CHAPTER 9

Testing the Difference Between Two Means, Two Variances, and Two Proportions

To Vaccinate or Not to Vaccinate? Small or Large?

Influenza is a serious disease among the elderly, especially those living in nursing homes. Those residents are more susceptible to influenza than elderly persons living in the community because the former are usually older and exposed more so than community residents to the virus. Three researchers decided to investigate the use of vaccine and its value in determining outbreaks of influenza in small nursing homes. They surveyed 83 licensed homes in seven counties in Michigan. Part of the study consisted of comparing the number of people being vaccinated in small nursing homes (100 or fewer beds) with the number in larger nursing homes (more than 100 beds). Unlike the statistical methods presented in Chapter 8, they used the techniques explained in this chapter to compare two sample proportions to see if there was a significant difference in the vaccination rates of patients in small nursing homes compared to those in larger nursing homes.

Objectives

- Test the difference between two large sample means using the *z* test.
- Test the difference between two variances or standard deviations.

Objectives (cont'd.)

- Test the difference between two means for small independent samples.
- Test the difference between two means for small dependent samples.
- Test the difference between two proportions.

The Difference Between Two Means

Suppose a researcher wishes to determine whether there is a difference in the average age of nursing students who enroll in a nursing program at a community college and those who enroll in a nursing program at a university. In this case, the researcher is not interested in the average age of all beginning nursing students; instead, he is interested in comparing the means of the two groups. His research question is: Does the mean age of nursing students who enroll at a community college differ from the mean age of nursing students who enroll at a university? Here, the hypotheses are

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

where

μ_1 = mean age of all beginning nursing students at the community college

The Difference Between Two Means (cont'd)

μ_2 = mean age of all beginning nursing students at the university.

Another way of stating the hypotheses for this situation is

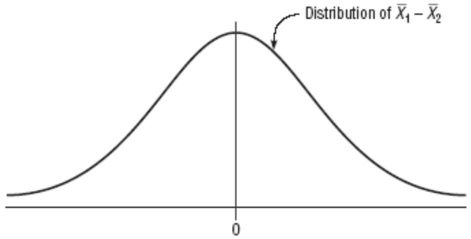
H_0: $\mu_1 - \mu_2 = 0$ and H_1: $\mu_1 - \mu_2 \neq 0$

If there is no difference in population means, subtracting them will give a difference of zero. If they are different, subtracting will give a number other than zero. Both methods of stating hypotheses are correct; however, the first method will be used in this chapter.

The theory behind testing the difference between two means is based on selecting pairs of samples and comparing the means of the pairs. The population means need not be known.

The Difference Between Two Means (cont'd)

All possible pairs of samples are taken from populations. The means for each pair of samples are computed and then subtracted, and the differences are plotted. If both populations have the same mean, then most of the differences will be zero or close to zero. Occasionally, there will be a few large differences due to chance alone, some positive and others negative. If the differences are plotted, the curve will be shaped like the normal distribution and have a mean of zero,



The Difference Between Two Means (cont'd)

- Assumptions for the test to determine the difference between two means:
  1. The samples must be independent of each other; that is, there can be no relationship between the subjects in each sample.
  2. The populations from which the samples were obtained must be normally distributed, and the standard deviations of the variable must be known, or the sample sizes must be greater than or equal to 30.

Formula for the z Test

- The variance of the difference $\overline{X}_1 - \overline{X}_2$ is equal to the sum of the individual variance of $\overline{X}_1$ and $\overline{X}_2$. That is

$$\sigma^2_{\overline{X}_1 - \overline{X}_2} = \sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2}$$

Where $\sigma^2_{\overline{X}_1} = \dfrac{\sigma_1^2}{n_1}$ and $\sigma^2_{\overline{X}_2} = \dfrac{\sigma_2^2}{n_2}$

- Formula for the z test for comparing two means from independent populations

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

Formula for the z Test (cont'd)

- In the formula, $\overline{X}_1 - \overline{X}_2$ is the observed difference, and the expected difference $\mu_1 - \mu_2$ is 0 when the null hypothesis is $\mu_1 = \mu_2$, since that is equivalent to $\mu_1 - \mu_2 = 0$.
- In the comparison of two sample means, the difference maybe due to chance, in which case the null hypothesis will not be rejected, and the researcher can assume that the means of the populations are basically the same.

z Test for Comparing Two Means from Independent Populations

These tests can also be one-tailed, using the following hypotheses:

Right-tailed			Left-tailed		
H_0: $\mu_1 \le \mu_2$ H_1: $\mu_1 > \mu_2$	or	H_0: $\mu_1 - \mu_2 \le 0$ H_1: $\mu_1 - \mu_2 > 0$	H_0: $\mu_1 \ge \mu_2$ H_1: $\mu_1 < \mu_2$	or	H_0: $\mu_1 - \mu_2 \ge 0$ H_1: $\mu_1 - \mu_2 < 0$

The same critical values used in the last chapter are used here. Again, they can be obtained from the Table of the Standard Normal Distribution.

z Test for Comparing Two Means from Independent Populations (cont'd)

If σ_1 and σ_2 are not known, the researcher can use the variance obtained from each sample s_1 and s_2, but the sample size must be 30 or more. The formula then is

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

provided that $n_1 \ge 30$ and $n_2 \ge 30$.

When one or both sample sizes are less than 30 and σ_1 and σ_2 are unknown, the t test must be used, as explained later in this chapter.

An Example

A survey found that the average hotel room rate in New Orleans is \$88.42 and the average room rate in Phoenix is \$80.61. Assume that the data were obtained from two samples of 50 hotels each and that the standard deviations were \$5.62 and \$4.83, respectively. At $\sigma = 0.05$, can it be concluded that there is a significant difference in the rates? (*Source: USA TODAY*)

Solution:

STEP 1 State the hypotheses and identify the claim.

H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$ (claim)

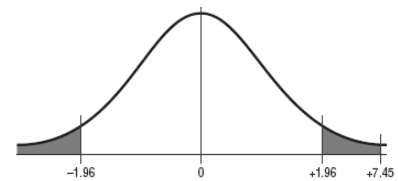
STEP 2 Find the critical values. Since $\alpha = 0.05$, the critical values are ± 1.96.

An Example (cont'd)

STEP 3 Compute the test value.

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(88.42 - 80.61) - 0}{\sqrt{\frac{5.62^2}{50} + \frac{4.83^2}{50}}} = 7.45$$

STEP 4 Make the decision. Reject the null hypothesis at $\alpha = 0.05$, since 7.45 > 1.96.



STEP 5 Summarise the results. There is enough evidence to support the claim that the means are not equal.

the *P*-value Method

- The *P*-values for this test can be determined by using the same procedure shown in the last chapter. For example, if the test value for a two-tailed test is 1.40, then the *P*-value obtained from the Table of the Standard Normal Distribution is 0.1616. This value is obtained by looking up the area for $z = 1.40$, which is 0.4192. Then 0.4192 is subtracted from 0.5000 to get 0.0808. Finally, this value is doubled to get 0.1616 since the test is two-tailed. If $\alpha = 0.05$, the decision would be to not reject the null hypothesis, since *P*-value > α.

An Example

A researcher hypothesizes that the average number of sports that colleges offer for males is greater than the average number of sports that colleges offer for females. A sample of the number of sports offered by colleges is shown. At $\alpha = 0.10$, is there enough evidence to support the claim?

Males					Females				
6	11	11	8	15	6	8	11	13	8
6	14	8	12	18	7	5	13	14	6
6	9	5	6	9	6	5	5	7	6
6	9	18	7	6	10	7	6	5	5
15	6	11	5	5	16	10	7	8	5
9	9	5	5	8	7	5	5	6	5
8	9	6	11	6	9	18	13	7	10
9	5	11	5	8	7	8	5	7	6
7	7	5	10	7	11	4	6	8	7
10	7	10	8	11	14	12	5	8	5

An Example (cont'd)

Solution:

STEP 1 State the hypotheses and identify the claim:

H_0: $\mu_1 \leq \mu_2$ and H_1: $\mu_1 > \mu_2$ (claim)

STEP 2 Compute the test value. Using the formulas in Chapter 3, find the mean and standard deviation for each data set.

For the males $\overline{X}_1 = 8.6$ and $s_1 = 3.3$

For the females $\overline{X}_2 = 7.9$ and $s_2 = 3.3$

Substitute in the formula

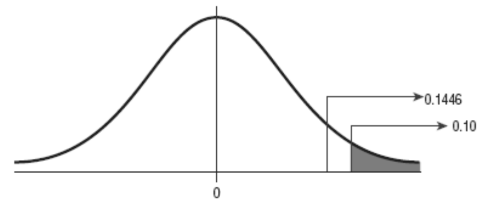
$$z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(8.6 - 7.9) - 0}{\sqrt{\frac{3.3^2}{50} + \frac{3.3^2}{50}}} = 1.06$$

An Example (cont'd)

STEP 3 Find the *P*-value. For $z = 1.06$, the area is 0.3554, and 0.5000 - 0.3554 = 0.1446 or a *P*-value of 0.1446.

STEP 4 Make the decision. Since the *P*-value is larger than α (that is, 0.1446 > 0.10), the decision is to not reject the null hypothesis.

STEP 5 Summarise the results. There is not enough evidence to support the claim that colleges offer more sports for males than they do for females.



Difference Between Two Means

- For Large Samples:

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

- When $n_1 \geq 30$ and $n_2 \geq 30$, s_1^2 and s_2^2 can be used in place of σ_1^2 and σ_2^2.

An Example

Find the 95% confidence interval for the difference between the means for the data in the example of the average room rate comparison.

Solution:

Substitute in the formula, using $z_{\alpha/2} = 1.96$.

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

$$(88.42 - 80.61) - 1.96\sqrt{\frac{5.62^2}{50} + \frac{4.83^2}{50}} < \mu_1 - \mu_2 < (88.42 - 80.61) + 1.96\sqrt{\frac{5.62^2}{50} + \frac{4.83^2}{50}}$$

$$5.76 < \mu_1 - \mu_2 < 9.86$$

Since the confidence interval does not contain zero, the decision is to reject the null hypothesis, which agrees with the previous result.

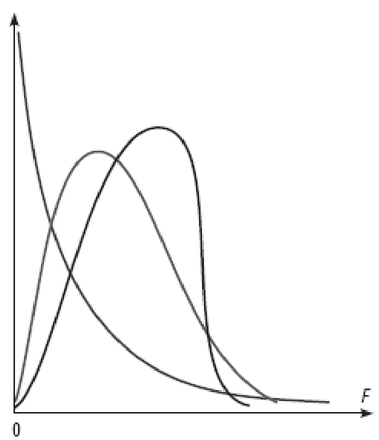
Testing the Difference Between Two Variances

- In addition to comparing two means, statisticians are interested in comparing two variances or standard deviations. For example, is the variation in the temperatures for a certain month for two cities different? In another situation, a researcher may be interested in comparing the variance of the cholesterol of men with the variance of the cholesterol of women. For the comparison of two variances or standard deviations, an ***F* test** is used. The *F* test should not be confused with the chi-square test, which compares a single sample variance to a specific population variance, as shown in Chapter 8.

F Distribution

- If two independent samples are selected from two normally distributed populations in which the variances are equal ($\sigma_1^2 = \sigma_2^2$) and if the variances s_1^2 and s_2^2 are compared as $\frac{s_1^2}{s_2^2}$, the sampling distribution of the ratio of the variances is called the *F distribution.*

F Distribution (cont'd)



Characteristics of the *F* Distribution

1. The values of *F* cannot be negative, because variances are always positive or zero.
2. The distribution is positively skewed.
3. The mean value of *F* is approximately equal to 1.
4. The *F* distribution is a family of curves based on the degrees of freedom of the variance of the numerator and the degrees of freedom of the variance of the denominator.

Using F test for Testing the Difference Between Two Variances

- The F test:

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

where s_1^2 is the larger of the two variances.

- The F test has two terms for the degrees of freedoms: that of the numerator, n_1-1, and that of the denominator, n_2-1, where n_1 is the sample size from which the larger variance was obtained.

Notes for the Use of the F test

1. The larger variance should always be designated as s_1^2 and be placed in the numerator of the formula.
2. For a two-tailed test, the α value must be divided by 2 and the critical value be placed on the right side of the F curve.

Notes for the Use of the F test (cont'd.)

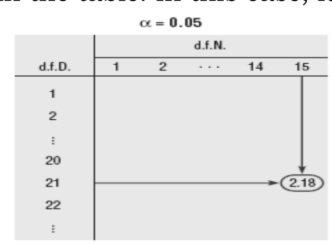
3. If the standard deviations instead of the variances are given in the problem, they must be squared for the formula for the F test.
4. When the degrees of freedom cannot be found in Table H, the closest value on the smaller side should be used.

Some Examples

1. Find the critical value for a right-tailed F test when $\alpha = 0.05$, the d.f. for the numerator (abbreviated d.f.N.) are 15, and the degrees of freedom for the denominator (d.f.D.) are 21.

Solution:

Since this test is right-tailed with $\alpha = 0.05$, use the 0.05 table. The d.f.N. is listed across the top, and the d.f.D. is listed in the left column. The critical value is found where the row and column intersect in the table. In this case, it is 2.18.



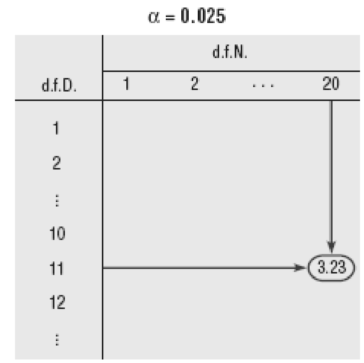
Some Examples (cont'd)

As noted previously, when the F test is used, the larger variance is always placed in the numerator of the formula. When one is conducting a two-tailed test, α is split; and even though there are two values, only the right tail is used. The reason is that the F test value is always greater than or equal to 1.

2. Find the critical value for a two-tailed F test with α = 0.05 when the sample size from which the variance for the numerator was obtained was 21 and the sample size from which the variance for the denominator was obtained was 12.

Some Examples (cont'd)

Since this is a two-tailed test with α = 0.05, the 0.05/2 = 0.025 table must be used. Here, d.f.N. = 21 - 1 = 20, and d.f.D. = 12 - 1 = 11; hence, the critical value is 3.23.



Using F test for Testing the Difference Between Two Variances

When the degrees of freedom values cannot be found in the table, the closest value on the smaller side should be used. For example, if d.f.N. = 14, this value is between the given table values of 12 and 15; therefore, 12 should be used, to be on the safe side. When one is testing the equality of two variances, these hypotheses are used:

Right-tailed	Left-tailed	Two-tailed
H_0: $\sigma_1^2 \leq \sigma_2^2$	H_0: $\sigma_1^2 \geq \sigma_2^2$	H_0: $\sigma_1^2 = \sigma_2^2$
H_1: $\sigma_1^2 > \sigma_2^2$	H_1: $\sigma_1^2 < \sigma_2^2$	H_1: $\sigma_1^2 \neq \sigma_2^2$

The Differences Between Two Variances

- The assumptions for testing the differences between two variances are:
  1. The populations from which the samples were obtained must be normally distributed. (Note: The test should not be used when the distributions depart from normality.)
  2. The samples must be independent of each other.

Some Examples

1. A medical researcher wishes to see whether the variance of the heart rates (in beats per minute) of smokers is different from the variance of heart rates of people who do not smoke. Two samples are selected, and the data are as shown. Using α = 0.05, is there enough evidence to support the claim?

Smokers	Nonsmokers
$n_1 = 26$	$n_2 = 18$
$s_1^2 = 36$	$s_2^2 = 10$

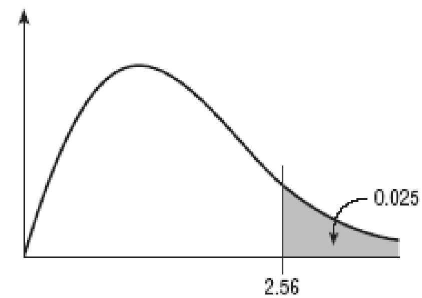
Solution:

STEP 1 State the hypotheses and identify the claim.

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{and} \quad H_0 : \sigma_1^2 \neq \sigma_2^2 \text{ (claim)}$$

Some Examples (cont'd)

STEP 2 Find the critical value. Use the 0.025 table in Table H since α = 0.05 andthis is a two-tailed test. Here, d.f.N. = 25, and d.f.D. = 17. The critical value is 2.56 (d.f.N. = 24 was used).



Some Examples (cont'd)

STEP 3 Compute the test value.

$$F = \frac{s_1^2}{s_2^2} = \frac{36}{10} = 3.6$$

STEP 4 Make the decision. Reject the null hypothesis, since 3.6 > 2.56.

STEP 5 Summarise the results. There is enough evidence to support the claim that the variance of the heart rates of smokers and nonsmokers is different.

Some Examples (cont'd)

2. An instructor hypothesises that the standard deviation of the final exam grades in her statistics class is larger for the male students than it is for the female students. The data from the final exam for the last semester are shown. Is there enough evidence to support her claim, using α = 0.01?

Males	Females
$n_1 = 16$	$n_2 = 18$
$s_1 = 4.2$	$s_2 = 2.3$

Solution:

STEP 1 State the hypotheses and identify the claim.

$$H_0: \sigma_1 \leq \sigma_2 \text{ and } H_1: \sigma_1 > \sigma_2 \text{ (claim)}$$

Some Examples (cont'd)

STEP 2 Find the critical value. Here, d.f.N. = 15, and d.f.D. = 17. From the 0.01 table, the critical value is 3.31.

STEP 3 Compute the test value.

$$F = \frac{s_1^2}{s_2^2} = \frac{4.2^2}{2.3^2} = 3.33$$

STEP 4 Make the decision. Reject the null hypothesis, since 3.33 > 3.31.

STEP 5 Summarise the results. There is enough evidence to support the claim that the standard deviation of the final exam grades for the male students is larger than that for the female students.

t Test

- A *t* test is used to test the difference between means when the two samples are *independent*, when the sample sizes are small, and when the samples are taken from two normally or approximately normally distributed populations.
- There are two different options for the use of *t* tests. One option is used when the variances of the populations are not equal, and the other option is used when the variances are equal.

Difference Between Two Means–Small Samples

When variances are assumed to be unequal:

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

where the degrees of freedom are equal to the smaller of $n_1 - 1$ or $n_2 - 1$.

Difference Between Two Means–Small Samples

- When variances are assumed to be equal:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where the degrees of freedom are equal to $n_1 + n_2 - 2$.

Pooled Estimate of Variance

- A *pooled estimate of the variance* is a weighted average of the variance using the two sample variances and the degrees of freedom of each variance as the weights.
- The pooled estimate of variance is used to calculate the standard error in the t test when the variances are equal.

Note on the t test

- To use the t test, first used the F test to determine whether the variances are equal. Then use the appropriate t test formula. This procedure involves two five-step processes.

Some Examples

1. The average size of a farm in Indiana County, Pennsylvania, is 191 acres. The average size of a farm in Greene County, Pennsylvania, is 199 acres. Assume the data were obtained from two samples with standard deviations of 38 acres and 12 acres, respectively, and sample sizes of 8 and 10, respectively. Can it be concluded at $\alpha = 0.05$ that the average size of the farms in the two counties is different? Assume the populations are normally distributed. (*Source: Pittsburgh Tribune-Review.*)

Solution:

Here we will use the F test to determine whether the variances are equal. The null hypothesis is that the variances are equal.

STEP 1 State the hypotheses and identify the claim.

H_0: $\sigma_1 = \sigma_2$ (claim) and H_1: $\sigma_1 \neq \sigma_2$

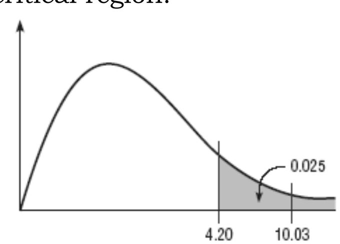
Some Examples (cont'd)

STEP 2 Find the critical value. The critical value for the F test found in the Table of the F Distribution for $\alpha = 0.05$ is 4.20, since there are 7 and 9 d.f. (Note: Use the 0.025 table.)

STEP 3 Compute the test value.

$$F = \frac{s_1^2}{s_2^2} = \frac{38^2}{12^2} = 10.03$$

STEP 4 Make the decision. Reject the null hypothesis since 10.03 falls in the critical region.



Some Examples (cont'd)

STEP 5 Summarize the results. It can be concluded that the variances are not equal.

Since the variances are not equal, the first formula will be used to test the equality of the means.

STEP 1 State the hypotheses and identify the claim for the means.

H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$ (claim)

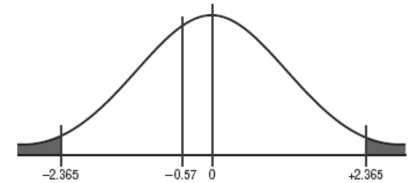
STEP 2 Find the critical values. Since the test is two-tailed, since α = 0.05, and since the variances are unequal, the d.f. are the smaller of n_1 - 1 or n_2 - 1. In this case, the d.f are 7. Hence, from the Table of the t Distribution, the critical values are ±2.365.

STEP 3 Compute the test value. Since the variances are unequal, use the first formula.

Some Examples (cont'd)

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(191 - 199) - 0}{\sqrt{\frac{38^2}{8} + \frac{12^2}{10}}} = -0.57$$

STEP 4 Make the decision. Do not reject the null hypothesis, since -0.57 < 2.365.



STEP 5 Summarise the results. There is not enough evidence to support the claim that the average size of the farms is different.

Some Examples (cont'd)

2. A researcher wishes to determine whether the salaries of professional nurses employed by private hospitals are higher than those of nurses employed by government-owned hospitals. She selects a sample of nurses from each type of hospital and calculates the means and standard deviations of their salaries. At α = 0.01, can she conclude that the private hospitals pay more than the government hospitals? Assume that the populations are approximately normally distributed. Use the P-value method.

Private	Government
$\bar{X}_1 = \$26,800$	$\bar{X}_2 = \$25,400$
$s_1 = \$600$	$s_2 = \$450$
$n_1 = 10$	$n_2 = 8$

Some Examples (cont'd)

Solution:

The F test will be used to determine whether the variances are equal. The null hypothesis is that the variances are equal.

STEP 1 State the hypotheses and identify the claim.

$H_0: \sigma_1^2 = \sigma_2^2$ and $H_1: \sigma_1^2 \neq \sigma_2^2$ (claim)

STEP 2 Compute the test value

$$F = \frac{s_1^2}{s_2^2} = \frac{600^2}{450^2} = 1.78$$

STEP 3 Find the P- value in the Table of F Distribution, using d.f.N. =9 and d.f.D. = 7. Since 1.78 <2.72, P-value > 0.20.

Some Examples (cont'd)

STEP 4 Make the decision. Do not reject the null hypothesis since P-value > 0.01 (the α value).

STEP 5 Summarise the results. There is not enough evidence to reject the claim that the variances are equal; therefore, the second formula is used to test the difference between the two means, as shown next.

STEP 1 State the hypotheses and identify the claim.

H_0: $\mu_1 \le \mu_2$ and H_1: $\mu_1 > \mu_2$ (claim)

STEP 2 Compute the test value. Use the second formula, since the variance are assumed to be equal

Some Examples (cont'd)

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \frac{(26{,}800 - 25{,}400) - 0}{\sqrt{\frac{(10 - 1)(600)^2 + (8 - 1)(450)^2}{10 + 8 - 2}}\sqrt{\frac{1}{10} + \frac{1}{8}}} = 5.47$$

Find the P-value, using the Table of the t Distribution. The P-value for t = 5.47 with d.f. = 16 (that is, 10 + 8 - 2) is P-value 0.005.

STEP 4 Make the decision. Since P-value < 0.01 (the α value), the decision is to reject the null hypothesis.

STEP 5 Summarise the results. There is enough evidence to support the claim that the salaries paid to nurses employed by private hospitals are higher than those paid to nurses employed by government-owned hospitals.

Confidence Intervals for the Difference of Two Means

Confidence intervals can also be found for the difference between two means with the following formulae.

- Variances unequal

$$(\overline{X}_1 - \overline{X}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

d.f. = smaller value of n_1-1 or n_2-1

- Variances equal

$$(\overline{X}_1 - \overline{X}_2) - t_{\alpha/2}\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2$$
$$< (\overline{X}_1 - \overline{X}_2) + t_{\alpha/2}\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

d.f. = $n_1 + n_2$-2

An Example

Find the 95% confidence interval for the data in Example 1 in the last section.

Solution:

Substitute in the formula

$$(\overline{X}_1 - \overline{X}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(191 - 199) - 2.365\sqrt{\frac{38^2}{8} + \frac{12^2}{10}} < \mu_1 - \mu_2 < (191 - 199) + 2.365\sqrt{\frac{38^2}{8} + \frac{12^2}{10}}$$

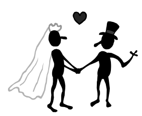
$$-41.02 < \mu_1 - \mu_2 < 25.02$$

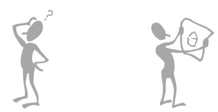
Since 0 is contained in the interval, the decision is not to reject the null hypothesis.

Independent and Dependent Samples

What we have introduced so far is to compared two sample means using the *t* test when the samples were independent. Next, a different version of the *t* test will be explained. This version is used when the samples are dependent.

- *Dependent Samples* are samples that are paired or matched in some way.
- *Independent Samples* are samples that are not related.





Examples of Dependent Samples

- Samples in which the same subjects are used in a pre-post situation are dependent.
- Another type of dependent samples are samples matched on the basis of variables extraneous to the study.

Two Notes of Caution

1. When subjects are matched according to one variable, the matching process does not eliminate the influence of other variables.
2. When the same subjects are used for a pre-post study, sometimes the knowledge that they are participating in a study can influence the results.

Special *t* test for Dependent Means

- Hypotheses:

Two-tailed	Left-tailed	Right-tailed
H_0: $\mu_D = 0$	H_0: $\mu_D \geq 0$	H_0: $\mu_D \leq 0$
H_1: $\mu_1 \neq 0$	H_1: $\mu_1 < 0$	H_1: $\mu_1 > 0$

- μ_D is the expected mean of the differences of the matched pairs.

General Procedure—Finding the Test Value

Step 1 Find the differences of the values of the pairs of data, D.

Step 2 Find the mean of the differences $\overline{D}$.

Step 3 Find the standard deviation of the differences, s_D .

Step 4 Find the estimated standard error of the differences, $s_{\overline{D}}$.

Step 5 Find the test value, t.

t Test Formulas

- The formula for the t test for dependent samples:

$$t = \frac{\overline{D} - \mu_D}{s_D / \sqrt{n}}$$

with d.f. = $n - 1$ and

where $$\overline{D} = \frac{\Sigma D}{n}$$

and $$s_D = \sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{n}}{n-1}}$$

Confidence Interval—Mean Difference

- The formula for calculating the confidence interval for the mean difference:

$$\overline{D} - t_{\alpha/2}\frac{s_D}{\sqrt{n}} < \mu_D < \overline{D} + t_{\alpha/2}\frac{s_D}{\sqrt{n}}$$

d.f. = $n - 1$

An Example

A physical education director claims by taking a special vitamin, a weight lifter can increase his strength. Eight athletes are selected and given a test of strength, using the standard bench press. After two weeks of regular training, supplemented with the vitamin, they are tested again. Test the effectiveness of the vitamin regimen at $\alpha = 0.05$. Each value in these data represents the maximum number of pounds the athlete can bench-press. Assume that the variable is approximately normally distributed.

Athlete	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Before (X_1)	210	230	182	205	262	253	219	216
After (X_2)	219	236	179	204	270	250	222	216

An Example (cont'd)

Solution

STEP 1 State the hypotheses and identify the claim. In order for the vitamin to be effective, the before weights must be significantly less than the after weights; hence, the mean of the differences must be less than zero.

H_0: $\mu_D \geq 0$ and H_1: $\mu_D < 0$ (claim)

STEP 2 Find the critical value. The degrees of freedom are n ┘ 1. In this case, d.f. = 7. The critical value for a left-tailed test with α = 0.05 is -1.895.

STEP 3 Compute the test value.

a. Make a table.

An Example (cont'd)

Before (X_1)	After (X_2)	A $D = X_1 - X_2$	B $D^2 = (X_1 - X_2)^2$
210	219		
230	236		
182	179		
205	204		
262	270		
253	250		
219	222		
216	216		

b. Find the differences and place the results in column A.

c. Find the mean of the differences.

$$\overline{D} = \frac{\sum D}{n} = \frac{-19}{8} = -2.375$$

An Example (cont'd)

d. Square the differences and place the result in column B.

The complete table is shown below.

Before (X_1)	After (X_2)	A $D = X_1 - X_2$	B $D^2 = (X_1 - X_2)^2$
210	219	−9	81
230	236	−6	36
182	179	+3	9
205	204	+1	1
262	270	−8	64
253	250	+3	9
219	222	−3	9
216	216	0	0
		$\Sigma D = -19$	$\Sigma D^2 = 209$

e. Find the standard deviation of the differences.

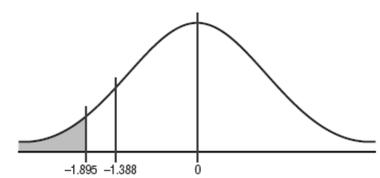
$$s_D = \sqrt{\frac{\sum D^2 - \frac{\left(\sum D\right)^2}{n}}{n-1}} = \sqrt{\frac{209 - \frac{(-19)^2}{8}}{8-1}} = 4.84$$

An Example (cont'd)

f. Find the test value

$$t = \frac{\overline{D} - \mu_D}{s_D / \sqrt{n}} = \frac{-2.375 - 0}{4.84 / \sqrt{8}} = -1.388$$

STEP 4 Make the decision. The decision is not to reject the null hypothesis at α = 0.05, since -1.388 > -1.895, as shown below:



STEP 5 Summarise the results. There is not enough evidence to support the claim that the vitamin increases the strength of weight lifters.

Testing the Difference Between Proportions

The *z* test with some modifications can be used to test the equality of two proportions. For example, a researcher might ask: Is the proportion of men who exercise regularly less than the proportion of women who exercise regularly? Is there a difference in the percentage of students who own a personal computer and the percentage of non-students who own one? Is there a difference in the proportion of college graduates who pay cash for purchases and the proportion of non-college graduates who pay cash?

Getting the Standard Error of Difference

- For population proportions, p_1 and p_2 the hypotheses can be stated as follows, if no difference between the proportions is hypothesised.

H_0: $p_1 = p_2$ or H_0: $p_1 - p_2 = 0$

H_1: $p_1 \neq p_2$ H_1: $p_1 - p_2 \neq 0$

- $\hat{p}_1 = X_1 / n_1$ is used to estimate p_1.
- $\hat{p}_2 = X_2 / n_2$ is used to estimate p_2.

The Standard Error of Difference

- The standard error of difference is

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

- Where $\sigma_{p_1}^2$ and $\sigma_{p_2}^2$ are the variances of the proportions, $q_1 = 1 - p_1$, $q_2 = 1 - p_2$, and n_1 and n_2 are the respective sample sizes.

A Weighted Estimate of *p*

- Since p_1 and p_2 are unknown, a weighted estimate of p can be computed by using the formula below.

$$\overline{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{X_1 + X_2}{n_1 + n_2}$$

This weighted estimate is based on the hypothesis that $p_1 = p_2$. Hence, $\overline{p}$ is better than both $\hat{p}_1$ and $\hat{p}_2$.

Standard Error of Difference

- The standard error of difference in terms of the weighted estimate is:

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

where

$$\bar{p} = \frac{X_1+X_2}{n_1+n_2} \qquad \hat{p}_1 = X_1/n_1$$

$$\bar{q} = 1-\bar{p} \qquad \hat{p}_2 = X_2/n_2$$

z Test

- The formula for the *z* test for comparing two proportions:

$$z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

- where

$$\bar{p} = \frac{X_1+X_2}{n_1+n_2} \qquad \hat{p}_1 = X_1/n_1$$

$$\bar{q} = 1-\bar{p} \qquad \hat{p}_2 = X_2/n_2$$

The Difference Between Two Proportions

- The confidence interval for the difference between two proportions can be calculated using the following formula:

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

$$< p_1-p_2 < (\hat{p}_1-\hat{p}_2)+z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}$$

An Example

In the nursing home study mentioned in the chapter-opening "Statistics Today," the researchers found that 12 out of 34 small nursing homes had a resident vaccination rate of less than 80%, while 17 out of 24 large nursing homes had a vaccination rate of less than 80%. At α = 0.05, test the claim that there is no difference in the proportions of the small and large nursing homes with a resident vaccination rate of less than 80%.

Source: Nancy Arden, Arnold S. Monto, and Suzanne E. Ohmit, "Vaccine Use and the Risk of Outbreaks in a Sample of Nursing Homes during an Influenza Epidemic," *American Journal of Public Health.*

Solution

Let $\hat{p}_1$ be the proportion of the small nursing homes with a vaccination rate of less than 80% and $\hat{p}_2$ be the proportion of the

An Example (cont'd)

large nursing homes with a vaccination rate of less than 80%. Then

$$\hat{p}_1 = \frac{X_1}{n_1} = \frac{12}{34} = 0.35 \qquad \hat{p}_2 = \frac{X_2}{n_2} = \frac{17}{24} = 0.71$$

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{12+17}{34+24} = 0.5$$

$$\bar{q} = 1 - \bar{p} = 0.5$$

Now, following the steps in hypothesis testing.

STEP 1 State the hypotheses and identify the claim

H_0: $p_1 = p_2$ (claim) and H_1: $p_1 \neq p_2$

An Example (cont'd)

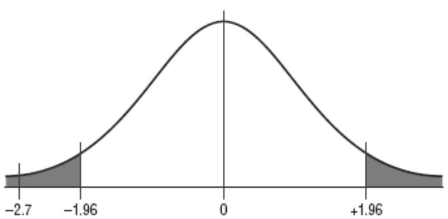
STEP 2 Find the critical values. Since $\alpha = 0.05$, the critical values are ± 1.96.

STEP 3 Compute the test value

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.35 - 0.71) - 0}{\sqrt{(0.5)(0.5)\left(\frac{1}{34} + \frac{1}{24}\right)}} = -2.7$$

STEP 4 Make the decision. Reject the hypothesis, since $-2.7 < -1.96$. See the graph in the next page.

An Example (cont'd)



STEP 5 Summarise the results. There is enough evidence to reject the claim that there is no difference in the proportions of small and large nursing homes with a resident vaccination rate of less than 80%.

Summary

- Means and Proportions are population parameters that are often compared.
- This comparison can be made with the *z* test if the samples are independent and the variances are known, or if the variances are unknown but both sample sizes are greater than or equal to 30.

Summary (cont'd.)

- If the variances are not know or one or both sample sizes are less than 30, then the *t* test must be used.
- For independent samples the *F* test must be used to determine whether or not the variances are equal.
- For dependent samples the dependent samples *t* test is used.
- A *z* test is used to compare two proportions.

Conclusions

- Special *z* and *t* tests allow researchers to compare population parameters, such as means or proportions.