Chapter 8

Hypothesis Testing: One Sample Case

Chapter Outline

8.1 Introduction
8.2 The Null Hypothesis
8.3 Two Types of Error
8.4 Levels of Significance and Critical Regions
8.5 Directional and Non-Directional Tests
8.6 Test of Significance for a Single Mean
8.7 Introduction to the $t$-distribution
8.8 Power of a Statistical Test
8.9 Measuring the Size of an Effect: COHEN’S $d$
8.10 Estimation and Confidence Intervals for One-Sample Case
8.11 Using Confidence Intervals in Making Decisions

Chapter Summary

Basic Terms and Concepts

Exercises
Chapter Objectives

1. Define hypothesis testing.
2. Define null hypothesis, alternative hypothesis, level of significance, test statistic, p value, and statistical significance.
3. Identify the four steps of hypothesis testing.
4. Define Type I error and Type II error.
5. Calculate the one-sample z test and interpret the results.
6. Calculate the one-sample t test and interpret the results.
7. Distinguish between non-directional and directional tests.
8. Calculate confidence intervals.
9. Calculate effect size.
10. Define power and identify factors that influence power.
11. Use Confidence intervals in making decisions.

8.1 Introduction

In chapters 6 and 7 we discussed probability and the distribution of sample means, focusing on the standard error of the mean (standard deviation of the distribution of sample means. We were able to find the probability of obtaining a specific sample mean from a specified population, considering the amount of error between the sample mean and the corresponding population. The question now arises as to the way in which inferences or conclusions about a population can be drawn from sample data. To illustrate, a social psychologist postulates that there are gender differences in self-esteem. He asked a sample of males and females to complete a self-esteem inventory, and calculates the mean self-esteem for each gender and found a mean difference in self-esteem scores between males and females. He knows that the observed difference in self-esteem is subject to sampling error. The observed gender difference in self-esteem may not exist at all, and the difference may have occurred because of sampling error. Is he justified in concluding that there are gender difference in self-esteem, or can the observed difference simply due to chance or sampling error? A decision is required between these two alternatives. The procedures associated with problems of this kind are generally known as hypothesis testing and the statistical techniques used in hypothesis testing in making decisions are called tests of significance.

Hypothesis testing is a statistical procedure that allows researchers to use sample data to draw inferences about the population of interest. In this chapter, we introduce the general procedure for a hypothesis test. We will combine the concepts of z-scores, probability, and the distribution of sample means to create a new statistical procedure known as a hypothesis test. We will also discuss concepts like alpha level, critical region, and errors in hypothesis test.

There are two kinds of hypotheses: scientific hypothesis (research hypothesis or experimental hypothesis) and a statistical hypothesis. A scientific hypothesis (research
hypothesis) is a knowledgeable statement that is tentatively advanced to account for particular facts. It is a testable idea or testable question on some phenomenon of interest. We call it a testable question because we believe that the truth or falsity of it can be investigated by recording facts (data) on the phenomenon of interest. If an experiment is used to investigate the question then the research / scientific hypothesis is referred to as an experimental hypothesis. It is usually expressed using an if-then structure (often referred to as the logical form of the general implication) such as: if aspirin is taken then pain will be reduced. Specifically, a scientific (research) hypothesis is a statement of an expected or predicted relationship between two or more variables. It’s what the experimenter believes will happen in his research study. To illustrate, a theoretical speculation may be advanced that a particular drug, PAZ, will reduce bad cholesterol levels in individuals with high cholesterol. This is a scientific hypothesis. This hypothesis may in turn lead to the conduct of a rigorous, and the collection of empirical data, usually in a numerical form. Statistical hypotheses are usually rigorously stated and testable.

Tests of statistical hypotheses are tests of population parameters. In an experiment to test the efficacy of a drug in reducing cholesterol levels of individuals, the scientific investigator is not necessarily interested in the efficacy of the drug for the current sample of individuals in his study, but rather in the efficacy of the drug when administered to a larger group or population of individuals with high cholesterol levels. He wishes to generalize from a sample of participants to a specified population. He is engaged in inductive reasoning or inductive inference. Testing a statistical hypothesis involves a number of ideas. These ideas have a wide range of applications in scientific research. There are many types of significance tests used in testing statistical hypotheses. For example, a test may be applied to test the differences between a single sample mean and a population mean. Other tests are available to test the difference in means between two population means or two treatments, and these tests can be extended to cover situations involving two or more means.

### 8.2 The Null Hypothesis

Statistical hypothesis are usually stated in a null form; that is, a statement is made that no difference exists or no relationship between $x$ and $y$. This is called the null hypothesis. For example, comparing a single sample mean $\bar{x}$ and a known population mean $\mu$, the null hypothesis is stated as

$$H_0: \mu - \bar{x} = 0$$

This is a hypothesis that no difference exists between a population parameter and a sample mean. This will be example about a hypothesis involving one sample case. However, there other research situations employing two experimental groups and the investigator wishes to ascertain if there is treatment effect or if the population means (treatment means) are the same. In this kind of situation, two treatment means are obtained. These means are estimates of the population means. In this example the null hypothesis is written as

$$H_0: \mu_1 - \mu_2 = 0$$

or

$$H_0: \mu_1 = \mu_2$$
This assertion simply states that no difference exists between the two population means. In general the null hypothesis when applied to two or more statistics is a tentative hypothesis asserting no differences between two or more population parameters.

A test of significance is a test of a null hypothesis. The scientific investigator wishes to decide between two alternatives. Can the observed difference be attributed to sampling error or does real difference exists between the parameters under consideration? This decision is based on a probability (chapter 6). A test of significance is simply a method for estimating a probability. Examples of such tests include a z-test, t-test, F-test, and Chi-Square statistic. Regardless of the test used, the end result is always the same.

The logical steps used in applying a test of significance are these. First, state the hypotheses and select the level of significance (alpha level) (to be discussed shortly). Second, state the criteria for a decision. Third, compute test statistic. Fourth, make a decision. In selecting the alpha level, a probability value, we are asking what is probability of obtaining a difference equal to or greater than the one observed in drawing samples at random from populations where the null hypothesis is assumed to be true? In step 4, if this probability is small, the observed difference or result being highly improbable on the basis of the null hypothesis, rejection of the null hypothesis is warranted. This means that the observed differences cannot be simply due to sample error (chance). These differences are real. To illustrate, a researcher states the probability that two groups are the same in the average reaction times is less than 5%. This basically means that the two groups under consideration are not the same in reaction times. Thus the result may be significant. Anytime the null hypothesis is rejected, we have a significant result or a significant difference, and this applies to all statistical tests.

In testing of any statistical hypothesis, it is necessary to state an alternative hypothesis. This alternative hypothesis is accepted if the null hypothesis is rejected. Thus in the testing of the null hypothesis $H_0: \mu_1 = \mu_2$, the alternative may be $H_1: \mu_1 \neq \mu_2$.

### 8.3 Two Types of Error

In making a decision about the null hypothesis $H_0$, two types of error may be made. Rejecting a true null hypothesis and accepting the alternative hypothesis is called a Type I error. In essence the researcher has concluded that there was an effect when there was none. Failing to reject a false null hypothesis is called a Type II error. In this case the researcher has concluded that there was no effect when there was one. The probabilities of making Type I and Type II errors are determined, respectively, by alpha and beta levels. Type I and Type II errors may be represented below:

<table>
<thead>
<tr>
<th>$H_0$ is true</th>
<th>$H_1$ is true</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept $H_1$</td>
<td>Type I error</td>
</tr>
<tr>
<td>Accept $H_0$</td>
<td>Correct decision</td>
</tr>
<tr>
<td></td>
<td>Type II error</td>
</tr>
</tbody>
</table>

The situation above is similar to that in the jury trial of a person charged with committing a crime. The accused is assumed innocent until found guilty. This is a null hypothesis. In making a
decision, the jury found the accused guilty when in fact he is innocent. This is a Type I error. If the jury found him not guilty when in fact the accused committed the crime, this is a Type II error. And of course, if the verdict was guilty when the accused is in fact guilty, the jury made the correct decision.

### 8.4 Levels of Significance and Critical Regions

The probability of a Type I error is called the level of significance or the alpha level. The alpha level is used to define the concept of “very unlikely” in a hypothesis test. Alpha values are very small probability values and it is a common convention to use levels of significance of either .05 or .01. An alpha level (level of significance) is used to establish critical region, which is composed of extreme sample values that are very unlikely to be obtained if the null hypothesis is true. Specifically, the critical region defines cut-off points for making decision regarding the null hypothesis. If a test statistic is beyond the cut-off points, assuming the alpha level is .05, the probability is equal to or less than .05, stating that there is a difference, say between the sample statistic and the corresponding population parameter. In this case, the difference is said to be significant. The .05 and .01 probability levels are descriptive of degree of confidence that a real difference exits, assuming that the null hypothesis was rejected, and the observed difference is not simply due to chance. It is important to note that a large alpha level will increase the probability of making a Type I error.

### 8.5 Directional and Non-Directional Tests

A researcher may wish to test the null hypothesis against the alternative hypothesis without regard to the direction of the difference. For example, are there gender differences in grade point average (GPA)? No assertion is made that males will have a higher GPA than females. This exemplifies a **non-directional test** or a two-tailed test, because if the normal distribution (i.e., z-score distribution), or the t-distribution is used, the critical region (cut-off points) will located in two tails or the two sides. Consider an alpha level of .05. If the sampling distribution is normal, 2.5 percent (.025) falls to the right of 1.96 standard deviation units above the mean, and 2.5 percent (.025) falls to the left of 1.96 standard deviation units below the mean. Thus for significance at the 5% level for a non-directional test, the observed statistic for testing the null hypothesis must be equal to or greater than 1.96. For significance at the 1% level a value of 2.58 is required for a non-directional test. Consider Figure 8.1, which shows alpha level of .05 and the critical values associated with it.
There are certain circumstances in which we may decide to make a decision about the direction of the difference. Tests for making specific direction of the difference or relationship are called directional or one-tailed tests. **Directional tests**, or **one-tailed tests**, are hypothesis tests where the alternative hypothesis is stated as greater than (>) or less than (<) a value stated in the null hypothesis. If the normal, or t, distribution is used, one-tail is employed to estimate the required probabilities. For example, with research hypothesis $\mu > 15$, the null hypothesis $H_0$ that $\mu \leq 15$ will be rejected if the sample mean is greater than 15 that its $z$ score falls beyond 1.65 in the upper right tail of the normal curve. The test is called an upper one-tailed test because the entire area of rejection, with area = $\alpha$, falls in the upper tail of the curve. Using the normal distribution, a test statistic value greater than 1.65 is required for significance at the .05 level. Consider Figure 8.2.

If the research hypothesis is $\mu < 15$, then the null hypothesis $H_0$: $\mu \geq 15$, $H_0$ will be rejected if the sample mean is less than 15, but this time the area of rejection (i.e., the critical region) will be on the left side of the normal curve. Hence, this is called a lower one-tailed test.
8.6 Test of Significance for a Single Mean

A test of significance may be applied to test whether a sample mean \( \bar{x} \) (i.e., \( M \)) is significantly different from a population mean \( \mu \). Two situations may be considered. First, the population mean and standard deviation may be known. Second, the population mean may be known, but not the standard deviation. In the first situation, where the population mean and the population standard deviation (population variance) are known or given, \textbf{z-test (z-statistic)} for the distribution of sample means is employed as a test of significance.

There are three assumptions underlying the use of a z-test for hypothesis testing. The following assumptions (conditions) should be considered before employing the z-test:

1. The population under consideration is normal or the sample selected from the population is reasonably large.
2. The sample selected has been drawn randomly from the population or the sample is thought to be highly representative of the population.
3. The population standard deviation or the population variance is known.

Assuming that the assumptions for using the z-test has been met, testing the significance for a single mean consists of the following steps:

1. State the hypotheses (null and alternative hypotheses)
2. Select a level of significance, alpha, \( \alpha \)
3. Compute or calculate the test statistic (in this case, z-statistic). Calculation of the z-statistic requires two steps: (1) standard error and (2) z-statistic
4. Find the probability of the obtained statistic by consulting the Unit Normal Table to determine whether the probability of the obtained z statistic is less than or greater than alpha.
5. Make a decision. If the probability is less than alpha (i.e., \( p < \alpha \)), reject Ho, if not, fail to reject Ho (i.e., do not reject Ho).

**Example of a Two-Tailed Test (Non-Direction Test)**

To illustrate the situation where the population mean and standard deviation are known, let us suppose that the mean GPA for a sample of 25 college students \( \bar{x} = 2.05 \) and the sample standard deviation, \( s \) is .35. The mean and standard deviation in the general population of students are \( \mu = 2 \) and \( \sigma = .40 \), respectively. The distribution is normal. Does the sample mean differ significantly from the population mean of 2.00? Since we are interested in any difference from 2, in either a positive or negative direction, we use a two-tailed test. In answering the question of whether the sample mean is significantly different from the population mean, we utilize the five steps in hypothesis testing outlined above: (1), state the hypotheses, (2) select an alpha level, (3), compute the test statistic, (4), find the probability of the obtained statistic, and (5), make a decision.
Step 1: State the hypotheses and select the alpha level. We begin by stating the value of a population mean in a null hypothesis, which we presume is true. For the college students’ example, we state the null hypothesis that college students in the general population have an average of GPA of 2.00. This can be written as

\[ H_0: \mu = 2.00 \text{ or } H_0: \mu - 2 = 0 \]

\[ H_1: \mu \neq 2.00 \text{ or } H_1: \mu - 2 \neq 0 \]

\[ \alpha = .05 \]

Step 2: Set the criteria for a decision. To set the criteria for a decision, we state the level of significance for a test. The likelihood or level of significance is typically set at 5% in behavioral research studies. When the probability of obtaining a sample mean is less than 5% if the null hypothesis were true, then we conclude that the sample we selected is too unlikely and so we reject the null hypothesis.

Since our distribution is normally distributed, the values required for significance at the .05 level for a non-directional test is plus minus 1.96, our cut-off points. Thus the null hypothesis may be rejected with probability less than .05. Hence \( \alpha = .05 \)

Step 3: Compute the test statistic. In the problem, it was stated that the average GPA for a sample of \( n = 25 \) students was 2.05. To make a decision, we need to evaluate how likely this sample outcome is, if the population mean stated by the null hypothesis \( \mu = 2.00 \), is true. We use a test statistic to determine this likelihood. Specifically, a test statistic tells us how far, or how many standard deviations, a sample mean is from the population mean. The larger the value of the test statistic, the further the distance, or number of standard deviations, a sample mean is from the population mean stated in the null hypothesis. The value of the test statistic is used to make a decision in Step 4. Computing the test statistic requires two steps: (1), standard error, \( \sigma_x \) and (2) a \( z \)-score. In drawing samples from a population with \( \mu = 2.00 \) and \( \sigma = .40 \) the standard deviation of the sampling distribution (standard error) is \( \sigma_x = \sigma / \sqrt{n} = .40 / \sqrt{25} = .08 \) and

\[ z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{2.05 - 2.00}{.08} = .63 \]

Step 4: Find the probability of the test statistic, \( z \). That is, we must find \( p(z < -.63 \text{ or } z > .63) \).

Consulting the Unit Normal Table, using the area beyond \( z \) in tail, column C, we find that \( p(z < -.63) = .2643 \) and \( p(z > .63) = .2643 \). Therefore, \( p(z < -.63 \text{ or } z > .63) = 2(.2643) = .5286 \), which is greater than .05.

Step 5: Make a decision. We use the value of the obtained probability value and the test statistic to make a decision about the null hypothesis. The decision is based on the probability of obtaining a sample mean, given that the value stated in the null hypothesis is true. If the \( \mu = 2.00 \), we expect the sample mean to be equal to the population mean \( \mu = 2 \). The test statistic, \( z = .63 \), less than our critical value of 1.96. Thus the actual probability is greater than .05. In conclusion,
we fail to reject the null hypothesis (accept the null hypothesis) and state that the sample mean is not significantly different from the population.

**Example of a Lower One-Tailed Test**
Mary Smith, a school psychologist, working with the Houston Independent School district (HISD), believes that the old math curriculum has lowered the math achievement test in the school district. She decides to test the hypothesis that the mean math achievement test in the district is less than the general population mean, 100. The population standard deviation is 15. Thus, Mary’s research hypothesis is \( \mu < 100 \). The appropriate test of significance is a lower one-tailed test. The steps for this test of significance are given below:

**Step 1**: State hypotheses (i.e., \( H_0 \) and \( H_a \))
- \( H_0: \mu \geq 100 \)
- \( H_a: \mu < 100 \)

(\*Note: the alternative hypothesis is Mary’s research hypothesis\*)

**Step 2**: Select alpha level, \( \alpha \)
For this test, we decide to use \( \alpha = 0.05 \)

**Step 3**: Collect data and calculate the test statistic, \( z \)
Mary administers a math achievement test to a random sample of 25 eight graders from HISD and finds the mean (M) = 94. As expected, the sample mean is less than 100. But is this mean of 94 significantly less than 100 to warrant rejection of the null hypothesis? To answer this question, we must calculate the \( z \)-score corresponding to 94, which entails two steps:

1. Standard error, \( \sigma_{\bar{X}} = \sigma/\sqrt{n} = 15/\sqrt{25} = 3.0 \)
2. \( z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = 94 - 100/3.0 = -2.00 \)

**Step 4**: Find the probability of the test statistic, \( z \)
Consulting the Unit Normal Table, Mary finds the \( p(z < -2.00) = 0.0228 \), which is less than 0.05.

**Step 5**: Make a decision. Since \( p < 0.05 \), Mary rejects the null hypothesis and concludes that the mean math achievement test score in the school district is significantly lowered. We can also use the critical value of \( z \) to make a decision. The \( z \)-critical value required for significance at the 0.05 level of significance is -1.65, and since our obtained \( z \) score of -2.00 falls beyond -1.65, we reject the null hypothesis.

**Example of an Upper One-Tailed Test**
John Williams has developed a program for teaching introductory statistic and believes that this program will increase student’s knowledge of basic statistics. He selects a random sample of students currently enrolled in a statistics course and the students are taught with this program for the entire semester. In the general population, midterm exam scores are normally distributed with a mean of 35 and a standard deviation of 8. He will use the \( z \)-test to test his hypothesis that the mean midterm exam score of students selected for the program will be greater than 35 (\( \mu > 35 \)).
Thus, John’s research hypothesis is $\mu > 35$. The appropriate test of significance is an upper one-tailed test. The steps for this test of significance are given below:

**Step 1**: State hypotheses (i.e., $H_0$ and $H_a$)
- $H_0$: $\mu \leq 100$
- $H_a$: $\mu > 100$

*(Note: the alternative hypothesis is John’s research hypothesis)*

**Step 2**: Select alpha level, $\alpha$
For this test, we decide to use $\alpha = .01$

**Step 3**: Collect data and calculate the test statistic, $z$
John administers the midterm exam to a random sample of 15 students. Their midterm exam scores are 40, 39, 37, 45, 50, 36, 49, 44, 45, 39, 40, 37, 44, 40, 38 and finds the mean ($M$) = 41.53. As expected, the sample mean is greater than 35. But is this mean of 41.53 significantly greater than 35 to warrant rejection of the null hypothesis? To answer this question, we must calculate the $z$-score corresponding to 94, which entails two steps:

1. Standard error, $\sigma_x$, $\sigma_x / \sqrt{n} = 8/\sqrt{15} = 2.07$
2. $z = \frac{X - \mu}{\sigma / \sqrt{n}} = 41.53 - 35 / 2.07 = 3.15$

**Step 4**: Find the probability of the test statistic, $z$
Consulting the Unit Normal Table, John finds the $p(z > 3.15) = .0008$, which is less than .01.

**Step 5**: Make a decision. Since $p < .01$, John rejects the null hypothesis and concludes that the program significantly increased the midterm exam scores. We can also use the critical value of $z$ to make a decision. The $z$-critical value required for significance at the .01 level of significance is 2.33, and since our obtained $z$ score of 3.15 falls beyond 2.33, we reject the null hypothesis.

An alternative method for testing significance can be simplified as follows:

**Step 1**: State hypotheses and select alpha level, $\alpha$.
**Step 2**: State the criteria for rejecting $H_0$
**Step 3**: Compute or calculate the test statistic
**Step 4**: Make a decision.

In the examples given above for testing the significance for a single mean, $z$-test was used. The **one sample $z$ test** is a statistical procedure used to test hypotheses concerning the mean in a single population with a known variance or standard deviation. However, situations arise where the population mean is known but not the population standard deviation. This parameter must then be estimated from the data. Here the $t$ distribution, and not the distribution of $z$, may be used.
8.7 Introduction to the t-distribution

In chapter 7, we discussed the sampling distribution of the sample means and stated that this distribution approaches the normal distribution as sample size increases, regardless of the shape of the population distribution. However, when means are expressed as z-scores (i.e., standard score), problems arise. Consequently, these problems lead to a discussion of an important theoretical distribution known as the t distribution. As previously stated, the z-statistic is used in situations where the population mean and standard deviation are known. In this situation, the z-statistic for one sample is written as

\[ z = \frac{X - \mu}{\sigma/\sqrt{n}} \]

The z statistic is a deviation of the sample mean from the population mean \( \mu \) divided by the standard error of the mean (i.e., the standard deviation of the sampling distribution). If each sample mean is subtracted from the population and divided by the standard deviation of the sampling, the resulting z values have a mean of 0 and a standard deviation of 1, and the shape of this new distribution is not altered, as transformation of sample means to z scores does not change the shape of the original distribution as discussed in chapter 5, z scores.

In practice, when testing hypotheses, seldom is the population variance or population standard deviation is known; these values must be estimated from the data. Consideration of this situation leads to the t statistic, given by

\[ t = \frac{M - \mu}{s_M} \]

where:

\[ s_M \text{ is the estimated standard deviation of the sampling distribution.} \]

\[ s_M = \sqrt{\frac{s_x^2}{n}} \]

The t ratio has two estimated values, M and \( s_M \). Both values vary with repeated sampling. The resulting distribution is not normal, but it approaches normality with increase in sample size.

The t distribution departs from the normal distribution for small sample sizes. To illustrate, if 100 samples of say 5 members were drawn from a normal population with mean \( \mu \) and variance, and 100 values of t obtained, the frequency distribution of these hundred values of t would not be normal. It would tend to be a symmetrical distribution with somewhat thicker tails than the normal distribution.

The theoretical sampling distribution of t is a symmetrical distribution with a mean of zero. It tapers off to infinity at the two extremities. It is variable and flatter than the normal standard distribution (i.e., z-score distribution). It is not a single distribution, like the standard normal distribution, but a family of t distributions. A different value of t exists for each number
of degrees of freedom. As the number of degrees of freedom increases the \( t \) distribution approaches the standard normal distribution. Figure 8.3 shows the distribution for various degrees of freedom.

![Figure 8.3](image)

**FIGURE 8.3**
Distribution of \( t \) for various degrees of freedom

A \( t \) distribution table can be used to obtain critical values required for significance.

**Example of a \( t \)-test for Significance Test for a Single Mean**

In the beginning of this section, we stated that situations arise where \( \mu \) is known but not the population standard deviation. Hence the \( t \) distribution is used. To illustrate, let us suppose that a widely used personality inventory is administered to a random sample of 25 participants. In this sample \( M = 47 \) and the sample standard deviation, \( s = 8 \). The mean and standard deviation in the population are \( \mu = 50 \) and \( SD = 10 \), respectively. The distribution is normal. Does the sample mean differ significantly from the population mean? Test at the .05 level of significance, two-tailed test.

**Note:** This is a two-tailed test, as no specific direction was stated in the problem.

We utilize the following steps in testing the hypothesis:

**Step 1:** State hypotheses (i.e., \( H_0 \) and \( H_a \))

\[
H_0: \mu = 50 \\
H_a: \mu \neq 50
\]

(Note: the alternative hypothesis is the research hypothesis)

**Step 2:** Select alpha level, \( \alpha \)

For this test, we decide to use \( \alpha = .05 \)

**Step 3:** Collect data and calculate the test statistic, \( z \)

The sample size is 25 and the sample mean given in the problem as 47 with \( s = 8 \). The population mean is 50. We want to determine whether the sample mean of 47 is significantly different from 50 to warrant rejection of the null hypothesis? To answer this question, we must calculate the \( t \)-statistic corresponding to 47, which entails two steps:

1. Estimated standard error, \( s_M = \sqrt{\frac{s^2}{n}} = s_M = \frac{8^2}{\sqrt{25}} = 1.6 \)
2. \[ t = \frac{M - \mu}{s_M} = \frac{47 - 50}{1.6} = -1.88 \]

**Step 4:** Find the probability of the test statistic, t
Consulting the t distribution Table, we find the \( p(t < -1.88) > .05 \)

**Step 5:** Make a decision. The number of degrees of freedom is 24 and the values of \( t \) required for significance at .05 level is 2.064 for a two-tailed test. Our obtained \( t \) statistic of -1.88 is not in the critical region of ± 2.064. Therefore, we do not have enough evidence to reject the null hypothesis, we fail to reject the null hypothesis.

### 8.8 Power of a Statistical Test

The probability of a test correctly rejecting a false \( H_0 \) is called power. Decisions regarding \( H_0 \) are frequently made without reference to Type II error. Here \( \beta \) is the probability of failing to reject \( H_0 \) when \( H_a \) is true. Mathematically speaking, the quantity \( 1 - \beta \) is called the power of a test. In other words, power is the probability of stating that \( H_a \) (i.e., the alternative hypothesis) is true when in fact this is true. The alternative hypothesis could be a statement of “there is an effect” or a statement of “there is a change or difference.” In an experimental analogy it is the probability that the independent variable has an effect when in fact this is so.

Power has many applications in research, ranging from determining adequate sample size to potential research outcomes. An illustration of how power \( (1 - \beta) \) is calculated shed light on our understanding of power. Consider the problem of comparing a single sample mean \( M \) based on a sample size of \( N \), with a known population mean \( \mu_0 \). The null hypothesis \( H_0: \mu = \mu_0 \).
Assume that the rejection leads to the rejection of \( H_0 \). How may \( \beta \) and \( 1 - \beta \) be calculated in this situation?

Consider Figure 8.5, which shows two overlapping normal distributions. The distribution on the left shows when \( H_0 \) is true and the right distribution shows when \( H_1 \) is true. The unshaded portion of the right distribution depicts \( 1 - \beta \).

![Figure 8.5](image)

**FIGURE 8.5**
Illustration of the power of a test
There are three factors that determine the power of a test: (1) the critical value \( \alpha \), (2) the difference between the two means, and (3) sample size \( N \). Large \( \alpha \) increases the power of a test, as the difference between means increases, the power of a test also increases. Power is a function of sample size \( N \). As \( N \) increases, the power of the test also increases. Sample size is inversely related to standard error. Small values for the standard error result in larger test statistic, which increases the likelihood of rejecting Ho.

### 8.9 Measuring Effect Sizes: Cohen’s \( d \) and Variance Accounted for

When a test of significance results in the rejection of the null hypothesis, in the context of an experiment, the researcher is concluding that there is an effect. By measuring the amount of an effect, an investigator gets additional information regarding the magnitude and strength of the effect. The most commonly used measure for an effect is the Cohen’s \( d \). This statistic can be easily calculated.

The computational formula for the Cohen’s \( d \) is given below for a normal standard distribution involving a single sample mean:

\[
Cohen's \ d = \frac{M - \mu}{SD}
\]

The numerator in the formula is a deviation value. \( M \) is the sample mean and \( \mu \) is the population mean. Because we are dealing with a standard normal distribution, the denominator is the population standard deviation.

#### Calculation of Cohen’s \( d \) for a Single Mean

Let us suppose that a researcher obtained the following from his data: \( M = 53, \mu = 50, SD = 9 \)

Calculate the Cohen’s \( d \).

Applying the formula above, we have

\[
Cohen's \ d = \frac{M - \mu}{SD} = \frac{53 - 50}{9} = .33
\]

Cohen suggested the following standards for evaluating effect values:

- \( d < 0.2 \), this is a small effect
- \( 0.2 < d < 0.8 \), this is a medium effect
- \( d > 0.8 \), this is a large effect

Another measure of an effect size is the amount of variance accounted for by treatment effect. For the \( t \) statistic hypothesis test, the variance attributed to treatment effect is given by:

\[
r^2 = \frac{t^2}{t^2 + df}
\]
Where:

\( t \) is the calculated statistic from sample data

\( df \), degrees of freedom, for a single sample, \( df = N - 1 \)

**Calculation of Amount of Variance Accounted for by Treatment Effect**

Consider the following numerical information from a research study:

\( N = 25, \ M = 110 \) and \( s = 14 \). The population mean \( \mu = 100 \). The standard error \( = 14/\sqrt{25} = 2.80 \). In calculating the \( t \) statistic, we obtain

\[
 t = \frac{110 - 100}{2.80} = 3.57
\]

The amount or percentage of variance accounted for can be calculated as

\[
 r^2 = \frac{t^2}{t^2 + df} = \frac{3.57^2}{3.57^2 + 24} = .35
\]

If this were to be an experiment, we would say that the independent variable explained 35% of the variability in the dependent variable.

**8.10 Estimation and Confidence Intervals for One-Sample Case**

There are two types of statistical inference: Tests of significance (i.e., hypothesis testing) and estimation. We have discussed tests of significance for a single mean and provided some examples. The procedure for inferring the probable value of the population mean from sample data is known as estimation. The estimation of population values from sample data by investigators of all kinds is very common in modern society. The print and electronic media are replete with estimates of all kinds, covering topics of all kinds. For example, the political process is influenced by estimates of public opinion on many subjects obtained by sample surveys. Some of the estimates obtained are subject to large errors. The focus of the present discussion is dealing with estimation in the presence of sampling error.

A distinction is made between point estimates and interval estimates. A point estimate is a value obtained by direct calculation. It is a precise value obtained from sample data. If for example, we calculated the mean GPA for a sample of 100 students and the sample mean happens to be 2.25, this is a point estimate of an unknown parameter value. If the sample variance is calculated to be 72.75, this is also a point estimate. A point estimate by itself provides no information about the error that attaches to it as an estimate of a population parameter. An alternative approach is to specify an interval within which we may assert with some known degree of confidence that the population mean lies. Thus, for example, instead of calculating the point estimate \( \bar{M} \), we may perform a simple calculation, described later in this section, and assert with 95 percent confidence that the population mean, whose point estimate is \( \bar{M} = 2.25 \), falls within the limits 1.87 and 2.08. These values are confidence limits, and the interval they bound is called a confidence interval. Such an interval incorporates information about the magnitude of error.
Example of Confidence Intervals for Means of Large Samples

The calculation of confidence intervals for a mean with a large sample is based on the standard normal distribution or the critical values required for significance at a desired level of confidence. For large samples the 95 and 99 percent confidence intervals for the mean are given, respectively by $M \pm 1.96(S.E)$ and $M \pm 2.58(S.E)$. To illustrate the calculation of confidence intervals, let the mean IQ of a random sample of 100 school children be 114 and the standard deviation 17. The standard deviation here is the square root of the unbiased variance estimate. Our estimate of the standard error of the mean is $S.E = 17/\sqrt{100} = 1.70$. The 95 percent confidence interval is then given by

$$
\mu = M \pm z \text{ critical (S.E)}
$$

Where $M = 114$, $z = 1.96$ (from the Unit Normal Table)

$S.E = 1.70$

Substituting these values into the formula, we obtain

$$
\mu = 114 \pm 1.96 (1.70)
$$

CI95 = 110.67, 117.33

In the above calculation, the upper limit of our interval is 117.33 and the lower limit is 110.67. Thus we may assert with 95 percent confidence that the population mean falls between 110.67 and 117.33.

What does it actually mean that we are 95% confident that the actual population mean falls within certain specific limits? It means that if we were to take many different samples from the same population, calculate the mean and the 95 percent confidence intervals for each sample, construct frequency distributions of the upper and lower limits (i.e., intervals), 95 percent of the intervals will include our population mean, but only 5 percent would not include our population mean.

Implicit in the discussion above is the assumption that the ratio $\left( \frac{M - \mu}{S_M} \right)$ is normally distributed. This ratio is not normally distributed when the sample size, $n$ is small, but approaches the normal distribution as $n$ increases in size. It is a common statistical convention to consider sample size larger than 30 to be large and sample size less than 30 to be small.

Example of Confidence Intervals for Means of Small Samples

The calculation of confidence intervals for a mean with a small sample is based on the $t$ distribution. For large samples the 95 and 99 percent confidence intervals for the mean are given, respectively by $M \pm 1.96(S.E)$ and $M \pm 2.58(S.E)$. For small samples an unbiased estimate of the population variance is used in estimating the standard error ($S.E$). The value of $t$ used in fixing the limits of the 95 and 99 percent intervals will vary, depending on the number of degrees of freedom. To illustrate the calculation of confidence intervals for small samples, consider an example where $M = 24.26$, $s^2 = 64$, $N = 16$, and $df = 16 – 1$. In consulting the $t$ distribution table (see Appendix B, Table B.2), we observe that for 15 degrees of freedom 95 percent (i.e., $\alpha = .05$)
of the area of the distribution falls within a \( t \) of \( \pm 2.13 \) from the mean. The standard error using the unbiased variance estimate is \( \frac{8}{\sqrt{16}} = 2 \). The 95 percent confidence interval is then given by

\[
\mu = M \pm t \text{ critical (S.E)}
\]

Where \( M = 24.26 \), \( t = 2.13 \) (from the \( t \) distribution Table)
\( \text{S.E} = 2 \)

Substituting these values into the formula, we obtain

\[
\mu = 24.26 \pm 2.13 (2)
\]
\( \text{CI95} = 20.00, 28.52 \)

Likewise, the 99% confidence interval is given by

\[
\mu = 24.26 \pm 2.95 (2)
\]
\( \text{CI95} = 18.36, 30.16 \)

8.11 Using Confidence Intervals in Making Decisions

Estimation as an inferential procedure can aid in decision making regarding the claim made by the null hypothesis. The procedure is straightforward. We note the claim made by the null hypothesis, collect data, calculate the sample mean (\( s \)), as point estimates of unknown parameter values, then we calculate our desired confidence intervals for estimating parameter values. Once these confidence intervals of the mean are collected, we evaluate the credibility of the claim made by the null hypothesis. If the claim made by the null hypothesis falls within the limits of the confidence intervals, the null hypothesis is not rejected.

Confidence intervals are important statistical inferences used in explaining the meaning of research result that goes beyond basic hypothesis test. Confidence intervals are closely related with the procedures used in two-tailed tests of hypothesis about a parameter value or \( \mu \). For example, if we construct a 95 percent confidence interval for the value of \( \mu \), which is based on a sample of \( N \) cases drawn randomly from a given population, 95 percent of the confidence intervals will cover the true value of the parameter value or \( \mu \). Any claim made by the null hypothesis for \( \mu \) that is covered by the confidence interval could not be rejected at \( \alpha = .05 \). To illustrate we consider two situations: confidence intervals with the standard normal distribution and confidence intervals with the \( t \) distribution.

Decision Making: Confidence Intervals with the Standard Normal Distribution

Let us imagine that a widely used intelligence test is administered to a class of 25 school children. In this sample \( M = 110 \) and \( s = 14 \). The mean and standard deviation in the population used in the standardization of the test are 100 and 15 respectively. Does the sample mean differ significantly from the population mean? Test at .05 level of significance. The null hypothesis is \( H_0: \mu = 100 \) or \( H_0: \mu - 100 = 0 \). In calculating the test statistic, \( z \), we obtain

\[
z = \frac{110 - 100}{15/\sqrt{25}} = 3.33
\]
The value required for significance at the .05 level for a two-tailed test is 1.96. Our decision is to reject the null hypothesis because our obtained statistic of 3.33 exceeds 1.96 and conclude the difference between the sample mean and the population mean is significant. Our estimate of the standard error of the mean is $S.E = \frac{14}{\sqrt{25}} = 2.8$. The 95 percent confidence interval for estimating $\mu$ is then given by

$$\mu = M \pm z \text{ critical } (S.E)$$

Where $M = 110$, $z = 1.96$ (from the Unit Normal Table)

$S.E = 2.8$

Substituting these values into the formula, we obtain

$$\mu = 110 \pm 1.96 (2.8)$$

$CI_{95} = 104.51, 115.49$

The probability is .95 that the true population mean $\mu$ falls between 104.51 and 115.49. This interval does not contain 100 nor does it contain the value of 0, indicating that the null hypothesis stated above is rejected.

**Decision Making: Confidence Intervals with the $t$ Distribution**

The problem above can be used to demonstrate confidence intervals with the $t$ distribution. If we apply the $t$ formula for testing the significance of a single mean, we obtain

$$t = \frac{M - 100}{S.E} = 3.57$$

Note: $S.E = 2.8$

Consulting the $t$ table, we find $t$ value required for significance at the .05 level is $\pm 2.064$. The obtained $t$ statistic of 3.57 exceeds 2.064 and hence $H_0$ is rejected. The 95 percent confidence interval for estimating $\mu$ is then given by

$$\mu = M \pm t \text{ critical } (S.E)$$

Where $M = 110$, $t = 2.06$ (from the $t$ distribution Table)

$S.E = 2.8$

Substituting these values into the formula, we obtain

$$\mu = 110 \pm 2.06 (2.8)$$

$CI_{95} = 104.23, 115.77$

The probability is .95 that the true population mean $\mu$ falls between 104.23 and 115.77. This interval does not contain 100 nor does it contain the value of 0, indicating that the null hypothesis stated above is rejected.
Chapter Summary

1. Hypothesis testing is a statistical procedure that allows researchers to use sample data to draw inferences about the population of interest. Tests of statistical hypotheses are tests of population parameters.

2. Statistical hypothesis are usually stated in a null form; that is, a statement is made that no difference exists or no relationship between x and y. This is called the null hypothesis.

3. The probabilities of making Type I and Type II errors are determined, respectively, by alpha and beta levels.

4. The probability of a Type I error is called the level of significance or the alpha level. The alpha level is used to define the concept of “very unlikely” in a hypothesis test. Alpha values are very small probability values and it is a common convention to use levels of significance of either .05 or .01.

5. There are four steps in hypothesis testing: (1), state the hypotheses and select an alpha level, (2), set the criteria for a decision, (3), compute test statistic, and (4), make a decision.

6. Hypothesis test with a z-statistic is used when the population variance or population standard deviation is known.

7. A t-statistic is used when the population variance is unknown.

8. Confidence intervals can be calculated with a t-distribution.

9. Power is the probability of test correctly rejecting a false Ho. There are several factors that influence power.

Basic Terms and Concepts

Statistical hypothesis
Test of significance
Null hypothesis
Type I error
Type II error
Significance level (alpha level)
Directional and non-directional tests
z-test for a single mean
Unit Normal Table
t-distribution
t ratio
t-test for a single mean
Degrees of freedom
Power of a test
Effect size
Statistical Inference
Estimation
Point Estimate
Interval Estimate (Confidence Intervals)
Unbiased
Distinction between large and small sample

Exercises

1. Define the critical region for a hypothesis test, and explain how the critical region is related to the alpha level.

2. The term error is used in two different ways in hypothesis testing: a Type I error (or Type II) and the standard error.
   a. What can a researcher do to influence the size of the standard error? Does this action have any effect on the probability of a Type I error?
   b. What can a researcher do to influence the probability of a Type I error? Does this action have any effect on the size of the standard error?

3. Some researchers claim that herbal supplements such as ginseng or ginkgo biloba enhance human memory. To test this claim, a researcher selects a sample of n = 25 college students. Each student is given a ginkgo biloba supplement daily for six weeks and then all the participants are given a standardized memory test. For the population, scores on the test are normally distributed with \( \mu = 70 \) and \( \sigma = 15 \). The sample of n = 25 students had a mean score of \( M = 75 \).
   a. Are the data sufficient to that the herb has a significant effect on memory? Use a two-tailed test with \( \alpha = .05 \).
   b. Compute Cohen’s d for this study.

4. A researcher would like to determine whether a new tax on cigarettes has had any effect on people’s behavior. During the year before the tax was imposed, stores located in rest areas on the state thruway reported selling an average of \( \mu = 410 \) packs per day with \( \sigma = 60 \). The distribution of daily sales was approximately normal. For a sample of n = 9 days following the new tax, the researcher found an average of \( M = 386 \) packs per day for the same stores.
   a. Is the sample mean sufficient to conclude that there was a significant change in cigarette purchases after the new tax. Use a two-tailed test with \( \alpha = .05 \).
   b. If the population standard deviation was \( \sigma = 30 \), is the result sufficient to conclude that there is a significant difference?
   c. Explain why the two tests lead to different outcomes.

5. A researcher is testing the effectiveness of a new herbal supplement that claims to improve physical fitness. A sample of n = 16 college students is obtained and each student takes the supplement daily for six weeks. At the end of the 6-week period, each student is given a standardized fitness test and the average score for the sample is \( M = 39 \). For the general population of college students, the distribution of test scores is normal with a mean of \( \mu = 35 \) and a standard deviation of \( \sigma = 12 \). Do students taking the supplement have significantly better fitness scores? Use a one-tailed test with \( \alpha = .05 \).
6. A researcher selects a sample of $n = 25$ from a normal population with $\mu = 40$ and $\sigma = 10$. If the treatment is expected to increase scores by 3 points, what is the power of a two-tailed hypothesis test using $\alpha = .05$?