統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 10: Inference About Means and Proportions with Two Populations

上課時間地點: 二 D56, 資訊 140306
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: http://www.hmwu.idv.tw

系級: ____________________ 學號: ____________________ 姓名: ____________________

Overview

1. Discuss the statistical inference (____________________ and ____________________) for two population means (three situations: population standard deviations known, unknown; match samples) and the two population proportions.

2. *Examples*:

   (a) Develop an interval estimate of the difference between the mean starting salary for a population of men and the mean starting salary for a population of women.

   (b) Conduct a hypothesis test to determine whether any difference is present between the proportion of defective parts in a population of parts produced by supplier A and the proportion of defective parts in a population of parts produced by supplier B.

10.1 Inferences About the Difference Between Two Population Means: σ_1 and σ_2 Known

1. ________ denote the mean of population 1 (2), we will focus on inferences about the difference between the means: ________.

2. A simple random sample of ________ units from population 1 (2). The two samples, taken separately and independently, are referred to as ________ simple random samples.

3. Assume the two population standard deviations, ________, can be assumed ________ to collecting the samples.

4. Question: how to compute a ________ and develop an ________ of the difference between the two population means when σ_1 and σ_2 are known.

Interval Estimation of $\mu_1 - \mu_2$

1. Example **Greystone Department Stores**, Inc., operates two stores in Buffalo, New York: One is in the inner city and the other is in a suburban shopping center. The regional manager noticed that products that sell well in one store do not always sell well in the other. The manager believes this situation may be attributable to differences in customer demographics at the two locations. Customers may differ in age, education, income, and so on. (對觀察事物提出問題)

2. Suppose the manager asks us to investigate the difference between the ________ of the customers who shop at the two stores. (針對問題收集資料)

3. Let us define population 1 as all customers who shop at the ________ and population 2 as all customers who shop at the ________.

   (a) ________: mean of population 1 (i.e., the mean age of all customers who shop at the inner-city store)

   (b) ________: 5 mean of population 2 (i.e., the mean age of all customers who shop at the suburban store)

4. The difference between the two population means is ________. To estimate $\mu_1 - \mu_2$, we will select a simple random sample of ______ customers from population 1 and a simple random sample of ______ customers from population 2.

5. We then compute the two sample means.

   (a) ______: sample mean age for the simple random sample of n_1 inner-city customers

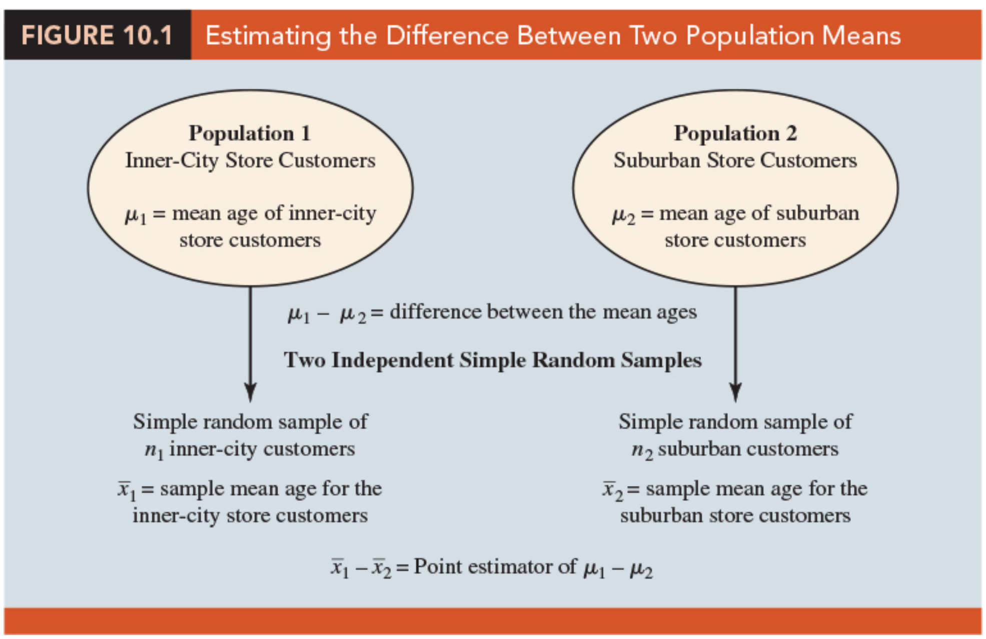
   (b) ______: sample mean age for the simple random sample of n_2 suburban customers

6. The point estimator of the difference between the two ________________ is the difference between the two ______________.

7. **Point Estimator of the Difference Between Two Population Means**

$$\rule{6em}{0.4pt} \qquad (10.1)$$

8. (Figure 10.1) the process used to estimate the difference between two population means based on two independent simple random samples.



9. The point estimator $\bar{x}_1 - \bar{x}_2$ has a standard error that describes the __________ in the sampling distribution of the estimator.

10. **Standard Error of $\bar{x}_1 - \bar{x}_2$** With two independent simple random samples, the standard error of $\bar{x}_1 - \bar{x}_2$ is as follows:

    ______________________ (10.2)

    (証明如下:) (Hint: $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$).

11. If both populations have a __________ distribution, or if the sample sizes are large enough that the ______________________ enables us to conclude that the sampling distributions of $\bar{x}_1$ and $\bar{x}_2$ can be approximated by a normal distribution, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will have a __________ distribution with mean given by __________. (Denoted by ______________________________)

12. In general, an interval estimate is given by a point estimate ± a margin of error. In the case of estimation of the difference between two population means, an interval estimate will take the following form:

    ____

13. With the sampling distribution of $\bar{x}_1 - \bar{x}_2$ having a normal distribution, we can write the margin of error as follows:

    Margin of error = _____________ = ________________ (10.3)

14. **Interval Estimate of the Difference Between Two Population Means: σ_1 and σ_2 Known**

    ___________________________ (10.4)

    where $1-\alpha$ is the confidence coefficient.

(公式說明如下:)

✎ Question .... (p485)

Example Greystone example. Based on data from previous customer demographic studies, the two population standard deviations are known with $\sigma_1 = 9$ years and $\sigma_2 = 10$ years. The data collected from the two independent simple random samples of Greystone customers provided the following results.

	Inner City Store	Suburban Store
Sample Size	$n_1 = 36$	$n_2 = 49$
Sample Mean	$\bar{x}_1 = 40$ years	$\bar{x}_2 = 35$ years

Find the margin of error and the 95% confidence interval estimate of the difference between the two population means.

sol:

Hypothesis Tests About $\mu_1 - \mu_2$

1. Let us consider hypothesis tests about the difference between two population means. Using _______ to denote the hypothesized difference between μ_1 and μ_2, the three forms for a hypothesis test are as follows:

Left-tailed test	Right-tailed test	Two-tailed test
H_0 : __________	H_0 : __________	H_0 : __________
H_a : __________	H_a : __________	H_a : __________

2. In many applications, __________. Using the two-tailed test as an example, when $D_0 = 0$ the null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$.

3. In this case, the null hypothesis is that μ_1 and μ_2 are equal. Rejection of H_0 leads to the conclusion that $H_a : \mu_1 - \mu_2 \neq 0$ is true; that is, μ_1 and μ_2 are not equal.

4. The general steps for conducting hypothesis tests: choose a ____________________, compute the value of the ______________, and find the __________ to ____________ whether the null hypothesis should be rejected.

5. With two independent simple random samples, we showed that the point estimator $\bar{x}_1 - \bar{x}_2$ has a standard error $\sigma_{\bar{x}_1 - \bar{x}_2}$ given by expression (10.2) and, when the sample sizes are large enough, the distribution of $\bar{x}_1 - \bar{x}_2$ can be described by a __________ distribution.

6. **Test Statistic for Hypothesis Tests About $\mu_1 - \mu_2$: σ_1 and σ_2 Known**

   (10.5)

7. We demonstrated a two-tailed hypothesis test about the difference between two population means. Lower tail and upper tail tests can also be considered. These tests use the ___________________ as given in equation (10.5). The procedure for computing the p-value and the rejection rules for these one-tailed tests are the same as those for hypothesis tests involving a single population mean and single population proportion.

✎ Question .... (p486)

As part of a study to evaluate differences in education quality between two training centers, a standardized examination is given to individuals who are trained at the centers. The difference between the mean examination scores is used to assess quality differences between the centers. The population means for the two centers are as follows. μ_1 is the mean examination score for the population of individuals trained at center A, μ_2 is the mean examination score for the population of individuals trained at center B. We begin with the tentative assumption that no difference exists between the training quality provided at the two centers. The standardized examination given previously in a variety of settings always resulted in an examination score standard deviation near 10 points. Thus, we will use this information to assume that the population standard deviations are known with $\sigma_1 = 10$ and $\sigma_2 = 10$. An $\alpha = 0.05$ level of significance is specified for the study. Independent simple random samples of $n_1 = 30$ individuals from training center A and $n_2 = 40$ individuals from training center B are taken. The respective sample means are $\bar{x}_1 = 82$ and $\bar{x}_2 = 78$. Do these data suggest a significant difference between the population means at the two training centers? State the null and alternative hypotheses for this two-tailed test, compute the test statistic, and state the decision rules based on the p-value approach and the critical value approach and make the decision.

sol:

Practical Advice

1. In most applications of the interval estimation and hypothesis testing procedures presented in this section, random samples with ____________ and ____________ are adequate.

2. In cases where either or both sample sizes are less than 30, the ________________ of the populations become important considerations.

3. In general, with smaller sample sizes, it is more important for the analyst to be satisfied that it is reasonable to assume that the distributions of the two populations

are at least ______________________.

😍 **EXERCISES 10.1**: 1, 2, 4, 6

10.2 10.2 Inferences About The Difference Between Two Population Means: σ_1 and σ_2 Unknown

1. Extend the discussion of inferences about the difference between two population means to the case when the two population standard deviations, ______ and ______, are ____________.

2. In this case, we will use the sample standard deviations, ______ and ______, to estimate the unknown population standard deviations.

3. When we use the sample standard deviations, the interval estimation and hypothesis testing procedures will be based on the _______________ rather than the standard normal distribution.

Interval Estimation of $\mu_1 - \mu_2$

1. Let us develop the margin of error and an interval estimate of the difference between these two population means. (Recall) The interval estimate for the case when the population standard deviations, σ_1 and σ_2, are known.

   ____

2. With σ_1 and σ_2 unknown, we will use the sample standard deviations s_1 and s_2 to estimate ____________ and replace $z_{\alpha/2}$ with _______.

3. **Interval Estimate of the Difference Between Two Population Means: σ_1 and σ_2 Unknown**

(10.6)

where $1-\alpha$ is the confidence coefficient.

4. In this expression, the use of the t distribution is an ________________, but it provides excellent results and is relatively easy to use. The only difficulty that we encounter in using expression (10.6) is determining the appropriate ____________________ for $t_{\alpha/2}$.

5. Statistical software packages compute the appropriate degrees of freedom automatically. The formula used is as follows:

Degrees of Freedom: t Distribution With Two Independent Random Samples

$$df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \qquad (10.7)$$

✎ Question .... (p490)

(**Clearwater National Bank example**) Clearwater National Bank is conducting a study designed to identify differences between checking account practices by customers at two of its branch banks. A simple random sample of 28 checking accounts is selected from the Cherry Grove Branch and an in dependent simple random sample of 22 checking accounts is selected from the Beechmont Branch. The current checking account balance is recorded for each of the checking accounts. A summary of the account balances follows:

	Cherry Grove	Beechmont
Sample Size	$n_1 = 28$	$n_2 = 22$
Sample Mean	$\bar{x}_1 = \$1025$	$\bar{x}_2 = \$910$
Sample Standard Deviation	$s_1 = \$150$	$s_2 = \$125$

Clearwater National Bank would like to estimate the difference between the mean checking account balance maintained by the population of Cherry Grove customers and the population of Beechmont customers. Compute a 95% confidence interval estimate of the difference between the population mean checking account balances at the two branch banks.

sol:

Hypothesis Tests About $\mu_1 - \mu_2$

1. (Recall) Letting D_0 denote the hypothesized difference between μ_1 and μ_2, the test statistic used for the case where σ_1 and σ_2 are known is as follows.

   The test statistic, z, follows ________________ the standard normal distribution.

2. When σ_1 and σ_2 are unknown, we use s_1 as an estimator of σ_1 and s_2 as an estimator of σ_2. Substituting these sample standard deviations for σ_1 and σ_2 provides the following test statistic when σ_1 and σ_2 are unknown.

3. **Test Statistic for Hypothesis Tests About $\mu_1 - \mu_2$: σ_1 and σ_2 Unknown**

   (10.8)

   The degrees of freedom ________________ for t are given by equation (10.7).

✐ Question .... (p491)

Consider a new computer software package developed to help systems analysts reduce the time required to design, develop, and implement an information system. To evaluate the benefits of the new software package, a random sample of 24 systems analysts is selected. Each analyst is given specifications for a hypothetical information system. Then 12 of the analysts are instructed to produce the information system by using current technology. The other 12 analysts are trained in the use of the new software package and then instructed to use it to produce the information system. This study involves two populations: a population of systems analysts using the current technology and a population of systems analysts using the new software package. In terms of the time required to complete the information system design project, the population means are as follows. μ_1 is the mean project completion time for systems analysts using the current technology and μ_2 is the mean project completion time for systems analysts using the new software package The researcher in charge of the new software evaluation project hopes to show that the new software package will provide a shorter mean project completion time. Thus, the researcher is looking for evidence to conclude that μ_2 is less than μ_1; in this case, the difference between the two population means, $\mu_1-\mu_2$, will be greater than zero. Suppose that the 24 analysts complete the study with the results shown in Table 10.1.

TABLE 10.1 Completion Time Data and Summary Statistics for the Software Testing Study

	Current Technology	New Software
	300	274
	280	220
	344	308
	385	336
	372	198
	360	300
	288	315
	321	258
	376	318
	290	310
	301	332
	283	263
Summary Statistics		
Sample size	$n_1 = 12$	$n_2 = 12$
Sample mean	$\bar{x}_1 = 325$ hours	$\bar{x}_2 = 286$ hours
Sample standard deviation	$s_1 = 40$	$s_2 = 44$

Let the level of significance be $\alpha = 0.05$. State the null and the alternative hypothesis, the test statistic, p-value, the rejection rule, make a decision and conclusion.

sol:

(Software Output)

TABLE 10.2 Output for the Hypothesis Test on the Difference Between the Current and New Software Technology

	Current	New
Mean	325	286
Variance	1600	1936
Observations	12	12
Hypothesized Mean Difference	0	
Degrees of Freedom	21	
Test Statistic	2.272	
One-Tail *p*-value	0.017	
One-Tail Critical Value	1.717	

Practical Advice

1. The interval estimation and hypothesis testing procedures presented in this section are __________ and can be used with __________________ sample sizes.

2. In most applications, equal or nearly equal sample sizes such that the total sample size ________________ can be expected to provide very good results even if the populations are not normal.

3. Larger sample sizes are recommended if the distributions of the populations are _________________ or contain ___________.

4. Smaller sample sizes should only be used if the analyst is satisfied that the distributions of the populations are at least ________________________.

Notes + Comments

1. How to make inferences about the difference between two population means when σ_1 and σ_2 are equal and unknown (_______________)?

2. Based on above assumption, the two sample standard deviations are combined to provide the following pooled sample variance:

____

3. The t test statistic becomes

   and has ____________ degrees of freedom. At this point, the computation of the p-value and the interpretation of the sample results are identical to the procedures discussed earlier in this section.

4. A difficulty with this procedure is that the assumption that the two population standard deviations are equal is usually difficult to ________. Unequal population standard deviations are frequently encountered.

5. Using the pooled procedure may not provide satisfactory results, especially if the sample sizes n_1 and n_2 are ______________.

6. The t procedure that we presented in this section does not require the assumption of equal population standard deviations and can be applied whether the population standard deviations are equal or not. It is a more general procedure and is recommended for most applications.

☺ **EXERCISES 10.2**: 9, 10, 13, 14, 15

10.3 Inferences About The Difference Between Two Population Means: Matched Samples

1. Example **Matched**.

   (a) Suppose employees at a manufacturing company can use two different methods to perform a production task. To maximize production output, the company wants to identify the method with the smaller population mean completion time.

(b) Let ______ denote the population mean completion time for production method 1 and ______ denote the population mean completion time for production method 2.

(c) With no preliminary indication of the preferred production method, we begin by tentatively assuming that the two production methods have the same population mean completion time. Thus, the null hypothesis is ________________.

(d) If this hypothesis is rejected, we can conclude that the population mean completion times differ. In this case, the method providing the smaller mean completion time would be recommended.

(e) The null and alternative hypotheses are written as follows.

____

2. In choosing the sampling procedure that will be used to collect production time data and test the hypotheses, we consider two alternative designs. One is based on ____________________ and the other is based on ________________.

   (a) *Independent sample design*: A simple random sample of workers is selected and each worker in the sample uses method 1. A second independent simple random sample of workers is selected and each worker in this sample uses method 2. The test of the difference between population means is based on the procedures in Section 10.2.

   (b) *Matched sample design*: One simple random sample of workers is selected. Each worker first uses one method and then uses the other method. The order of the two methods is assigned randomly to the workers, with some workers performing method 1 first and others performing method 2 first. Each worker provides ____________________, one value for method 1 and another value for method 2.

3. In the matched sample design the two production methods are tested under similar conditions (i.e., with the same workers); hence this design often leads to a ______________________ than the independent sample design. The primary reason is that in a matched sample design, ________________________ is eliminated because the same workers are used for both production methods.

4. Assuming the analysis of a matched sample design is the method used to test the difference between population means for the two production methods. The key to the analysis of the matched sample design is to realize that we consider only ______________________________.

5. Therefore, we have six data values ($0.6, -0.2, 0.5, 0.3, 0.0$, and 0.6) that will be used to analyze the difference between population means of the two production methods.

6. Let _______ is the mean of the difference in values for the population of workers. With this notation, the null and alternative hypotheses are rewritten as follows.

   ____

7. Assume the population of ______________ has a ___________ distribution. This assumption is necessary so that we may use the ________________ for hypothesis testing and interval estimation procedures. Based on this assumption, the following test statistic has a t distribution with ________ degrees of freedom.

8. **Test Statistic for Hypothesis Tests Involving Matched Samples**

   ____

   where

   ______________, and ______________________ (10.9)

Question .... (p498)

(Table 10.3) (Matched Example). A random sample of six workers is used. The data on completion times for the six workers are given in Table 10.3. Note that each worker provides a pair of data values, one for each production method. Also note that the last column contains the difference in completion times di for each worker in the sample. Assume that the population of differences has a normal distribution. Test the hypotheses $H_0 : \mu_d = 0$ and $H_a : \mu_d \neq 0$, using $\alpha = 0.05$. Compute the test statistic, the p-value and draw a conclusion. Compute the 95% confidence interval for the difference between the population means of the two production methods. If H_0 is rejected, we can conclude that the population mean completion times differ.

TABLE 10.3 Task Completion Times for a Matched Sample Design

Worker	Completion Time for Method 1 (minutes)	Completion Time for Method 2 (minutes)	Difference in Completion Times (d_i)
1	6.0	5.4	.6
2	5.0	5.2	−.2
3	7.0	6.5	.5
4	6.2	5.9	.3
5	6.0	6.0	.0
6	6.4	5.8	.6

sol:

Area in Upper Tail	0.20	0.10	0.05	0.025	0.01	0.005
t-Value (5 df)	0.920	1.476	2.015	2.571	3.365	4.032

☺ **EXERCISES 10.3**: 19, 23, 24

10.4 Inferences About The Difference Between Two Population Proportions

1. Letting ______ denote the proportion for population 1 and ______ denote the proportion for population 2.

2. Consider inferences about the difference between the two population proportions: _________.

3. To make an inference about this difference, we will select two independent random samples consisting of n_1 units from population 1 and n_2 units from population 2.

Interval Estimation of $p_1 - p_2$

1. Example **Tax Preparation Firm**
   A tax preparation firm is interested in comparing the quality of work at two of its regional offices. By randomly selecting samples of tax returns prepared at each office and verifying the sample returns' accuracy, the firm will be able to estimate the proportion of erroneous returns prepared at each office. Of particular interest is the difference between these proportions.

   (a) p_1: proportion of erroneous returns for population 1 (office 1)

   (b) p_2: proportion of erroneous returns for population 2 (office 2)

   (c) ______: sample proportion for a simple random sample from population 1

   (d) ______: sample proportion for a simple random sample from population 2

2. **Point Estimator of the Difference Between Two Population Proportions**

   __________________ (10.10)

3. Thus, the point estimator of the difference between two ____________ proportions is the difference between the __________ proportions of two independent simple random samples.

4. As with other point estimators, the point estimator $\bar{p}_1 - \bar{p}_2$ has a sampling distribution that reflects the possible values of $\bar{p}_1 - \bar{p}_2$ if we repeatedly took two independent

random samples. The mean of this sampling distribution is __________ and the standard error of $\bar{p}_1 - \bar{p}_2$ is:

Standard Error of $\bar{p}_1 - \bar{p}_2$

____________________________________ (10.11)

5. If the sample sizes are large enough that ________, ____________, ________, and ____________ are all greater than or equal to ______, the sampling distribution of $\bar{p}_1 - \bar{p}_2$ can be approximated by a ___________ distribution.
6. With the sampling distribution of $\bar{p}_1 - \bar{p}_2$ approximated by a normal distribution, we would like to use _____________ as the margin of error.
7. However, $\sigma_{\bar{p}_1 - \bar{p}_2}$ given by equation (10.11) cannot be used directly because the two population proportions, p_1 and p_2, are unknown. Using the sample proportion $\bar{p}_1$ to estimate $\bar{p}_1$ and the sample proportion $\bar{p}_2$ to estimate p_2, the margin of error is:

   Margin of error = _______________________________ (10.12)

8. **Interval Estimate of the Difference Between Two Population Proportions**

   ____ (10.13)

   where $1 - \alpha$ is the confidence coefficient.

✎ Question .... (p504)

(Tax Preparation Example) We find that independent simple random samples from the two offices provide the following information.

Office	1	2
n_i	250	300
Number of returns with errors	35	27

Find a margin of error and interval estimate of the difference between the two population proportions. and 90% confidence interval.

sol:

Hypothesis Tests About $p_1 - p_2$

1. Let us now consider hypothesis tests about no difference between the proportions of two populations. In this case, the three forms for a hypothesis test are as follows:

$$\begin{array}{lll} H_0 : p_1 - p_2 \geq 0, & H_0 : p_1 - p_2 \leq 0, & H_0 : p_1 - p_2 = 0 \\ H_a : p_1 - p_2 < 0 & H_a : p_1 - p_2 > 0 & H_a : p_1 - p_2 \neq 0 \end{array}$$

2. When we assume ________________________, we have $p_1 - p_2 = 0$, which is the same as saying that the population proportions are equal, $p_1 = p_2$.

3. Under the assumption H_0 is true as an equality, the population proportions are equal and ______________. In this case, $\sigma_{\bar{p}_1 - \bar{p}_2}$ becomes **Standard Error of** $\bar{p}_1 - \bar{p}_2$ **when** $p_1 = p_2 = p$

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \tag{10.14}$$

4. With p unknown, we pool, or combine, the point estimators from the two samples ($\bar{p}_1$ and $\bar{p}_2$) to obtain a single point estimator of p as follows:
Pooled Estimator of p **When** $p_1 = p_2 = p$

$$\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \tag{10.15}$$

This pooled estimator of p is a weighted average of $\bar{p}_1$ and $\bar{p}_2$.

5. Substituting $\bar{p}$ for p in equation (10.14), we obtain an estimate of the standard error of $\bar{p}_1 - \bar{p}_2$. This estimate of the standard error is used in the test statistic.

6. The general form of the test statistic for hypothesis tests about the difference between two population proportions is the point estimator divided by the estimate of $\sigma_{\bar{p}_1 - \bar{p}_2}$.

7. **Test Statistic for Hypothesis Tests About $p_1 - p_2$**

   (10.16)

   This test statistic applies to large sample situations where $n_1 p_1$, $n_1(1-p_1)$, $n_2 p_2$, and $n_2(1-p_2)$ are all greater than or equal to 5.

✎ Question .... (p506)

(**Tax Preparation Firm Example**) Assume that the firm wants to use a hypothesis test to determine whether the error proportions differ between the two offices. A two-tailed test is required. Use $\alpha = 0.10$ as the level of significance. State the null and alternative hypotheses. Compute the test statistic, and the p-value for this two-tailed test. State the decison rule and draw a conclusion.

sol:

😍 **EXERCISES 10.4**: 28, 29, 31, 34

😍 **SUPPLEMENTARY EXERCISES**: 38, 39, 44