10-3.2 on the CD.

Table 10-1 Catalyst Yield Data, Example 10-5

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
	$\bar{x}_1 = 92.255$	$\bar{x}_2 = 92.733$
	$s_1 = 2.39$	$s_2 = 2.98$

3. H_1: $\mu_1 \neq \mu_2$

4. $\alpha = 0.05$

5. The test statistic is

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

6. Reject H_0 if $t_0 > t_{0.025,14} = 2.145$ or if $t_0 < -t_{0.025,14} = -2.145$.

7. Computations: From Table 10-1 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

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

8. Conclusions: Since $-2.145 < t_0 = -0.35 < 2.145$, 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.

A *P*-value could also be used for decision making in this example. From Appendix Table IV we find that $t_{0.40,14} = 0.258$ and $t_{0.25,14} = 0.692$. Therefore, since $0.258 < 0.35 < 0.692$, we conclude that lower and upper bounds on the *P*-value are $0.50 < P < 0.80$. Therefore, since the *P*-value exceeds $\alpha = 0.05$, the null hypothesis cannot be rejected.

The Minitab two-sample t-test and confidence interval procedure for Example 10-5 follows:

Two-Sample T-Test and CI: Cat 1, Cat 2

Two-sample T for Cat 1 vs Cat 2

	N	Mean	StDev	SE Mean
Cat 1	8	92.26	2.39	0.84
Cat 2	8	92.73	2.99	1.1

Difference = mu Cat 1 − mu Cat 2
Estimate for difference: −0.48
95% CI for difference: (−3.37, 2.42)
T-Test of difference = 0 (vs not =): T-Value = −0.35 P-Value = 0.730 DF = 14
Both use Pooled StDev = 2.70

Notice that the numerical results are essentially the same as the manual computations in Example 10-5. The P-value is reported as $P = 0.73$. The two-sided CI on $\mu_1 - \mu_2$ is also reported. We will give the computing formula for the CI in Section 10-3.3. Figure 10-2 shows the normal probability plot of the two samples of yield data and comparative box plots. The normal probability plots indicate that there is no problem with the normality assumption. Furthermore, both straight lines have similar slopes, providing some verification of the assumption of equal variances. The comparative box plots indicate that there is no obvious difference in the two catalysts, although catalyst 2 has slightly greater sample variability.

Case 2: $\sigma_1^2 \neq \sigma_2^2$

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, the statistic

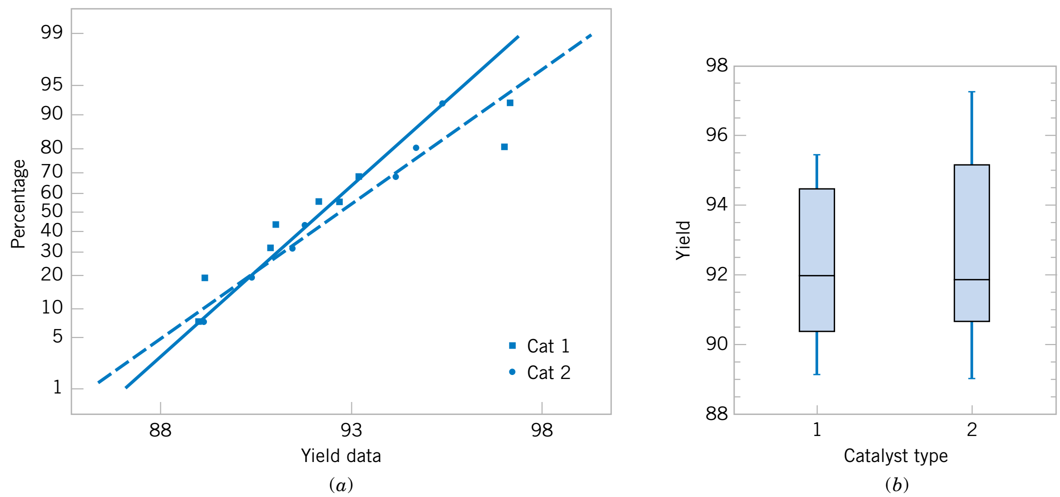

Figure 10-2 Normal probability plot and comparative box plot for the catalyst yield data in Example 10-5. (a) Normal probability plot, (b) Box plots.

$$T_0^* = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \qquad (10\text{-}15)$$

is distributed approximately as t with degrees of freedom given by

$$v = \frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_2 - 1}} \qquad (10\text{-}16)$$

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.

EXAMPLE 10-6

Arsenic concentration in public drinking water supplies is a potential health risk. An article in the *Arizona Republic* (Sunday, May 27, 2001) reported drinking water arsenic concentrations in parts per billion (ppb) for 10 methropolitan Phoenix communities and 10 communities in rural Arizona. The data follow:

Metro Phoenix ($\bar{x}_1 = 12.5, s_1 = 7.63$)	Rural Arizona ($\bar{x}_2 = 27.5, s_2 = 15.3$)
Phoenix, 3	Rimrock, 48
Chandler, 7	Goodyear, 44
Gilbert, 25	New River, 40
Glendale, 10	Apachie Junction, 38
Mesa, 15	Buckeye, 33
Paradise Valley, 6	Nogales, 21
Peoria, 12	Black Canyon City, 20
Scottsdale, 25	Sedona, 12
Tempe, 15	Payson, 1
Sun City, 7	Casa Grande, 18

We wish to determine it there is any difference in mean arsenic concentrations between metropolitan Phoenix communities and communities in rural Arizona. Figure 10-3 shows a normal probability plot for the two samples of arsenic concentration. The assumption of normality appears quite reasonable, but since the slopes of the two straight lines are very different, it is unlikely that the population variances are the same.

Applying the eight-step procedure gives the following:

1. The parameters of interest are the mean arsenic concentrations for the two geographic regions, say, μ_1 and μ_2, and we are interested in determining whether $\mu_1 - \mu_2 = 0$.
2. H_0: $\mu_1 - \mu_2 = 0$, or H_0: $\mu_1 = \mu_2$

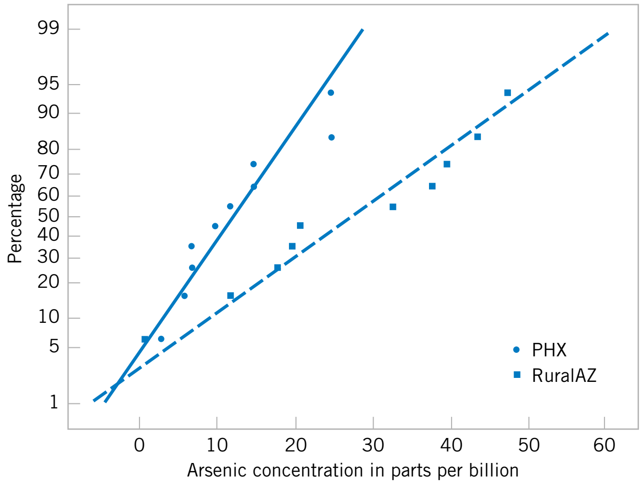

Figure 10-3 Normal probability plot of the arsenic concentration data from Example 10-6.

3. H_1: $\mu_1 \neq \mu_2$

4. $\alpha = 0.05$ (say)

5. The test statistic is

$$t_0^* = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

6. The degrees of freedom on t_0^* are found from Equation 10-16 as

$$v = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{\left[\dfrac{(7.63)^2}{10} + \dfrac{(15.3)^2}{10}\right]^2}{\dfrac{[(7.63)^2/10]^2}{9} + \dfrac{[(15.3)^2/10]^2}{9}} = 13.2 \simeq 13$$

Therefore, using $\alpha = 0.05$, we would reject H_0: $\mu_1 = \mu_2$ if $t_0^* > t_{0.025,13} = 2.160$ or if $t_0^* < -t_{0.025,13} = -2.160$

7. Computations: Using the sample data we find

$$t_0^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{12.5 - 27.5}{\sqrt{\dfrac{(7.63)^2}{10} + \dfrac{(15.3)^2}{10}}} = -2.77$$

8. Conclusions: Because $t_0^* = -2.77 < t_{0.025,13} = -2.160$, we reject the null hypothesis. Therefore, there is evidence to conclude that mean arsenic concentration in the drinking water in rural Arizona is different from the mean arsenic concentration in metropolitan Phoenix drinking water. Furthermore, the mean arsenic concentration is higher in rural Arizona communities. The P-value for this test is approximately $P = 0.016$.

The Minitab output for this example follows:

Two-Sample T-Test and CI: PHX, RuralAZ

Two-sample T for PHX vs RuralAZ

	N	Mean	StDev	SE Mean
PHX	10	12.50	7.63	2.4
RuralAZ	10	27.5	15.3	4.9

Difference = mu PHX − mu RuralAZ
Estimate for difference: −15.00
95% CI for difference: (−26.71, −3.29)
T-Test of difference = 0 (vs not =): T-Value = −2.77 P-Value = 0.016 DF = 13

The numerical results from Minitab exactly match the calculations from Example 10-6. Note that a two-sided 95% CI on $\mu_1 - \mu_2$ is also reported. We will discuss its computation in Section 10-3.4; however, note that the interval does not include zero. Indeed, the upper 95% of confidence limit is -3.29 ppb, well below zero, and the mean observed difference is $\bar{x}_1 - \bar{x}_2 = 12 - 5 - 17.5 = -15$ ppb.

10-3.2 More about the Equal Variance Assumption (CD Only)

10-3.3 Choice of Sample Size

The operating characteristic curves in Appendix Charts VI*e*, VI*f*, VI*g*, and VI*h* are used to evaluate the type II error for the case where $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Unfortunately, when $\sigma_1^2 \neq \sigma_2^2$, the distribution of T_0^* is unknown if the null hypothesis is false, and no operating characteristic curves are available for this case.

For the two-sided alternative H_1: $\mu_1 - \mu_2 = \Delta \neq \Delta_0$, when $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and $n_1 = n_2 = n$, Charts VI*e* and VI*f* are used with

$$d = \frac{|\Delta - \Delta_0|}{2\sigma} \qquad (10\text{-}17)$$

where Δ is the true difference in means that is of interest. To use these curves, they must be entered with the sample size $n^* = 2n - 1$. For the one-sided alternative hypothesis, we use Charts VI*g* and VI*h* and define d and Δ as in Equation 10-17. It is noted that the parameter d is a function of σ, which is unknown. As in the single-sample t-test, we may have to rely on a prior estimate of σ or use a subjective estimate. Alternatively, we could define the differences in the mean that we wish to detect relative to σ.

EXAMPLE 10-7

Consider the catalyst experiment in Example 10-5. Suppose that, if catalyst 2 produces a mean yield that differs from the mean yield of catalyst 1 by 4.0%, we would like to reject the null hypothesis with probability at least 0.85. What sample size is required?

Using $s_p = 2.70$ as a rough estimate of the common standard deviation σ, we have $d = |\Delta|/2\sigma = |4.0|/[(2)(2.70)] = 0.74$. From Appendix Chart VI*e* with $d = 0.74$ and $\beta = 0.15$, we find $n^* = 20$, approximately. Therefore, since $n^* = 2n - 1$,

$$n = \frac{n^* + 1}{2} = \frac{20 + 1}{2} = 10.5 \simeq 11(\text{say})$$

and we would use sample sizes of $n_1 = n_2 = n = 11$.

Minitab will also perform power and sample size calculations for the two-sample t-test (equal variances). The output from Example 10-7 is as follows:

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

Difference	Sample Size	Target Power	Actual Power
4	10	0.8500	0.8793

The results agree fairly closely with the results obtained from the O.C. curve.

10-3.4 Confidence Interval on the Difference in Means

Case 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

To develop the confidence interval for the difference in means $\mu_1 - \mu_2$ when both variances are equal, note that the distribution of the statistic

$$T = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{10-18}$$

is the t distribution with $n_1 + n_2 - 2$ degrees of freedom. Therefore $P(-t_{\alpha/2,n_1+n_2-2} \le T \le t_{\alpha/2,n_1+n_2-2}) = 1 - \alpha$. Now substituting Equation 10-18 for T and manipulating the quantities inside the probability statement will lead to the $100(1 - \alpha)\%$ confidence interval on $\mu_1 - \mu_2$.

Definition

If $\overline{x}_1, \overline{x}_2, s_1^2$ and s_2^2 are the sample means and variances of two random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown but equal variances, then a **$100(1 - \alpha)\%$ confidence interval on the difference in means $\mu_1 - \mu_2$ is**

$$\overline{x}_1 - \overline{x}_2 - t_{\alpha/2,n_1+n_2-2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\le \mu_1 - \mu_2 \le \overline{x}_1 - \overline{x}_2 + t_{\alpha/2,n_1+n_2-2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{10-19}$$

where $s_p = \sqrt{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)}$ is the pooled estimate of the common population standard deviation, and $t_{\alpha/2,n_1+n_2-2}$ is the upper $\alpha/2$ percentage point of the t distribution with $n_1 + n_2 - 2$ degrees of freedom.

EXAMPLE 10-8 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 $s_1 = 5.0$, while 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 $s_2 = 4.0$.

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 10-12 as follows:

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

Therefore, the pooled standard deviation estimate is $s_p = \sqrt{19.52} = 4.4$. The 95% confidence interval is found using Equation 10-19:

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

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

$$
\begin{aligned}
90.0 - 87.0 - 2.069(4.4)\sqrt{\frac{1}{10} + \frac{1}{15}} &\le \mu_1 - \mu_2 \\
&\le 90.0 - 87.0 + 2.069(44)\sqrt{\frac{1}{10} + \frac{1}{15}}
\end{aligned}
$$

which reduces to

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

Notice that the 95% confidence interval 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.

Case 2: $\sigma_1^2 \neq \sigma_2^2$

In many situations it is not reasonable to assume that $\sigma_1^2 = \sigma_2^2$. When this assumption is unwarranted, we may still find a $100(1 - \alpha)\%$ confidence interval on $\mu_1 - \mu_2$ using the fact that $T^* = [\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)]/\sqrt{S_1^2/n_1 + S_2^2/n_2}$ is distributed approximately as t with degrees of freedom v given by Equation 10-16. The CI expression follows.

Definition

If $\bar{x}_1$, $\bar{x}_2$, s_1^2, and s_2^2 are the means and variances of two random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown and unequal variances, an approximate $100(1-\alpha)\%$ confidence interval on the difference in means $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (10\text{-}20)$$

where ν is given by Equation 10-16 and $t_{\alpha/2,\nu}$ is the upper $\alpha/2$ percentage point of the t distribution with ν degrees of freedom.

EXERCISES FOR SECTION 10-3

10-17. The diameter of steel rods manufactured on two different extrusion machines is being investigated. Two random samples of sizes $n_1 = 15$ and $n_2 = 17$ are selected, and the sample means and sample variances are $\bar{x}_1 = 8.73$, $s_1^2 = 0.35$, $\bar{x}_2 = 8.68$, and $s_2^2 = 0.40$, respectively. Assume that $\sigma_1^2 = \sigma_2^2$ and that the data are drawn from a normal distribution.

(a) Is there evidence to support the claim that the two machines produce rods with different mean diameters? Use $\alpha = 0.05$ in arriving at this conclusion.
(b) Find the *P*-value for the *t*-statistic you calculated in part (a).
(c) Construct a 95% confidence interval for the difference in mean rod diameter. Interpret this interval.

10-18. An article in *Fire Technology* investigated two different foam expanding agents that can be used in the nozzles of fire-fighting spray equipment. A random sample of five observations with an aqueous film-forming foam (AFFF) had a sample mean of 4.7 and a standard deviation of 0.6. A random sample of five observations with alcohol-type concentrates (ATC) had a sample mean of 6.9 and a standard deviation 0.8. Find a 95% confidence interval on the difference in mean foam expansion of these two agents. Can you draw any conclusions about which agent produces the greatest mean foam expansion? Assume that both populations are well represented by normal distributions with the same standard deviations.

10-19. Two catalysts may be used in a batch chemical process. Twelve batches were prepared using catalyst 1, resulting in an average yield of 86 and a sample standard deviation of 3. Fifteen batches were prepared using catalyst 2, and they resulted in an average yield of 89 with a standard deviation of 2. Assume that yield measurements are approximately normally distributed with the same standard deviation.

(a) Is there evidence to support a claim that catalyst 2 produces a higher mean yield than catalyst 1? Use $\alpha = 0.01$.
(b) Find a 95% confidence interval on the difference in mean yields.

10-20. The deflection temperature under load for two different types of plastic pipe is being investigated. Two random samples of 15 pipe specimens are tested, and the deflection temperatures observed are as follows (in °F):

Type 1: 206, 188, 205, 187, 194, 193, 207, 185, 189, 213, 192, 210, 194, 178, 205.

Type 2: 177, 197, 206, 201, 180, 176, 185, 200, 197, 192, 198, 188, 189, 203, 192.

(a) Construct box plots and normal probability plots for the two samples. Do these plots provide support of the assumptions of normality and equal variances? Write a practical interpretation for these plots.
(b) Do the data support the claim that the deflection temperature under load for type 2 pipe exceeds that of type 1? In reaching your conclusions, use $\alpha = 0.05$.
(c) Calculate a *P*-value for the test in part (b).
(d) Suppose that if the mean deflection temperature for type 2 pipe exceeds that of type 1 by as much as 5°F, it is important to detect this difference with probability at least 0.90. Is the choice of $n_1 = n_2 = 15$ in part (a) of this problem adequate?

10-21. In semiconductor manufacturing, wet chemical etching is often used to remove silicon from the backs of wafers prior to metalization. The etch rate is an important characteristic in this process and known to follow a normal distribution. Two different etching solutions have been compared, using two random samples of 10 wafers for each solution. The observed etch rates are as follows (in mils per minute):

Solution 1		Solution 2	
9.9	10.6	10.2	10.0
9.4	10.3	10.6	10.2
9.3	10.0	10.7	10.7
9.6	10.3	10.4	10.4
10.2	10.1	10.5	10.3

(a) Do the data support the claim that the mean etch rate is the same for both solutions? In reaching your conclusions, use $\alpha = 0.05$ and assume that both population variances are equal.
(b) Calculate a P-value for the test in part (a).
(c) Find a 95% confidence interval on the difference in mean etch rates.
(d) Construct normal probability plots for the two samples. Do these plots provide support for the assumptions of normality and equal variances? Write a practical interpretation for these plots.

10-22. Two suppliers manufacture a plastic gear used in a laser printer. The impact strength of these gears measured in foot-pounds is an important characteristic. A random sample of 10 gears from supplier 1 results in $\bar{x}_1 = 290$ and $s_1 = 12$, while another random sample of 16 gears from the second supplier results in $\bar{x}_2 = 321$ and $s_2 = 22$.
(a) Is there evidence to support the claim that supplier 2 provides gears with higher mean impact strength? Use $\alpha = 0.05$, and assume that both populations are normally distributed but the variances are not equal.
(b) What is the P-value for this test?
(c) Do the data support the claim that the mean impact strength of gears from supplier 2 is at least 25 foot-pounds higher than that of supplier 1? Make the same assumptions as in part (a).

10-23. Reconsider the situation in Exercise 10-22, part (a). Construct a confidence interval estimate for the difference in mean impact strength, and explain how this interval could be used to answer the question posed regarding supplier-to-supplier differences.

10-24. A photoconductor film is manufactured at a nominal thickness of 25 mils. The product engineer wishes to increase the mean speed of the film, and believes that this can be achieved by reducing the thickness of the film to 20 mils. Eight samples of each film thickness are manufactured in a pilot production process, and the film speed (in microjoules per square inch) is measured. For the 25-mil film the sample data result is $\bar{x}_1 = 1.15$ and $s_1 = 0.11$, while for the 20-mil film, the data yield $\bar{x}_2 = 1.06$ and $s_2 = 0.09$. Note that an increase in film speed would lower the value of the observation in microjoules per square inch.
(a) Do the data support the claim that reducing the film thickness increases the mean speed of the film? Use $\alpha = 0.10$ and assume that the two population variances are equal and the underlying population of film speed is normally distributed.
(b) What is the P-value for this test?
(c) Find a 95% confidence interval on the difference in the two means.

10-25. The melting points of two alloys used in formulating solder were investigated by melting 21 samples of each material. The sample mean and standard deviation for alloy 1 was $\bar{x}_1 = 420°F$ and $s_1 = 4°F$, while for alloy 2 they were $\bar{x}_2 = 426°F$, and $s_2 = 3°F$. Do the sample data support the claim that both alloys have the same melting point? Use $\alpha = 0.05$ and assume that both populations are normally distributed and have the same standard deviation. Find the P-value for the test.

10-26. Referring to the melting point experiment in Exercise 10-25, suppose that the true mean difference in melting points is 3°F. How large a sample would be required to detect this difference using an $\alpha = 0.05$ level test with probability at least 0.9? Use $\sigma_1 = \sigma_2 = 4$ as an initial estimate of the common standard deviation.

10-27. Two companies manufacture a rubber material intended for use in an automotive application. The part will be subjected to abrasive wear in the field application, so we decide to compare the material produced by each company in a test. Twenty-five samples of material from each company are tested in an abrasion test, and the amount of wear after 1000 cycles is observed. For company 1, the sample mean and standard deviation of wear are $\bar{x}_1 = 20$ milligrams/1000 cycles and $s_1 = 2$ milligrams/1000 cycles, while for company 2 we obtain $\bar{x}_2 = 15$ milligrams/1000 cycles and $s_2 = 8$ milligrams/1000 cycles.
(a) Do the data support the claim that the two companies produce material with different mean wear? Use $\alpha = 0.05$, and assume each population is normally distributed but that their variances are not equal.
(b) What is the P-value for this test?
(c) Do the data support a claim that the material from company 1 has higher mean wear than the material from company 2? Use the same assumptions as in part (a).

10-28. The thickness of a plastic film (in mils) on a substrate material is thought to be influenced by the temperature at which the coating is applied. A completely randomized experiment is carried out. Eleven substrates are coated at 125°F, resulting in a sample mean coating thickness of $\bar{x}_1 = 103.5$ and a sample standard deviation of $s_1 = 10.2$. Another 13 substrates are coated at 150°F, for which $\bar{x}_2 = 99.7$ and $s_2 = 20.1$ are observed. It was originally suspected that raising the process temperature would reduce mean coating thickness. Do the data support this claim? Use $\alpha = 0.01$ and assume that the two population standard deviations are not equal. Calculate an approximate P-value for this test.

10-29. Reconsider the coating thickness experiment in Exercise 10-28. How could you have answered the question posed regarding the effect of temperature on coating thickness by using a confidence interval? Explain your answer.

10-30. Reconsider the abrasive wear test in Exercise 10-27. Construct a confidence interval that will address the questions in parts (a) and (c) in that exercise.

10-31. The overall distance traveled by a golf ball is tested by hitting the ball with Iron Byron, a mechanical golfer with a swing that is said to emulate the legendary champion, Byron Nelson. Ten randomly selected balls of two different brands are tested and the overall distance measured. The data follow:

Brand 1: 275, 286, 287, 271, 283, 271, 279, 275, 263, 267
Brand 2: 258, 244, 260, 265, 273, 281, 271, 270, 263, 268

(a) Is there evidence that overall distance is approximately normally distributed? Is an assumption of equal variances justified?
(b) Test the hypothesis that both brands of ball have equal mean overall distance. Use $\alpha = 0.05$.
(c) What is the P-value of the test statistic in part (b)?
(d) What is the power of the statistical test in part (b) to detect a true difference in mean overall distance of 5 yards?
(e) What sample size would be required to detect a true difference in mean overall distance of 3 yards with power of approximately 0.75?
(f) Construct a 95% two-sided CI on the mean difference in overall distance between the two brands of golf balls.

10-32. In Example 9-6 we described how the "spring-like effect" in a golf club could be determined by measuring the coefficient of restitution (the ratio of the outbound velocity to the inbound velocity of a golf ball fired at the clubhead). Twelve randomly selected drivers produced by two clubmakers are tested and the coefficient of restitution measured. The data follow:

Club 1: 0.8406, 0.8104, 0.8234, 0.8198, 0.8235, 0.8562, 0.8123, 0.7976, 0.8184, 0.8265, 0.7773, 0.7871
Club 2: 0.8305, 0.7905, 0.8352, 0.8380, 0.8145, 0.8465, 0.8244, 0.8014, 0.8309, 0.8405, 0.8256, 0.8476

(a) Is there evidence that coefficient of restitution is approximately normally distributed? Is an assumption of equal variances justified?
(b) Test the hypothesis that both brands of ball have equal mean coefficient of restitution. Use $\alpha = 0.05$.
(c) What is the P-value of the test statistic in part (b)?
(d) What is the power of the statistical test in part (b) to detect a true difference in mean coefficient of restitution of 0.2?
(e) What sample size would be required to detect a true difference in mean coefficient of restitution of 0.1 with power of approximately 0.8?
(f) Construct a 95% two-sided CI on the mean difference in coefficient of restitution between the two brands of golf clubs.

10-4 PAIRED t-TEST

A special case of the two-sample t-tests of Section 10-3 occurs when the observations on the two populations of interest are collected in **pairs.** Each pair of observations, say (X_{1j}, X_{2j}), is taken under homogeneous conditions, but these conditions may change from one pair to another. For example, suppose that we are interested in comparing two different types of tips for a hardness-testing machine. This machine presses the tip into a metal specimen with a known force. By measuring the depth of the depression caused by the tip, the hardness of the specimen can be determined. If several specimens were selected at random, half tested with tip 1, half tested with tip 2, and the pooled or independent t-test in Section 10-3 was applied, the results of the test could be erroneous. The metal specimens could have been cut from bar stock that was produced in different heats, or they might not be homogeneous in some other way that might affect hardness. Then the observed difference between mean hardness readings for the two tip types also includes hardness differences between specimens.

A more powerful experimental procedure is to collect the data in pairs—that is, to make two hardness readings on each specimen, one with each tip. The test procedure would then consist of analyzing the *differences* between hardness readings on each specimen. If there is no difference between tips, the mean of the differences should be zero. This test procedure is called the **paired t-test.**

Let $(X_{11}, X_{21}), (X_{12}, X_{22}), \ldots, (X_{1n}, X_{2n})$ be a set of n paired observations where we assume that the mean and variance of the population represented by X_1 are μ_1 and σ_1^2, and the mean and variance of the population represented by X_2 are μ_2 and σ_2^2. Define the differences between each pair of observations as $D_j = X_{1j} - X_{2j}$, $j = 1, 2, \ldots, n$. The D_j's are assumed to be normally distributed with mean

$$\mu_D = E(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1 - \mu_2$$

and variance σ_D^2, so testing hypotheses about the difference between μ_1 and μ_2 can be accomplished by performing a one-sample t-test on μ_D. Specifically, testing H_0: $\mu_1 - \mu_2 = \Delta_0$ against H_1: $\mu_1 - \mu_2 \neq \Delta_0$ is equivalent to testing

$$H_0: \mu_D = \Delta_0$$
$$H_1: \mu_D \neq \Delta_0 \qquad (10\text{-}21)$$

The test statistic is given below.

The Paired *t*-Test

Null hypothesis: H_0: $\mu_D = \Delta_0$

Test statistic: $$T_0 = \frac{\overline{D} - \Delta_0}{S_D/\sqrt{n}} \qquad (10\text{-}22)$$

Alternative Hypothesis	Rejection Region
H_1: $\mu_D \neq \Delta_0$	$t_0 > t_{\alpha/2,n-1}$ or $t_0 < -t_{\alpha/2,n-1}$
H_1: $\mu_D > \Delta_0$	$t_0 > t_{\alpha,n-1}$
H_1: $\mu_D < \Delta_0$	$t_0 < -t_{\alpha,n-1}$

In Equation 10-22, $\overline{D}$ is the sample average of the n differences $D_1, D_2, \ldots, D_n$, and S_D is the sample standard deviation of these differences.

EXAMPLE 10-9

An article in the *Journal of Strain Analysis* (1983, Vol. 18, No. 2) compares several methods for predicting the shear strength for steel plate girders. Data for two of these methods, the Karlsruhe and Lehigh procedures, when applied to nine specific girders, are shown in Table 10-2. We wish to determine whether there is any difference (on the average) between the two methods.

The eight-step procedure is applied as follows:

1. The parameter of interest is the difference in mean shear strength between the two methods, say, $\mu_D = \mu_1 - \mu_2 = 0$.
2. H_0: $\mu_D = 0$

Table 10-2 Strength Predictions for Nine Steel Plate Girders (Predicted Load/Observed Load)

Girder	Karlsruhe Method	Lehigh Method	Difference d_j
S1/1	1.186	1.061	0.119
S2/1	1.151	0.992	0.159
S3/1	1.322	1.063	0.259
S4/1	1.339	1.062	0.277
S5/1	1.200	1.065	0.138
S2/1	1.402	1.178	0.224
S2/2	1.365	1.037	0.328
S2/3	1.537	1.086	0.451
S2/4	1.559	1.052	0.507

3. H_1: $\mu_D \neq 0$

4. $\alpha = 0.05$

5. The test statistic is

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

6. Reject H_0 if $t_0 > t_{0.025,8} = 2.306$ or if $t_0 < -t_{0.025,8} = -2.306$.

7. Computations: The sample average and standard deviation of the differences d_j are $\bar{d} = 0.2736$ and $s_D = 0.1356$, so the test statistic is

$$t_0 = \frac{\bar{d}}{s_D/\sqrt{n}} = \frac{0.2736}{0.1356/\sqrt{9}} = 6.05$$

8. Conclusions: Since $t_0 = 6.05 > 2.306$, we conclude that the strength prediction methods yield different results. Specifically, the data indicate that the Karlsruhe method produces, on the average, higher strength predictions than does the Lehigh method. The P-value for $t_0 = 6.05$ is $P = 0.0002$, so the test statistic is well into the critical region.

Paired Versus Unpaired Comparisons

In performing a comparative experiment, the investigator can sometimes choose between the paired experiment and the two-sample (or unpaired) experiment. If n measurements are to be made on each population, the two-sample t-statistic is

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p\sqrt{\dfrac{1}{n} + \dfrac{1}{n}}}$$

which would be compared to t_{2n-2}, and of course, the paired t-statistic is

$$T_0 = \frac{\bar{D} - \Delta_0}{S_D/\sqrt{n}}$$

which is compared to t_{n-1}. Notice that since

$$\bar{D} = \sum_{j=1}^{n} \frac{D_j}{n} = \sum_{j=1}^{n} \frac{(X_{1j} - X_{2j})}{n} = \sum_{j=1}^{n} \frac{X_{1j}}{n} - \sum_{j=1}^{n} \frac{X_{2j}}{n} = \bar{X}_1 - \bar{X}_2$$

the numerators of both statistics are identical. However, the denominator of the two-sample t-test is based on the assumption that X_1 and X_2 are *independent*. In many paired experiments, a strong positive correlation ρ exists between X_1 and X_2. Then it can be shown that

$$\begin{aligned} V(\bar{D}) &= V(\bar{X}_1 - \bar{X}_2 - \Delta_0) \\ &= V(\bar{X}_1) + V(\bar{X}_2) - 2\,\text{cov}(\bar{X}_1, \bar{X}_2) \\ &= \frac{2\sigma^2(1-\rho)}{n} \end{aligned}$$

assuming that both populations X_1 and X_2 have identical variances σ^2. Furthermore, S_D^2/n estimates the variance of $\overline{D}$. Whenever there is positive correlation within the pairs, the denominator for the paired t-test will be smaller than the denominator of the two-sample t-test. This can cause the two-sample t-test to considerably understate the significance of the data if it is incorrectly applied to paired samples.

Although pairing will often lead to a smaller value of the variance of $\overline{X}_1 - \overline{X}_2$, it does have a disadvantage—namely, the paired t-test leads to a loss of $n - 1$ degrees of freedom in comparison to the two-sample t-test. Generally, we know that increasing the degrees of freedom of a test increases the power against any fixed alternative values of the parameter.

So how do we decide to conduct the experiment? Should we pair the observations or not? Although there is no general answer to this question, we can give some guidelines based on the above discussion.

1. If the experimental units are relatively homogeneous (small σ) and the correlation within pairs is small, the gain in precision attributable to pairing will be offset by the loss of degrees of freedom, so an independent-sample experiment should be used.
2. If the experimental units are relatively heterogeneous (large σ) and there is large positive correlation within pairs, the paired experiment should be used. Typically, this case occurs when the experimental units are the *same* for both treatments; as in Example 10-9, the same girders were used to test the two methods.

Implementing the rules still requires judgment, because σ and ρ are never known precisely. Furthermore, if the number of degrees of freedom is large (say, 40 or 50), the loss of $n - 1$ of them for pairing may not be serious. However, if the number of degrees of freedom is small (say, 10 or 20), losing half of them is potentially serious if not compensated for by increased precision from pairing.

A Confidence Interval for μ_D

To construct the confidence interval for $\mu_D = \mu_1 - \mu_2$, note that

$$T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}}$$

follows a t distribution with $n - 1$ degrees of freedom. Then, since $P(-t_{\alpha/2,n-1} \le T \le t_{\alpha/2,n-1}) = 1 - \alpha$, we can substitute for T in the above expression and perform the necessary steps to isolate $\mu_D = \mu_1 - \mu_2$ between the inequalities. This leads to the following $100(1 - \alpha)\%$ confidence interval on $\mu_1 - \mu_2$.

Definition

If $\overline{d}$ and s_D are the sample mean and standard deviation of the difference of n random pairs of normally distributed measurements, a **$100(1 - \alpha)\%$ confidence interval on the difference in means $\mu_D = \mu_1 - \mu_2$** is

$$\overline{d} - t_{\alpha/2,n-1} s_D/\sqrt{n} \le \mu_D \le \overline{d} + t_{\alpha/2,n-1} s_D/\sqrt{n} \tag{10-23}$$

where $t_{\alpha/2,n-1}$ is the upper $\alpha/2\%$ point of the t-distribution with $n - 1$ degrees of freedom.

Table 10-3 Time in Seconds to Parallel Park Two Automobiles

	Automobile		**Difference**
Subject	$1(x_{1j})$	$2(x_{2j})$	(d_j)
1	37.0	17.8	19.2
2	25.8	20.2	5.6
3	16.2	16.8	−0.6
4	24.2	41.4	−17.2
5	22.0	21.4	0.6
6	33.4	38.4	−5.0
7	23.8	16.8	7.0
8	58.2	32.2	26.0
9	33.6	27.8	5.8
10	24.4	23.2	1.2
11	23.4	29.6	−6.2
12	21.2	20.6	0.6
13	36.2	32.2	4.0
14	29.8	53.8	−24.0

This confidence interval is also valid for the case where $\sigma_1^2 \neq \sigma_2^2$, because s_D^2 estimates $\sigma_D^2 = V(X_1 - X_2)$. Also, for large samples (say, $n \geq 30$ pairs), the explicit assumption of normality is unnecessary because of the central limit theorem.

EXAMPLE 10-10 The journal *Human Factors* (1962, pp. 375-380) reports a study in which $n = 14$ subjects were asked to parallel park two cars having very different wheel bases and turning radii. The time in seconds for each subject was recorded and is given in Table 10-3. From the column of observed differences we calculate $\bar{d} = 1.21$ and $s_D = 12.68$. The 90% confidence interval for $\mu_D = \mu_1 - \mu_2$ is found from Equation 9-24 as follows:

$$\bar{d} - t_{0.05,13}s_D/\sqrt{n} \leq \mu_D \leq \bar{d} + t_{0.05,13}s_D/\sqrt{n}$$

$$1.21 - 1.771(12.68)/\sqrt{14} \leq \mu_D \leq 1.21 + 1.771(12.68)/\sqrt{14}$$

$$-4.79 \leq \mu_D \leq 7.21$$

Notice that the confidence interval on μ_D includes zero. This implies that, at the 90% level of confidence, the data do not support the claim that the two cars have different mean parking times μ_1 and μ_2. That is, the value $\mu_D = \mu_1 - \mu_2 = 0$ is not inconsistent with the observed data.

EXERCISES FOR SECTION 10-4

10-33. Consider the shear strength experiment described in Example 10-9. Construct a 95% confidence interval on the difference in mean shear strength for the two methods. Is the result you obtained consistent with the findings in Example 10-9? Explain why.

10-34. Reconsider the shear strength experiment described in Example 10-9. Do each of the individual shear strengths have to be normally distributed for the paired *t*-test to be appropriate, or is it only the difference in shear strengths that must be normal? Use a normal probability plot to investigate the normality assumption.

10-35. Consider the parking data in Example 10-10. Use the paired *t*-test to investigate the claim that the two types of cars have different levels of difficulty to parallel park. Use $\alpha = 0.10$. Compare your results with the confidence interval constructed in Example 10-10 and comment on why they are the same or different.

10-36. Reconsider the parking data in Example 10-10. Investigate the assumption that the differences in parking times are normally distributed.

10-37. The manager of a fleet of automobiles is testing two brands of radial tires and assigns one tire of each brand at random to the two rear wheels of eight cars and runs the cars until the tires wear out. The data (in kilometers) follow. Find a 99% confidence interval on the difference in mean life. Which brand would you prefer, based on this calculation?

Car	Brand 1	Brand 2
1	36,925	34,318
2	45,300	42,280
3	36,240	35,500
4	32,100	31,950
5	37,210	38,015
6	48,360	47,800
7	38,200	37,810
8	33,500	33,215

10-38. A computer scientist is investigating the usefulness of two different design languages in improving programming tasks. Twelve expert programmers, familiar with both languages, are asked to code a standard function in both languages, and the time (in minutes) is recorded. The data follow:

	Time	
Programmer	Design Language 1	Design Language 2
1	17	18
2	16	14
3	21	19
4	14	11
5	18	23
6	24	21
7	16	10
8	14	13
9	21	19
10	23	24
11	13	15
12	18	20

(a) Find a 95% confidence interval on the difference in mean coding times. Is there any indication that one design language is preferable?

(b) Is the assumption that the difference in coding time is normally distributed reasonable? Show evidence to support your answer.

10-39. Fifteen adult males between the ages of 35 and 50 participated in a study to evaluate the effect of diet and exercise on blood cholesterol levels. The total cholesterol was measured in each subject initially and then three months after participating in an aerobic exercise program and switching to a low-fat diet. The data are shown in the accompanying table. Do the data support the claim that low-fat diet and aerobic exercise are of value in producing a mean reduction in blood cholesterol levels? Use $\alpha = 0.05$.

	Blood Cholesterol Level	
Subject	Before	After
1	265	229
2	240	231
3	258	227
4	295	240
5	251	238
6	245	241
7	287	234
8	314	256
9	260	247
10	279	239
11	283	246
12	240	218
13	238	219
14	225	226
15	247	233

10-40. An article in the *Journal of Aircraft* (Vol. 23, 1986, pp. 859–864) describes a new equivalent plate analysis method formulation that is capable of modeling aircraft structures such as cranked wing boxes and that produces results similar to the more computationally intensive finite element analysis method. Natural vibration frequencies for the cranked wing box structure are calculated using both methods, and results for the first seven natural frequencies follow:

Freq.	Finite Element Cycle/s	Equivalent Plate, Cycle/s
1	14.58	14.76
2	48.52	49.10
3	97.22	99.99
4	113.99	117.53
5	174.73	181.22
6	212.72	220.14
7	277.38	294.80

(a) Do the data suggest that the two methods prove the same mean value for natural vibration frequency? Use $\alpha = 0.05$.

(b) Find a 95% confidence interval on the mean difference between the two methods.

10-41. Ten individuals have participated in a diet-modification program to stimulate weight loss. Their weight both before and after participation in the program is shown in the following list. Is there evidence to support the claim that this particular diet-modification program is effective in producing a mean weight reduction? Use $\alpha = 0.05$.

Subject	Before	After
1	195	187
2	213	195
3	247	221
4	201	190
5	187	175
6	210	197
7	215	199
8	246	221
9	294	278
10	310	285

10-42. Two different analytical tests can be used to determine the impurity level in steel alloys. Eight specimens are tested using both procedures, and the results are shown in the following tabulation. Is there sufficient evidence to conclude that both tests give the same mean impurity level, using $\alpha = 0.01$?

Specimen	Test 1	Test 2
1	1.2	1.4
2	1.3	1.7
3	1.5	1.5
4	1.4	1.3
5	1.7	2.0
6	1.8	2.1
7	1.4	1.7
8	1.3	1.6

10-43. Consider the weight-loss data in Exercise 10-41. Is there evidence to support the claim that this particular diet-modification program will result in a mean weight loss of at least 10 pounds? Use $\alpha = 0.05$.

10-44. Consider the weight-loss experiment in Exercise 10-41. Suppose that, if the diet-modification program results in mean weight loss of at least 10 pounds, it is important to detect this with probability of at least 0.90. Was the use of 10 subjects an adequate sample size? If not, how many subjects should have been used?

10-5 INFERENCES ON THE VARIANCES OF TWO NORMAL POPULATIONS

We now introduce tests and confidence intervals for the two population variances shown in Fig. 10-1. We will assume that both populations are normal. Both the hypothesis-testing and confidence interval procedures are relatively sensitive to the normality assumption.

10-5.1 The *F* Distribution

Suppose that two independent normal populations are of interest, where the population means and variances, say, μ_1, σ_1^2, μ_2, and σ_2^2, are unknown. We wish to test hypotheses about the equality of the two variances, say, H_0: $\sigma_1^2 = \sigma_2^2$. Assume that two random samples of size n_1 from population 1 and of size n_2 from population 2 are available, and let S_1^2 and S_2^2 be the sample variances. We wish to test the hypotheses

$$H_0\colon \sigma_1^2 = \sigma_2^2$$
$$H_1\colon \sigma_1^2 \neq \sigma_2^2 \qquad (10\text{-}24)$$

The development of a test procedure for these hypotheses requires a new probability distribution, the *F* distribution. The random variable *F* is defined to be the ratio of two independent chi-square random variables, each divided by its number of degrees of freedom. That is,

$$F = \frac{W/u}{Y/v} \tag{10-25}$$

where W and Y are independent chi-square random variables with u and v degrees of freedom, respectively. We now formally state the sampling distribution of F.

Definition

Let W and Y be independent chi-square random variables with u and v degrees of freedom, respectively. Then the ratio

$$F = \frac{W/u}{Y/v} \tag{10-26}$$

has the probability density function

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x + 1\right]^{(u+v)/2}}, \qquad 0 < x < \infty \tag{10-27}$$

and is said to follow the F distribution with u degrees of freedom in the numerator and v degrees of freedom in the denominator. It is usually abbreviated as $F_{u,v}$.

The mean and variance of the F distribution are $\mu = v/(v-2)$ for $v > 2$, and

$$\sigma^2 = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)}, \qquad v > 4$$

Two F distributions are shown in Fig. 10-4. The F random variable is nonnegative, and the distribution is skewed to the right. The F distribution looks very similar to the chi-square distribution; however, the two parameters u and v provide extra flexibility regarding shape.

The percentage points of the F distribution are given in Table V of the Appendix. Let $f_{\alpha,u,v}$ be the percentage point of the F distribution, with numerator degrees of freedom u and denominator degrees of freedom v such that the probability that the random variable F exceeds this value is

$$P(F > f_{\alpha,u,v}) = \int_{f_{\alpha,u,v}}^{\infty} f(x)\,dx = \alpha$$

This is illustrated in Fig. 10-5. For example, if $u = 5$ and $v = 10$, we find from Table V of the Appendix that

$$P(F > f_{0.05,5,10}) = P(F_{5,10} > 3.33) = 0.05$$

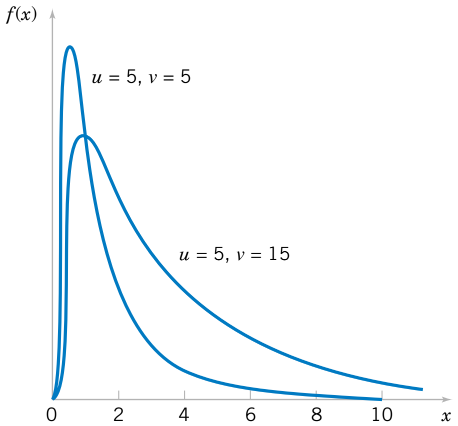

Figure 10-4 Probability density functions of two F distributions.

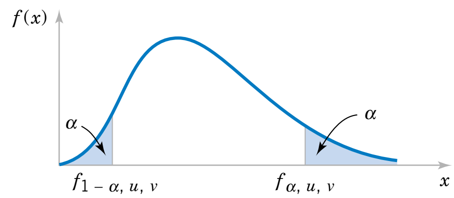

Figure 10-5 Upper and lower percentage points of the F distribution.

That is, the upper 5 percentage point of $F_{5,10}$ is $f_{0.05,5,10} = 3.33$.

Table V contains only upper-tail percentage points (for selected values of $f_{\alpha,u,v}$ for $\alpha \leq 0.25$) of the F distribution. The lower-tail percentage points $f_{1-\alpha,u,v}$ can be found as follows.

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,v,u}} \qquad (10\text{-}28)$$

For example, to find the lower-tail percentage point $f_{0.95,5,10}$, note that

$$f_{0.95,5,10} = \frac{1}{f_{0.05,10,5}} = \frac{1}{4.74} = 0.211$$

10-5.2 Development of the *F* Distribution (CD Only)

10-5.3 Hypothesis Tests on the Ratio of Two Variances

A hypothesis-testing procedure for the equality of two variances is based on the following result.

Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be a random sample from a normal population with mean μ_1 and variance σ_1^2, and let $X_{21}, X_{22}, \ldots, X_{2n_2}$ be a random sample from a second normal population with mean μ_2 and variance σ_2^2. Assume that both normal populations are independent. Let S_1^2 and S_2^2 be the sample variances. Then the ratio

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom.

This result is based on the fact that $(n_1 - 1)S_1^2/\sigma_1^2$ is a chi-square random variable with $n_1 - 1$ degrees of freedom, that $(n_2 - 1)S_2^2/\sigma_2^2$ is a chi-square random variable with $n_2 - 1$ degrees of freedom, and that the two normal populations are independent. Clearly under the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ the ratio $F_0 = S_1^2/S_2^2$ has an F_{n_1-1,n_2-1} distribution. This is the basis of the following test procedure.

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

Test statistic:
$$F_0 = \frac{S_1^2}{S_2^2} \tag{10-29}$$

Alternative Hypotheses	Rejection Criterion
H_1: $\sigma_1^2 \neq \sigma_2^2$	$f_0 > f_{\alpha/2,n_1-1,n_2-1}$ or $f_0 < f_{1-\alpha/2,n_1-1,n_2-1}$
H_1: $\sigma_1^2 > \sigma_2^2$	$f_0 > f_{\alpha,n_1-1,n_2-1}$
H_1: $\sigma_1^2 < \sigma_2^2$	$f_0 < f_{1-\alpha, n_1-1,n_2-1}$

EXAMPLE 10-11

Oxide layers on semiconductor wafers are etched in a mixture of gases to achieve the proper thickness. The variability in the thickness of these oxide layers is a critical characteristic of the wafer, and low variability is desirable for subsequent processing steps. Two different mixtures of gases are being studied to determine whether one is superior in reducing the variability of the oxide thickness. Twenty wafers are etched in each gas. The sample standard deviations of oxide thickness are $s_1 = 1.96$ angstroms and $s_2 = 2.13$ angstroms, respectively. Is there any evidence to indicate that either gas is preferable? Use $\alpha = 0.05$.

The eight-step hypothesis-testing procedure may be applied to this problem as follows:

1. The parameters of interest are the variances of oxide thickness σ_1^2 and σ_2^2. We will assume that oxide thickness is a normal random variable for both gas mixtures.
2. H_0: $\sigma_1^2 = \sigma_2^2$
3. H_1: $\sigma_1^2 \neq \sigma_2^2$
4. $\alpha = 0.05$
5. The test statistic is given by Equation 10-29:
$$f_0 = \frac{s_1^2}{s_2^2}$$
6. Since $n_1 = n_2 = 20$, we will reject H_0: $\sigma_1^2 = \sigma_2^2$ if $f_0 > f_{0.025,19,19} = 2.53$ or if $f_0 < f_{0.975,19,19} = 1/f_{0.025,19,19} = 1/2.53 = 0.40$.
7. Computations: Since $s_1^2 = (1.96)^2 = 3.84$ and $s_2^2 = (2.13)^2 = 4.54$, the test statistic is
$$f_0 = \frac{s_1^2}{s_2^2} = \frac{3.84}{4.54} = 0.85$$
8. Conclusions: Since $f_{0.975,19,19} = 0.40 < f_0 = 0.85 < f_{0.025,19,19} = 2.53$, we cannot reject the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ at the 0.05 level of significance. Therefore, there is no strong evidence to indicate that either gas results in a smaller variance of oxide thickness.

We may also find a P-value for the F-statistic in Example 10-11. Since $f_{0.50,19,19} = 1.00$, the computed value of the test statistic $f_0 = s_1^2/s_2^2 = 3.84/4.54 = 0.85$ is nearer the lower tail of the F distribution than the upper tail. The probability that an F-random variable with 19 numerator and denominator degrees of freedom is less than 0.85 is 0.3634. Since it is arbitrary which population is identified as "one," we could have computed the test statistic as $f_0 = 4.54/3.84 = 1.18$. The probability that an F-random variable with 19 numerator and denominator degrees of freedom exceeds 1.18 is 0.3610. Therefore, the P-value for the test statistic $f_0 = 0.85$ is the sum of these two probabilities, or $P = 0.3634 + 0.3610 = 0.7244$. Since the P-value exceeds 0.05, the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ cannot be rejected. (The probabilities given above were computed using a hand-held calculator.)

10-5.4 β-Error and Choice of Sample Size

Appendix Charts VI*o*, VI*p*, VI*q*, and VI*r* provide operating characteristic curves for the F-test given in Section 10-5.1 for $\alpha = 0.05$ and $\alpha = 0.01$, assuming that $n_1 = n_2 = n$. Charts VI*o* and VI*p* are used with the two-sided alternate hypothesis. They plot β against the abscissa parameter

$$\lambda = \frac{\sigma_1}{\sigma_2} \tag{10-30}$$

for various $n_1 = n_2 = n$. Charts VI*q* and VI*r* are used for the one-sided alternative hypotheses.

EXAMPLE 10-12 For the semiconductor wafer oxide etching problem in Example 10-11, suppose that one gas resulted in a standard deviation of oxide thickness that is half the standard deviation of oxide thickness of the other gas. If we wish to detect such a situation with probability at least 0.80, is the sample size $n_1 = n_2 = 20$ adequate?

Note that if one standard deviation is half the other,

$$\lambda = \frac{\sigma_1}{\sigma_2} = 2$$

By referring to Appendix Chart VI*o* with $n_1 = n_2 = n = 20$ and $\lambda = 2$, we find that $\beta \simeq 0.20$. Therefore, if $\beta = 0.20$, the power of the test (which is the probability that the difference in standard deviations will be detected by the test) is 0.80, and we conclude that the sample sizes $n_1 = n_2 = 20$ are adequate.

10-5.5 Confidence Interval on the Ratio of Two Variances

To find the confidence interval on σ_1^2/σ_2^2, recall that the sampling distribution of

$$F = \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2}$$

is an F with $n_2 - 1$ and $n_1 - 1$ degrees of freedom. Therefore, $P(f_{1-\alpha/2, n_2-1, n_1-1} \le F \le f_{\alpha/2, n_2-1, n_1-1}) = 1 - \alpha$. Substitution for F and manipulation of the inequalities will lead to the $100(1 - \alpha)\%$ confidence interval for σ_1^2/σ_2^2.

Definition

If s_1^2 and s_2^2 are the sample variances of random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown variances σ_1^2 and σ_2^2, then a **100(1 − α)% confidence interval on the ratio σ_1^2/σ_2^2** is

$$\frac{s_1^2}{s_2^2} f_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} f_{\alpha/2,n_2-1,n_1-1} \qquad (10\text{-}31)$$

where $f_{\alpha/2,n_2-1,n_1-1}$ and $f_{1-\alpha/2,n_2-1,n_1-1}$ are the upper and lower $\alpha/2$ percentage points of the F distribution with $n_2 - 1$ numerator and $n_1 - 1$ denominator degrees of freedom, respectively. A confidence interval on the ratio of the standard deviations can be obtained by taking square roots in Equation 10-31.

EXAMPLE 10-13 A company manufactures impellers for use in jet-turbine engines. One of the operations involves grinding a particular surface finish on a titanium alloy component. Two different grinding processes can be used, and both processes can produce parts at identical mean surface roughness. The manufacturing engineer would like to select the process having the least variability in surface roughness. A random sample of $n_1 = 11$ parts from the first process results in a sample standard deviation $s_1 = 5.1$ microinches, and a random sample of $n_2 = 16$ parts from the second process results in a sample standard deviation of $s_2 = 4.7$ microinches. We will find a 90% confidence interval on the ratio of the two standard deviations, σ_1/σ_2.

Assuming that the two processes are independent and that surface roughness is normally distributed, we can use Equation 10-31 as follows:

$$\frac{s_1^2}{s_2^2} f_{0.95,15,10} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} f_{0.05,15,10}$$

$$\frac{(5.1)^2}{(4.7)^2} 0.39 \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{(5.1)^2}{(4.7)^2} 2.85$$

or upon completing the implied calculations and taking square roots,

$$0.678 \le \frac{\sigma_1}{\sigma_2} \le 1.887$$

Notice that we have used Equation 10-28 to find $f_{0.95,15,10} = 1/f_{0.05,10,15} = 1/2.54 = 0.39$. Since this confidence interval includes unity, we cannot claim that the standard deviations of surface roughness for the two processes are different at the 90% level of confidence.

EXERCISES FOR SECTION 10-5

10-45. For an F distribution, find the following:
(a) $f_{0.25,5,10}$ (b) $f_{0.10,24,9}$
(c) $f_{0.05,8,15}$ (d) $f_{0.75,5,10}$
(e) $f_{0.90,24,9}$ (f) $f_{0.95,8,15}$

10-46. For an F distribution, find the following:
(a) $f_{0.25,7,15}$ (b) $f_{0.10,10,12}$
(c) $f_{0.01,20,10}$ (d) $f_{0.75,7,15}$
(e) $f_{0.90,10,12}$ (f) $f_{0.99,20,10}$

10-47. Two chemical companies can supply a raw material. The concentration of a particular element in this material is important. The mean concentration for both suppliers is the same, but we suspect that the variability in concentration may differ between the two companies. The standard deviation of concentration in a random sample of $n_1 = 10$ batches produced by company 1 is $s_1 = 4.7$ grams per liter, while for company 2, a random sample of $n_2 = 16$ batches yields $s_2 = 5.8$ grams per liter. Is there sufficient evidence to conclude that the two population variances differ? Use $\alpha = 0.05$.

10-48. Consider the etch rate data in Exercise 10-21. Test the hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 \neq \sigma_2^2$ using $\alpha = 0.05$, and draw conclusions.

10-49. Consider the etch rate data in Exercise 10-21. Suppose that if one population variance is twice as large as the other, we want to detect this with probability at least 0.90 (using $\alpha = 0.05$). Are the sample sizes $n_1 = n_2 = 10$ adequate?

10-50. Consider the diameter data in Exercise 10-17. Construct the following:

(a) A 90% two-sided confidence interval on σ_1/σ_2.

(b) A 95% two-sided confidence interval on σ_1/σ_2. Comment on the comparison of the width of this interval with the width of the interval in part (a).

(c) A 90% lower-confidence bound on σ_1/σ_2.

10-51. Consider the foam data in Exercise 10-18. Construct the following:

(a) A 90% two-sided confidence interval on σ_1^2/σ_2^2.

(b) A 95% two-sided confidence interval on σ_1^2/σ_2^2. Comment on the comparison of the width of this interval with the width of the interval in part (a).

(c) A 90% lower-confidence bound on σ_1/σ_2.

10-52. Consider the film speed data in Exercise 10-24. Test H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$ using $\alpha = 0.02$.

10-53. Consider the gear impact strength data in Exercise 10-22. Is there sufficient evidence to conclude that the variance of impact strength is different for the two suppliers? Use $\alpha = 0.05$.

10-54. Consider the melting point data in Exercise 10-25. Do the sample data support a claim that both alloys have the same variance of melting point? Use $\alpha = 0.05$ in reaching your conclusion.

10-55. Exercise 10-28 presented measurements of plastic coating thickness at two different application temperatures. Test H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 \neq \sigma_2^2$ using $\alpha = 0.01$.

10-56. A study was performed to determine whether men and women differ in their repeatability in assembling components on printed circuit boards. Random samples of 25 men and 21 women were selected, and each subject assembled the units. The two sample standard deviations of assembly time were $s_{\text{men}} = 0.98$ minutes and $s_{\text{women}} = 1.02$ minutes. Is there evidence to support the claim that men and women differ in repeatability for this assembly task? Use $\alpha = 0.02$ and state any necessary assumptions about the underlying distribution of the data.

10-57. Reconsider the assembly repeatability experiment described in Exercise 10-56. Find a 98% confidence interval on the ratio of the two variances. Provide an interpretation of the interval.

10-58. Reconsider the film speed experiment in Exercise 10-24. Suppose that one population standard deviation is 50% larger than the other. Is the sample size $n_1 = n_2 = 8$ adequate to detect this difference with high probability? Use $\alpha = 0.01$ in answering this question.

10-59. Reconsider the overall distance data for golf balls in Exercise 10-31. Is there evidence to support the claim that the standard deviation of overall distance is the same for both brands of balls (use $\alpha = 0.05$)? Explain how this question can be answered with a 95% confidence interval on σ_1/σ_2.

10-60. Reconsider the coefficient of restitution data in Exercise 10-32. Do the data suggest that the standard deviation is the same for both brands of drivers (use $\alpha = 0.05$)? Explain how to answer this question with a confidence interval on σ_1/σ_2.

10-6 INFERENCE ON TWO POPULATION PROPORTIONS

We now consider the case where there are two binomial parameters of interest, say, p_1 and p_2, and we wish to draw inferences about these proportions. We will present large-sample hypothesis testing and confidence interval procedures based on the normal approximation to the binomial.

10-6.1 Large-Sample Test for H_0: $p_1 = p_2$

Suppose that 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 the estimators of the population proportions

$\hat{P}_1 = X_1/n_1$ and $\hat{P}_2 = X_2/n_2$ have approximate normal distributions. We are interested in testing the hypotheses

$$H_0: p_1 = p_2$$
$$H_1: p_1 \neq p_2$$

The statistic

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1 - p_1)}{n_1} + \dfrac{p_2(1 - p_2)}{n_2}}} \qquad (10\text{-}32)$$

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, 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

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

This leads to the test procedures described below.

Null hypothesis: H_0: $p_1 = p_2$

Test statistic: $$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1 - \hat{P})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \qquad (10\text{-}33)$$

Alternative Hypotheses	Rejection Criterion
H_1: $p_1 \neq p_2$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
H_1: $p_1 > p_2$	$z_0 > z_\alpha$
H_1: $p_1 < p_2$	$z_0 < -z_\alpha$

EXAMPLE 10-14 Extracts of St. John's Wort are widely used to treat depression. An article in the April 18, 2001 issue of the *Journal of the American Medical Association* ("Effectiveness of St. John's Wort on Major Depression: A Randomized Controlled Trial") compared the efficacy of a standard extract of St. John's Wort with a placebo in 200 outpatients diagnosed with major depression. Patients were randomly assigned to two groups; one group received the St. John's Wort, and the other received the placebo. After eight weeks, 19 of the placebo-treated patients showed improvement, whereas 27 of those treated with St. John's Wort improved. Is there any reason to believe that St. John's Wort is effective in treating major depression? Use $\alpha = 0.05$.

The eight-step hypothesis testing procedure leads to the following results:

1. The parameters of interest are p_1 and p_2, the proportion of patients who improve following treatment with St. John's Wort (p_1) or the placebo (p_2).
2. H_0: $p_1 = p_2$
3. H_1: $p_1 \neq p_2$
4. $\alpha = 0.05$
5. The test statistic is

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p}_1 = 27/100 = 0.27, \hat{p}_2 = 19/100 = 0.19, n_1 = n_2 = 100,$ and

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{19 + 27}{100 + 100} = 0.23$$

6. Reject H_0: $p_1 = p_2$ if $z_0 > z_{0.025} = 1.96$ or if $z_0 < -z_{0.025} = -1.96$.
7. Computations: The value of the test statistic is

$$z_0 = \frac{0.27 - 0.19}{\sqrt{0.23(0.77)\left(\frac{1}{100} + \frac{1}{100}\right)}} = 1.35$$

8. Conclusions: Since $z_0 = 1.35$ does not exceed $z_{0.025}$, we cannot reject the null hypothesis. Note that the P-value is $P \simeq 0.177$. There is insufficient evidence to support the claim that St. John's Wort is effective in treating major depression.

The following box shows the Minitab two-sample hypothesis test and CI procedure for proportions. Notice that the 95% CI on $p_1 - p_2$ includes zero. The equation for constructing the CI will be given in Section 10-6.4.

Test and CI for Two Proportions

Sample	X	N	Sample p
1	27	100	0.270000
2	19	100	0.190000

Estimate for p(1) − p(2): 0.08
95% CI for p(1) − p(2): (−0.0361186, 0.196119)
Test for p(1) − p(2) = 0 (vs not = 0): Z = 1.35 P-Value = 0.177

10-6.2 Small-Sample Test for H_0: $p_1 = p_2$ (CD Only)

10-6.3 β-Error and Choice of Sample Size

The computation of the β-error for the large-sample test of H_0: $p_1 = p_2$ is somewhat more involved than in the single-sample case. The problem is that the denominator of the test statistic Z_0 is an estimate of the standard deviation of $\hat{P}_1 - \hat{P}_2$ under the assumption that $p_1 = p_2 = p$. When H_0: $p_1 = p_2$ is false, the standard deviation of $\hat{P}_1 - \hat{P}_2$ is

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{10-34}$$

If the alternative hypothesis is two sided, the β-error is

$$\beta = \Phi\left[\frac{z_{\alpha/2}\sqrt{\bar{p}\bar{q}(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}}\right] - \Phi\left[\frac{-z_{\alpha/2}\sqrt{\bar{p}\bar{q}(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}}\right] \tag{10-35}$$

where

$$\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \quad \text{and} \quad \bar{q} = \frac{n_1(1-p_1) + n_2(1-p_2)}{n_1 + n_2}$$

and $\sigma_{\hat{P}_1 - \hat{P}_2}$ is given by Equation 10-34.

If the alternative hypothesis is H_1: $p_1 > p_2$,

$$\beta = \Phi\left[\frac{z_{\alpha}\sqrt{\bar{p}\bar{q}(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}}\right] \tag{10-36}$$

and if the alternative hypothesis is H_1: $p_1 < p_2$,

$$\beta = 1 - \Phi\left[\frac{-z_{\alpha}\sqrt{\bar{p}\bar{q}(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}}\right] \tag{10-37}$$

For a specified pair of values p_1 and p_2, we can find the sample sizes $n_1 = n_2 = n$ required to give the test of size α that has specified type II error β.

For the two-sided alternative, the common sample size is

$$n = \frac{\left[z_{\alpha/2}\sqrt{(p_1 + p_2)(q_1 + q_2)/2} + z_\beta\sqrt{p_1q_1 + p_2q_2}\right]^2}{(p_1 - p_2)^2} \tag{10-38}$$

where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

For a one-sided alternative, replace $z_{\alpha/2}$ in Equation 10-38 by z_α.

10-6.4 Confidence Interval for $p_1 - p_2$

The confidence interval for $p_1 - p_2$ can be found directly, since we know that

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

is a standard normal random variable. Thus $P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) \simeq 1 - \alpha$, so we can substitute for Z in this last expression and use an approach similar to the one employed previously to find an approximate $100(1 - \alpha)\%$ two-sided confidence interval for $p_1 - p_2$.

Definition

If $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions of observation in two independent random samples of sizes n_1 and n_2 that belong to a class of interest, **an approximate two-sided $100(1 - \alpha)\%$ confidence interval on the difference in the true proportions $p_1 - p_2$ is**

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$\le p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \tag{10-39}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

EXAMPLE 10-15 Consider the process manufacturing crankshaft bearings described in Example 8-6. Suppose that a modification is made in the surface finishing process and that, subsequently, a second random sample of 85 axle shafts is obtained. The number of defective shafts in this second sample is 8. Therefore, since $n_1 = 85$, $\hat{p}_1 = 0.12$, $n_2 = 85$, and $\hat{p}_2 = 8/85 = 0.09$, we can obtain an approximate 95% confidence interval on the difference in the proportion of defective bearings produced under the two processes from Equation 10-39 as follows:

$$\hat{p}_1 - \hat{p}_2 - z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \leq p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

or

$$0.12 - 0.09 - 1.96\sqrt{\frac{0.12(0.88)}{85} + \frac{0.09(0.91)}{85}} \leq p_1 - p_2 \leq 0.12 - 0.09 + 1.96\sqrt{\frac{0.12(0.88)}{85} + \frac{0.09(0.91)}{85}}$$

This simplifies to

$$-0.06 \leq p_1 - p_2 \leq 0.12$$

This confidence interval includes zero, so, based on the sample data, it seems unlikely that the changes made in the surface finish process have reduced the proportion of defective crankshaft bearings being produced.

EXERCISES FOR SECTION 10-6

10-61. Two different types of injection-molding machines are used to form plastic parts. A part is considered defective if it has excessive shrinkage or is discolored. Two random samples, each of size 300, are selected, and 15 defective parts are found in the sample from machine 1 while 8 defective parts are found in the sample from machine 2. Is it reasonable to conclude that both machines produce the same fraction of defective parts, using $\alpha = 0.05$? Find the P-value for this test.

10-62. Two different types of polishing solution are being evaluated for possible use in a tumble-polish operation for manufacturing interocular lenses used in the human eye following cataract surgery. Three hundred lenses were tumble-polished using the first polishing solution, and of this number 253 had no polishing-induced defects. Another 300 lenses were tumble-polished using the second polishing solution, and 196 lenses were satisfactory upon completion. Is there any reason to believe that the two polishing solutions differ? Use $\alpha = 0.01$. Discuss how this question could be answered with a confidence interval on $p_1 - p_2$.

10-63. Consider the situation described in Exercise 10-61. Suppose that $p_1 = 0.05$ and $p_2 = 0.01$.
(a) With the sample sizes given here, what is the power of the test for this two-sided alternate?
(b) Determine the sample size needed to detect this difference with a probability of at least 0.9. Use $\alpha = 0.05$.

10-64. Consider the situation described in Exercise 10-61. Suppose that $p_1 = 0.05$ and $p_2 = 0.02$.
(a) With the sample sizes given here, what is the power of the test for this two-sided alternate?
(b) Determine the sample size needed to detect this difference with a probability of at least 0.9. Use $\alpha = 0.05$.

10-65. A random sample of 500 adult residents of Maricopa County found that 385 were in favor of increasing the highway speed limit to 75 mph, while another sample of 400 adult residents of Pima County found that 267 were in favor of the increased speed limit. Do these data indicate that there is a difference in the support for increasing the speed limit between the residents of the two counties? Use $\alpha = 0.05$. What is the P-value for this test?

10-66. Construct a 95% confidence interval on the difference in the two fractions defective for Exercise 10-61.

10-67. Construct a 95% confidence interval on the difference in the two proportions for Exercise 10-65. Provide a practical interpretation of this interval.

10-7 SUMMARY TABLE FOR INFERENCE PROCEDURES FOR TWO SAMPLES

The table in the end papers of the book summarizes all of the two-sample inference procedures given in this chapter. The table contains the null hypothesis statements, the test statistics, the criteria for rejection of the various alternative hypotheses, and the formulas for constructing the $100(1 - \alpha)\%$ confidence intervals.

Supplemental Exercises

10-68. A procurement specialist has purchased 25 resistors from vendor 1 and 35 resistors from vendor 2. Each resistor's resistance is measured with the following results:

Vendor 1

96.8	100.0	100.3	98.5	98.3	98.2
99.6	99.4	99.9	101.1	103.7	97.7
99.7	101.1	97.7	98.6	101.9	101.0
99.4	99.8	99.1	99.6	101.2	98.2
98.6					

Vendor 2

106.8	106.8	104.7	104.7	108.0	102.2
103.2	103.7	106.8	105.1	104.0	106.2
102.6	100.3	104.0	107.0	104.3	105.8
104.0	106.3	102.2	102.8	104.2	103.4
104.6	103.5	106.3	109.2	107.2	105.4
106.4	106.8	104.1	107.1	107.7	

(a) What distributional assumption is needed to test the claim that the variance of resistance of product from vendor 1 is not significantly different from the variance of resistance of product from vendor 2? Perform a graphical procedure to check this assumption.

(b) Perform an appropriate statistical hypothesis-testing procedure to determine whether the procurement specialist can claim that the variance of resistance of product from vendor 1 is significantly different from the variance of resistance of product from vendor 2.

10-69. An article in the *Journal of Materials Engineering* (1989, Vol. 11, No. 4, pp. 275–282) reported the results of an experiment to determine failure mechanisms for plasma-sprayed thermal barrier coatings. The failure stress for one particular coating (NiCrAlZr) under two different test conditions is as follows:

Failure stress ($\times 10^6$ Pa) after nine 1-hour cycles: 19.8, 18.5, 17.6, 16.7, 16.7, 14.8, 15.4, 14.1, 13.6

Failure stress ($\times 10^6$ Pa) after six 1-hour cycles: 14.9, 12.7, 11.9, 11.4, 10.1, 7.9

(a) What assumptions are needed to construct confidence intervals for the difference in mean failure stress under the two different test conditions? Use normal probability plots of the data to check these assumptions.

(b) Find a 99% confidence interval on the difference in mean failure stress under the two different test conditions.

(c) Using the confidence interval constructed in part (b), does the evidence support the claim that the first test conditions yield higher results, on the average, than the second? Explain your answer.

10-70. Consider Supplemental Exercise 10-69.

(a) Construct a 95% confidence interval on the ratio of the variances, σ_1/σ_2, of failure stress under the two different test conditions.

(b) Use your answer in part (b) to determine whether there is a significant difference in variances of the two different test conditions. Explain your answer.

10-71. A liquid dietary product implies in its advertising that use of the product for one month results in an average weight loss of at least 3 pounds. Eight subjects use the product for one month, and the resulting weight loss data are reported below. Use hypothesis-testing procedures to answer the following questions.

Subject	Initial Weight (lb)	Final Weight (lb)
1	165	161
2	201	195
3	195	192
4	198	193
5	155	150
6	143	141
7	150	146
8	187	183

(a) Do the data support the claim of the producer of the dietary product with the probability of a type I error set to 0.05?

(b) Do the data support the claim of the producer of the dietary product with the probability of a type I error set to 0.01?

(c) In an effort to improve sales, the producer is considering changing its claim from "at least 3 pounds" to "at least 5 pounds." Repeat parts (a) and (b) to test this new claim.

10-72. The breaking strength of yarn supplied by two manufacturers is being investigated. We know from experience with the manufacturers' processes that $\sigma_1 = 5$ psi and $\sigma_2 = 4$ psi. A random sample of 20 test specimens from each manufacturer results in $\bar{x}_1 = 88$ psi and $\bar{x}_2 = 91$ psi, respectively.

(a) Using a 90% confidence interval on the difference in mean breaking strength, comment on whether or not there is evidence to support the claim that manufacturer 2 produces yarn with higher mean breaking strength.

(b) Using a 98% confidence interval on the difference in mean breaking strength, comment on whether or not there is evidence to support the claim that manufacturer 2 produces yarn with higher mean breaking strength.

(c) Comment on why the results from parts (a) and (b) are different or the same. Which would you choose to make your decision and why?

10-73. The Salk polio vaccine experiment in 1954 focused on the effectiveness of the vaccine in combatting paralytic polio. Because it was felt that without a control group of children there would be no sound basis for evaluating the efficacy of the Salk vaccine, the vaccine was administered to one group, and a placebo (visually identical to the vaccine but known to have no effect) was administered to a second group. For ethical reasons, and because it was suspected that knowledge of vaccine administration would affect subsequent diagnoses, the experiment was conducted in a double-blind fashion. That is, neither the subjects nor the administrators knew who received the vaccine and who received the placebo. The actual data for this experiment are as follows:

Placebo group: $n = 201{,}299$: 110 cases of polio observed

Vaccine group: $n = 200{,}745$: 33 cases of polio observed

(a) Use a hypothesis-testing procedure to determine if the proportion of children in the two groups who contracted paralytic polio is statistically different. Use a probability of a type I error equal to 0.05.

(b) Repeat part (a) using a probability of a type I error equal to 0.01.

(c) Compare your conclusions from parts (a) and (b) and explain why they are the same or different.

10-74. Consider Supplemental Exercise 10-72. Suppose that prior to collecting the data, you decide that you want the error in estimating $\mu_1 - \mu_2$ by $x_1 - x_2$ to be less than 1.5 psi. Specify the sample size for the following percentage confidence:

(a) 90%

(b) 98%

(c) Comment on the effect of increasing the percentage confidence on the sample size needed.

(d) Repeat parts (a)–(c) with an error of less than 0.75 psi instead of 1.5 psi.

(e) Comment on the effect of decreasing the error on the sample size needed.

10-75. A random sample of 1500 residential telephones in Phoenix in 1990 found that 387 of the numbers were unlisted. A random sample in the same year of 1200 telephones in Scottsdale found that 310 were unlisted.

(a) Find a 95% confidence interval on the difference in the two proportions and use this confidence interval to determine if there is a statistically significant difference in proportions of unlisted numbers between the two cities.

(b) Find a 90% confidence interval on the difference in the two proportions and use this confidence interval to determine if there is a statistically significant difference in proportions of unlisted numbers between the two cities.

(c) Suppose that all the numbers in the problem description were doubled. That is, 774 residents out of 3000 sampled in Phoenix and 620 residents out of 2400 in Scottsdale had unlisted phone numbers. Repeat parts (a) and (b) and comment on the effect of increasing the sample size without changing the proportions on your results.

10-76. In a random sample of 200 Phoenix residents who drive a domestic car, 165 reported wearing their seat belt regularly, while another sample of 250 Phoenix residents who drive a foreign car revealed 198 who regularly wore their seat belt.

(a) Perform a hypothesis-testing procedure to determine if there is a statistically significant difference in seat belt usage between domestic and foreign car drivers. Set your probability of a type I error to 0.05.

(b) Perform a hypothesis-testing procedure to determine if there is a statistically significant difference in seat belt usage between domestic and foreign car drivers. Set your probability of a type I error to 0.1.

(c) Compare your answers for parts (a) and (b) and explain why they are the same or different.

(d) Suppose that all the numbers in the problem description were doubled. That is, in a random sample of 400 Phoenix residents who drive a domestic car, 330 reported wearing their seat belt regularly, while another sample of 500 Phoenix residents who drive a foreign car revealed 396 who regularly wore their seat belt. Repeat parts (a) and (b) and comment on the effect of increasing the sample size without changing the proportions on your results.

10-77. Consider the previous exercise, which summarized data collected from drivers about their seat belt usage.

(a) Do you think there is a reason not to believe these data? Explain your answer.

(b) Is it reasonable to use the hypothesis-testing results from the previous problem to draw an inference about the difference in proportion of seat belt usage

  (i) of the spouses of these drivers of domestic and foreign cars? Explain your answer.

  (ii) of the children of these drivers of domestic and foreign cars? Explain your answer.

  (iii) of all drivers of domestic and foreign cars? Explain your answer.

  (iv) of all drivers of domestic and foreign trucks? Explain your answer.

10-78. Consider the situation described in Exercise 10-62.

(a) Redefine the parameters of interest to be the proportion of lenses that are unsatisfactory following tumble polishing with polishing fluids 1 or 2. Test the hypothesis that the two polishing solutions give different results using $\alpha = 0.01$.

(b) Compare your answer in part (a) with that for Exercise 10-62. Explain why they are the same or different.

10-79. Consider the situation of Exercise 10-62, and recall that the hypotheses of interest are H_0: $p_1 = p_2$ versus H_1: $p_1 \neq p_2$. We wish to use $\alpha = 0.01$. Suppose that if $p_1 = 0.9$ and $p_2 = 0.6$, we wish to detect this with a high probability, say, at least 0.9. What sample sizes are required to meet this objective?

10-80. A manufacturer of a new pain relief tablet would like to demonstrate that its product works twice as fast as the competitor's product. Specifically, the manufacturer would like to test

$$H_0: \mu_1 = 2\mu_2$$
$$H_1: \mu_1 > 2\mu_2$$

where μ_1 is the mean absorption time of the competitive product and μ_2 is the mean absorption time of the new product. Assuming that the variances σ_1^2 and σ_2^2 are known, develop a procedure for testing this hypothesis.

10-81. Suppose that we are testing H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$, and we plan to use equal sample sizes from the two populations. Both populations are assumed to be normal with unknown but equal variances. If we use $\alpha = 0.05$ and if the true mean $\mu_1 = \mu_2 + \sigma$, what sample size must be used for the power of this test to be at least 0.90?

10-82. Consider the fire-fighting foam expanding agents investigated in Exercise 10-18, in which five observations of each agent were recorded. Suppose that, if agent 1 produces a mean expansion that differs from the mean expansion of agent 1 by 1.5, we would like to reject the null hypothesis with probability at least 0.95.

(a) What sample size is required?

(b) Do you think that the original sample size in Exercise 10-18 was appropriate to detect this difference? Explain your answer.

10-83. A fuel-economy study was conducted for two German automobiles, Mercedes and Volkswagen. One vehicle of each brand was selected, and the mileage performance was observed for 10 tanks of fuel in each car. The data are as follows (in miles per gallon):

Mercedes		Volkswagen	
24.7	24.9	41.7	42.8
24.8	24.6	42.3	42.4
24.9	23.9	41.6	39.9
24.7	24.9	39.5	40.8
24.5	24.8	41.9	29.6

(a) Construct a normal probability plot of each of the data sets. Based on these plots, is it reasonable to assume that they are each drawn from a normal population?

(b) Suppose that it was determined that the lowest observation of the Mercedes data was erroneously recorded and should be 24.6. Furthermore, the lowest observation of the Volkswagen data was also mistaken and should be 39.6. Again construct normal probability plots of each of the data sets with the corrected values. Based on these new plots, is it reasonable to assume that they are each drawn from a normal population?

(c) Compare your answers from parts (a) and (b) and comment on the effect of these mistaken observations on the normality assumption.

(d) Using the corrected data from part (b) and a 95% confidence interval, is there evidence to support the claim that the variability in mileage performance is greater for a Volkswagen than for a Mercedes?

10-84. Reconsider the fuel-economy study in Supplemental Exercise 10-83. Rework part (d) of this problem using an appropriate hypothesis-testing procedure. Did you get the same answer as you did originally? Why?

10-85. An experiment was conducted to compare the filling capability of packaging equipment at two different wineries. Ten bottles of pinot noir from Ridgecrest Vineyards were randomly selected and measured, along with 10 bottles of pinot noir from Valley View Vineyards. The data are as follows (fill volume is in milliliters):

Ridgecrest				Valley View			
755	751	752	753	756	754	757	756
753	753	753	754	755	756	756	755
752	751			755	756		

(a) What assumptions are necessary to perform a hypothesis-testing procedure for equality of means of these data? Check these assumptions.

(b) Perform the appropriate hypothesis-testing procedure to determine whether the data support the claim that both wineries will fill bottles to the same mean volume.

10-86. Consider Supplemental Exercise 10-85. Suppose that the true difference in mean fill volume is as much as 2 fluid ounces; did the sample sizes of 10 from each vineyard provide good detection capability when $\alpha = 0.05$? Explain your answer.

10-87. A Rockwell hardness-testing machine presses a tip into a test coupon and uses the depth of the resulting depression to indicate hardness. Two different tips are being compared to determine whether they provide the same Rockwell C-scale hardness readings. Nine coupons are tested, with both tips being tested on each coupon. The data are shown in the accompanying table.

Coupon	Tip 1	Tip 2	Coupon	Tip 1	Tip 2
1	47	46	6	41	41
2	42	40	7	45	46
3	43	45	8	45	46
4	40	41	9	49	48
5	42	43			

(a) State any assumptions necessary to test the claim that both tips produce the same Rockwell C-scale hardness readings. Check those assumptions for which you have the information.

(b) Apply an appropriate statistical method to determine if the data support the claim that the difference in Rockwell C-scale hardness readings of the two tips is significantly different from zero

(c) Suppose that if the two tips differ in mean hardness readings by as much as 1.0, we want the power of the test to be at least 0.9. For an $\alpha = 0.01$, how many coupons should have been used in the test?

10-88. Two different gauges can be used to measure the depth of bath material in a Hall cell used in smelting aluminum. Each gauge is used once in 15 cells by the same operator.

Cell	Gauge 1	Gauge 2	Cell	Gauge 1	Gauge 2
1	46 in.	47 in.	9	52	51
2	50	53	10	47	45
3	47	45	11	49	51
4	53	50	12	45	45
5	49	51	13	47	49
6	48	48	14	46	43
7	53	54	15	50	51
8	56	53			

(a) State any assumptions necessary to test the claim that both gauges produce the same mean bath depth readings. Check those assumptions for which you have the information.

(b) Apply an appropriate statistical procedure to determine if the data support the claim that the two gauges produce different mean bath depth readings.

(c) Suppose that if the two gauges differ in mean bath depth readings by as much as 1.65 inch, we want the power of the test to be at least 0.8. For $\alpha = 0.01$, how many cells should have been used?

10-89. An article in the *Journal of the Environmental Engineering Division* ("Distribution of Toxic Substances in Rivers," 1982, Vol. 108, pp. 639–649) investigates the concentration of several hydrophobic organic substances in the Wolf River in Tennessee. Measurements on hexachlorobenzene (HCB) in nanograms per liter were taken at different depth downstream of an abandoned dump site. Data for two depths follow:

Surface: 3.74, 4.61, 4.00, 4.67, 4.87, 5.12, 4.52, 5.29, 5.74, 5.48

Bottom: 5.44, 6.88, 5.37, 5.44, 5.03, 6.48, 3.89, 5.85, 6.85, 7.16

(a) What assumptions are required to test the claim that mean HCB concentration is the same at both depths? Check those assumptions for which you have the information.

(b) Apply an appropriate procedure to determine if the data support the claim in part a.

(c) Suppose that the true difference in mean concentrations is 2.0 nanograms per liter. For $\alpha = 0.05$, what is the power of a statistical test for H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$?

(d) What sample size would be required to detect a difference of 1.0 nanograms per liter at $\alpha = 0.05$ if the power must be at least 0.9?

MIND-EXPANDING EXERCISES

10-90. Three different pesticides can be used to control infestation of grapes. It is suspected that pesticide 3 is more effective than the other two. In a particular vineyard, three different plantings of pinot noir grapes are selected for study. The following results on yield are obtained:

Pesticide	$\bar{x}_i$ (Bushels/ Plant)	s_i	n_i (Number of Plants)
1	4.6	0.7	100
2	5.2	0.6	120
3	6.1	0.8	130

If μ_i is the true mean yield after treatment with the ith pesticide, we are interested in the quantity

$$\mu = \frac{1}{2}(\mu_1 + \mu_2) - \mu_3$$

which measures the difference in mean yields between pesticides 1 and 2 and pesticide 3. If the sample sizes n_i are large, the estimator (say, $\hat{\mu}$) obtained by replacing each individual μ_i by $\bar{X}_i$ is approximately normal.

(a) Find an approximate $100(1 - \alpha)\%$ large-sample confidence interval for μ.

MIND-EXPANDING EXERCISES

(b) Do these data support the claim that pesticide 3 is more effective than the other two? Use $\alpha = 0.05$ in determining your answer.

10-91. Suppose that we wish to test H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$, where σ_1^2 and σ_2^2 are known. The total sample size N is to be determined, and the allocation of observations to the two populations such that $n_1 + n_2 = N$ is to be made on the basis of cost. If the cost of sampling for populations 1 and 2 are C_1 and C_2, respectively, find the minimum cost sample sizes that provide a specified variance for the difference in sample means.

10-92. Suppose that we wish to test the hypothesis H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$, where both variances σ_1^2 and σ_2^2 are known. A total of $n_1 + n_2 = N$ observations can be taken. How should these observations be allocated to the two populations to maximize the probability that H_0 will be rejected if H_1 is true and $\mu_1 - \mu_2 = \Delta \neq 0$?

10-93. Suppose that we wish to test H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$, where the population is normal with known σ. Let $0 < \epsilon < \alpha$, and define the critical region so that we will reject H_0 if $z_0 > z_\epsilon$ or if $z_0 < -z_{\alpha-\epsilon}$, where z_0 is the value of the usual test statistic for these hypotheses.

(a) Show that the probability of type I error for this test is α.

(b) Suppose that the true mean is $\mu_1 = \mu_0 + \Delta$. Derive an expression for β for the above test.

10-94. Construct a data set for which the paired t-test statistic is very large, indicating that when this analysis is used the two population means are different, but t_0 for the two-sample t-test is very small so that the incorrect analysis would indicate that there is no significant difference between the means.

10-95. In some situations involving proportions, we are interested in the ratio $\theta = p_1/p_2$ rather than the difference $p_1 - p_2$. Let $\hat{\theta} = \hat{p}_1/\hat{p}_2$. We can show that $\ln(\hat{\theta})$ has an approximate normal distribution with the mean (n/θ) and variance $[(n_1 - x_1)/(n_1x_1) + (n_2 - x_2)/(n_2x_2)]^{1/2}$.

(a) Use the information above to derive a large-sample confidence interval for $\ln \theta$.

(b) Show how to find a large-sample CI for θ.

(c) Use the data from the St. John's Wort study in Example 10-14, and find a 95% CI on $\theta = p_1/p_2$. Provide a practical interpretation for this CI.

10-96. Derive an expression for β for the test of the equality of the variances of two normal distributions. Assume that the two-sided alternative is specified.

IMPORTANT TERMS AND CONCEPTS

In the E-book, click on any term or concept below to go to that subject.

- Comparative experiments
- Critical region for a test statistic
- Identifying cause and effect
- Null and alternative hypotheses
- One-sided and two-sided alternative hypotheses
- Operating characteristic curves
- Paired *t*-test
- Pooled *t*-test
- *P*-value
- Reference distribution for a test statistic
- Sample size determination for hypothesis tests and confidence intervals
- Statistical hypotheses
- Test statistic

CD MATERIAL

- Fisher-Irwin test on two proportions

10-3.2 More About the Equal Variance Assumption (CD Only)

In practice, one often has to choose between case 1 and case 2 of the two-sample t-test. In case 1, we assume that $\sigma_1^2 = \sigma_2^2$ and use the pooled t-test. On the surface this test would seem to have some advantages. It is a likelihood ratio test, whereas the case 2 test with $\sigma_1^2 \neq \sigma_2^2$ is not. Furthermore, it is an exact test (if the assumptions of normality, independence, and equal variances are correct), whereas the case 2 test is an approximate procedure. However, the pooled t-test can be very sensitive to the assumption of equal variances, especially when the sample sizes are not equal. To help see this, consider the denominator of the test statistic for the pooled t-test:

$$S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\,\frac{n_1+n_2}{n_1n_2}} \simeq \sqrt{\frac{S_1^2}{n_2}+\frac{S_2^2}{n_1}}$$

Because the variances are divided (approximately) by the wrong sample sizes, use of the pooled t-test when the variances are unequal and when $n_1 \neq n_2$ can lead to very frequent erroneous conclusions. This is why using $n_1 = n_2$ is a good idea in general, and especially when we are in doubt about the validity of the equal variance assumption.

It would, of course, be possible to perform a test of H_0: $\sigma_1^2 = \sigma_2^2$ versus H_1: $\sigma_1^2 \neq \sigma_2^2$ and then use the pooled t-test if the null hypothesis is not rejected. This test is discussed in Section 10-5. However, the test on variances is much more sensitive to the normality assumption than are t-tests. A conservative approach would be to always use the case 2 procedure. Alternatively, one can use the normal probability plot both as a check of the normality assumption and as a check for equality of variance. If there is a noticable difference in the slopes of the two straight lines on the normal probability plot, the case 2 procedure would be preferred, especially when $n_1 \neq n_2$.

10-5.2 Development of the F Distribution (CD Only)

We now give a formal development of the F distribution. The development makes use of the material in Section 5-8 (CD Only).

Theorem: The F-Distribution

> Let U_1 and U_2 be independent chi-square random variables with v_1 and v_2 degrees of freedom, respectively. Then the ratio
>
> $$F = \frac{U_1/\nu_1}{U_2/\nu_2}$$
>
> has the probability density function
>
> $$f(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{\nu_1/2-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left[\left(\frac{\nu_1}{\nu_2}\right)x+1\right]^{(\nu_1+\nu_2)/2}}, \qquad 0 < x < \infty$$
>
> This is the F-distribution with v_1 degrees of freedom in the numerator and v_2 degrees of freedom in the denominator.

Proof Since U_1 and U_2 are independent chi-square random variables, their joint probability distribution is

$$f(u_1, u_2) = \frac{u_1^{\nu_1/2-1} u_2^{\nu_2/2-1}}{2^{\nu_1/2}\Gamma\left(\frac{\nu_1}{2}\right)2^{\nu_2/2}\Gamma\left(\frac{\nu_2}{2}\right)} e^{-(u_1+u_2)/2}, \qquad 0 < u_1, u_2 < \infty$$

Using the method in Equation S5-4, define the new random variable $M = U_2$. The inverse solutions of

$$x = \left(\frac{u_1}{\nu_1}\right)\Big/\left(\frac{u_2}{\nu_2}\right) \quad \text{and} \quad m = u_2$$

are

$$u_1 = \frac{\nu_1}{\nu_2} mx \quad \text{and} \quad u_2 = m$$

Therefore, the Jacobian is

$$J = \begin{vmatrix} \frac{\nu_1}{\nu_2} m & \frac{\nu_1}{\nu_2} x \\ 0 & 1 \end{vmatrix} = \frac{\nu_1}{\nu_2} m$$

Thus, the joint probability density function of X and M is

$$f(x, m) = \frac{\left(\frac{\nu_1}{\nu_2} mx\right)^{\nu_1/2-1} m^{\nu_2/2-1}\, e^{-(1/2)[(\nu_1/\nu_2)mx+m]}\left(\frac{\nu_1}{\nu_2}\right)m}{2^{\nu_1/2}\Gamma\left(\frac{\nu_1}{2}\right)2^{\nu_2/2}\,\Gamma\left(\frac{\nu_2}{2}\right)}, \qquad 0 < x, m < \infty$$

The probability density function of F is

$$\begin{aligned} f(x) &= \int_0^\infty f(x, m)\, dm \\ &= \frac{\left(\frac{\nu_1}{\nu_2} x\right)^{\nu_1/2-1}\left(\frac{\nu_1}{\nu_2}\right)}{2^{(\nu_1+\nu_2)/2}\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \int_0^\infty m^{(\nu_1+\nu_2)/2-1} e^{-(m/2)[(\nu_1/\nu_2)x+1]}\, dm \end{aligned}$$

Substituting $z = \frac{m}{z}\left(\frac{\nu_1}{\nu_2} x + 1\right)$ and $dm = 2\left(\frac{\nu_1}{\nu_2} x + 1\right)^{-1} dz$, we obtain

$$f(x) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{\nu_1/2-1}}{2^{(\nu_1+\nu_2)/2}\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \int_0^\infty \left(\frac{2z}{\frac{\nu_1}{\nu_2} x + 1}\right)^{(\nu_1+\nu_2)/2-1} e^{-z}\, 2\left(\frac{\nu_1}{\nu_2} x + 1\right)^{-1} dz$$

$$= \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{\nu_1/2-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}x + 1\right)^{(\nu_1+\nu_2)/2}} \int_0^\infty z^{(\nu_1+\nu_2)/2-1} e^{-z}\, dz$$

$$= \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{\nu_1/2-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}x + 1\right)^{(\nu_1+\nu_2)/2}}, \qquad 0 < x < \infty$$

which is the probability density function in the theorem on page 10-1.

10-6.2 Small-Sample Test for H_0: $p_1 = p_2$ (CD Only)

Many problems involving the comparison of proportions p_1 and p_2 have relatively large sample sizes, so the procedure based on the normal approximation to the binomial is widely used in practice. However, occasionally, a small-sample-size problem is encountered. In such cases, the Z-tests are inappropriate and an alternative procedure is required. In this section we describe a procedure based on the hypergeometric distribution.

Suppose that X_1 and X_2 are the number of successes in two random samples of size n_1 and n_2, respectively. The test procedure requires that we view the total number of successes as fixed at the value $X_1 + X_2 = Y$. Now consider the hypotheses

$$H_0\colon p_1 = p_2$$
$$H_1\colon p_1 > p_2$$

Given that $X_1 + X_2 = Y$, large values of X_1 support H_1, and small moderate values of X_1 support H_0. Therefore, we will reject H_0 whenever X_1 is sufficiently large.

Since the combined sample of $n_1 + n_2$ observations contains $X_1 + X_2 = Y$ total successes, if H_0: $p_1 = p_2$ the successes are no more likely to be concentrated in the first sample than in the second. That is, all the ways in which the $n_1 + n_2$ responses can be divided into one sample of n_1 responses and a second sample of n_2 responses are equally likely. The number of ways of selecting X_1 successes for the first sample leaving $Y - X_1$ successes for the second is

$$\binom{Y}{X_1}\binom{n_1 + n_2 - Y}{n_1 - X_1}$$

Because outcomes are equally likely, the probability of exactly X_1 successes in sample 1 is the ratio of the number of sample 1 outcomes having X_1 successes to the total number of outcomes, or

$$P(X_1 = x_1 \mid Y \text{ success in } n_1 + n_2 \text{ responses}) = \frac{\binom{Y}{x_1}\binom{n_1 + n_2 - Y}{n_1 - x_1}}{\binom{n_1 + n_2}{n_1}} \tag{S10-1}$$

given that H_0: $p_1 = p_2$ is true. We recognize Equation S10-1 as a hypergeometric distribution.

To use Equation S10-1 for hypothesis testing, we would compute the probability of finding a value of X_1 at least as extreme as the observed value of X_1. Note that this probability is a

P-value. If this P-value is sufficiently small, the null hypothesis is rejected. This approach could also be applied to lower-tailed and two-tailed alternatives.

EXAMPLE S10-1

Insulating cloth used in printed circuit boards is manufactured in large rolls. The manufacturer is trying to improve the process *yield*, that is, the number of defect-free rolls produced. A sample of 10 rolls contains exactly 4 defect-free rolls. From analysis of the defect types, process engineers suggest several changes in the process. Following implementation of these changes, another sample of 10 rolls yields 8 defect-free rolls. Do the data support the claim that the new process is better than the old one, using $\alpha = 0.10$?

To answer this question, we compute the P-value. In our example, $n_1 = n_2 = 10$, $y = 8 + 4 = 12$, and the observed value of $x_1 = 8$. The values of x_1 that are more extreme than 8 are 9 and 10. Therefore

$$P(X_1 = 8|12 \text{ successes}) = \frac{\binom{12}{8}\binom{2}{2}}{\binom{20}{10}} = .0750$$

$$P(X_1 = 9|12 \text{ successes}) = \frac{\binom{12}{9}\binom{8}{1}}{\binom{20}{10}} = .0095$$

$$P(X_1 = 10|12 \text{ successes}) = \frac{\binom{12}{10}\binom{8}{0}}{\binom{20}{10}} = .0003$$

The P-value is $P = .0750 + .0095 + .0003 = .0848$. Thus, at the level $\alpha = 0.10$, the null hypothesis is rejected and we conclude that the engineering changes have improved the process yield.

This test procedure is sometimes called the **Fisher-Irwin test.** Because the test depends on the assumption that $X_1 + X_2$ is fixed at some value, some statisticians argue against use of the test when $X_1 + X_2$ is not actually fixed. Clearly $X_1 + X_2$ is not fixed by the sampling procedure in our example. However, because there are no other better competing procedures, the Fisher-Irwin test is often used whether or not $X_1 + X_2$ is actually fixed in advance.