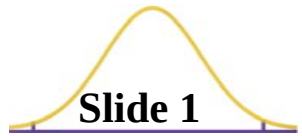
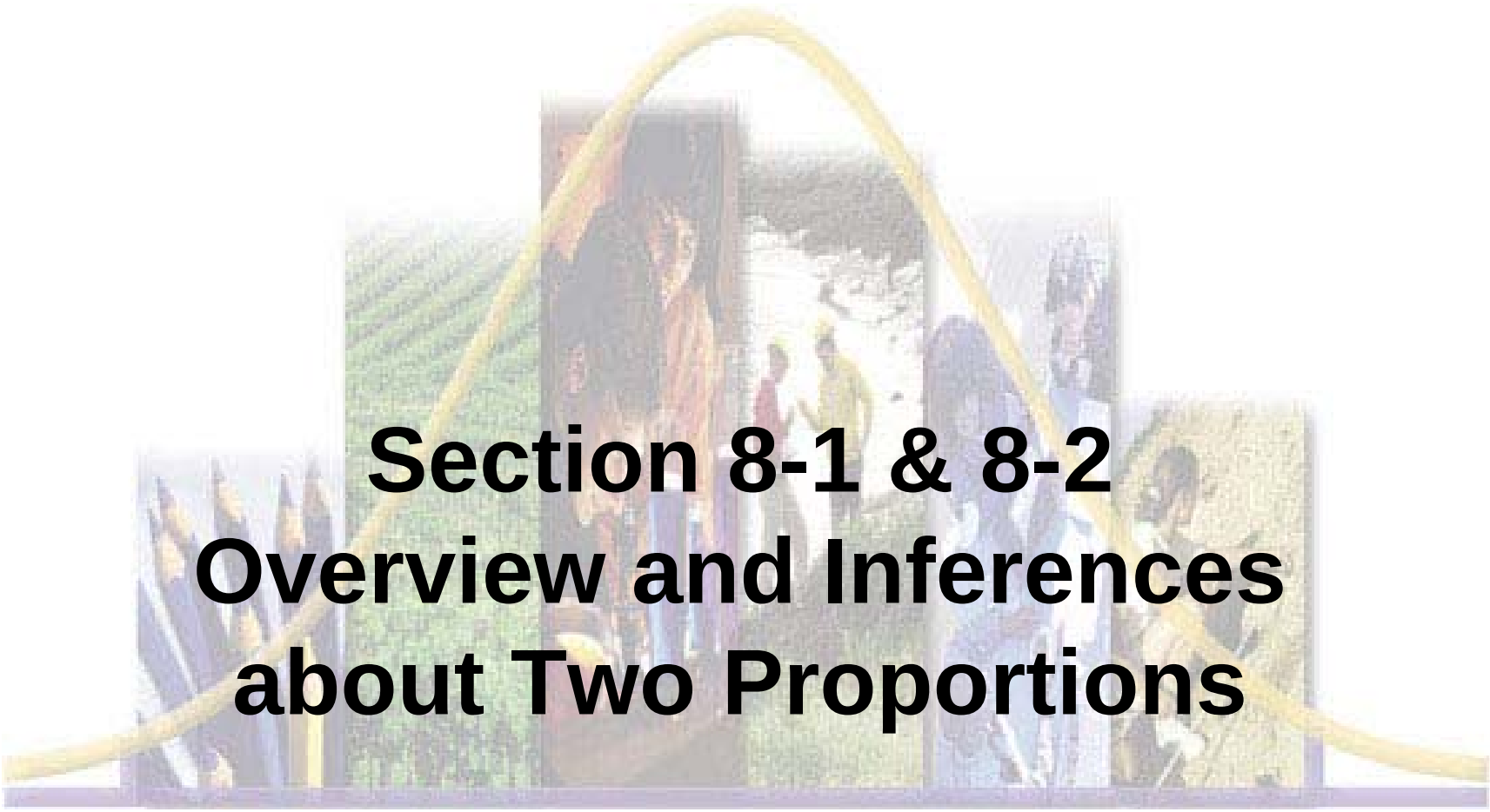
Chapter 8
Inferences from Two Samples





Section 8-1 & 8-2 Overview and Inferences about Two Proportions

Created by Erin Hodgess, Houston, Texas

Overview (p.438)

There are many important and meaningful situations in which it becomes necessary to compare two sets of sample data.

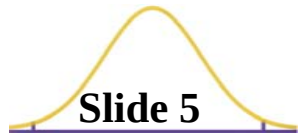
Inferences about Two Proportions

Assumptions (p.439)

1. We have proportions from two **independent** simple random samples.

2. For both samples, the conditions $np \geq 5$ and $nq \geq 5$ are satisfied.

Notation for Two Proportions



For population 1, we let:

p_1 = population proportion

n_1 = size of the sample

x_1 = number of successes in the sample

$\hat{p}_1 = \frac{x_1}{n_1}$ (the *sample* proportion)

$\hat{q}_1 = 1 - \hat{p}_1$

Corresponding meanings are attached to p_2, n_2, x_2, $\hat{p}_2$. and $\hat{q}_2$, which come from population 2.

Pooled Estimate of p_1 and p_2

- The **pooled estimate of p_1 and p_2** is denoted by $\bar{p}$.

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

- $\bar{q} = 1 - \bar{p}$

Test Statistic for Two Proportions (p.441)

For $H_0: p_1 = p_2$, $H_0: p_1 = p_2$, $H_0: p_1 = p_2$

$H_1: p_1 \neq p_2$, $H_1: p_1 < p_2$, $H_1: p_1 > p_2$

where $p_1 - p_2 = 0$ **(assumed in the null hypothesis)**

$$\hat{p}_1 = \frac{x_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{x_2}{n_2}$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \quad \text{and} \quad \bar{q} = 1 - \bar{p}$$

Test Statistic for Two Proportions (p.441)

For $H_0: p_1 = p_2$, $H_0: p_1 = p_2$, $H_0: p_1 = p_2$

$H_1: p_1 \neq p_2$, $H_1: p_1 < p_2$, $H_1: p_1 > p_2$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

Example: **For the sample data listed in Table 8-1, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped. (p.441)**

Table 8-1 Racial Profiling Data

	Race and Ethnicity	
	Black and Non-Hispanic	White and Non-Hispanic
Drivers stopped by police	24	147
Total number of observed drivers	200	1400
Percent Stopped by Police	**12.0%**	**10.5%**

Example: **For the sample data listed in Table 8-1, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped.**

$n_1 = 200$

$x_1 = 24$

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{24}{200} = 0.120$

$n_2 = 1400$

$x_2 = 147$

$\hat{p}_2 = \frac{x_2}{n_2} = \frac{147}{1400} = 0.105$

$H_0: p_1 = p_2, H_1: p_1 > p_2$

$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{24 + 147}{200+1400} = 0.106875$

$\overline{q} = 1 - 0.106875 = 0.893125.$

Example: For the sample data listed in Table 8-1, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped.

$n_1 = 200$

$x_1 = 24$

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{24}{200} = 0.120$

$n_2 = 1400$

$x_2 = 147$

$\hat{p}_2 = \frac{x_2}{n_2} = \frac{147}{1400} = 0.105$

$$z = \frac{(0.120 - 0.105) - 0}{\sqrt{\frac{(0.106875)(0.893125)}{200} + \frac{(0.106875)(0.893125)}{1400}}}$$

$z = 0.64$

Example: For the sample data listed in Table 8-1, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped.

$n_1 = 200$

$x_1 = 24$

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{24}{200} = 0.120$

$n_2 = 1400$

$x_2 = 147$

$\hat{p}_2 = \frac{x_2}{n_2} = \frac{147}{1400} = 0.105$

$(0.120 - 0.105) - 0.040 < (p_1 - p_2) < (0.120 - 0.105) + 0.040$

$-0.025 < (p_1 - p_2) < 0.055$

Example: **For the sample data listed in Table 8-1, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped.**

$n_1 = 200$

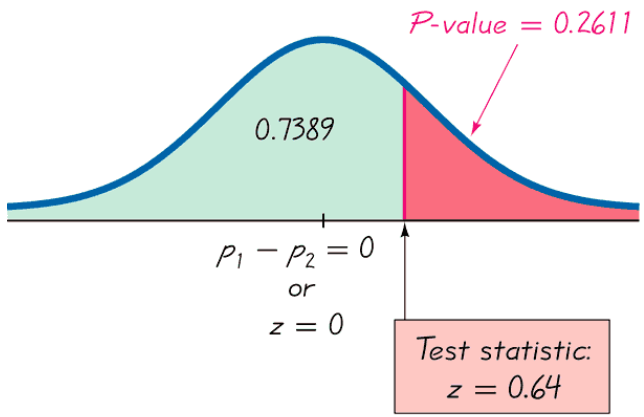
$x_1 = 24$

$\hat{p}_1 = \dfrac{x_1}{n_1} = \dfrac{24}{200} = 0.120$

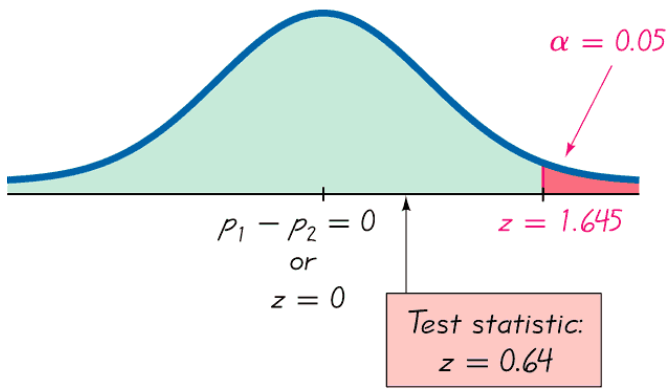
$n_2 = 1400$

$x_2 = 147$

$\hat{p}_2 = \dfrac{x_2}{n_2} = \dfrac{147}{1400} = 0.105$



(a) P-Value Method



(b) Traditional Method

Example: **For the sample data listed in Table 8-1, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped.**

$n_1 = 200$

$x_1 = 24$

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{24}{200} = 0.120$

$n_2 = 1400$

$x_2 = 147$

$\hat{p}_2 = \frac{x_2}{n_2} = \frac{147}{1400} = 0.105$

$z = 0.64$

This is a right-tailed test, so the P-value is the area to the right of the test statistic $z = 0.64$. The P-value is 0.2611.

Because the P-value of 0.2611 is greater than the significance level of $\alpha = 0.05$, we fail to reject the null hypothesis.

Example: **For the sample data listed in Table 8-1, use a 0.05 significance level to test the claim that the proportion of black drivers stopped by the police is greater than the proportion of white drivers who are stopped.**

$n_1 = 200$

$x_1 = 24$

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{24}{200} = 0.120$

$n_2 = 1400$

$x_2 = 147$

$\hat{p}_2 = \frac{x_2}{n_2} = \frac{147}{1400} = 0.105$

$z = 0.64$

Because we fail to reject the null hypothesis, we conclude that there is not sufficient evidence to support the claim that the proportion of black drivers stopped by police is greater than that for white drivers. This does *not* mean that racial profiling has been disproved. The evidence might be strong enough with more data.

Confidence Interval

Estimate of $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E$$

where $$E = Z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Example: **For the sample data listed in Table 8-1, find a 90% confidence interval estimate of the difference between the two population proportions. (p.444)**

$n_1 = 200$

$x_1 = 24$

$\hat{p}_1 = \frac{x_1}{n_1} = \frac{24}{200} = 0.120$

$n_2 = 1400$

$x_2 = 147$

$\hat{p}_2 = \frac{x_2}{n_2} = \frac{147}{1400} = 0.105$

$$E = Z_{\alpha/2}\sqrt{\frac{\hat{p}_1\,\hat{q}_1}{n_1} + \frac{\hat{p}_2\,\hat{q}_2}{n_2}}$$

$$E = 1.645\sqrt{\frac{(.12)(.88)}{200} + \frac{(0.105)(0.895)}{1400}}$$

$$E = 0.400$$

Section 8-3 Inferences about Two Means: Independent Samples

Created by Erin Hodgess, Houston, Texas

Definitions

Two Samples: Independent

The sample values selected from one population are not related or somehow paired with the sample values selected from the other population.

If the values in one sample are related to the values in the other sample, the samples are dependent. Such samples are often referred to as matched pairs or paired samples.

Assumptions (p.453)

1. **The two samples are *independent*.**

2. **Both samples are *simple random samples.***

3. **Either or both of these conditions are satisfied: The two sample sizes are both *large* (with $n_1 > 30$ and $n_2 > 30$) or both samples come from populations having normal distributions.**

Hypothesis Tests

Test Statistic for Two Means:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Hypothesis Tests

Test Statistic for Two Means:

Degrees of freedom: In this book we use this estimate: *df* = smaller of $n_1 - 1$ and $n_2 - 1$.

***P*-value:** Refer to Table A-3. Use the procedure summarized in Figure 7-6.

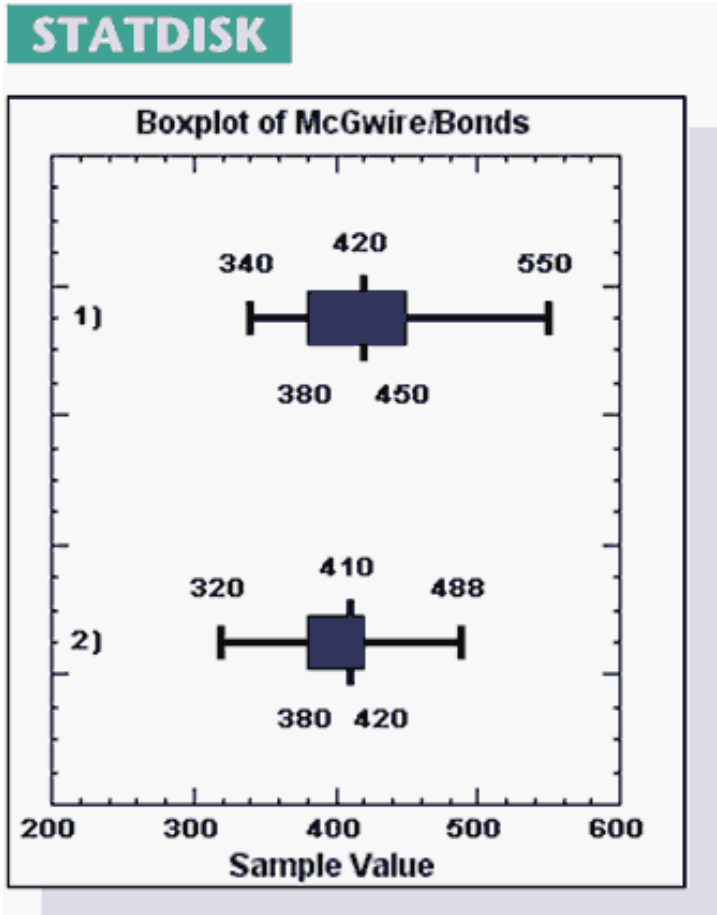
Critical values: Refer to Table A-3.

McGwire Versus Bonds (p.455)

Data Set 30 in Appendix B includes the distances of the home runs hit in record-setting seasons by Mark McGwire and Barry Bonds. Sample statistics are shown. Use a 0.05 significance level to test the claim that the distances come from populations with different means.

	McGwire	Bonds
n	70	73
$\overline{x}$	418.5	403.7
s	45.5	30.6

McGwire Versus Bonds



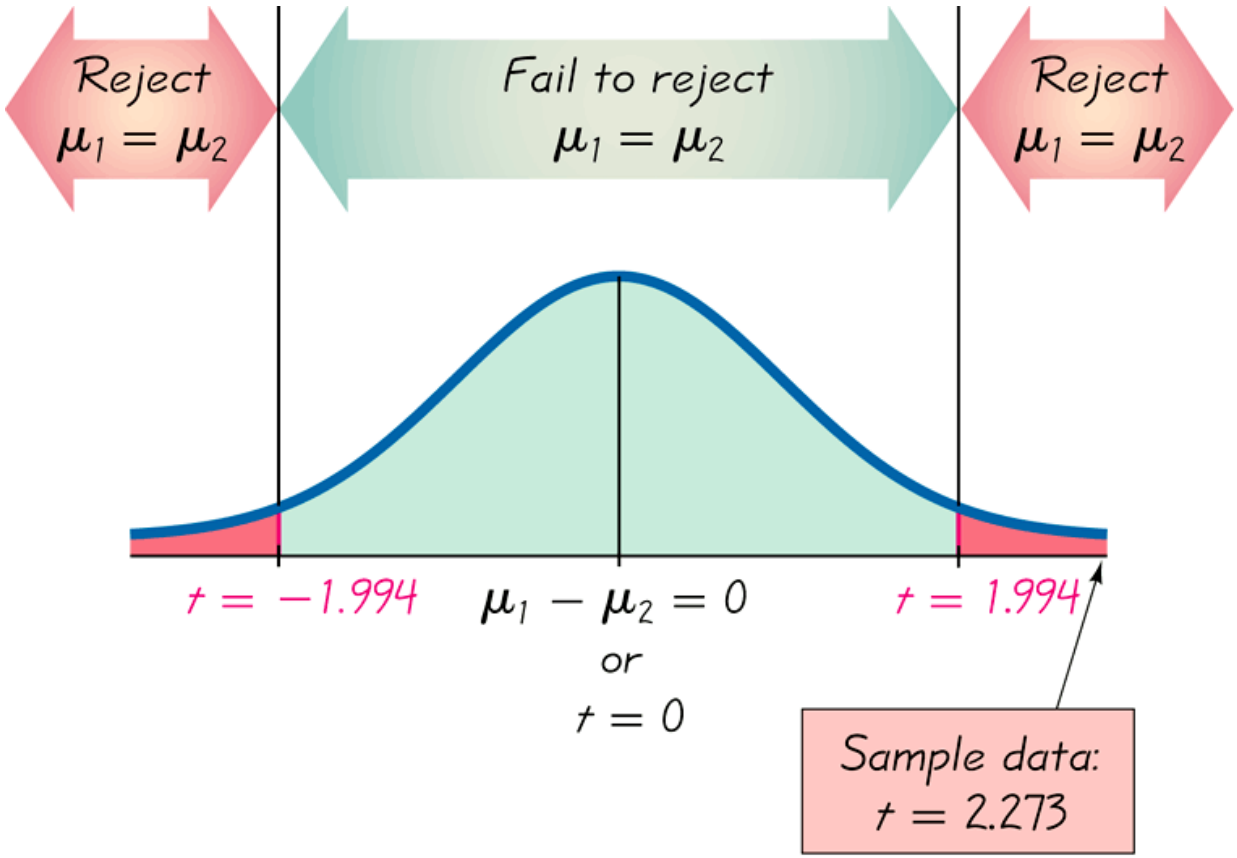
McGwire Versus Bonds

Claim: $\mu_1 \neq \mu_2$

$H_o : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$\alpha = 0.05$



$n_1 - 1 = 69$

$n_2 - 1 = 72$

$df = 69$

$t_{.025} = 1.994$

McGwire Versus Bonds

Test Statistic for Two Means:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

McGwire Versus Bonds

Test Statistic for Two Means:

$$t = \frac{(418.5 - 403.7) - 0}{\sqrt{\frac{45.5^2}{70} + \frac{30.6^2}{73}}}$$

$$= 2.273$$

McGwire Versus Bonds

Claim: $\mu_1 \neq \mu_2$

$H_o : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$\alpha = 0.05$

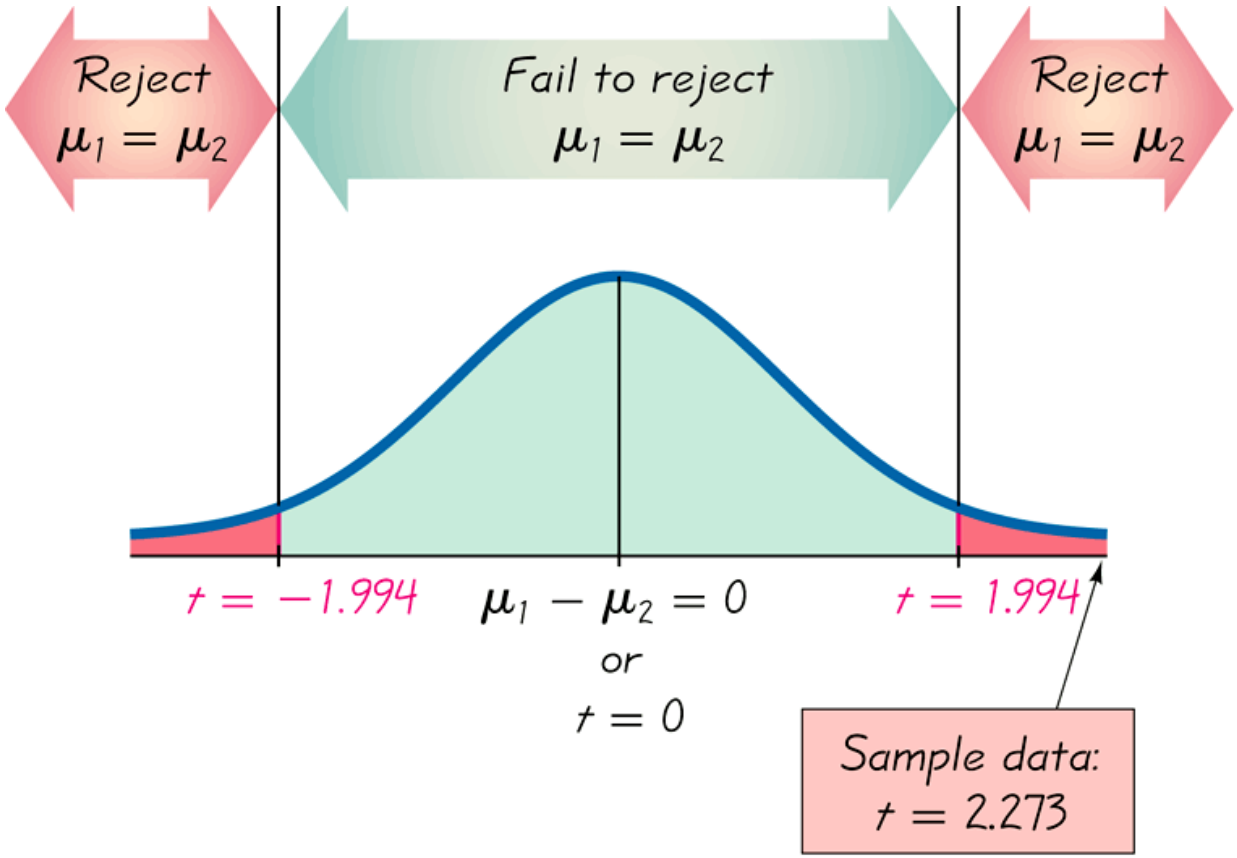


Figure 8-2

McGwire Versus Bonds

Claim: $\mu_1 \neq \mu_2$

$H_o : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$\alpha = 0.05$

There is significant evidence to support the claim that there is a difference between the mean home run distances of Mark McGwire and Barry Bonds.

Reject Null

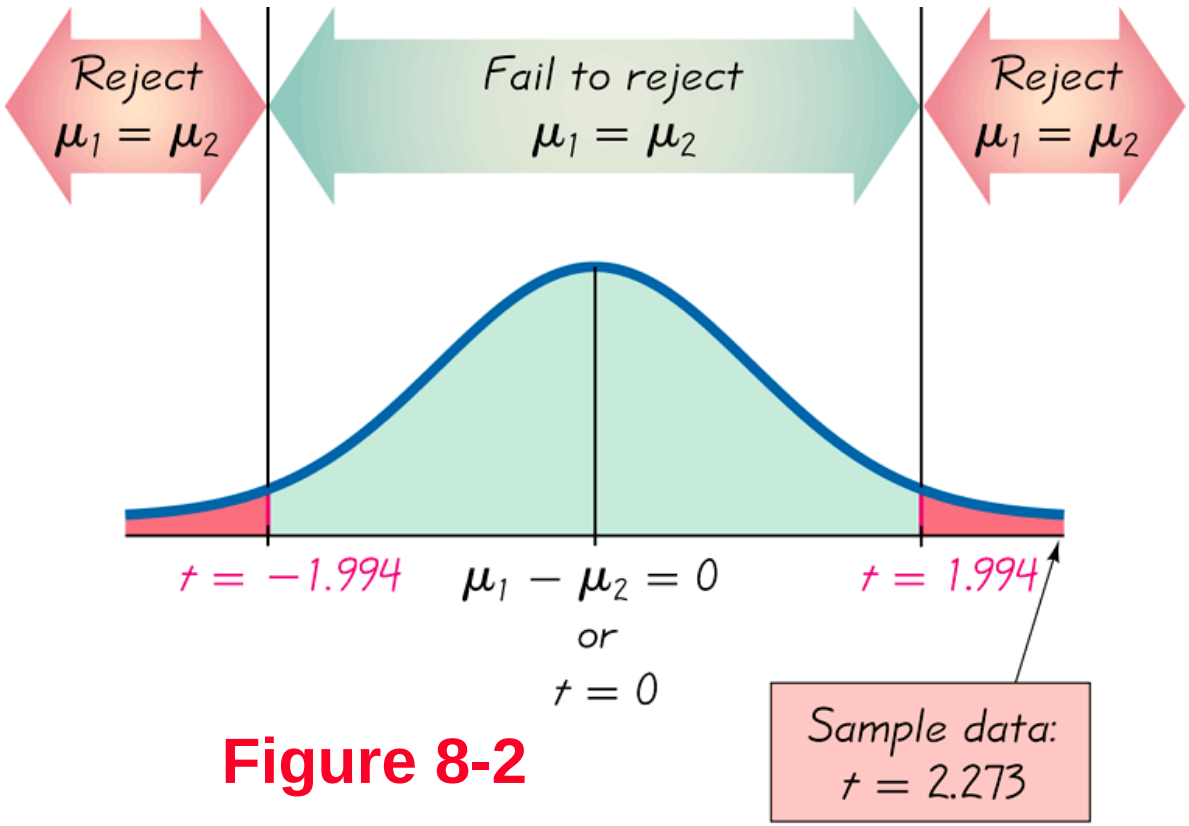


Figure 8-2

Confidence Intervals

$$(\bar{x}_1 - \bar{x}_2) - E < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + E$$

where $$E = t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

McGwire Versus Bonds (p.457)

Using the sample data given in the preceding example, construct a 95%confidence interval estimate of the difference between the mean home run distances of Mark McGwire and Barry Bonds.

$$E = t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$E = 1.994 \sqrt{\frac{45.5^2}{70} + \frac{30.6^2}{73}}$$

$$E = 13.0$$

McGwire Versus Bonds

Using the sample data given in the preceding example, construct a 95%confidence interval estimate of the difference between the mean home run distances of Mark McGwire and Barry Bonds.

$$(418.5 - 403.7) - 13.0 < (\mu_1 - \mu_2) < (418.5 - 403.7) + 13.0$$
$$1.8 < (\mu_1 - \mu_2) < 27.8$$

We are 95% confident that the limits of 1.8 ft and 27.8 ft actually do contain the difference between the two population means.

Section 8-4
Inferences from Matched Pairs

Created by Erin Hodgess, Houston, Texas

Assumptions (p.467)

1. The sample data consist of matched pairs.

2. The samples are simple random samples.

3. Either or both of these conditions is satisfied: The number of matched pairs of sample data is ($n > 30$) or the pairs of values have differences that are from a population having a distribution that is approximately normal.

Notation for Matched Pairs

μ_d = mean value of the differences d for the **population** of paired data

$\bar{d}$ = mean value of the differences d for the paired **sample** data (equal to the mean of the $x - y$ values)

s_d = standard deviation of the differences d for the paired **sample** data

n = number of **pairs** of data.

Test Statistic for Matched Pairs of Sample Data (p.467)

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

where degrees of freedom = $n - 1$

P-values and Critical Values

Use Table A-3 (t-distribution) on p.736.

Confidence Intervals

$$\bar{d} - E < \mu_d < \bar{d} + E$$

where $E = t_{\alpha/2} \frac{s_d}{\sqrt{n}}$

degrees of freedom = n –1

Are Forecast Temperatures Accurate?

Using Table A-2 consists of five actual low temperatures and the corresponding low temperatures that were predicted five days earlier. The data consist of matched pairs, because each pair of values represents the same day. Use a 0.05 significant level to test the claim that there is a difference between the actual low temperatures and the low temperatures that were forecast five days earlier. (p.468)

Are Forecast Temperatures Accurate?

Table 8-2 Actual and Forecast Temperature

Actual low	1	−5	−5	23	9
Low forecast five days earlier	16	16	20	22	15
Difference d = actual − predicted	−15	−21	−25	1	−6

Are Forecast Temperatures Accurate?

$\bar{d} = -13.2$

$s = 10.7$

$n = 5$

$t_{\alpha/2} = 2.776$ (found from Table A-3 with 4 degrees of freedom and 0.05 in two tails)

Are Forecast Temperatures Accurate?

$H_0: \mu_d = 0$

$H_1: \mu_d \neq 0$

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{-13.2 - 0}{\frac{10.7}{\sqrt{5}}} = -2.759$$

Are Forecast Temperatures Accurate?

$H_0: \mu_d = 0$

$H_1: \mu_d \neq 0$

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{-13.2 - 0}{\frac{10.7}{\sqrt{5}}} = -2.759$$

Because the test statistic does not fall in the critical region, we fail to reject the null hypothesis.

Are Forecast Temperatures Accurate?

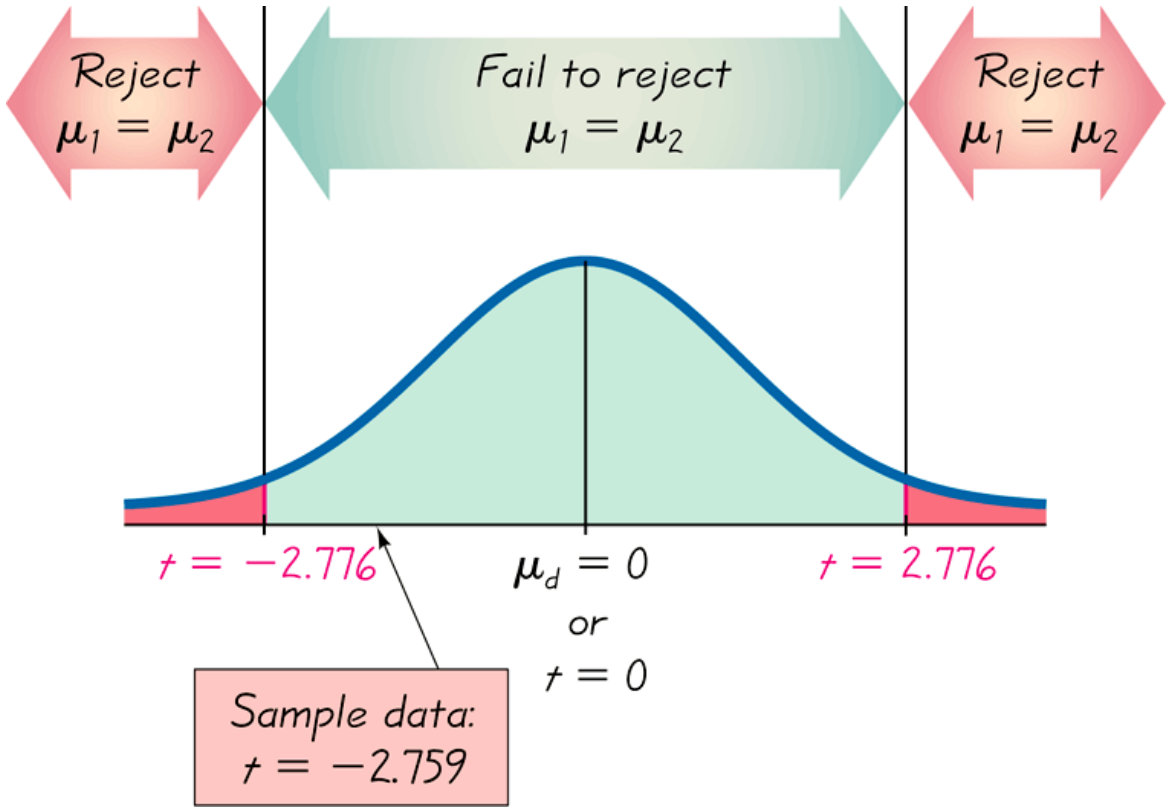
$H_0: \mu_d = 0$

$H_1: \mu_d \neq 0$

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{-13.2 - 0}{\frac{10.7}{\sqrt{5}}} = -2.759$$

The sample data in Table 8-2 do not provide sufficient evidence to support the claim that actual and five-day forecast low temperatures are different.

Are Forecast Temperatures Accurate? (p.469)



Are Forecast Temperatures Accurate? (p.470)

Using the same sample matched pairs in Table 8-2, construct a 95% confidence interval estimate of μ_d, which is the mean of the differences between actual low temperatures and five-day forecasts.

Are Forecast Temperatures Accurate?

$$E = t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

$$E = (2.776)\left(\frac{10.7}{\sqrt{5}}\right)$$

$$= 13.3$$

Are Forecast Temperatures Accurate?

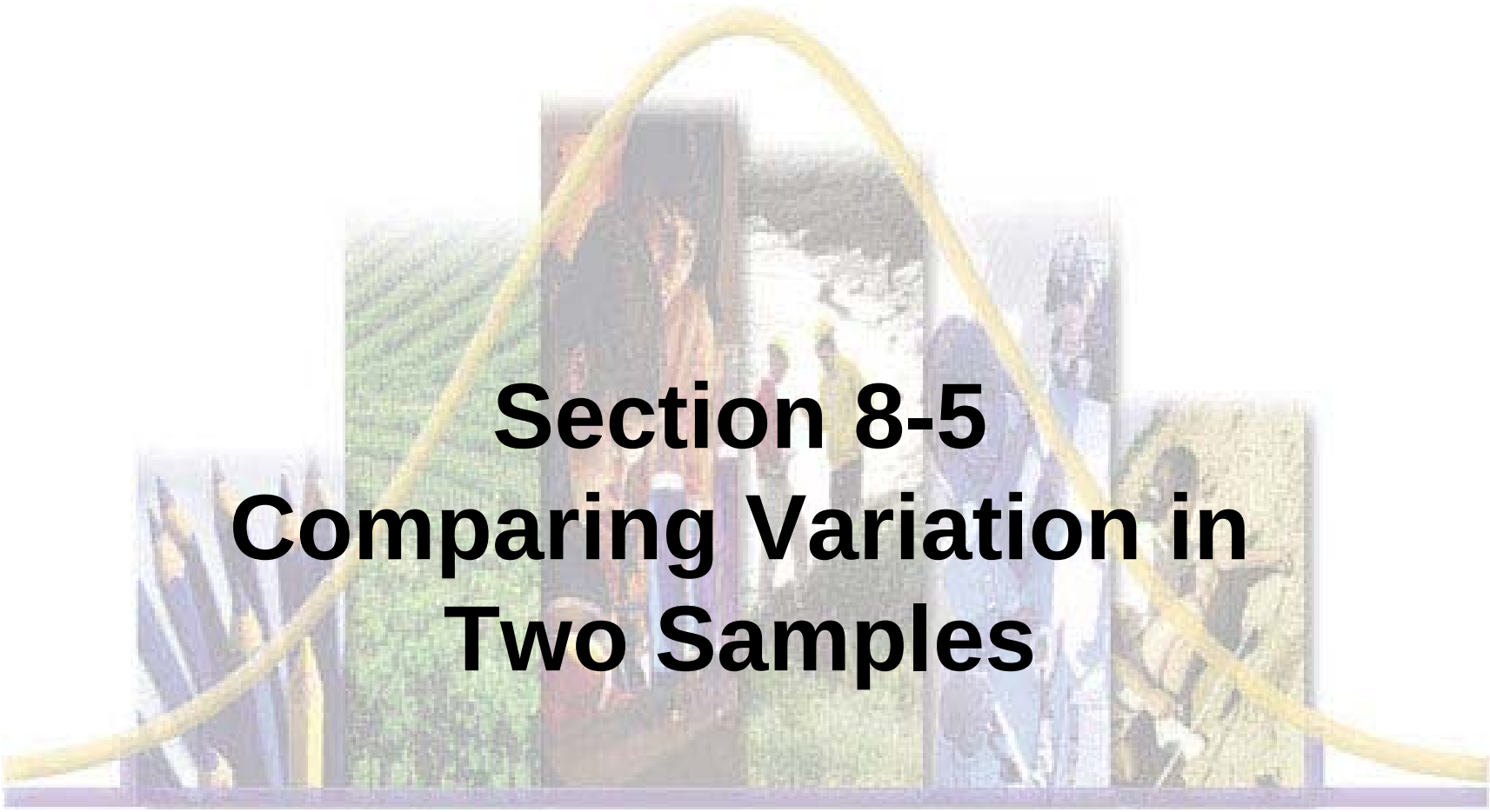
$$\bar{d} - E < \mu_d < \bar{d} + E$$

$$-13.2 - 13.3 < \mu_d < -13.2 + 13.3$$

$$-26.5 < \mu_d < 0.1$$

Are Forecast Temperatures Accurate? (p.470)

In the long run, 95% of such samples will lead to confidence intervals that actually do contain the true population mean of the differences.



Section 8-5
Comparing Variation in Two Samples

Created by Erin Hodgess, Houston, Texas

Measures of Variation (p.476)

s = standard deviation of sample

σ = standard deviation of population

s^2 = variance of sample

σ^2 = variance of population

Assumptions

1. The two populations are *independent* of each other.

2. The two populations are each normally distributed.

Notation for Hypothesis Tests with Two Variances

s_1^2 = ***larger*** **of the two sample variances**

n_1 = **size of the sample with the *larger* variance**

σ_1^2 = **variance of the population from which the sample with the *larger* variance was drawn**

The symbols s_2^2 , n_2 , and σ_2^2 are used for the other sample and population.

Test Statistic for Hypothesis Tests with Two Variances

$$F = \frac{s_1^2}{s_2^2}$$

Critical Values: **Using Table A-5, we obtain critical *F* values that are determined by the following three values:**

1. **The significance level α.**
2. **Numerator degrees of freedom (df_1) = $n_1 - 1$**
3. **Denominator degrees of freedom (df_2) = $n_2 - 1$**

- **All one-tailed tests will be right-tailed.**

- **All two-tailed tests will need only the critical value to the right.**

- **When degrees of freedom is not listed exactly, use the critical values on either side as an interval. Use interpolation only if the test statistic falls within the interval.**

If the two populations do have equal variances, then $F = \frac{s_1^2}{s_2^2}$ will be close to 1 because s_1^2 and s_2^2 are close in value. (p.478)

If the two populations have radically different variances, then *F* will be a large number.

Remember, the larger sample variance will be s_1^2.

Consequently, a value of *F* near 1 will be evidence in favor of the conclusion that $\sigma_1^2 = \sigma_2^2$.

But a large value of *F* will be evidence against the conclusion of equality of the population variances.

Coke Versus Pepsi (p.480)

Data Set 17 in Appendix B includes the weights (in pounds) of samples of regular Coke and regular Pepsi. Sample statistics are shown. Use the 0.05 significance level to test the claim that the weights of regular Coke and the weights of regular Pepsi have the same standard deviation.

	Regular Coke	Regular Pepsi
n	36	36
$\bar{x}$	0.81682	0.82410
s	0.007507	0.005701

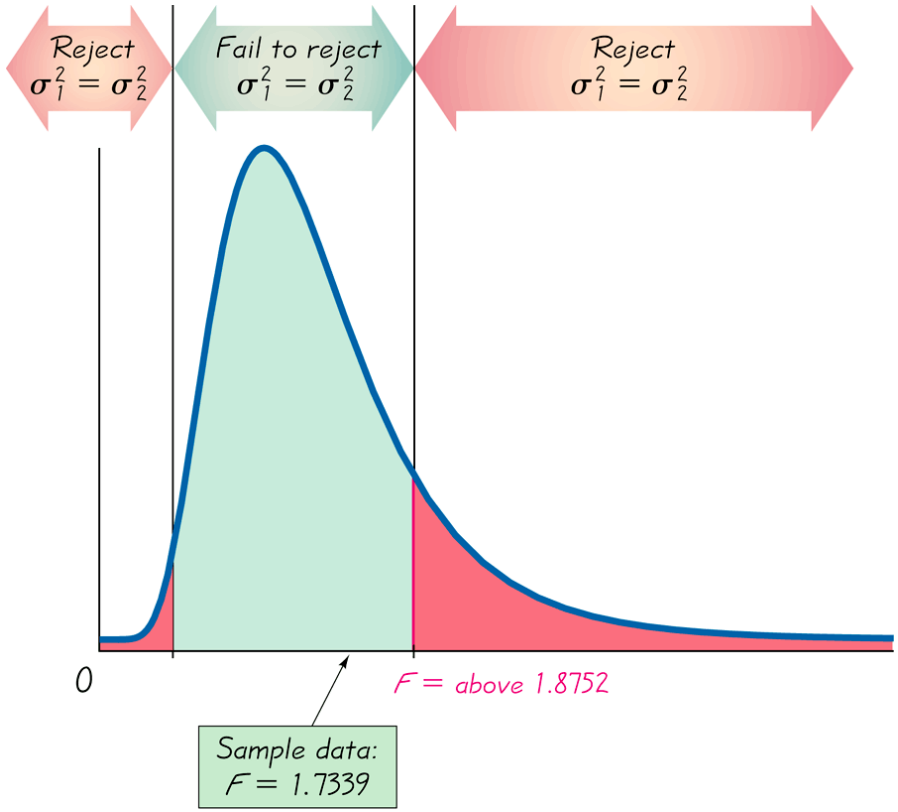
Coke Versus Pepsi

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

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

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

$\alpha = 0.05$



$$\text{Value of } F = \frac{s_1^2}{s_2^2}$$

$$= \frac{0.007507^2}{0.005701^2}$$

$$= 1.7339$$

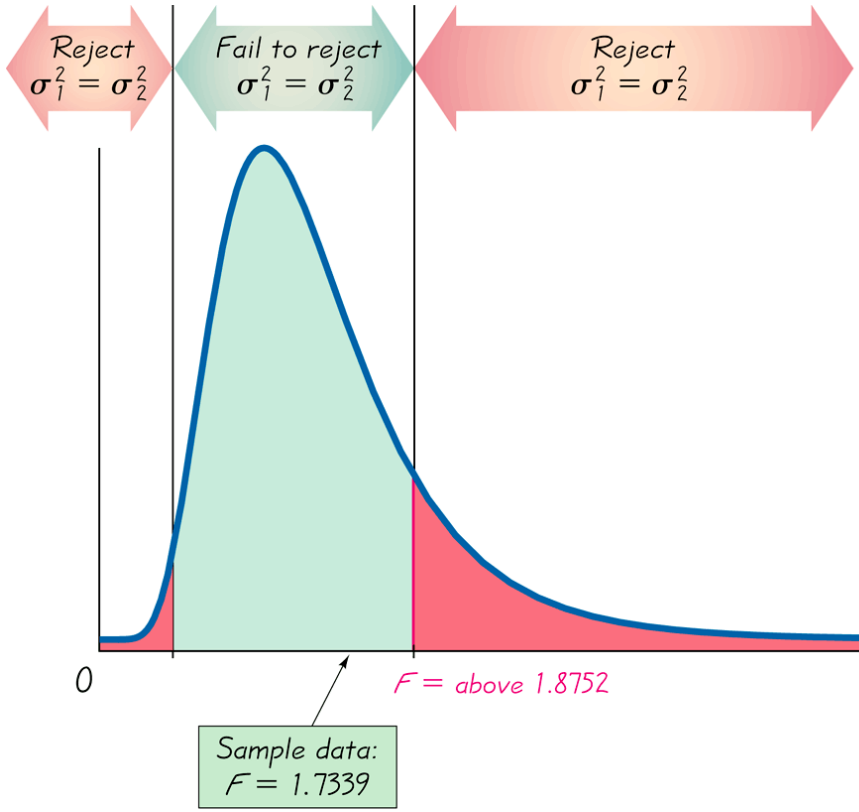
Coke Versus Pepsi

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

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

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

$\alpha = 0.05$



There is not sufficient evidence to warrant rejection of the claim that the two variances are equal.