Chapter 7

Testing Hypotheses

Chapter Learning Objectives

- Understanding the assumptions of statistical hypothesis testing
- Defining and applying the components in hypothesis testing: the research and null hypotheses, sampling distribution, and test statistic
- Understanding what it means to reject or fail to reject a null hypothesis
- Applying hypothesis testing to two sample cases, with means or proportions

In the past, the increase in the price of gasoline could be attributed to major national or global event, such as the Lebanon and Israeli war or Hurricane Katrina. However, in 2005, the price for a gallon of regular gasoline reached $3.00 and remained high for a long time afterward. The impact of unpredictable fuel prices is still felt across the nation, but the burden is greater among distinct social economic groups and geographic areas.

Lower-income Americans spend eight times more of their disposable income on gasoline than wealthier Americans do.[1] For example, in Wilcox, Alabama, individuals spend 12.72% of their income to fuel one vehicle, while in Hunterdon Co., New Jersey, people spend 1.52%. Nationally, Americans spend 3.8% of their income fueling one vehicle.

The first state to reach the $3.00-per-gallon milestone was California in 2005. California's drivers were especially hit hard by the rising price of gas, due in part to their reliance on automobiles, especially for work commuters. Analysts predicted that gas prices would continue to rise nationally. Declines in consumer spending and confidence in the economy have been attributed in part to the high (and rising) cost of gasoline.

In 2010, gasoline prices have remained higher for states along the West Coast, particularly in Alaska and California. Let's say we drew a random sample of California gas stations ($N = 100$) and calculated the mean price for a gallon of regular gas. Based on consumer information,[2] we also know that nationally the mean price of a gallon was $2.86, with a standard deviation of 0.17 for the same week. We can thus compare the mean price of gas in California with the mean price of all gas stations in April 2010. By comparing these means, we are asking whether it is reasonable to consider our random sample of California gas as representative of the population of gas stations in the United States. Actually, we

expect to find that the average price of gas from a sample of California gas stations will be unrepresentative of the population of gas stations because we assume higher gas prices in the state.

The mean price for our sample is $3.11. This figure is higher than $2.86, the mean price per gallon across the nation. But is the observed gap of 25 cents ($3.11 − $2.86) large enough to convince us that the sample of California gas stations is not representative of the population?

There is no easy answer to this question. The sample mean of $3.11 is higher than the population mean, but it is an estimate based on a single sample. Thus, it could mean one of two things: (1) The average price of gas in California is indeed higher than the national average or (2) the average price of gas in California is about the same as the national average, and this sample happens to show a particularly high mean.

How can we decide which of these explanations makes more sense? Because most estimates are based on single samples and different samples may result in different estimates, sampling results cannot be used directly to make statements about a population. We need a procedure that allows us to evaluate hypotheses about population parameters based on sample statistics. In Chapter 6, we saw that population parameters can be estimated from sample statistics. In this chapter, we will learn how to use sample statistics to make decisions about population parameters. This procedure is called **statistical hypothesis testing**.

> ***Statistical hypothesis testing*** A procedure that allows us to evaluate hypotheses about population parameters based on sample statistics.

ASSUMPTIONS OF STATISTICAL HYPOTHESIS TESTING

Statistical hypothesis testing requires several assumptions. These assumptions include considerations of the level of measurement of the variable, the method of sampling, the shape of the population distribution, and the sample size. The specific assumptions may vary, depending on the test or the conditions of testing. However, without exception, *all* statistical tests assume random sampling. Tests of hypotheses about means also assume interval-ratio level of measurement and require that the population under consideration be normally distributed or that the sample size be larger than 50.

Based on our data, we can test the hypothesis that the average price of gas in California is higher than the average national price of gas. The test we are considering meets these conditions:

1. The sample of California gas stations was randomly selected.
2. The variable *price per gallon* is measured at the interval-ratio level.
3. We cannot assume that the population is normally distributed. However, because our sample size is sufficiently large ($N > 50$), we know, based on the central limit theorem, that the sampling distribution of the mean will be approximately normal.

STATING THE RESEARCH AND NULL HYPOTHESES

Hypotheses are usually defined in terms of interrelations between variables and are often based on a substantive theory. Earlier, we defined *hypotheses* as tentative answers to research questions. They

are tentative because we can find evidence for them only after being empirically tested. The testing of hypotheses is an important step in this evidence-gathering process.

The Research Hypothesis (H_1)

Our first step is to formally express the hypothesis in a way that makes it amenable to a statistical test. The substantive hypothesis is called the **research hypothesis** and is symbolized as H_1. Research hypotheses are always expressed in terms of population parameters because we are interested in making statements about population parameters based on our sample statistics.

> ***Research hypothesis (H_1)*** A statement reflecting the substantive hypothesis. It is always expressed in terms of population parameters, but its specific form varies from test to test.

In our research hypothesis (H_1), we state that the average price of gas in California is higher than the average price of gas nationally. Symbolically, we use μ_Y to represent the population mean; our hypothesis can be expressed as

$$H_1: \mu_Y > \$2.86$$

In general, the research hypothesis (H_1) specifies that the population parameter is one of the following:

1. Not equal to some specified value: $\mu_Y \neq$ some specified value
2. Greater than some specified value: $\mu_Y >$ some specified value
3. Less than some specified value: $\mu_Y <$ some specified value

The Null Hypothesis (H_0)

Is it possible that in the population there is no real difference between the mean price of gas in California and the mean price of gas in the nation and that the observed difference of 0.25 is actually due to the fact that this particular sample happened to contain California gas stations with higher prices? Since statistical inference is based on probability theory, it is not possible to prove or disprove the research hypothesis directly. We can, at best, estimate the *likelihood* that it is true or false.

To assess this likelihood, statisticians set up a hypothesis that is counter to the research hypothesis. The **null hypothesis**, symbolized as H_0, contradicts the research hypothesis and usually states that there is no difference between the population mean and some specified value. It is also referred to as the hypothesis of "no difference." Our null hypothesis can be stated symbolically as

$$H_0: \mu_Y = \$2.86$$

Rather than directly testing the substantive hypothesis (H_1) that there is a difference between the mean price of gas in California and the mean price nationally, we test the null hypothesis (H_0) that there

is no difference in prices. In hypothesis testing, we hope to reject the null hypothesis to provide support for the research hypothesis. Rejection of the null hypothesis will strengthen our belief in the research hypothesis and increase our confidence in the importance and utility of the broader theory from which the research hypothesis was derived.

Null hypothesis (H_0) A statement of "no difference" that contradicts the research hypothesis and is always expressed in terms of population parameters.

More About Research Hypotheses: One- and Two-Tailed Tests

In a **one-tailed test**, the research hypothesis is directional; that is, it specifies that a population mean is either less than (<) or greater than (>) some specified value. We can express our research hypothesis as either

$$H_1: \mu_Y < \text{some specified value}$$

or

$$H_1: \mu_Y > \text{some specified value}$$

The research hypothesis we've stated for the average price of a gallon of regular gas in California is a one-tailed test.

When a one-tailed test specifies that the population mean is *greater than* some specified value, we call it a **right-tailed test** because we will evaluate the outcome at the right tail of the sampling distribution. If the research hypothesis specifies that the population mean is *less than* some specified value, it is called a **left-tailed test** because the outcome will be evaluated at the left tail of the sampling distribution. Our example is a right-tailed test because the research hypothesis states that the mean gas prices in California are higher than $2.86. (Refer to Figure 7.1 on page 163.)

Sometimes, we have some theoretical basis to believe that there is a difference between groups, but we cannot anticipate the direction of that difference. For example, we may have reason to believe that the average price of California gas is *different* from that of the general population, but we may not have enough research or support to predict whether it is *higher* or *lower*. When we have no theoretical reason for specifying a direction in the research hypothesis, we conduct a **two-tailed test**. The research hypothesis specifies that the population mean is not equal to some specified value. For example, we can express the research hypothesis about the mean price of gas as

$$H_1: \mu_Y \neq \$2.86$$

With both one- and two-tailed tests, our null hypothesis of no difference remains the same. It can be expressed as

$$H_0: \mu_Y = \text{some specified value}$$

One-tailed test A type of hypothesis test that involves a directional research hypothesis. It specifies that the values of one group are either larger or smaller than some specified population value.

Right-tailed test A one-tailed test in which the sample outcome is hypothesized to be at the right tail of the sampling distribution.

Left-tailed test A one-tailed test in which the sample outcome is hypothesized to be at the left tail of the sampling distribution.

Two-tailed test A type of hypothesis test that involves a nondirectional research hypothesis. We are equally interested in whether the values are less than or greater than one another. The sample outcome may be located at both the lower and the higher ends of the sampling distribution.

DETERMINING WHAT IS SUFFICIENTLY IMPROBABLE: PROBABILITY VALUES AND ALPHA

Now let's put all our information together. We're assuming that our null hypothesis ($\mu_Y = \$2.86$) is true, and we want to determine whether our sample evidence casts doubt on that assumption, suggesting that there is evidence for research hypothesis, $\mu_Y > \$2.86$. What are the chances that we would have randomly selected a sample of California gas stations such that the average price per gallon is higher than \$2.86, the average for the nation? We can determine the chances or probability because of what we know about the sampling distribution and its properties. We know, based on the central limit theorem, that if our sample size is larger than 50, the sampling distribution of the mean is approximately normal, with a mean $\mu_{\bar{Y}} = \mu_Y$ and a standard deviation (standard error) of

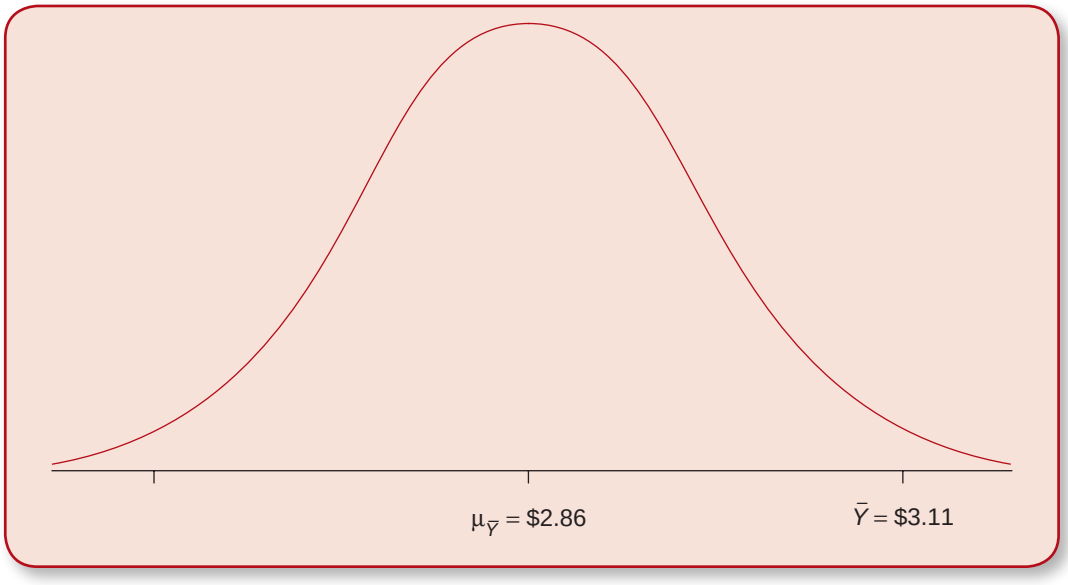
$$\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{N}}$$

We are going to assume that the null hypothesis is true and then see if our sample evidence casts doubt on that assumption. We have a population mean $\mu_Y = \$2.86$ and a standard deviation $\sigma_Y = 0.17$. Our sample size is $N = 100$, and the sample mean is \$3.11. We can assume that the distribution of means of all possible samples of size $N = 100$ drawn from this distribution would be approximately normal, with a mean of \$2.86 and a standard deviation of

$$\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{N}} = \frac{0.17}{\sqrt{100}} = 0.017$$

This sampling distribution is shown in Figure 7.1. Also shown in Figure 7.1 is the mean gas price we observed for our sample of California gas stations.

Figure 7.1 Sampling Distribution of Sample Means Assuming H_0 Is True for a Sample $N = 100$



Because this distribution of sample means is normal, we can use Appendix A to determine the probability of drawing a sample mean of $3.11 or higher from this population. We will translate our sample mean into a Z score so that we can determine its location relative to the population mean. In Chapter 5, we learned how to translate a raw score into a Z score by using Formula 5.1:

$$Z = \frac{Y - \bar{Y}}{S_Y}$$

Because we are dealing with a sampling distribution in which our raw score is $\bar{Y}$, the mean, and the standard deviation (standard error) is $\sigma_Y/\sqrt{N}$, we need to modify the formula somewhat:

$$Z = \frac{\bar{Y} - \mu_Y}{\sigma_Y/\sqrt{N}} \tag{7.1}$$

Converting the sample mean to a Z-score equivalent is called computing the *test statistic*. The Z value we obtain is called the ***Z* statistic (obtained)**. The obtained Z gives us the number of standard deviations (standard errors) that our sample is from the hypothesized value (μ_Y or $\mu_{\bar{Y}}$), assuming the null hypothesis is true. For our example, the obtained Z is

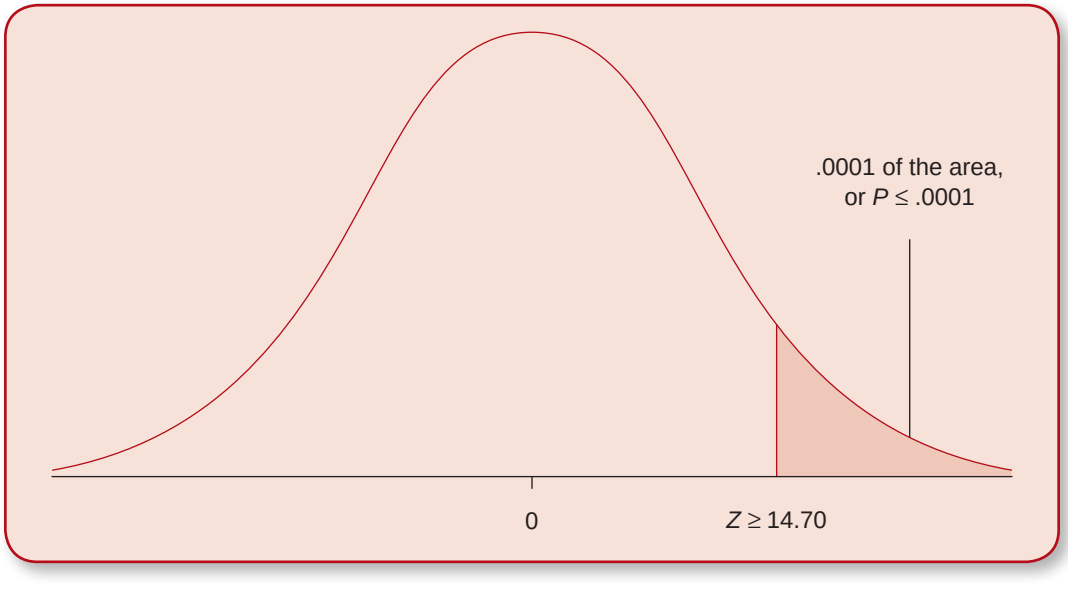
$$Z = \frac{3.11 - 2.86}{0.17/\sqrt{100}} = 14.70$$

Z statistic (obtained) The test statistic computed by converting a sample statistic (such as the mean) to a Z score. The formula for obtaining Z varies from test to test.

Before we determine the probability of our obtained *Z* statistic, let's determine whether it is consistent with our research hypothesis. Recall that we defined our research hypothesis as a right-tailed test ($\mu_Y > \$2.86$), predicting that the difference would be assessed on the right tail of the sampling distribution. The positive value of our obtained *Z* statistic confirms that we will be evaluating the difference on the right tail. (If we had a negative obtained *Z*, it would mean the difference would have to be evaluated at the left tail of the distribution, contrary to our research hypothesis.)

To determine the probability of observing a *Z* value of 14.70, assuming that the null hypothesis is true, look up the value in Appendix A to find the area to the right of (above) the *Z* of 14.70. Our calculated *Z* value is not listed in Appendix A, so we'll need to rely on the last *Z* value reported in the table, 4.00. Recall from Chapter 5, where we calculated *Z* scores and their probability, that the *Z* values are located in Column A. The *P* value is the probability to the right of the obtained *Z*, or the "area beyond *Z*" in Column C. This area includes the proportion of all sample means that are \$3.11 or higher. The proportion is less than 0.0001 (Figure 7.2). This value is the probability of getting a result as extreme as the sample result if the null hypothesis is true; it is symbolized as *P*. Thus, for our example, $P \leq .0001$.

Figure 7.2 The Probability (*P*) Associated With $Z \geq 14.70$



A ***P* value** can be defined as the actual probability associated with the obtained value of *Z*. It is a measure of how unusual or rare our obtained statistic is compared with what is stated in our null hypothesis. The smaller the *P* value, the more evidence we have that the null hypothesis should be rejected in favor of the research hypothesis.

P *value* The probability associated with the obtained value of *Z*.

Researchers usually define in advance what a sufficiently improbable Z value is by specifying a cutoff point below which P must fall to reject the null hypothesis. This cutoff point, called **alpha** and denoted by the Greek letter α, is customarily set at the .05, .01, or .001 level. Let's say that we decide to reject the null hypothesis if $P \leq .05$. The value .05 is referred to as alpha (α); it defines for us what result is sufficiently improbable to allow us to take the risk and reject the null hypothesis. An alpha (α) of .05 means that even if the obtained Z statistic is due to sampling error, so that the null hypothesis is true, we would allow a 5% risk of rejecting it. Alpha values of .01 and .001 are more cautionary levels of risk. The difference between P and alpha is that P is the *actual probability* associated with the obtained value of Z, whereas alpha is the level of probability *determined in advance* at which the null hypothesis is rejected. The null hypothesis is rejected when $P \leq \alpha$.

alpha (α) The level of probability at which the null hypothesis is rejected. It is customary to set alpha at the .05, .01, or .001 level.

We have already determined that our obtained Z has a probability value less than .0001. Since our observed P is less than .05 ($P = .0001 < \alpha = .05$), we reject the null hypothesis. The value of .0001 means that fewer than 1 out of 10,000 samples drawn from this population are likely to have a mean that is 14.70 Z scores above the hypothesized mean of \$2.86. Another way to say it is as follows: There is only 1 chance out of 10,000 (or .0001%) that we would draw a random sample with a $Z \geq 14.70$ if the mean price of California gas were equal to the national mean price.

Based on the P value, we can also make a statement regarding the "significance" of the results. If the P value is equal to or less than our alpha level, our obtained Z statistic is considered *statistically significant*—that is to say, it is very unlikely to have occurred by random chance or sampling error. We can state that the difference between the average price of gas in California and nationally is significantly different at the .05 level, or we can specify the actual level of significance by saying that the level of significance is less than .0001.

Recall that our hypothesis was a one-tailed test ($\mu_Y >$ \$2.86). In a two-tailed test, sample outcomes may be located at both the higher and the lower ends of the sampling distribution. Thus, the null hypothesis will be rejected if our sample outcome falls either at the left or right tail of the sampling distribution. For instance, a .05 alpha or P level means that H_0 will be rejected if our sample outcome falls among either the lowest or the highest 5% of the sampling distribution.

Suppose we had expressed our research hypothesis about the mean price of gas as

$$H_1: \mu_Y \neq \$2.86$$

The null hypothesis to be directly tested still takes the form $H_0: \mu_Y =$ \$2.86 and our obtained Z is calculated using the same formula (7.1) as was used with a one-tailed test. To find P for a two-tailed test, look up the area in Column C of Appendix A that corresponds to your obtained Z (as we did earlier) and then multiply it by 2 to obtain the two-tailed probability. Thus, the two-tailed P value for $Z =$ 14.70 is $.0001 \times 2 = .0002$. This probability is less than our stated alpha (.05), and thus, we reject the null hypothesis.

THE FIVE STEPS IN HYPOTHESIS TESTING: A SUMMARY

Regardless of the particular application or problem, statistical hypothesis testing can be organized into five basic steps. Let's summarize these steps:

1. Making assumptions
2. Stating the research and null hypotheses and selecting alpha
3. Selecting the sampling distribution and specifying the test statistic
4. Computing the test statistic
5. Making a decision and interpreting the results

Making Assumptions. Statistical hypothesis testing involves making several assumptions regarding the level of measurement of the variable, the method of sampling, the shape of the population distribution, and the sample size. In our example, we made the following assumptions:

1. A random sample was used.
2. The variable *price per gallon* is measured on an interval-ratio level of measurement.
3. Because $N > 50$, the assumption of normal population is not required.

Stating the Research and Null Hypotheses and Selecting Alpha. The substantive hypothesis is called the *research hypothesis* and is symbolized as H_1. Research hypotheses are always expressed in terms of population parameters because we are interested in making statements about population parameters based on sample statistics. Our research hypothesis was

$$H_1: \mu_Y > \$2.86$$

The *null hypothesis*, symbolized as H_0, contradicts the research hypothesis in a statement of no difference between the population mean and our hypothesized value. For our example, the null hypothesis was stated symbolically as

$$H_0: \mu_Y = \$2.86$$

We set alpha at .05, meaning that we would reject the null hypothesis if the probability of our obtained Z was less than or equal to .05.

Selecting the Sampling Distribution and Specifying the Test Statistic. The normal distribution and the Z statistic are used to test the null hypothesis.

Computing the Test Statistic. Based on Formula 7.1, our Z statistic is 14.70.

Making a Decision and Interpreting the Results. We confirm that our obtained Z is on the right tail of the distribution, consistent with our research hypothesis. Based on our obtained Z statistic of 14.70, we

determine that its *P* value is less than .0001, less than our .05 alpha levels. We have evidence to reject the null hypothesis of no difference between the mean price of California gas and the mean price of gas nationally. We thus conclude that the price of California gas is, on average, significantly higher than the national average.

ERRORS IN HYPOTHESIS TESTING

We should emphasize that because our conclusion is based on sample data, we will never really know if the null hypothesis is true or false. In fact, as we have seen, there is a 0.01% chance that the null hypothesis is true and that we are making an error by rejecting it.

The null hypothesis can be either true or false, and in either case, it can be rejected or not rejected. If the null hypothesis is true and we reject it nonetheless, we are making an incorrect decision. This type of error is called a **Type I error**. Conversely, if the null hypothesis is false but we fail to reject it, this incorrect decision is a **Type II error**.

Type I error The probability associated with rejecting a null hypothesis when it is true.

Type II error The probability associated with failing to reject a null hypothesis when it is false.

In Table 7.1, we show the relationship between the two types of errors and the decisions we make regarding the null hypothesis. The probability of a Type I error—rejecting a true hypothesis—is equal to the chosen alpha level. For example, when we set alpha at the .05 level, we know that the probability that the null hypothesis is in fact true is .05 (or 5%).

Table 7.1 Type I and Type II Errors

	True State of Affairs	
Decision Made	**H_0 *Is True***	**H_0 *Is False***
Reject H_0	Type I error (α)	Correct decision
Do not reject H_0	Correct decision	Type II error

We can control the risk of rejecting a true hypothesis by manipulating alpha. For example, by setting alpha at .01, we are reducing the risk of making a Type I error to 1%. Unfortunately, however, Type I and Type II errors are inversely related; thus, by reducing alpha and lowering the risk of making a Type I error, we are increasing the risk of making a Type II error (Table 7.1).

As long as we base our decisions on sample statistics and not population parameters, we have to accept a degree of uncertainty as part of the process of statistical inference.

✓ *Learning Check*

The implications of research findings are not created equal. For example, researchers might hypothesize that eating spinach increases the strength of weight lifters. Little harm will be done if the null hypothesis that eating spinach has no effect on the strength of weight lifters is rejected in error. The researchers would most likely be willing to risk a high probability of a Type I error, and all weight lifters would eat spinach. However, when the implications of research have important consequences, the balancing act between Type I and Type II errors becomes more important. Can you think of some examples where researchers would want to minimize Type I errors? When might they want to minimize Type II errors?

The *t* Statistic and Estimating the Standard Error

The Z statistic we have calculated (Formula 7.1) to test the hypothesis involving a sample of California gas stations assumes that the population standard deviation σ_Y is known. The value of σ_Y is required to calculate the standard error

$$\frac{\sigma_Y}{\sqrt{N}}$$

In most situations, σ_Y will not be known, and we will need to estimate it using the sample standard deviation S_Y. We then use the t statistic instead of the Z statistic to test the null hypothesis. The formula for computing the t statistic is

$$t = \frac{\bar{Y} - \mu_Y}{S_Y/\sqrt{N}} \tag{7.2}$$

The t value we calculate is called the ***t* statistic (obtained)**. The obtained t represents the number of standard deviation units (or standard error units) that our sample mean is from the hypothesized value of μ_Y, assuming that the null hypothesis is true.

t *statistic (obtained)* The test statistic computed to test the null hypothesis about a population mean when the population standard deviation is unknown and is estimated using the sample standard deviation.

The *t* Distribution and Degrees of Freedom

To understand the t statistic, we should first be familiar with its distribution. The ***t* distribution** is actually a family of curves, each determined by its *degrees of freedom*. The concept of degrees of freedom is used in calculating several statistics, including the t statistic. The **degrees of freedom (*df*)** represent the number of scores that are free to vary in calculating each statistic.

t *distribution* A family of curves, each determined by its degrees of freedom *(df)*. It is used when the population standard deviation is unknown and the standard error is estimated from the sample standard deviation.

***Degrees of freedom* (df)** The number of scores that are free to vary in calculating a statistic.

To calculate the degrees of freedom, we must know the sample size and whether there are any restrictions in calculating that statistic. The number of restrictions is then subtracted from the sample size to determine the degrees of freedom. When calculating the *t* statistic for a one-sample test, we start with the sample size *N* and lose 1 degree of freedom for the population standard deviation we estimate.[3] Note that the degrees of freedom will increase as the sample size increases. In the case of a single-sample mean, the *df* is calculated as follows:

$$df = N - 1 \tag{7.3}$$

Comparing the *t* and *Z* Statistics

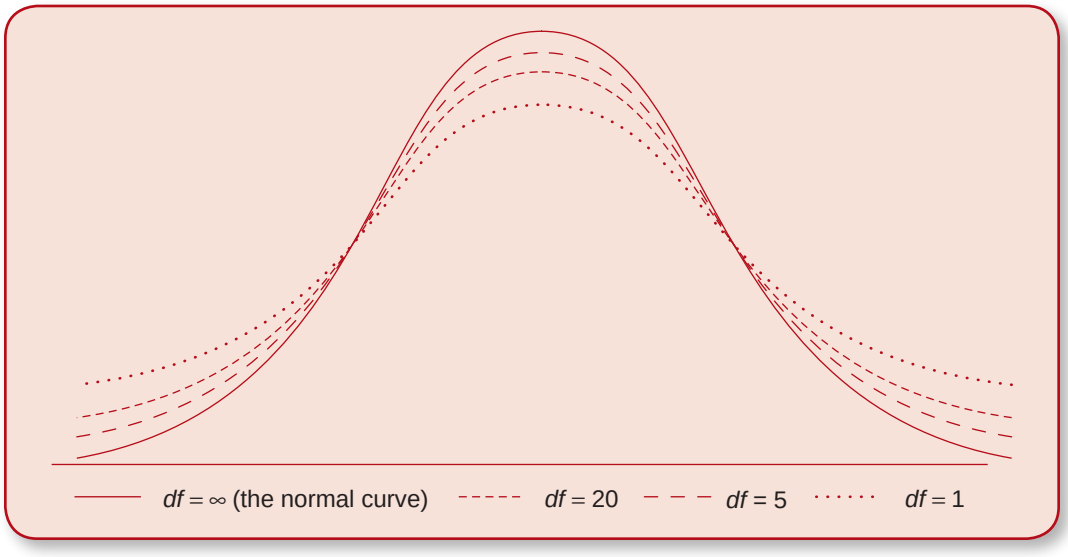
Notice the similarities between the formulas for the *t* and *Z* statistics. The only apparent difference is in the denominator. The denominator of *Z* is the standard error based on the population standard deviation σ_Y. For the denominator of *t*, we replace $\sigma_Y/\sqrt{N}$ with $S_Y/\sqrt{N}$, the estimated standard error based on the sample standard deviation.

However, there is another important difference between the *Z* and *t* statistics*:* Because it is estimated from sample data, the denominator of the *t* statistic is subject to sampling error. The sampling distribution of the test statistic is not normal, and the standard normal distribution cannot be used to determine probabilities associated with it.

In Figure 7.3, we present the *t* distribution for several *df*s. Like the standard normal distribution, the *t* distribution is bell shaped. The *t* statistic, similar to the *Z* statistic, can have positive and negative values. A positive *t* statistic corresponds to the right tail of the distribution; a negative value corresponds to the left tail. Note that when the *df* is small, the *t* distribution is much flatter than the normal curve. But as the degrees of freedom increase, the shape of the *t* distribution gets closer to the normal distribution, until the two are almost identical when *df* is greater than 120.

Appendix B summarizes the *t* distribution. We have reproduced a small part of this appendix in Table 7.2. Note that the *t* table differs from the normal (*Z*) table in several ways. First, the column on the left side of the table shows the degrees of freedom. The *t* statistic will vary depending on the degrees of freedom, which must first be computed ($df = N - 1$). Second, the probabilities or alpha, denoted as significance levels, are arrayed across the top of the table in two rows, the first for a one-tailed and the second for a two-tailed test. Finally, the values of *t*, listed as the entries of this table, are a function of (1) the degrees of freedom, (2) the level of significance (or probability), and (3) whether the test is a one- or a two-tailed test.

Figure 7.3 The Normal Distribution and *t* Distributions for 1, 5, 20, and ∞ Degrees of Freedom



To illustrate the use of this table, let's determine the probability of observing a *t* value of 2.021 with 40 degrees of freedom and a two-tailed test. Locating the proper row ($df = 40$) and column (two-tailed test), we find the *t* statistic of 2.021 corresponding to the .05 level of significance. Restated, we can say that the probability of obtaining a *t* statistic of 2.021 is .05, or that there are less than 5 chances out of 100 that we would have drawn a random sample with an obtained *t* of 2.021 if the null hypothesis were correct.

Statistics in Practice: The Earnings of White Women

We drew a 2002 General Social Survey (GSS) sample ($N = 371$) of white females who worked full-time. We found their mean earnings to be \$28,889, with a standard deviation $S_Y = \$21{,}071$. Based on the Current Population Survey,[4] we also know that the 2002 mean earnings nationally for all women was $\mu_Y = \$24{,}146$. However, we do not know the value of the population standard deviation. We want to determine whether the sample of white women was representative of the population of all full-time women workers in 2002. Although we suspect that white American women experienced a relative advantage in earnings, we are not sure enough to predict that their earnings were indeed higher than the earnings of all women nationally. Therefore, the statistical test is two-tailed.

Let's apply the five-step model to test the hypothesis that the average earnings of white women differed from the average earnings of all women working full-time in the United States in 2002.

Making Assumptions. Our assumptions are as follows:

1. A random sample is selected.
2. Because $N > 50$, the assumption of normal population is not required.
3. The level of measurement of the variable *income* is interval ratio.

Table 7.2 Values of the *t* Distribution

	Level of Significance for One-Tailed Test					
	.10	*.05*	*.025*	*.01*	*.005*	*.0005*
	Level of Significance for Two-Tailed Test					
df	*.20*	*.10*	*.05*	*.02*	*.01*	*.001*
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
10	1.372	1.812	2.228	2.764	3.169	4.587
15	1.341	1.753	2.131	2.602	2.947	4.073
20	1.325	1.725	2.086	2.528	2.845	3.850
25	1.316	1.708	2.060	2.485	2.787	3.725
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
80	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

Source: Abridged from R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research,* Table 111. Copyright © R. A. Fisher and F. Yates, 1963. Published by Pearson Education Limited.

Stating the Research and the Null Hypotheses and Selecting Alpha. The research hypothesis is

$$H_1: \mu_Y \neq \$24{,}146$$

and the null hypothesis is

$$H_0: \mu_Y = \$24{,}146$$

We'll set alpha at .05, meaning that we will reject the null hypothesis if the probability of our obtained statistic is less than or equal to .05.

Selecting the Sampling Distribution and Specifying the Test Statistic. We use the *t* distribution and the *t* statistic to test the null hypothesis.

Computing the Test Statistic. We first calculate the *df* associated with our test:

$$df = (N - 1) = (371 - 1) = 370$$

We need to calculate the obtained *t* statistic by using Formula 7.2:

$$t = \frac{\bar{Y} - \mu_Y}{S_Y/\sqrt{N}} = \frac{28{,}889 - 24{,}146}{21{,}071/\sqrt{371}} = 4.33$$

Making a Decision and Interpreting the Results. Given our research hypothesis, we will conduct a two-tailed test. To determine the probability of observing a *t* value of 4.33 with 370 degrees of freedom, let's refer to Table 7.2. From the first column, we can see that 370 degrees of freedom is not listed, so we'll have to use the last row, $df = \infty$, to locate our obtained *t* statistic.

Though our obtained *t* statistic of 4.33 is not listed in the last row of *t* statistics, in fact, it is greater than the last value listed in the row, 3.291. The *t* statistic of 3.291 corresponds to the .001 level of significance for two-tailed tests. Restated, we can say that our obtained *t* statistic of 4.33 is greater than 3.291 or the probability of obtaining a *t* statistic of 4.33 is less than .001 ($P < .001$). This *P* value is below our .05 alpha level. The probability of obtaining the difference of \$4,743 (\$28,889 − \$24,146) between the income of white women and the national average for all women, if the null hypothesis were true, is extremely low. We have sufficient evidence to reject the null hypothesis and conclude that the average earnings of white women in 2002 were significantly different from the average earnings of all women. The difference of \$4,743 is significant at the .05 level. We can also say that the level of significance is less than .001.

TESTING HYPOTHESES ABOUT TWO SAMPLES

In practice, social scientists are often more interested in situations involving two (sample) parameters than those involving one. For example, we may be interested in finding out whether the average years of education for one racial/ethnic group is the same, lower, or higher than another group.

U.S. data on educational attainment reveal that Asians and Pacific Islanders have more years of education than any other racial/ethnic groups; this includes the percentage of those earning a high school degree or higher or a college degree or higher. Though years of education have steadily increased for blacks and Hispanics since 1990, their numbers remain behind Asians and Pacific Islanders and whites.

Using data from the 2008 GSS, we examine the difference in black and Hispanic educational attainment. From the GSS sample, black respondents reported an average of 12.80 years of education and Hispanics an average of 10.63 years as shown in Table 7.3. These sample averages could mean either (1) the average number of years of education for blacks is higher than the average for Hispanics or (2) the average for blacks is actually about the same as for Hispanics, but our sample just happens to indicate a higher average for blacks. What we are applying here is a bivariate analysis (for more information, refer to Chapter 8), a method to detect and describe the relationship between two variables—race/ethnicity and educational attainment.

The statistical procedures discussed in the following sections allow us to test whether the differences that we observe between two samples are large enough for us to conclude that the populations from which these samples are drawn are different as well. We present tests for the significance of the differences between two groups. Primarily, we consider differences between sample means and differences between sample proportions.

Table 7.3 Years of Education for Black and Hispanic Men and Women, GSS 2008

	Blacks (Sample 1)	*Hispanics (Sample 2)*
Mean	12.80	10.63
Standard deviation	2.55	3.76
Variance	6.50	14.14
N	279	82

The Assumption of Independent Samples

With a two-sample case, we assume that the samples are independent of each other. The choice of sample members from one population has no effect on the choice of sample members from the second population. In our comparison of blacks and Hispanics, we are assuming that the selection of blacks is independent of the selection of Hispanics. (The requirement of independence is also satisfied by selecting one sample randomly, then dividing the sample into appropriate subgroups. For example, we could randomly select a sample and then divide it into groups based on gender, religion, income, or any other attribute that we are interested in.)

Stating the Research and Null Hypotheses

With two-sample tests, we compare two population parameters.

Our research hypothesis (H_1) is that the average years of education for blacks is not equal to the average years of education for Hispanic respondents. We are stating a hypothesis about the relationship between race/ethnicity and education in the general population by comparing the mean educational attainment of blacks with the mean educational attainment of Hispanics. Symbolically, we use μ to represent the population mean; the subscript 1 refers to our first sample (blacks) and subscript 2 to our second sample (Hispanics). Our research hypothesis can then be expressed as

$$H_1: \mu_1 \neq \mu_2$$

Because H_1 specifies that the mean education for blacks is not equal to the mean education for Hispanics, it is a nondirectional hypothesis. Thus, our test will be a two-tailed test. Alternatively, if there were sufficient basis for deciding which population mean score is larger (or smaller), the research hypothesis for our test would be a one-tailed test:

$$H_1: \mu_1 < \mu_2 \text{ or } H_1: \mu_1 > \mu_2$$

In either case, the null hypothesis states that there are no differences between the two population means:

$$H_0: \mu_1 = \mu_2$$

We are interested in finding evidence to reject the null hypothesis of no difference so that we have sufficient support for our research hypothesis.

✓ *Learning Check*

For the following research situations, state your research and null hypotheses:

- *There is a difference between the mean statistics grades of social science majors and the mean statistics grades of business majors.*
- *The average number of children in two-parent black families is lower than the average number of children in two-parent nonblack families.*
- *Grade point averages are higher among girls who participate in organized sports than among girls who do not.*

THE SAMPLING DISTRIBUTION OF THE DIFFERENCE BETWEEN MEANS

The sampling distribution allows us to compare our sample results with all possible sample outcomes and estimate the likelihood of their occurrence. Tests about differences between two sample means are based on the **sampling distribution of the difference between means**. The sampling distribution of the difference between two sample means is a theoretical probability distribution that would be obtained by calculating all the possible mean differences $(\bar{Y}_1 - \bar{Y}_2)$ by drawing all possible independent random samples of size N_1 and N_2 from two populations.

Sampling distribution of the difference between means A theoretical probability distribution that would be obtained by calculating all the possible mean differences $(\bar{Y}_1 - \bar{Y}_2)$ that would be obtained by drawing all the possible independent random samples of size N_1 and N_2 from two populations where N_1 and N_2 are both greater than 50.

The properties of the sampling distribution of the difference between two sample means are determined by a corollary to the central limit theorem. This theorem assumes that our samples are independently drawn from normal populations, but that with sufficient sample size ($N_1 > 50, N_2 > 50$) the sampling distribution of the difference between means will be approximately normal, even if the original populations are not normal. This sampling distribution has a mean $\mu_{\bar{Y}_1} - \mu_{\bar{Y}_2}$ and a standard deviation (standard error)

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{\sigma^2_{Y_1}}{N_1} + \frac{\sigma^2_{Y_2}}{N_2}} \tag{7.4}$$

which is based on the variances in each of the two populations ($\sigma^2_{Y_1}$ and $\sigma^2_{Y_2}$).

Estimating the Standard Error

Formula 7.4 assumes that the population variances are known and that we can calculate the standard error $\sigma_{\bar{Y}_1 - \bar{Y}_2}$ (the standard deviation of the sampling distribution). However, in most situations, the only data we have are based on sample data, and we do not know the true value of the population variances, $\sigma^2_{Y_1}$ and $\sigma^2_{Y_2}$. Thus, we need to estimate the standard error from the sample variances, $S^2_{Y_1}$ and $S^2_{Y_2}$. The estimated standard error of the difference between means is symbolized as $S_{\bar{Y}_1 - \bar{Y}_2}$ (instead of $\sigma_{\bar{Y}_1 - \bar{Y}_2}$).

Calculating the Estimated Standard Error

When we can assume that the two population variances are equal, we combine information from the two sample variances to calculate the estimated standard error.

$$S_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{(N_1 - 1)S^2_{Y_1} + (N_2 - 1)S^2_{Y_2}}{(N_1 + N_2) - 2}} \sqrt{\frac{N_1 + N_2}{N_1 N_2}} \tag{7.5}$$

where $S_{\bar{Y}_1 - \bar{Y}_2}$ is the estimated standard error of the difference between means, and $S^2_{Y_1}$ and $S^2_{Y_2}$ are the variances of the two samples. As a rule of thumb, when either sample variance is more than *twice* as large as the other, we can no longer assume that the two population variances are equal and would need to use Formula 7.8.

The *t* Statistic

As with single sample means, we use the *t* distribution and the *t* statistic whenever we estimate the standard error for a difference between means test. The *t* value we calculate is the obtained *t*. It represents the number of standard deviation units (or standard error units) that our mean difference $(\bar{Y}_1 - \bar{Y}_2)$ is from the hypothesized value of $\mu_1 - \mu_2$, assuming that the null hypothesis is true.

The formula for computing the *t* statistic for a difference between means test is

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{\bar{Y}_1 - \bar{Y}_2}} \tag{7.6}$$

where $S_{\bar{Y}_1 - \bar{Y}_2}$ is the estimated standard error.

Calculating the Degrees of Freedom for a Difference Between Means Test

To use the *t* distribution for testing the difference between two sample means, we need to calculate the degrees of freedom. As we saw earlier, the degrees of freedom (*df*) represent the number of scores that are free to vary in calculating each statistic. When calculating the *t* statistic for the two-sample test, we lose 2 degrees of freedom, one for every population variance we estimate. When population variances are assumed to be equal or if the size of both samples is greater than 50, the *df* is calculated as follows:

$$df = (N_1 + N_2) - 2 \tag{7.7}$$

When we cannot assume that the population variances are equal and when the size of one or both samples is equal to or less than 50, we use Formula 7.9 to calculate the degrees of freedom.

POPULATION VARIANCES ARE ASSUMED TO BE UNEQUAL

If the variances of the two samples ($S^2_{Y_1}$ and $S^2_{Y_2}$) are very different (one variance is twice as large as the other), the formula for the estimated standard error becomes

$$S_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{S^2_{Y_1}}{N_1} + \frac{S^2_{Y_2}}{N_2}} \tag{7.8}$$

When the population variances are unequal and the size of one or both samples is equal to or less than 50, we use another formula to calculate the degrees of freedom associated with the t statistic:[5]

$$df = \frac{\left(S^2_{Y_1}/N_1 + S^2_{Y_2}/N_2\right)^2}{\left(S^2_{Y_1}/N_1\right)^2/(N_1 - 1) + \left(S^2_{Y_2}/N_2\right)^2/(N_2 - 1)} \tag{7.9}$$

THE FIVE STEPS IN HYPOTHESIS TESTING ABOUT DIFFERENCE BETWEEN MEANS: A SUMMARY

As with single-sample tests, statistical hypothesis testing involving two sample means can be organized into five basic steps. Let's summarize these steps:

1. Making assumptions
2. Stating the research and null hypotheses and selecting alpha
3. Selecting the sampling distribution and specifying the test statistic
4. Computing the test statistic
5. Making a decision and interpreting the results

Making Assumptions. In our example, we made the following assumptions:

1. Independent random samples are used.
2. The variable *years of education* is measured at an interval-ratio level of measurement.
3. Because $N_1 > 50$ and $N_2 > 50$, the assumption of normal population is not required.
4. The population variances are assumed to be equal.

Stating the Research and Null Hypotheses and Selecting Alpha. Our research hypothesis is that the mean education of blacks is different from the mean education of Hispanics, indicating a two-tailed test. Symbolically, the research hypothesis is expressed as

$$H_1: \mu_1 \neq \mu_2$$

with μ_1 representing the mean education of blacks and μ_2 the mean education of Hispanics.

The null hypothesis states that there are no differences between the two population means, or

$$H_0: \mu_1 = \mu_2$$

We are interested in finding evidence to reject the null hypothesis of no difference so that we have sufficient support for our research hypothesis. We will reject the null hypothesis if the probability of *t* (obtained) is less than or equal to .05 (our alpha value).

Selecting the Sampling Distribution and Specifying the Test Statistic. The *t* distribution and the *t* statistic are used to test the significance of the difference between the two sample means.

Computing the Test Statistic. To test the null hypothesis about the differences between the mean education of blacks and Hispanics, we need to translate the ratio of the observed differences to its standard error into a *t* statistic (based on data presented in Table 7.3). The obtained *t* statistic is calculated using Formula 7.6:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{\bar{Y}_1 - \bar{Y}_2}}$$

where $S_{\bar{Y}_1 - \bar{Y}_2}$ is the estimated standard error of the sampling distribution. Because the population variances are assumed to be equal, *df* is $(N_1 + N_2) - 2 = (279 + 82) - 2 = 359$ and we can combine information from the two sample variances to estimate the standard error (Formula 7.5):

$$S_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{(279 - 1)2.55^2 + (82 - 1)3.76^2}{(279 + 82) - 2}} \sqrt{\frac{279 + 82}{279(82)}} = (2.87)(0.13) = 0.37$$

We substitute this value into the denominator for the *t* statistic (Formula 7.6):

$$t = \frac{12.80 - 10.63}{0.37} = 5.86$$

Making a Decision and Interpreting the Results. We confirm that our obtained *t* is on the right tail of the distribution. Since our obtained *t* statistic of 5.86 is greater than $t = 3.291$ ($df = \infty$, two-tailed; see Appendix B), we can state that its probability is less than .001. This is less than our .05 alpha level, and we can reject the null hypothesis of no difference between the educational attainment of blacks and Hispanics. We conclude that black men and women, on average, have significantly higher years of education than Hispanic men and women do.

TESTING THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO SAMPLE PROPORTIONS

Numerous variables in the social sciences are measured at a nominal or an ordinal level. These variables are often described in terms of proportions. For example, we might be interested in comparing the proportion of those who support immigrant policy reform among Hispanics and non-Hispanics or the proportion of men and women who supported the Democratic candidate during the last

presidential election. In this section, we present statistical inference techniques to test for significant differences between two sample proportions.

Hypothesis testing with two sample proportions follows the same structure as the statistical tests presented earlier: The assumptions of the test are stated, the research and null hypotheses are formulated, the sampling distribution and the test statistic are specified, the test statistic is calculated, and a decision is made whether or not to reject the null hypothesis.

Statistics in Practice: Political Party Affiliation and Confidence in the Executive Branch

Pollsters often collect data to measure public opinions on current social and political issues. Differences are often categorized by income, race/ethnicity, gender, region, or political groups. For example, do Republicans and Democrats have the same confidence in the executive branch of government? We can use data from the 2008 GSS to test the null hypothesis that the proportion of Republicans and Democrats who have a "great deal" of confidence in the executive branch of the government is equal. The proportion of Republicans who reported a great deal of confidence was 0.21 (p_1); the proportion of Democrats with the same response was lower at 0.06 (p_2). A total of 141 Republicans (N_1) and 267 Democratic respondents (N_2) answered this question.

Making Assumptions. Our assumptions are as follows:

1. Independent random samples of $N_1 > 50$ and $N_2 > 50$ are used.
2. The level of measurement of the variable is nominal.

Stating the Research and Null Hypotheses and Selecting Alpha. We propose a two-tailed test that the population proportions for Republicans and Democrats are not equal.

$$H_1: \pi_1 \neq \pi_2$$
$$H_0: \pi_1 = \pi_2$$

We decide to set alpha at .05.

Selecting the Sampling Distribution and Specifying the Test Statistic. The population distributions of dichotomies are not normal. However, based on the central limit theorem, we know that the sampling distribution of the difference between sample proportions is normally distributed when the sample size is large (when $N_1 > 50$ and $N_2 > 50$), with mean $\mu_{p_1 - p_2}$ and the estimated standard error $S_{p_1 - p_2}$. Therefore, we can use the normal distribution as the sampling distribution and we can calculate Z as the test statistic.[6]

The formula for computing the Z statistic for a difference between proportions test is

$$Z = \frac{p_1 - p_2}{S_{p_1 - p_2}} \tag{7.10}$$

where p_1 and p_2 are the sample proportions for Republicans and Democrats, and $S_{p_1 - p_2}$ is the estimated standard error of the sampling distribution of the difference between sample proportions.

The estimated standard error is calculated using the following formula:

$$S_{p_1-p_2} = \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}} \tag{7.11}$$

Calculating the Test Statistic. We calculate the standard error using Formula 7.11:

$$S_{p_1-p_2} = \sqrt{\frac{0.21(1-0.21)}{141} + \frac{0.06(1-0.06)}{267}} = .037 = 0.04$$

Substituting this value into the denominator of Formula 7.10, we get

$$Z = \frac{0.21 - 0.06}{0.04} = 3.75$$

Making a Decision and Interpreting the Results. Our obtained Z of 3.75 indicates that the difference between the two proportions will be evaluated at the right tail of the Z distribution. To determine the probability of observing a Z value of 3.75 if the null hypothesis is true, look up the value in Appendix A (Column C) to find the area to the right of (above) the obtained Z.

The P value corresponding to a Z score of 3.75 is .0001. However, for a two-tailed test, we'll have to multiply P by 2 ($.0001 \times 2 = .0002$). The probability of 3.75 for a two-tailed test is less than our alpha level of .05 ($.0002 < .05$).

Thus, we reject the null hypothesis of no difference and conclude that there is a significant political party difference in confidence in the executive branch of the government. Republicans are more likely than Democrats to report a great deal of confidence in the executive branch.

IS THERE A SIGNIFICANT DIFFERENCE?

The news media made note of a 2010 Centers for Disease Control (CDC) study that examined the difference in length of marriage between couples in 2005 who first cohabited before marriage and couples who did not cohabit before marriage. Several news services released stories noting the "troubles" associated with living together. As reported, the percentage of marriages surviving to the 10th anniversary, among those who cohabited before marriage, was lower than those who did not cohabit before their first marriage. A closer look at the report reveals important (overlooked) details.

CDC researchers Goodwin, Mosher, and Chandra (2010) reported that previous cohabitation experience was significantly associated with marriage survival probabilities for men. On the other hand, though the probability that a woman's marriage would last at least 10 years was lower for those who cohabited before marriage (60%) than for women who did not (66%), the researchers wrote, "however, in the 2002 data, the difference was not significant at the 5% level" (p. 7).[7]

Throughout this chapter, we've assessed the difference between two means and two proportions, attempting to determine whether the difference between them is due to real effects in the population or due to sampling error. A significant difference is one that confirms that effects of the independent variable, such as cohabiting before marriage, are real. As in the case of marriage survival rate, cohabitation before marriage makes a significant difference in marital outcomes for men, but not for women in the CDC

sample. Take caution in accepting comparative statements that fail to mention significance. There may be a difference, but you have to ask, is it a significant difference?

READING THE RESEARCH LITERATURE: REPORTING THE RESULTS OF STATISTICAL HYPOTHESIS TESTING

Robert Emmet Jones and Shirley A. Rainey (2006) examined the relationship between race, environmental attitudes, and perceptions about environmental health and justice.[8] Researchers have documented how people of color and the poor are more likely than whites and more affluent groups to live in areas with poor environmental quality and protection, exposing them to greater health risks. Yet little is known about how this disproportional exposure and risk are perceived by those affected. Jones and Rainey studied black and white residents from the Red River community in Tennessee, collecting data from interviews and a mail survey during 2001 to 2003.

They created a series of index scales measuring residents' attitudes pertaining to environmental problems and issues. The Environmental Concern (EC) Index measures public concern for specific environmental problems in the neighborhood. It includes questions on drinking water quality, landfills, loss of trees, lead paint and poisoning, the condition of green areas, and stream and river conditions. EC-II measures public concern (very unconcerned to very concerned) for the overall environmental quality in the neighborhood. EC-III measures the seriousness (not serious at all to very serious) of environmental problems in the neighborhood. Higher scores on all EC indicators indicate greater concern for environmental problems in their neighborhood. The Environmental Health (EH) Index measures public perceptions of certain physical side effects, such as headaches, nervous disorders, significant weight loss or gain, skin rashes, and breathing problems. The EH Index measures the likelihood (very unlikely to very likely) that the person believes that he or she or a household member experienced health problems due to exposure to environmental contaminants in his or her neighborhood. Higher EH scores reflect a greater likelihood that respondents believe that they have experienced health problems from exposure to environmental contaminants. Finally, the Environmental Justice (EJ) Index measures public perceptions about environmental justice, measuring the extent to which they agreed (or disagreed) that public officials had informed residents about environmental problems, enforced environmental laws, or held meetings to address residents' concerns. A higher mean EJ score indicates a greater likelihood that respondents think public officials failed to deal with environmental problems in their neighborhood. Index score comparisons between black and white respondents are presented in Table 7.4.

Let's examine the table carefully. Each row represents a single index measurement, reporting means and standard deviations separately for black and white residents. Obtained *t*-test statistics are reported in the second to last column. The probability of each *t* test is reported in the last column ($P < .001$), indicating a significant difference in responses between the two groups. All index score comparisons are significant at the .001 level.

While not referring to specific differences in index scores or to *t*-test results, Jones and Rainey use data from this table to summarize the differences between black and white residents on the three environmental index measurements:

> The results presented [in Table 1] suggest that as a group, Blacks are significantly more concerned than Whites about local environmental conditions (EC Index). . . . The results . . . also indicate that

Table 7.4 Environmental Concern (EC), Environmental Health (EH), and Environmental Justice (EJ)

Indicator	*Group*	*Mean*	*Standard Deviation*	t	*Significance (one-tailed)*
EC Index	Blacks	56.2	13.7	6.2	<0.001
	Whites	42.6	15.5		
EC-II	Blacks	4.4	1.0	5.6	<0.001
	Whites	3.5	1.3		
EC-III	Blacks	3.4	1.1	6.7	<0.001
	Whites	2.3	1.0		
EH Index	Blacks	23.0	10.5	5.1	<0.001
	Whites	16.0	7.3		
EJ Index	Blacks	31.0	7.3	3.8	<0.001
	Whites	27.2	6.3		

Source: Robert E. Jones and Shirley A. Rainey, "Examining Linkages Between Race, Environmental Concern, Health and Justice in a Highly Polluted Community of Color," *Journal of Black Studies* 36, no. 4 (2006): 473–496.

Note: $N = 78$ blacks, 113 whites.

> as a group, Blacks believe they have suffered more health problems from exposure to poor environmental conditions in their neighborhood than Whites (EH Index). . . . [T]here is greater likelihood that Blacks feel local public agencies and officials failed to deal with environmental problems in their neighborhood in a fair, just, and effective manner (EJ Index). (p. 485)

MAIN POINTS

- Statistical hypothesis testing is a decision-making process that enables us to determine whether a particular sample result falls within a range that can occur by an acceptable level of chance. The process of statistical hypothesis testing consists of five steps: (1) making assumptions, (2) stating the research and null hypotheses and selecting alpha, (3) selecting a sampling distribution and a test statistic, (4) computing the test statistic, and (5) making a decision and interpreting the results.

- Statistical hypothesis testing may involve a comparison between a sample mean and a population mean or a comparison between two sample means. If we know the population variance(s) when testing for differences between means, we can use the Z statistic and the normal distribution. However, in practice, we are unlikely to have this information.

- When testing for differences between means when the population variance(s) are unknown, we use the t statistic and the t distribution.

- Tests involving differences between proportions follow the same procedure as tests for differences between means when population variances are known. The test statistic is Z, and the sampling distribution is approximated by the normal distribution.

KEY TERMS

alpha (α)
degrees of freedom (*df*)
left-tailed test
null hypothesis (H_0)
one-tailed test
P value
research hypothesis (H_1)
right-tailed test
sampling distribution of the difference between means
statistical hypothesis testing
t distribution
t statistic (obtained)
two-tailed test
Type I error
Type II error
Z statistic (obtained)

ON YOUR OWN

Log on to the web-based student study site at **www.sagepub.com/ssdsessentials** for additional study questions, web quizzes, web resources, flashcards, codebooks and datasets, web exercises, appendices, and links to social science journal articles reflecting the statistics used in this chapter.

CHAPTER EXERCISES

1. It is known that, nationally, doctors working for health maintenance organizations (HMOs) average 13.5 years of experience in their specialties, with a standard deviation of 7.6 years. The executive director of an HMO in a western state is interested in determining whether or not its doctors have less experience than the national average. A random sample of 150 doctors from HMOs shows a mean of only 10.9 years of experience.
   a. State the research and the null hypotheses to test whether or not doctors in this HMO have less experience than the national average.
   b. Using an alpha level of .01, make this test.

2. Consider the problem facing security personnel at a military facility in the Southwest. Their job is to detect infiltrators (spies trying to break in). The facility has an alarm system to assist the security officers. However, sometimes the alarm doesn't work properly, and sometimes the officers don't notice a real alarm. In general, the security personnel must decide between these two alternatives at any given time:

   H_0: Everything is fine; no one is attempting an illegal entry.

   H_1: There are problems; someone is trying to break into the facility.

   Based on this information, fill in the blanks in these statements:
   a. A "missed alarm" is a Type ____ error, and its probability of occurrence is denoted as ____.
   b. A "false alarm" is a Type ____ error.

3. For each of the following situations determine whether a one- or a two-tailed test is appropriate. Also, state the research and the null hypotheses.
   a. You are interested in finding out if the average household income of residents in your state is different from the national average household. According to the U.S. Census, for 2010, the national average household income is $50,303.
   b. You believe that students in small liberal arts colleges attend more parties per month than students nationwide. It is known that, nationally, undergraduate students attend an average of 3.2 parties per month. The average number of parties per month will be calculated from a random sample of students from small liberal arts colleges.
   c. A sociologist believes that the average income of elderly women is lower than the average income of elderly men.

d. Is there a difference in the amount of study time on-campus and off-campus students devote to their schoolwork during an average week? You prepare a survey to determine the average number of study hours for each group of students.
e. Reading scores for a group of third graders enrolled in an accelerated reading program are predicted to be higher than the scores for nonenrolled third graders.
f. Stress (measured on an ordinal scale) is predicted to be lower for adults who own dogs (or other pets) than for non–pet owners.

4. a. For each situation in Exercise 3, describe the Type I and Type II errors that could occur.
   b. What are the general implications of making a Type I error? Of making a Type II error?
   c. When would you want to minimize Type I error? Type II error?

5. One way to check on how representative a survey is of the population from which it was drawn is to compare various characteristics of the sample with the population characteristics. A typical variable used for this purpose is age. The 2008 GSS of the American adult population found a mean age of 47.71 and a standard deviation of 17.35 for its sample of 2,013 adults. Assume that we know from census data that the mean age of all American adults is 37.7. Use this information to answer these questions.
   a. State the research and the null hypotheses for a two-tailed test.
   b. Calculate the *t* statistic and test the null hypothesis at the .001 significance level. What did you find?
   c. What is your decision about the null hypothesis? What does this tell us about how representative the sample is of the American adult population?

6. Using data on average school grade for first-time cigarette use from MTF 2008, use the *t* test to conduct a one-tailed test of the null hypothesis, assuming that the average grade is higher for males than for females. Set alpha at .05. What can you conclude? Would your conclusions have been different if you had used a two-tailed test?

	Males	***Females***
Grade first tried cigarettes	Mean = 4.90 $S_Y = 1.74$ $N = 150$	Mean = 4.76 $S_Y = 1.73$ $N = 169$

Source: Monitoring the Future, 2008.

Note: Only valid responses are included in the table. Students who indicated that they have never used the drug are not included.

7. In this exercise, we will examine the attitudes of liberals and conservatives toward affirmative action policies in the workplace. Data from the 2008 GSS reveal that 10% of conservatives ($N = 424$) and 28% of liberals ($N = 336$) indicate that they "strongly support" or "support" affirmative action policies for African Americans in the workplace.
   a. What is the appropriate test statistic? Why?
   b. Test the null hypothesis with a one-tailed test (conservatives are less likely to support affirmative action policies than liberals); $\alpha = .05$. What do you conclude about the difference in attitudes between conservatives and liberals?
   c. If you conducted a two-tailed test with $\alpha = .05$, would your decision have been different?

8. Let's continue our analysis of liberals and conservatives, taking a look this time at differences in their educational attainment. We obtain the following information from the 2008 GSS—the average educational attainment for liberals is 13.90 years ($S_Y = 3.27$) and the average educational attainment for conservatives is 13.55 years ($S_Y = 2.82$). Data are based on 187 liberals and 227 conservative responses.
   a. Test the research hypothesis that there is a difference in level of education between liberals and conservatives; set alpha at .01.
   b. Would your decision have been different if alpha were set at .05?

9. During the 2008 Democratic presidential campaign, gender was considered more of an issue for Hillary Clinton's campaign than for her opponent, Barack Obama. During the campaign, pollsters consistently reported how Clinton's supporters were mostly (older) women, with men less likely to support her candidacy. In surveys conducted during December 2007 and January 2008, the Pew Research Center reported that among 240 men, 41% indicated that Senator Clinton was their first-choice candidate. Among 381 women, 49% reported the same. Do these differences reflect a gender gap among Clinton supporters?
   a. If you wanted to test the research hypothesis that the proportion of male voters identifying Senator Clinton as their first-choice candidate is less than female voters, would you conduct a one- or a two-tailed test?
   b. Test the null hypothesis at the .05 alpha level. What do you conclude?
   c. If alpha were changed to .01, would your decision remain the same?

10. Data from the MTF 2008 survey reveal that 75.7% (493 out of 651) of males and 70.4% (501 out of 712) of females reported trying alcohol. You wonder whether there is any difference between males and females in the population trying alcohol (this variable does not measure regular use, only if the student had ever tried alcohol). Use a test of the difference between proportions when answering these questions.
   a. What is the research hypothesis? Should you conduct a one- or a two-tailed test? Why?
   b. Test your hypothesis at the .05 level. What do you conclude?

11. Marcelline Fusilier, Subhash Durlabhji, Alain Cucchi, and Michael Collins (2005) examined Internet usage among college students in the United States and in India.[9] Usage was divided into two categories, for personal use and course-related use. We compare average hours between the U.S. and Indian students in the following table.

Internet Use Means and Standard Deviations for U.S. and Indian Students Sample

	Course Work Hours	*Personal Hours*
U.S. students	$\bar{Y} = 1.76$	$\bar{Y} = 2.08$
$N = 149$	$S_Y = 1.52$	$S_Y = 1.91$
Indian students	$\bar{Y} = 0.73$	$\bar{Y} = 0.87$
$N = 306$	$S_Y = 0.79$	$S_Y = 0.78$

Source: Marcelline Fusilier, Subhash Durlabhji, Alain Cucchi, and Michael Collins, "A four country investigation of factors facilitating student Internet use," *CyberPsychology and Behavior* 8(5): 454–464. Copyright © Mary Ann Liebert, Inc. Publishers.

Exercises

a. Determine whether U.S. students have significantly higher Internet use for course work than the Indian students. Test at the .05 alpha level.
b. Test whether there is a significant difference in Internet use for personal use between the U.S. and Indian students. Test at the .01 alpha level.

12. Do men and women have different beliefs on the ideal number of children in a family? Based on the following GSS 2008 data and obtained *t* statistic, what would you conclude? (*Hint:* Assume a two-tailed test; $\alpha = .05$.)

	Men	*Women*
Mean ideal number of children	3.06	3.22
Standard deviation	1.92	1.99
N	604	678
Obtained *t* statistic	−1.450	

13. Does access to cocaine vary by sex? Data from the MTF 2008 survey indicate that among male students ($N = 670$) 18% say that it would be "very easy" to obtain cocaine compared with 14% of female students ($N = 723$). Is there a significant difference in proportion of easy cocaine access between the two student groups? Set alpha at .01. What can you conclude?

14. We recalculated our comparison of ideal number of children, this time only for women, separating them into two groups, those who indicated that they were "very happy" or "not too happy" in general. Results, based on the GSS 2008, are presented below.

	Very Happy	*Not Too Happy*
Mean ideal number of children	3.49	2.73
Standard deviation	2.24	1.35
N	228	90
Obtained *t* statistic	2.98	

Based on a two-tailed test, $\alpha = .05$. What do you conclude?