openstax™

Statistics

HIGH SCHOOL

8 | CONFIDENCE INTERVALS

Figure 8.1 Have you ever wondered what the average number of chocolate candies in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: comedy_nose/flickr)

Introduction

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Calculate and interpret confidence intervals for estimating a population mean and a population proportion
- Interpret the Student's *t* probability distribution as the sample size changes
- Discriminate between problems applying the normal and the Student's *t*-distributions
- Calculate the sample size required to estimate a population mean and a population proportion, given a desired confidence level and margin of error

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempt. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. *The sample data help us to make an estimate of a population* **parameter**. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called *confidence intervals*.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-*t*, and how it is used with those intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from an internet music store. If so, you could conduct a survey and calculate the sample mean, $\bar{x}$, and the sample standard deviation, *s*. You would use $\bar{x}$ to estimate the population mean and *s* to estimate the population standard deviation. The sample mean, $\bar{x}$, is the **point estimate** for the population mean, μ. The sample standard deviation, *s*, is the point estimate for the population standard deviation, σ.

Each instance of $\bar{x}$ and *s* is called a *statistic*.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose, for the internet music example, we do not know the population mean, μ, but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1.$$

The **empirical rule**, which applies to bell-shaped distributions, says that in approximately 95 percent of the samples, the sample mean, $\bar{x}$, will be within two standard deviations of the population mean, μ. For our internet music example, two standard deviations would be calculated as (2)(0.1) = 0.2. The sample mean, $\bar{x}$, is likely to be within 0.2 units of μ.

In this example, we do not know the true population mean μ (because we do not have information from all the internet music users!), but we can compute the sample mean $\bar{x}$ based on our sample of 100 individuals. Because the sample mean is likely to be within 0.2 units of the true population mean 95 percent of the times that we take a sample of 100 users, we can say with 95 percent confidence that μ is within 0.2 units of $\bar{x}$. In other words, μ is somewhere between $\bar{x} - 0.2$ and $\bar{x} + 0.2$.

Suppose that from the sample of 100 internet music customers, we compute a sample mean download of $\bar{x} = 2$ songs per month. Since we know that the population standard deviation is $\sigma - 1$, according to the central limit theorem, the standard deviation for the sample means is $\sigma = \frac{1}{\sqrt{100}} = 0.1$.

We know that there is a 95 percent chance that the true population mean value μ is between two standard deviations from the sample mean. That is, with 95 percent confidence we can say that μ is between $\bar{x} - 2\times\frac{\sigma}{\sqrt{n}}$ and $\bar{x} - 2\times\frac{\sigma}{\sqrt{n}}$.

Replacing the symbols for their values in this example, we say that we are **95 percent confident** that the true average number of songs downloaded from an internet music store per month is between $\bar{x} - 2\times\frac{\sigma}{\sqrt{n}} = 2 - 2\times\frac{\sigma}{\sqrt{100}} = 2 - .02 = 1.8$, and

$$\bar{x} + 2\times\frac{\sigma}{\sqrt{n}} = 2 + 2\times\frac{\sigma}{\sqrt{100}} = 2 + .02 = 2.2.$$

The 95 percent confidence interval for μ is (1.8, 2.2).

The 95 percent confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean, μ, or our sample produced an $\bar{x}$ that is not within 0.2 units of the true mean μ. The second possibility happens for only 5 percent of all the samples (95–100 percent).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ.

Confidence intervals for some parameters have the form

(point estimate – margin of error, point estimate + margin of error).

The margin of error depends on the confidence level or percentage of confidence and the standard error of the mean.

When you read newspapers and journals, you might notice that some reports use the phrase *margin of error*. Other reports will not use that phrase, but include a confidence interval as the point estimate plus or minus the margin of error. Those are two ways of expressing the same concept.

NOTE

Although the text covers only symmetrical confidence intervals, there are non-symmetrical confidence intervals (for example, a confidence interval for the standard deviation).

Collaborative Exercise

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be three meals. Construct an approximate 95 percent confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. Let $\sigma = 3$ and n = the number of students surveyed.
3. Construct the interval. $\left(\bar{x} - 2\left(\frac{\sigma}{\sqrt{n}}\right)\right), \left(\bar{x} + 2\left(\frac{\sigma}{\sqrt{n}}\right)\right)$

We say we are approximately 95 percent confident that the true mean number of meals that students eat out in a week is between _____ and _____.

8.1 | A Single Population Mean Using the Normal Distribution

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$ and we have constructed the 90 percent confidence interval (5, 15), where the margin of error = 5.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean, μ, *where the population standard deviation is known*, we need $\bar{x}$ as an estimate for μ, and we need the margin of error. Here, the margin of error is called the **error bound for a population mean** (*EBM*) is called the **margin of error for a population mean** (***EBM***). The sample mean, $\bar{x}$, is the **point estimate** of the unknown population mean, μ.

The confidence interval (*CI*) estimate will have the form:

(point estimate – error bound, point estimate + error bound) or, in symbols, ($\bar{x} - EBM$, $\bar{x} + EBM$).

The margin of error (*EBM*) depends on the **confidence level** (***CL***). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percentage of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, the person constructing the confidence interval will choose a confidence level of 90 percent or higher, because that person wants to be reasonably certain of his or her conclusions.

Another probability, which is called alpha (α) is related to the confidence level, *CL*. Alpha is the probability that the confidence interval does not contain the unknown population parameter. Mathematically, alpha can be computed as

$\alpha = 1 - CL$.

Example 8.1

Suppose we have collected data from a sample. We know the sample mean, but we do not know the mean for the entire population.

The sample mean is seven, and the error bound for the mean is 2.5.

$$\bar{x} \text{ and } EBM = 2.5.$$

The confidence interval is (7 – 2.5, 7 + 2.5), and calculating the values gives (4.5, 9.5).

If the confidence level is 95 percent, then we say, "We estimate with 95 percent confidence that the true value of the population mean is between 4.5 and 9.5."

Try It

8.1 Suppose we have data from a sample. The sample mean is 15, and the error bound for the mean is 3.2.

What is the confidence interval estimate for the population mean?

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $\bar{x} = 10$, and we have constructed the 90 percent confidence interval (5, 15) where $EBM = 5$.

To get a 90 percent confidence interval, we must include the central 90 percent of the probability of the normal distribution. If we include the central 90 percent, we leave out a total of $\alpha = 10$ percent in both tails, or 5 percent in each tail, of the normal distribution.

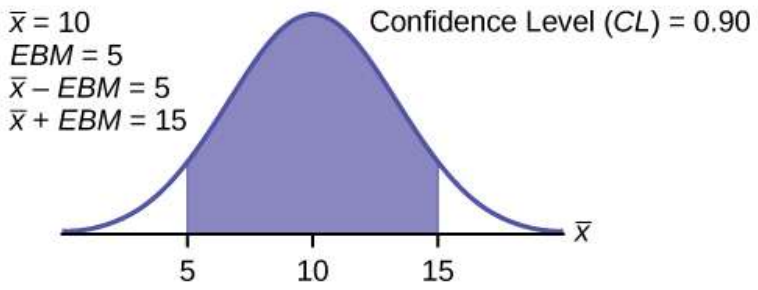


Figure 8.2

The critical value *1.645* is the **z**-*score* in a standard normal probability distribution that puts an area of *0.90* in the center, an area of *0.05* in the far left tail, and an area of *0.05* in the far right tail. To capture the central 90 percent, we must go out *1.645 standard deviations* on either side of the calculated sample mean. The critical value will change depending on the confidence level of the interval.

It is important that the *standard deviation* used be appropriate for the parameter we are estimating, so in this section, we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. The fraction $\frac{\sigma}{\sqrt{n}}$ is commonly called the *standard error of the mean* in order to distinguish clearly the standard deviation for a mean from the population standard deviation, σ.

In summary, as a result of the central limit theorem, the following statements apply:

- $\bar{X}$ is normally distributed, that is, $\bar{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$.

- **When the population standard deviation σ is known, we use a normal distribution to calculate the error bound.**

Calculating the Confidence Interval

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are as follows:

- Calculate the sample mean, $\bar{x}$, from the sample data. Remember, in this section, we already know the population standard deviation, σ.
- Find the z-score that corresponds to the confidence level.
- Calculate the error bound EBM.
- Construct the confidence interval.
- If we denote the critical z-score by $z_{\frac{\alpha}{2}}$, and the sample size by n, then the formula for the confidence interval with confidence level $Cl = 1 - \alpha$, is given by $(\bar{x} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}})$.
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail and then illustrate the process with some examples.

Finding the *z*-Score for the Stated Confidence Level

When we know the population standard deviation, σ, we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0, 1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$, so α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$.

For example, when $CL = 0.95$, $\alpha = 0.05$, and $\frac{\alpha}{2} = 0.025$, we write $z_{\frac{\alpha}{2}} = z_{0.025}$.

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$.

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a calculator, computer, or standard normal probability table.

Normal table (see appendices) shows that the probability for 0 to 1.96 is 0.47500, and so the probability to the right tail of the critical value 1.96 is $0.5 - 0.475 = 0.025$

Using the TI-83, 83+, 84, 84+ Calculator

`invNorm`(0.975, 0, 1) = 1.96. In this command, the value 0.975 is the total area to the left of the critical value that we are looking to calculate. The parameters 0 and 1 are the mean value and the standard deviation of the standard normal distribution Z.

NOTE

Remember to use the area to the LEFT of $z_{\frac{\alpha}{2}}$. In this chapter, the last two inputs in the invNorm command are 0, 1, because you are using a standard normal distribution Z with mean 0 and standard deviation 1.

Calculating the Margin of Error *EBM*

The error bound formula for an unknown population mean, μ, when the population standard deviation, σ, is known is

Margin of error = $\left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$.

Constructing the Confidence Interval

The confidence interval estimate has the format sample mean plus or minus the margin of error.

The graph gives a picture of the entire situation

$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1$.

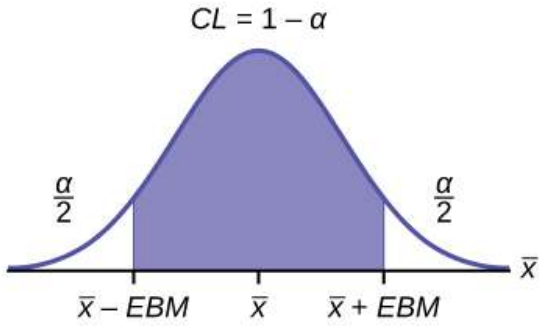


Figure 8.3

Writing the Interpretation

The interpretation should clearly state the confidence level (*CL*), explain which population parameter is being estimated (here, a **population mean**), and state the confidence interval (both endpoints): "We estimate with ___percent confidence that the true population mean (include the context of the problem) is between ___ and ___ (include appropriate units)."

Example 8.2

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of three points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Find a 90 percent confidence interval for the true (population) mean of statistics exam scores.

Solution 8.2

- You can use technology to calculate the confidence interval directly.
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+, and 84+ calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, $\bar{x}$, and the *EBM*.

- $\bar{x} = 68$

$$EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\sigma = 3;\ n = 36;$$

- The confidence level is 90 percent ($CL = 0.90$).

$$CL = 0.90, \text{ so } \alpha = 1 - CL = 1 - 0.90 = 0.10.$$

$$\frac{\alpha}{2} = 0.05,\ z_{\frac{\alpha}{2}} = z_{0.05}$$

The area to the right of $z_{0.05}$ is 0.05 and the area to the left of $z_{0.05}$ is 1 – 0.05 = 0.95.

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

using invNorm(0.95, 0, 1) on the TI-83,83+, and 84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.

$EBM = (1.645)\left(\frac{3}{\sqrt{36}}\right) = 0.8225$

$\bar{x} - EBM = 68 - 0.8225 = 67.1775$

$\bar{x} + EBM = 68 + 0.8225 = 68.8225$

The 90 percent confidence interval is **(67.1775, 68.8225).**

Solution 8.2

Solution B

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS.
Arrow down to 7:ZInterval.
Press ENTER.
Arrow to Stats and press ENTER.
Arrow down and enter 3 for σ, 68 for $\bar{x}$, 36 for n, and .90 for C-level.
Arrow down to Calculate and press ENTER.
The confidence interval is (to three decimal places)(67.178, 68.822).

Interpretation

We estimate with 90 percent confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90 percent Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Try It

8.2 Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of 6 minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 min.

Find a 90 percent confidence interval estimate for the population mean delivery time.

Example 8.3

The specific absorption rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. For certification from the Federal Communications Commission for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. **Table 8.1** shows the highest SAR level for a random selection of cell phone models of a random cell phone company.

Phone Model #	SAR	Phone Model #	SAR	Phone Model #	SAR
800	1.11	1800	1.36	2800	0.74
900	1.48	1900	1.34	2900	0.5
1000	1.43	2000	1.18	3000	0.4
1100	1.3	2100	1.3	3100	0.867
1200	1.09	2200	1.26	3200	0.68
1300	0.455	2300	1.29	3300	0.51
1400	1.41	2400	0.36	3400	1.13
1500	0.82	2500	0.52	3500	0.3
1600	0.78	2600	1.6	3600	1.48
1700	1.25	2700	1.39	3700	1.38

Table 8.1

Find a 98 percent confidence interval for the true (population) mean of the SARs for cell phones. Assume that the population standard deviation is $\sigma = 0.337$.

Solution 8.3

Solution A

To find the confidence interval, start by finding the point estimate: the sample mean,

$$\bar{x} = 1.024.$$

This is calculated by adding the specific absorption rate for the 30 cell phones in the sample, and dividing the result by 30.

Next, find the *EBM*. Because you are creating a 98 percent confidence interval, $CL = 0.98$.

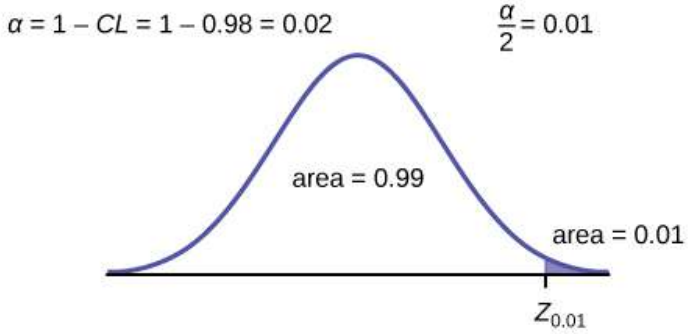


Figure 8.4

You need to find $z_{0.01}$, having the property that the area under the normal density curve to the right of $z_{0.01}$ is 0.01 and the area to the left is 0.99. Use your calculator, a computer, or a probability table for the standard normal distribution to find $z_{0.01} = 2.326$.

$$EBM = (z_{0.01})\frac{\sigma}{\sqrt{n}} = (2.326)\frac{0.337}{\sqrt{30}} = 0.1431$$

To find the 98 percent confidence interval, find $\bar{x} \pm EBM$.

$$\bar{x} - EBM = 1.024 - 0.1431 = 0.8809$$
$$\bar{x} + EBM = 1.024 + 0.1431 = 1.1671$$

We estimate with 98 percent confidence that the true SAR mean for the population of cell phones in the United States is between 0.8809 and 1.1671 watts per kilogram.

Solution 8.3

Solution B

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS.
Arrow down to 7:ZInterval.
Press ENTER.
Arrow to Stats and press ENTER.
Arrow down and enter the following values:
σ: 0.337
$\bar{x}$: 1.024
n: 30
C-level: 0.98
Arrow down to Calculate and press ENTER.
The confidence interval is (to three decimal places) (0.881, 1.167).

Try It

8.3 **Table 8.2** shows a different random sampling of 20 cell phone models. Use these data to calculate a 93 percent confidence interval for the true mean SAR for cell phones certified for use in the United States. As previously, assume that the population standard deviation is $\sigma = 0.337$.

Phone Model	SAR	Phone Model	SAR
450	1.48	1450	1.53
550	0.8	1550	0.68
650	1.15	1650	1.4
750	1.36	1750	1.24
850	0.77	1850	0.57
950	0.462	1950	0.2
1050	1.36	2050	0.51
1150	1.39	2150	0.3
1250	1.3	2250	0.73
1350	0.7	2350	0.869

Table 8.2

Notice the difference in the confidence intervals calculated in **Example 8.3** and the following **Try It** exercise. These intervals are different for several reasons: they are calculated from different samples, the samples are different sizes, and the intervals are calculated for different levels of confidence. Even though the intervals are different, they do not yield conflicting information. The effects of these kinds of changes are the subject of the next section in this chapter.

Changing the Confidence Level or Sample Size

Example 8.4

Suppose we change the original problem in **Example 8.2** by using a 95 percent confidence level. Find a 95 percent confidence interval for the true (population) mean statistics exam score.

Solution 8.4

To find the confidence interval, you need the sample mean, $\bar{x}$, and the *EBM*.

- $\bar{x} = 68$

$$EBM = (z_{\frac{\alpha}{2}})(\frac{\sigma}{\sqrt{n}})$$

$$\sigma = 3;\ n = 36$$

- The confidence level is 95 percent ($CL = 0.95$).

$$CL = 0.95, \text{ so } \alpha = 1 - CL = 1 - 0.95 = 0.05.$$

$$\frac{\alpha}{2} = 0.025 \quad z_{\frac{\alpha}{2}} = z_{0.025}$$

The area to the right of $z_{0.025}$ is 0.025, and the area to the left of $z_{0.025}$ is 1 – 0.025 = 0.975.

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96,$$

when using invnorm(0.975,0,1) on the TI-83, 83+, or 84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the standard normal distribution.)

$$EBM = (1.96)\left(\frac{3}{\sqrt{36}}\right) = 0.98$$

$$\bar{x} - EBM = 68 - 0.98 = 67.02$$

$$\bar{x} + EBM = 68 + 0.98 = 68.98$$

Notice that the *EBM* is larger for a 95 percent confidence level in the original problem.

Interpretation

We estimate with 95 percent confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95 percent Confidence Level

95 percent of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the Results

The 90 percent confidence interval is (67.18, 68.82). The 95 percent confidence interval is (67.02, 68.98). The 95 percent confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95 percent confidence interval is wider. For more certainty that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider.

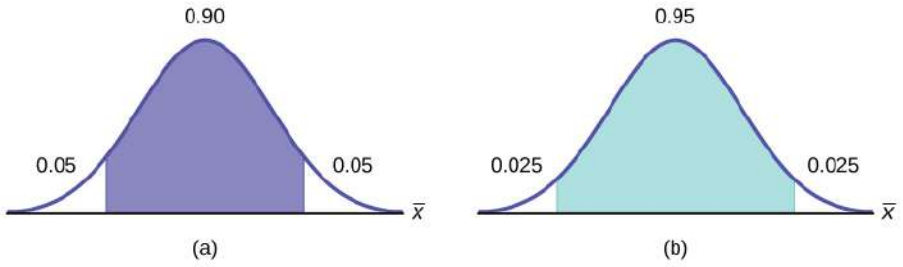


Figure 8.5

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Try It

8.4 Refer back to the pizza-delivery **Try It** exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95 percent confidence interval estimate for the true mean pizza-delivery time.

Example 8.5

Suppose we change the original problem in **Example 8.2** to see what happens to the error bound if the sample size is changed.

Leave everything the same except the sample size. Use the original 90 percent confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use $n = 100$ instead of $n = 36$? What happens if we decrease the sample size to $n = 25$ instead of $n = 36$?

- $\bar{x} = 68$
- $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right)$
- $\sigma = 3$, the confidence level is 90 percent ($CL = 0.90$), $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$.

Solution 8.5

Solution A

If we *increase* the sample size n to 100, we *decrease* the margin of error.

When $n = 100$, $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{100}}\right) = 0.4935$.

Solution 8.5

Solution B

If we *decrease* the sample size n to 25, we *increase* the error bound.

When $n = 25$, $EBM = \left(z_{\frac{\alpha}{2}}\right)\left(\frac{\sigma}{\sqrt{n}}\right) = (1.645)\left(\frac{3}{\sqrt{25}}\right) = 0.987$.

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

Try It Σ

8.5 Refer back to the pizza-delivery **Try It** exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90 percent confidence interval estimate for the population mean delivery time.

Working Backward to Find the Error Bound or Sample Mean

When we calculate a confidence interval, we find the sample mean, calculate the error bound, and use them to calculate the confidence interval. However, sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backward to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean,
- Or, from the upper value for the interval, subtract the lower value. Then divide the difference by 2.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval,

- Or, average the upper and lower endpoints of the confidence interval.

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example 8.6

Suppose we know that a confidence interval is (67.18, 68.82) and we want to find the error bound. We may know that the sample mean is 68, or perhaps our source only gives the confidence interval and does not tell us the value of the sample mean.

Calculate the error bound:

- If we know that the sample mean is 68, $EBM = 68.82 - 68 = 0.82$.
- If we do not know the sample mean, $EBM = \frac{(68.82 - 67.18)}{2} = 0.82$. The margin of error is the quantity that we add and subtract from the sample mean to obtain the confidence interval. Therefore, the margin of error is half of the length of the interval.

Calculate the sample mean:

- If we know the error bound, $\bar{x} = 68.82 - 0.82 = 68$.
- If we do not know the error bound, $\bar{x} = \frac{(67.18 + 68.82)}{2} = 68$.

Try It

8.6 Suppose we know that a confidence interval is (42.12, 47.88). Find the error bound and the sample mean.

Calculating the Sample Size *n*

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size. In this situation, we are given the desired margin of error, *EBM*, and we need to compute the sample size *n*.

The formula for sample size is $n = \frac{z^2\sigma^2}{EBM^2}$, found by solving the error bound formula for *n*. Always round up the value of *n* to the closest integer.

In this formula, *z* is the critical value $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example 8.7

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95 percent confident that the sample mean age is within two years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

From the problem, we know that $\sigma = 15$ and $EBM = 2$.

$z = z_{0.025} = 1.96$, because the confidence level is 95 percent.

$n = \frac{z^2\sigma^2}{EBM^2} = \frac{(1.96)^2(15)^2}{2^2} = 216.09$ using the sample size equation.

Use $n = 217$. Always round the answer up to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95 percent confident that we are within two years of the true population mean age of Foothill College students.

Try It

8.7 The population standard deviation for the height of high school basketball players is three inches. If we want to be 95 percent confident that the sample mean height is within one inch of the true population mean height, how many randomly selected students must be surveyed?

8.2 | A Single Population Mean Using the Student's t-Distribution

In practice, we rarely know the **population standard deviation**. In the past, when the sample size was large, this unknown number did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a **confidence interval** with close-enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Gosset (1876–1937) of the Guinness brewery in Dublin, Ireland, ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to *discover* what is called the **Student's *t*-distribution**. The name comes from the fact that Gosset wrote under the pen name *Student*.

Up until the mid-1970s, some statisticians used the **normal distribution** approximation for large sample sizes and used the Student's t-distribution only for sample sizes of at most 30. With graphing calculators and computers, the practice now is to use the Student's t-distribution whenever s is used as an estimate for σ.

If you draw a simple random sample of size n from a population that has an approximately normal distribution with mean μ and unknown population standard deviation σ and calculate the t-score $t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the t-scores follow a *Student's t-distribution with n – 1 degrees of freedom*. The t-score has the same interpretation as the **z-score**: It measures how far $\bar{x}$ is from its mean μ. For each sample size n, there is a different Student's t-distribution.

The **degrees of freedom (*df*)**, $n - 1$, are the sample size minus 1.

Properties of the Student's *t*-distribution

- The graph for the Student's t-distribution is similar to the standard normal curve.
- The mean for the Student's t-distribution is zero, and the distribution is symmetric about zero.
- The Student's t-distribution has more probability in its tails than the standard normal distribution. **Figure 8.6** shows the graphs of the student t-distribution for 1, 2 and 5 degrees of freedom: (v), compare to the standard normal distribution (in black).

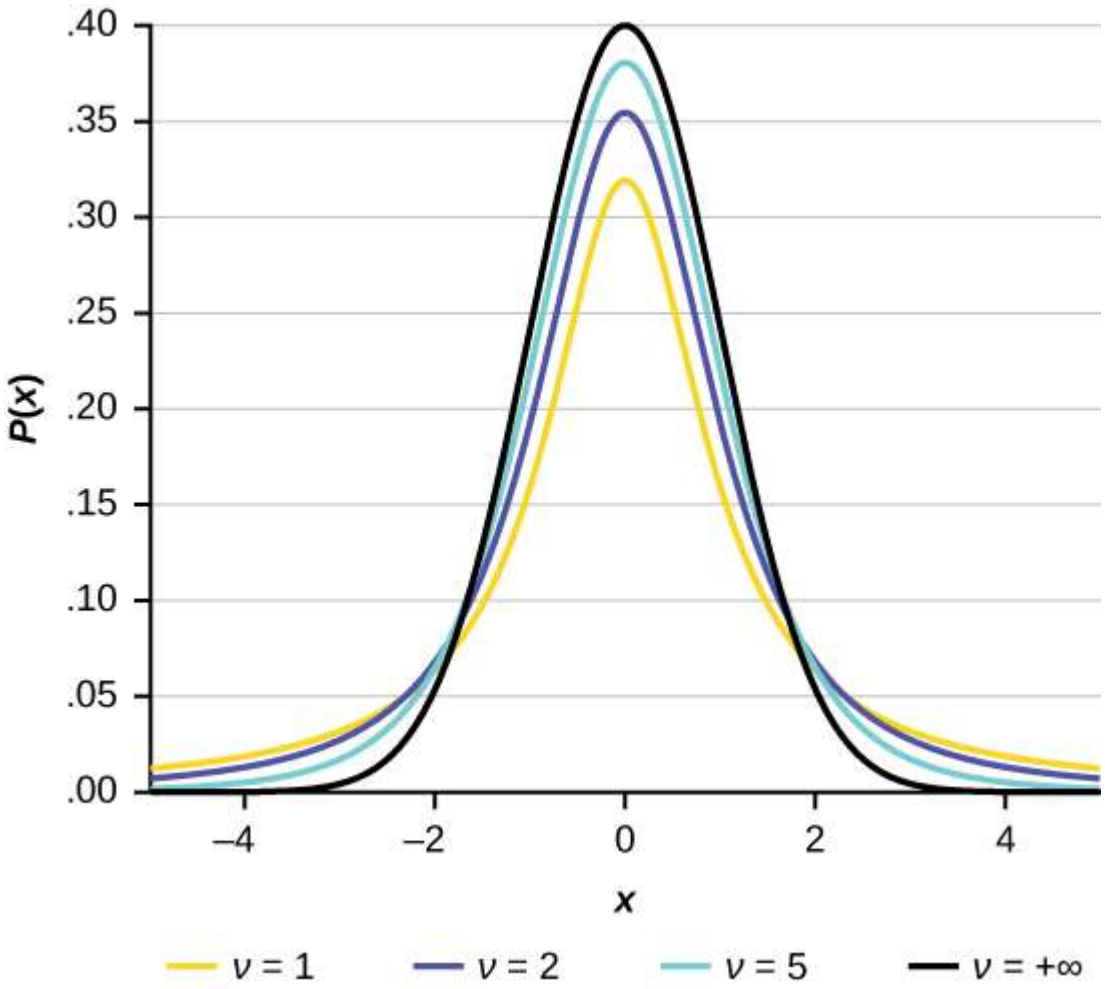


Figure 8.6

- The exact shape of the Student's *t*-distribution depends on the degrees of freedom. As the degrees of freedom increase, the graph of the Student's *t*-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. The size of the underlying population is generally not relevant unless it is very small. If it is bell-shaped (normal), then the assumption is met and does not need discussion. Random sampling is assumed, but that is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's *t*-probabilities. The TI-83, 83+, and 84+ have a tcdf function to find the probability for given values of *t*. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However, for confidence intervals, we need to use inverse probability to find the value of *t* when we know the probability.

For the TI-84+, you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command requires two inputs: invT(area to the left, degrees of freedom). The output is the *t*-score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's *t*-distribution can also be used. The table gives critical *t*-values that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator, you need to use a probability table for the Student's *t*-distribution.) When using a *t*-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's *t*-table (see **Appendix H**) gives *t*-scores given the degrees of freedom and the right-tailed probability. The table is very limited. *Calculators and computers can easily calculate any Student's t-probabilities.*

If the population standard deviation is not known, the error bound for a population mean is

- $EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$,
- $t_{\frac{\sigma}{2}}$ is the *t*-score with area to the right equal to $\frac{\alpha}{2}$,
- use $df = n - 1$ degrees of freedom, and
- s = sample standard deviation.

The format for the confidence interval is

$$(\bar{x} - EBM,\ \bar{x} + EBM).$$

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the confidence interval directly, do the following:
Press STAT.
Arrow over to TESTS.
Arrow down to 8:TInterval and press ENTER (or just press 8).

Example 8.8

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given. Use the sample data to construct a 95 percent confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.
The solution is shown step-by-step and by using the TI-83, 83+, or 84+ calculators.

8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

Solution 8.8

- The first solution is step-by-step (Solution A).
- The second solution uses the TI-83+ and TI-84 calculators (Solution B).

To find the confidence interval, you need the sample mean, $\bar{x}$, and the *EBM*.

$$\bar{x} = \frac{8.6 + 9.4 + 7.9 + 6.8 + 8.3 + 7.3 + 9.2 + 9.6 + 8.7 + 11.4 + 10.3 + 5.4 + 8.1 + 5.5 + 6.9}{15} = 8.2267;$$

$$s = \sqrt{\frac{(8.6 - \bar{x})^2 + (9.4 - \bar{x})^2 + \cdots + (5.5 - \bar{x})^2 + (6.9 - \bar{x})^2}{14}} = 1.6722;$$

$n = 15$

$df = 15 - 1 = 14$ *CL*, so $\alpha = 1 - CL = 1 - 0.95 = 0.05$

$\frac{\alpha}{2} = 0.025$; $t_{\frac{\alpha}{2}} = t_{0.025}$

The area to the right of $t_{0.025}$ is 0.025, and the area to the left of $t_{0.025}$ is $1 - 0.025 = 0.975$.

$t_{\frac{\alpha}{2}} = t_{0.025} = 2.14$ using invT(.975,14) on the TI-84+ calculator.

$$EBM = \left(t_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$

$$EBM = (2.14)\left(\frac{1.6722}{\sqrt{15}}\right) = 0.924$$

$$\bar{x} - EBM = 8.2267 - 0.9240 = 7.3$$

$$\bar{x} + EBM = 8.2267 + 0.9240 = 9.15$$

The 95 percent confidence interval is (7.30, 9.15).

We estimate with 95 percent confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution 8.8

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS.
Arrow down to 8:TInterval and press ENTER (or you can just press 8).
Arrow to Data and press ENTER.
Arrow down to List and enter the list name where you put the data.
There should be a 1 after Freq.
Arrow down to C-level and enter 0.95.
Arrow down to Calculate and press ENTER.
The 95 percent confidence interval is (7.3006, 9.1527).

NOTE

When calculating the error bound, you can also use a probability table for the Student's *t*-distribution to find the value of *t*. The table gives *t*-scores that correspond to the confidence level (column) and degrees of freedom (row); the *t*-score is found where the row and column intersect in the table.

Try It

8.8 You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95 percent confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2, 9.1, 7.7, 8.6, 6.9, 11.2, 10.1, 9.9, 8.9, 9.2, 7.5, 10.5

Example 8.9

A group of researchers is working to understand the scope of industrial pollution in the human body. Industrial chemicals may enter the body through pollution or as ingredients in consumer products. In October 2008, the scientists tested cord-blood samples for 20 newborn infants in the United States. The cord blood of the *in utero/ newborn* group was tested for 430 industrial compounds, pollutants, and other chemicals, including chemicals linked to brain and nervous-system toxicity, immune-system toxicity, reproductive toxicity, and fertility problems. There are health concerns about the effects of some chemicals on the brain and nervous system. **Table 8.2** shows how many of the targeted chemicals were found in each infant's cord blood.

79	145	147	160	116	100	159	151	156	126
137	83	156	94	121	144	123	114	139	99

Table 8.3

Use this sample data to construct a 90 percent confidence interval for the mean number of targeted industrial chemicals to be found in an infant's blood.

Solution 8.9

Solution A

From the sample data, you can calculate $\bar{x} = \frac{79 + 145 + \cdots + 139 + 99}{20} = 127.45$

$s = \sqrt{\frac{(79 - \bar{x})^2 + (145 - \bar{x})^2 + \cdots + (139 - \bar{x})^2 + (99 - \bar{x})^2}{19}} = 25.965.$ There are 20 infants in the sample, so $n = 20$, and $df = 20 - 1 = 19$.

You are asked to calculate a 90 percent confidence interval: $CL = 0.90$, so $\alpha = 1 - CL = 1 - 0.90 = 0.10$. $\frac{\alpha}{2} = 0.05, t_{\frac{\alpha}{2}} = t_{0.05}$

By definition, the area to the right of $t_{0.05}$ is 0.05, and so the area to the left of $t_{0.05}$ is $1 - 0.05 = 0.95$.

Use a table, calculator, or computer to find that $t_{0.05} = 1.729$.

$$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 1.729\left(\frac{25.965}{\sqrt{20}}\right) \approx 10.038$$

$$\bar{x} - EBM = 127.45 - 10.038 = 117.412$$
$$\bar{x} + EBM = 127.45 + 10.038 = 137.488$$

We estimate with 90 percent confidence that the mean number of all targeted industrial chemicals found in cord blood in the United States is between 117.412 and 137.488.

Solution 8.9

Solution B

Using the TI-83, 83+, 84, 84+ Calculator

Enter the data as a list.
Press STAT and arrow over to TESTS.
Arrow down to 8:TInterval and press ENTER (or you can just press 8). Arrow to Data and press ENTER.
Arrow down to List and enter the list name where you put the data.
Arrow down to Freq and enter 1.
Arrow down to C-level and enter 0.90.
Arrow down to Calculate and press ENTER.
The 90 percent confidence interval is (117.41, 137.49).

Try It

8.9 A random sample of statistics students was asked to estimate the total number of hours they spend watching television in an average week. The responses are recorded in **Table 8.4**. Use the following sample data to construct a 98 percent confidence interval for the mean number of hours statistics students will spend watching television in one

week.

0	3	1	20	9
5	10	1	10	4
14	2	4	4	5

Table 8.4

8.3 | A Population Proportion

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40 percent of the vote within 3 percentage points (if the sample is large enough). Often, election polls are calculated with 95 percent confidence, so the pollsters would be 95 percent confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43 (0.40 – 0.03, 0.40 + 0.03).

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the **error bound for a population (*EBP*)**, and the **confidence level** for a proportion is similar to that for the population mean, but the formulas are different.

How do you know you are dealing with a proportion problem? First, the data that you are collecting is categorical, consisting of two categories: Success or Failure, Yes or No. Examples of situations where you are the following trying to estimate the true population proportion are the following: What proportion of the population smoke? What proportion of the population will vote for candidate A? What proportion of the population has a college-level education?

The distribution of the sample proportions (based on samples of size n) is denoted by P' (read “P prime”).

The central limit theorem for proportions asserts that the sample proportion distribution P' follows a normal distribution with mean value p, and standard deviation $\sqrt{\frac{p \bullet q}{n}}$, where p is the population proportion and $q = 1 - p$.

The confidence interval has the form $(p' - EBP, p' + EBP)$. EBP is error bound for the proportion.

$$p' = \frac{x}{n}$$

p' = the *estimated proportion* of successes (p' is a *point estimate* for p, the true proportion.)

x = the *number* of successes

n = the size of the sample

The error bound for a proportion is

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\left(\sqrt{\frac{p'q'}{n}}\right), \text{ where } q' = 1 - p'.$$

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is $\sqrt{\frac{pq}{n}}$.

However, in the error bound formula, we use $\sqrt{\frac{p'q'}{n}}$ as the standard deviation, instead of $\sqrt{\frac{pq}{n}}$.

In the error bound formula, the *sample proportions p′ and q′, are estimates of the unknown population proportions p and q*. The estimated proportions p' and q' are used because p and q are not known. The sample proportions p' and q' are calculated

from the data: p' is the estimated proportion of successes, and q' is the estimated proportion of failures.

The confidence interval can be used only if the number of successes np' and the number of failures nq' are both greater than five.

That is, in order to use the formula for confidence intervals for proportions, you need to verify that both $np' \geq 5$ and $nq' \geq 5$.

Example 8.10

Suppose that a market research firm is hired to estimate the percentage of adults living in a large city who have cell phones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes, they own cell phones. Using a 95 percent confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.

Solution 8.10

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Let X = the number of people in the sample who have cell phones. X is binomial. $X \sim B\left(500, \frac{421}{500}\right)$.

To calculate the confidence interval, you must find p', q', and EBP.

$n = 500$

x = the number of successes = 421

$$p' = \frac{x}{n} = \frac{421}{500} = 0.842$$

$p' = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$$q' = 1 - p' = 1 - 0.842 = 0.158$$

Because $CL = 0.95$, then $\alpha = 1 - CL = 1 - 0.95 = 0.05$ $\left(\frac{\alpha}{2}\right) = 0.025$.

Then, $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$.

Use the TI-83, 83+, or 84+ calculator command invNorm(0.975,0,1) to find $z_{0.025}$. Remember that the area to the right of $z_{0.025}$is 0.025, and the area to the left of $z_{0.025}$is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{(0.842)(0.158)}{500}} = 0.032$$

$$p' - EBP = 0.842 - 0.032 = 0.81$$

$$p' + EBP = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP) = (0.810, 0.874)$.

Interpretation

We estimate with 95 percent confidence that between 81 percent and 87.4 percent of all adult residents of this city have cell phones.

Explanation of 95 percent Confidence Level

Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution 8.10

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.
Arrow down to *x* and enter 421.
Arrow down to *n* and enter 500.
Arrow down to C-Level and enter .95.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.81003, 0.87397).

Try It

8.10 Suppose 250 randomly selected people are surveyed to determine whether they own tablets. Of the 250 surveyed, 98 reported owning tablets. Using a 95 percent confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Example 8.11

For a class project, a political science student at a large university wants to estimate the percentage of students who are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90 percent confidence interval for the true percentage of students who are registered voters, and interpret the confidence interval.

Solution 8.11

- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+, or 84 calculators (Solution B).

Solution A

$$x = 300 \text{ and } n = 500$$

$$p' = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$q' = 1 - p' = 1 - 0.600 = 0.400$$

Because $CL = 0.90$, then $\alpha = 1 - CL = 1 - 0.90 = 0.10$ $\left(\frac{\alpha}{2}\right) = 0.05$.

$$z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$$

Use the TI-83, 83+, or 84+ calculator command invNorm(0.95,0,1) to find $z_{0.05}$. Remember that the area to the right of $z_{0.05}$ is 0.05, and the area to the left of $z_{0.05}$ is 0.95. This can also be found using appropriate commands on other calculators, using a computer, or using a standard normal probability table.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.645)\sqrt{\frac{(0.60)(0.40)}{500}} = 0.036$$

$$p' - EBP = 0.60 - 0.036 = 0.564$$

$$p' + EBP = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is ($p' - EBP$, $p' + EBP$) = (0.564, 0.636).

Interpretation

- We estimate with 90 percent confidence that the true percentage of all students who are registered voters is between 56.4 percent and 63.6 percent.
- Alternate wording: We estimate with 90 percent confidence that between 56.4 percent and 63.6 percent of all students are registered voters.

Explanation of 90 percent Confidence Level

Ninety percent of all confidence intervals constructed in this way contain the true value for the population percentage of students who are registered voters.

Solution 8.11

Solution B

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.
Arrow down to *x* and enter 300.
Arrow down to *n* and enter 500.
Arrow down to C-Level and enter 0.90.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.564, 0.636).

Try It

8.11 A student polls her school to determine whether students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90 percent confidence interval for the true percentage of students who are against the new legislation, and interpret the confidence interval.

b. In a sample of 300 students, 68 percent said they own an iPod and a smartphone. Compute a 97 percent confidence interval for the true percentage of students who own an iPod and a smartphone.

Plus-Four Confidence Interval for *p*

There is a certain amount of error introduced into the process of calculating a confidence interval for a proportion. Because we do not know the true proportion for the population, we are forced to use point estimates to calculate the appropriate standard deviation of the sampling distribution. Studies have shown that the resulting estimation of the standard deviation can be flawed.

Fortunately, there is a simple adjustment that allows us to produce more accurate confidence intervals: We simply pretend that we have four additional observations. Two of these observations are successes, and two are failures. The new sample size, then, is $n + 4$, and the new count of successes is $x + 2$.

Computer studies have demonstrated the effectiveness of the **plus-four confidence interval for *p* method**. It should be used when the confidence level desired is at least 90 percent and the sample size is at least ten.

Example 8.12

A random sample of 25 statistics students was asked: "Have you used a product in the past week?" Six students reported using the product within the past week. Use the plus-four method to find a 95 percent confidence interval for the true proportion of statistics students who use the product weekly.

Solution 8.12

Six students out of 25 reported using a product within the past week, so $x = 6$ and $n = 25$. Because we are using the plus-four method, we will use $x = 6 + 2 = 8$, and $n = 25 + 4 = 29$.

$$p' = \frac{x}{n} = \frac{8}{29} \approx 0.276$$

$$q' = 1 - p' = 1 - 0.276 = 0.724$$

Because $CL = 0.95$, we know $\alpha = 1 - 0.95 = 0.05$, and $\frac{\alpha}{2} = 0.025$.

$$z_{0.025} = 1.96$$

$$EPB = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} = (1.96)\sqrt{\frac{0.276(0.724)}{29}} \approx 0.163$$

$$p' - EPB = 0.276 - 0.163 = 0.113$$

$$p' + EPB = 0.276 + 0.163 = 0.439$$

We are 95 percent confident that the true proportion of all statistics students who use the product is between 0.113 and 0.439.

Solution 8.12

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.

Arrow down to *x* and enter 8.
Arrow down to *n* and enter 29.
Arrow down to C-Level and enter 0.95.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.113, 0.439).

REMINDER

Remember that the plus-four method assumes an additional four trials: two successes and two failures. You do not need to change the process for calculating the confidence interval; simply update the values of *x* and *n* to reflect these additional trials.

Try It

8.12 Out of a random sample of 65 freshmen at State University, 31 students have declared their majors. Use the plus-four method to find a 96 percent confidence interval for the true proportion of freshmen at State University who have declared their majors.

Example 8.13

A group of researchers recently conducted a study analyzing the privacy management habits of teen internet users. In a group of 50 teens, 13 reported having more than 500 friends on a social media site. Use the plus four method to find a 90 percent confidence interval for the true proportion of teens who would report having more than 500 online friends.

Solution 8.13

Using plus-four, we have x = 13 + 2 = 15, and n = 50 + 4 = 54.

$$p' = \frac{15}{54} \approx 0.278$$

$$q' = 1 - p' = 1 - 0.278 = 0.722$$

Because CL = 0.90, we know α = 1 – 0.90 = 0.10, and $\frac{\alpha}{2}$ = 0.05.

$$z_{0.05} = 1.645$$

$$EPB = (z_{\frac{\alpha}{2}})\left(\sqrt{\frac{p'q'}{n}}\right) = (1.645)\left(\sqrt{\frac{(0.278)(0.722)}{54}}\right) \approx 0.100$$

$$p' - EPB = 0.278 - 0.100 = 0.178$$

$$p' + EPB = 0.278 + 0.160 = 0.378$$

We are 90 percent confident that between 17.8 percent and 37.8 percent of all teens would report having more than 500 friends on a social media site.

Solution 8.13

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS.
Arrow down to A:1-PropZint. Press ENTER.
Arrow down to x and enter 15.
Arrow down to n and enter 54.
Arrow down to C-Level and enter 0.90.
Arrow down to Calculate and press ENTER.
The confidence interval is (0.178, 0.378).

Try It

8.13 The research group referenced in **Example 8.13** talked to teens in smaller focus groups but also interviewed additional teens over the phone. When the study was complete, 588 teens had answered the question about their social media site friends, with 159 saying that they have more than 500 friends. Use the plus-four method to find a 90 percent confidence interval for the true proportion of teens who would report having more than 500 online friends based on this larger sample. Compare the results to those in **Example 8.13**.

Calculating the Sample Size *n*

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The margin of error formula for a population proportion is

- $EBP = z_{\frac{\alpha}{2}} \times \sqrt{\frac{p' \bullet q'}{n}}$, where p' is the sample proportion, $q' = 1 - p'$, and n is the sample size.
- Solving for n gives you an equation for the sample size.
- $n = \frac{\left(z_{\frac{\alpha}{2}}\right)^2 (p' q')}{EBP^2}$. This formula tells us that we can compute the sample size n required for a confidence level of $Cl = 1 - \alpha$ by taking the square of the critical value $z_{\frac{\alpha}{2}}$, multiplying by the point estimate p', and by $q' = 1 - p'$ and finally dividing the result by the square of the margin of error. *Always remember to round up the value of n.*

Example 8.14

Suppose a mobile phone company wants to determine the current percentage of customers ages 50+ who use text messaging on their cell phones. How many customers ages 50+ should the company survey in order to be 90 percent confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of customers ages 50+ who use text messaging on their cell phones? Assume that $p' = 0.5$.

Solution 8.14

From the problem, we know that $EBP = 0.03$ (3 percent=0.03), and $z_{\frac{\alpha}{2}}$ $z_{0.05} = 1.645$ because the confidence level is 90 percent.

To calculate the sample size n, use the formula and make the substitutions.

$$n = \frac{z^2 p' q'}{EBP^2} \; gives \; n = \frac{1.645^2 (0.5)(0.5)}{0.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers ages 50+ in order to be 90 percent confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of all customers ages 50+ who use text messaging on their cell phones.

Try It

8.14 An internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90 percent confident that the estimated proportion is within 5 percentage points of the true population proportion of customers who click on ads on their smartphones? Assume that the sample proportion p' is 0.50.

8.4 | Confidence Interval (Home Costs)

Stats Lab

8.1 Confidence Interval (Home Costs)

Student Learning Outcomes

- The student will calculate the 90 percent confidence interval for the mean cost of a home in the area in which this school is located.
- The student will interpret confidence intervals.
- The student will determine the effects of changing conditions on the confidence interval.

Collect the Data

Check the Real Estate section in your local newspaper. Record the sale prices for 35 randomly selected homes recently listed in the county.

NOTE

Many newspapers list them only one day per week. Also, we will assume that homes come up for sale randomly.

1. Complete the following table:

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 8.5

Describe the Data

1. Compute the following:
   a. $\bar{x}$ = _____
   b. s_x = _____
   c. n = _____
2. In words, define the random variable $\bar{X}$.
3. State the estimated distribution to use. Use both words and symbols.

Find the Confidence interval

1. Calculate the confidence interval and the error bound.
   a. Confidence interval: _____
   b. Error Bound: _____
2. How much area is in both tails (combined)? α = _____

3. How much area is in each tail? $\frac{\alpha}{2}$ = _____

4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample mean.

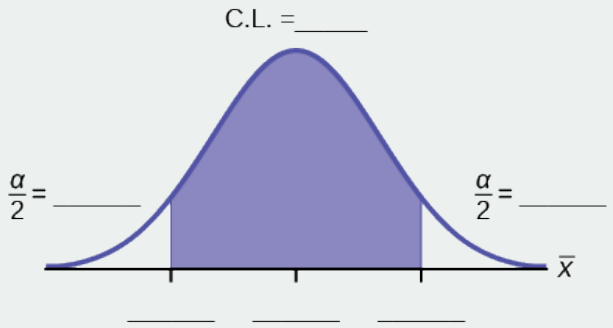


Figure 8.7

5. Some students think that a 90 percent confidence interval contains 90 percent of the data. Use the list of data on the first page and count how many of the data values lie within the confidence interval. What percentage is this? Is this percentage close to 90 percent? Explain why this percentage should or should not be close to 90 percent.

Describe the Confidence Interval

1. In two to three complete sentences, explain what a confidence interval means (in general), as if you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this confidence interval means for this particular study.

Use the Data to Construct Confidence Intervals

1. Using the given information, construct a confidence interval for each confidence level given.

Confidence Level	*EBM*/Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

Table 8.6

2. What happens to the *EBM* as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

8.5 | Confidence Interval (Place of Birth)

Stats Lab

8.2 Confidence Interval (Place of Birth)

Student Learning Outcomes

- The student will calculate the 90 percent confidence interval of the proportion of students in this school who were born in this state.
- The student will interpret confidence intervals.
- The student will determine the effects of changing conditions on the confidence interval.

Collect the Data

1. Survey the students in your class, asking them whether they were born in this state. Let X = the number who were born in this state.
   a. n = _____
   b. x = _____
2. In words, define the random variable P'.
3. State the estimated distribution to use.

Find the Confidence interval and Error bound

1. Calculate the confidence interval and the error bound.
   a. Confidence interval: _____
   b. Error Bound: _____
2. How much area is in both tails (combined)? α = _____
3. How much area is in each tail? $\frac{\alpha}{2}$ = _____
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample proportion.

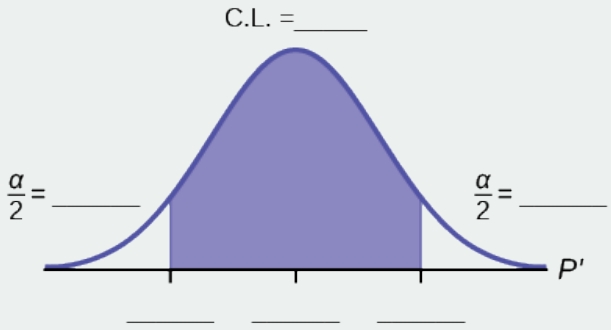


Figure 8.8

Describe the Confidence Interval

1. In two to three complete sentences, explain what a confidence interval means (in general), as though you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this confidence interval means for this particular study.
3. Construct a confidence interval for each confidence level given.

Confidence Level	*EBP*/Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

Table 8.7

4. What happens to the *EBP* as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

8.6 | Confidence Interval (Women's Heights)

Stats Lab

8.3 Confidence Interval (Women's Heights)

Student Learning Outcomes

- The student will calculate a 90 percent confidence interval using the given data.
- The student will determine the relationship between the confidence level and the percentage of constructed intervals that contain the population mean.

Given:

59.4	71.6	69.3	65.0	62.9	66.5	61.7	55.2
67.5	67.2	63.8	62.9	63.0	63.9	68.7	65.5
61.9	69.6	58.7	63.4	61.8	60.6	69.8	60.0
64.9	66.1	66.8	60.6	65.6	63.8	61.3	59.2
64.1	59.3	64.9	62.4	63.5	60.9	63.3	66.3
61.5	64.3	62.9	60.6	63.8	58.8	64.9	65.7
62.5	70.9	62.9	63.1	62.2	58.7	64.7	66.0
60.5	64.7	65.4	60.2	65.0	64.1	61.1	65.3
64.6	59.2	61.4	62.0	63.5	61.4	65.5	62.3
65.5	64.7	58.8	66.1	64.9	66.9	57.9	69.8
58.5	63.4	69.2	65.9	62.2	60.0	58.1	62.5
62.4	59.1	66.4	61.2	60.4	58.7	66.7	67.5
63.2	56.6	67.7	62.5				

Table 8.8 Heights of 100 Women (in Inches)

1. **Table 8.8** lists the heights of 100 women. Use a random number generator to select 10 data values randomly.
2. Calculate the sample mean and the sample standard deviation. Assume that the population standard deviation is known to be 3.3 in. With these values, construct a 90 percent confidence interval for your sample of 10 values. Write the confidence interval you obtained in the first space of **Table 8.9**.
3. Now write your confidence interval on the board. As others in the class write their confidence intervals on the board, copy them into **Table 8.9**.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Table 8.9 90 percent Confidence Intervals

Discussion Questions

1. The actual population mean for the 100 heights given in **Table 8.8** is μ = 63.4. Using the class listing of confidence intervals, count how many of them contain the population mean μ; i.e., for how many intervals does the value of μ lie between the endpoints of the confidence interval?
2. Divide this number by the total number of confidence intervals generated by the class to determine the percentage of confidence intervals that contain the mean μ. Write that percentage here: _____.
3. Is the percentage of confidence intervals that contain the population mean μ close to 90 percent?
4. Suppose we had generated 100 confidence intervals. What do you think would happen to the percentage of confidence intervals that contained the population mean?
5. When we construct a 90 percent confidence interval, we say that we are *90 percent confident that the true population mean lies within the confidence interval.* Using complete sentences, explain what we mean by this phrase.
6. Some students think that a 90 percent confidence interval contains 90 percent of the data. Use the list of data given (the heights of women) and count how many of the data values lie within the confidence interval that you generated based on that data. How many of the 100 data values lie within your confidence interval? What percentage is this? Is this percentage close to 90 percent?
7. Explain why it does not make sense to count data values that lie in a confidence interval. Think about the random variable that is being used in the problem.
8. Suppose you obtained the heights of 10 women and calculated a confidence interval from this information. Without knowing the population mean μ, would you have any way of knowing *for certain* whether your interval actually contained the value of μ? Explain.

KEY TERMS

binomial distribution a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, n, of independent trials
Independent means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances, the binomial RV X is defined as the number of successes in n trials. The notation is $X \sim B(\mathbf{n},\mathbf{p})$. The mean is $\mu = np$, and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

confidence interval (*CI*) an interval estimate for an unknown population parameter.
This depends on the following:

- the desired confidence level,
- information that is known about the distribution (for example, known standard deviation), and
- the sample and its size.

confidence level (*CL*) the percentage expression for the probability that the confidence interval contains the true population parameter; for example, if the CL = 90 percent, then in 90 out of 100 samples, the interval estimate will enclose the true population parameter

degrees of freedom (*df*) the number of objects in a sample that are free to vary

error bound for a population mean (*EBM*) the margin of error; depends on the confidence level, sample size, and known or estimated population standard deviation

error bound for a population proportion (*EBP*) the margin of error; depends on the confidence level, the sample size, and the estimated (from the sample) proportion of successes

inferential statistics also called statistical inference or inductive statistics; this facet of statistics deals with estimating a population parameter based on a sample statistic
For example, if four out of the 100 calculators sampled are defective, we might infer that 4 percent of the production is defective.

normal distribution a bell-shaped continuous random variable X, with center at the mean value (μ) and distance from the center to the inflection points of the bell curve given by the standard deviation (σ).
We write $X \sim N(\mu, \sigma)$. If the mean value is 0 and the standard deviation is 1, the random variable is called the standard normal distribution, and it is denoted with the letter Z

parameter a numerical characteristic of a population

plus-four confidence interval plus-four confidence interval when you add two imaginary successes and two imaginary failures (four overall) to your sample

point estimate a single number computed from a sample and used to estimate a population parameter

standard deviation a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation

Student's *t*-distribution investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student
the major characteristics of the random variable (RV) are as follows:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n get larger.
- There is a family of t-distributions: Each representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data.

CHAPTER REVIEW

8.1 A Single Population Mean Using the Normal Distribution

In this module, we learned how to calculate the confidence interval for a single population mean where the population standard deviation is known. When estimating a population mean, the margin of error is called the error bound for a population mean (*EBM*). A confidence interval has the general form

(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*).

The calculation of *EBM* depends on the size of the sample and the level of confidence desired. The confidence level is the percentage of all possible samples that can be expected to include the true population parameter. As the confidence level increases, the corresponding *EBM* increases as well. As the sample size increases, the *EBM* decreases. By the central limit theorem,

$$EBM = z\frac{\sigma}{\sqrt{n}.}$$

Given a confidence interval, you can work backward to find the error bound (*EBM*) or the sample mean. To find the error bound, find the difference of the upper bound of the interval and the mean. If you do not know the sample mean, you can find the error bound by calculating half of the difference of the upper and lower bounds. To find the sample mean given a confidence interval, find the difference of the upper bound and the error bound. If the error bound is unknown, then average the upper and lower bounds of the confidence interval to find the sample mean.

Sometimes researchers know in advance that they want to estimate a population mean within a specific margin of error for a given level of confidence. In that case, solve the *EBM* formula for *n* to discover the size of the sample that is needed to achieve this goal:

$$n = \frac{z^2\sigma^2}{EBM^2}$$

8.2 A Single Population Mean Using the Student's t-Distribution

In many cases, the researcher does not know the population standard deviation, σ, of the measure being studied. In these cases, it is common to use the sample standard deviation, *s*, as an estimate of σ. The normal distribution creates accurate confidence intervals when σ is known, but it is not as accurate when *s* is used as an estimate. In this case, the Student's *t*-distribution is much better. Define a *t*-score using the following formula:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The *t*-score follows the Student's *t*-distribution with $n - 1$ degrees of freedom. The confidence interval under this distribution is calculated with $EBM = \left(t_{\frac{\alpha}{2}}\right)\frac{s}{\sqrt{n}},$ where $t_{\frac{\alpha}{2}}$ is the *t*-score with area to the right equal to $\frac{\alpha}{2}$, *s* is the sample standard deviation, and *n* is the sample size. Use a table, calculator, or computer to find $t_{\frac{\alpha}{2}}$ for a given α.

8.3 A Population Proportion

Some statistical measures, like many survey questions, measure qualitative rather than quantitative data. In this case, the population parameter being estimated is a proportion. It is possible to create a confidence interval for the true population proportion by following procedures similar to those used in creating confidence intervals for population means. The formulas are slightly different, but they follow the same reasoning.

Let p' represent the sample proportion, x/n, where *x* represents the number of successes, and *n* represents the sample size. Let $q' = 1 - p'$. Then the confidence interval for a population proportion is given by the following formula:

$$\text{(lower bound, upper bound)} = (p' - EBP,\ p' + EBP) = (p' - z\sqrt{\frac{p'q'}{n}},\ p' + z\sqrt{\frac{p'q'}{n}}).$$

The **plus–four method** for calculating confidence intervals is an attempt to balance the error introduced by using estimates of the population proportion when calculating the standard deviation of the sampling distribution. Simply imagine four

additional trials in the study; two are successes and two are failures. Calculate $p' = \frac{x+2}{n+4}$, and proceed to find the confidence interval. When sample sizes are small, this method has been demonstrated to provide more accurate confidence intervals than the standard formula used for larger samples.

FORMULA REVIEW

8.1 A Single Population Mean Using the Normal Distribution

$\bar{X} \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$ The distribution of sample means is normally distributed with mean equal to the population mean and standard deviation given by the population standard deviation divided by the square root of the sample size.

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by
(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*)

$$= (\bar{x} - EBM, \bar{x} + EBM)$$
$$= \left(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}}\right).$$

$EBM = z\frac{\sigma}{\sqrt{n}}$ = the error bound for the mean, or the margin of error for a single population mean; this formula is used when the population standard deviation is known.

CL = confidence level, or the proportion of confidence intervals created that is expected to contain the true population parameter

$\alpha = 1 - CL$ = the proportion of confidence intervals that will not contain the population parameter

$z_{\frac{\alpha}{2}}$ = the *z*-score with the property that the area to the right of the *z*-score is $\frac{\alpha}{2}$; this is the *z*-score, used in the calculation of *EBM*, where $\alpha = 1 - CL$.

$n = \frac{z^2\sigma^2}{EBM^2}$ = the formula used to determine the sample size (*n*) needed to achieve a desired margin of error at a given level of confidence

General form of a confidence interval

(lower value, upper value) = (point estimate error bound, point estimate + error bound)

To find the error bound when you know the confidence interval,

error bound = upper value point estimate **or** error bound = $\frac{\text{upper value} - \text{lower value}}{2}$.

Single population mean, known standard deviation, normal distribution

Use the normal distribution for means; population standard deviation is known: $EBM = z\frac{\alpha}{2} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format ($\bar{x} - EBM$, $\bar{x} + EBM$).

8.2 A Single Population Mean Using the Student's t-Distribution

s = the standard deviation of sample values

$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ is the formula for the *t*-score, which measures how far away a measure is from the population mean in the Student's *t*-distribution.

$df = n - 1$; the degrees of freedom for a Student's *t*-distribution, where *n* represents the size of the sample

$T \sim t_{df}$ the random variable, *T*, has a Student's *t*-distribution with *df* degrees of freedom

$EBM = t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$ = the error bound for the population mean when the population standard deviation is unknown

$t_{\frac{\alpha}{2}}$ is the *t*-score in the Student's *t*-distribution with area to the right equal to $\frac{\alpha}{2}$.

The general form for a confidence interval for a single mean, population standard deviation unknown, Student's *t* is given by
(lower bound, upper bound) = (point estimate – *EBM*, point estimate + *EBM*)

$$= \left(\bar{x} - \frac{ts}{\sqrt{n}}, \bar{x} + \frac{ts}{\sqrt{n}}\right).$$

8.3 A Population Proportion

$p' = x/n$, where *x* represents the number of successes and *n* represents the sample size. The variable p' is the sample proportion and serves as the point estimate for the true population proportion.

$$q' = 1 - p'$$

$p' \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$ The variable p' has a **binomial distribution** that can be approximated with the normal distribution shown here,

$$EBP = \text{the error bound for a proportion} = z_{\frac{\alpha}{2}}\sqrt{\frac{p'q'}{n}}.$$

Confidence interval for a proportion:

$$\text{(lower bound, upper bound)} = (p' - EBP,\ p' + EBP) = \left(p' - z\sqrt{\frac{p'q'}{n}},\quad p' + z\sqrt{\frac{p'q'}{n}}\right)$$

$n = \frac{z_{\frac{\alpha}{2}}^{2} p'q'}{EBP^2}$ provides the number of participants needed to estimate the population proportion with confidence $1 - \alpha$ and margin of error EBP.

Use the normal distribution for a single population proportion $p' = \frac{x}{n}$.

$$EBP = \left(z_{\frac{\alpha}{2}}\right)\sqrt{\frac{p'q'}{n}} \quad p' + q' = 1$$

The confidence interval has the format ($p' - EBP$, $p' + EBP$).

$\bar{x}$ is a point estimate for μ.

p' is a point estimate for ρ.

s is a point estimate for σ.

PRACTICE

8.1 A Single Population Mean Using the Normal Distribution

Use the following information to answer the next five exercises: The standard deviation of the weights of elephants is known to be approximately 15 lb. We wish to construct a 95 percent confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 lb. The sample standard deviation is 11 lb.

1. Identify the following:

a. $\bar{x}$ = _____
b. σ = _____
c. n = _____

2. In words, define the random variables X and $\bar{X}$.

3. Which distribution should you use for this problem?

4. Construct a 95 percent confidence interval for the population mean weight of newborn elephants. State the confidence interval, sketch the graph, and calculate the error bound.

5. What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?

Use the following information to answer the next seven exercises: The U.S. Census Bureau conducts a study to determine the time needed to complete the short form. The bureau surveys 200 people. The sample mean is 8.2 minutes. There is a known standard deviation of 2.2 minutes. The population distribution is assumed to be normal.

6. Identify the following:

a. $\bar{x}$ = _____
b. σ = _____
c. n = _____

7. In words, define the random variables X and $\bar{X}$.

8. Which distribution should you use for this problem?

9. Construct a 90 percent confidence interval for the population mean time to complete the forms. State the confidence interval, sketch the graph, and calculate the error bound.

10. If the Census wants to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

11. If the Census did another survey, kept the error bound the same, and surveyed only 50 people instead of 200, what would happen to the level of confidence? Why?

12. Suppose the Census needed to be 98 percent confident of the population mean length of time. Would the Census have to survey more people? Why or why not?

Use the following information to answer the next 10 exercises: A sample of 20 heads of lettuce was selected. Assume that the population distribution of head weight is normal. The weight of each head of lettuce was then recorded. The mean weight was 2.2 lb, with a standard deviation of 0.1 lb. The population standard deviation is known to be 0.2 lb.

13. Identify the following:

a. $\bar{x}$ = _____
b. σ = _____
c. n = _____

14. In words, define the random variable X.

15. In words, define the random variable $\bar{X}$.

16. Which distribution should you use for this problem?

17. Construct a 90 percent confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

18. Construct a 95 percent confidence interval for the population mean weight of the heads of lettuce. State the confidence interval, sketch the graph, and calculate the error bound.

19. In complete sentences, explain why the confidence interval in **Exercise 8.17** is larger than in **Exercise 8.18**.

20. In complete sentences, give an interpretation of what the interval in **Exercise 8.18** means.

21. What would happen if 40 heads of lettuce were sampled instead of 20 and the error bound remained the same?

22. What would happen if 40 heads of lettuce were sampled instead of 20 and the confidence level remained the same?

Use the following information to answer the next 14 exercises: The mean age for all Foothill College students for a recent fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that 25 winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for winter Foothill College students. Let X = the age of a winter Foothill College student.

23. $\bar{x}$ = _____

24. n = _____

25. _____ = 15

26. In words, define the random variable $\bar{X}$.

27. What is $\bar{x}$ estimating?

28. Is σ_x known?

29. As a result of your answer to **Exercise 8.26**, state the exact distribution to use when calculating the confidence interval.

Construct a 95 percent confidence interval for the true mean age of winter Foothill College students by working out and then answering the next eight exercises.

30. How much area is in both tails (combined)? α =_____

31. How much area is in each tail? $\frac{\alpha}{2}$ =_____

32. Identify the following specifications:

a. lower limit
b. upper limit
c. error bound

33. The 95 percent confidence interval is _____.

34. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample mean.

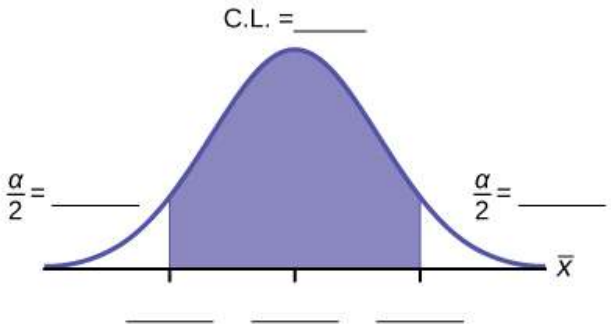


Figure 8.9

35. In one complete sentence, explain what the interval means.

36. Using the same mean, standard deviation, and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

37. Using the same mean, standard deviation, and sample size, how would the error bound change if the confidence level were reduced to 90 percent? Why?

8.2 A Single Population Mean Using the Student's t-Distribution

Use the following information to answer the next five exercises: A hospital is trying to cut down on emergency room wait times. It is interested in the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean was 1.5 hr, with a sample standard deviation of 0.5 hr.

38. Identify the following:

a. $\bar{x}$ =_____
b. s_x =_____
c. n =_____
d. $n - 1$ =_____

39. Define the random variables X and $\bar{X}$ in words.

40. Which distribution should you use for this problem?

41. Construct a 95 percent confidence interval for the population mean time spent waiting. State the confidence interval, sketch the graph, and calculate the error bound.

42. Explain in complete sentences what the confidence interval means.

Use the following information to answer the next six exercises: One hundred eight Americans were surveyed to determine the number of hours they spend watching television each month. It was revealed that they watch an average of 151 hours each month, with a standard deviation of 32 hours. Assume that the underlying population distribution is normal.

43. Identify the following:

a. $\bar{x}$ =_____
b. s_x =_____
c. n =_____
d. $n - 1$ =_____

44. Define the random variable X in words.

45. Define the random variable $\bar{X}$ in words.

46. Which distribution should you use for this problem?

47. Construct a 99 percent confidence interval for the population mean hours spent watching television per month. State the confidence interval, sketch the graph, and calculate the error bound.

48. Why would the error bound change if the confidence level were lowered to 95 percent?

Use the following information to answer the next 13 exercises: The data in **Table 8.10** are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let X = the number of colors on a national flag.

X	Freq.
1	1
2	7
3	18
4	7
5	6

Table 8.10

49. Calculate the following:

a. $\bar{x}$ =_____

b. s_x =_____

c. n =_____

50. Define the random variable $\bar{X}$ in words.

51. What is $\bar{x}$ estimating?

52. Is σ_x known?

53. As a result of your answer to **Exercise 8.52**, state the exact distribution to use when calculating the confidence interval.

Construct a 95 percent confidence interval for the true mean number of colors on national flags.

54. How much area is in both tails (combined)?

55. How much area is in each tail?

56. Calculate the following:

a. lower limit

b. upper limit

c. error bound

57. The 95 percent confidence interval is_____.

58. Fill in the blanks on the graph with the areas, the upper and lower limits of the confidence interval, and the sample mean.

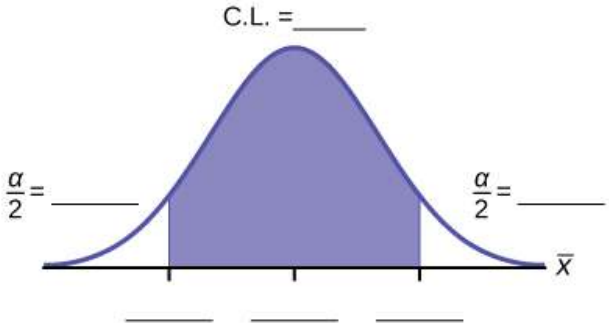


Figure 8.10

59. In one complete sentence, explain what the interval means.

60. Using the same $\bar{x}$, s_x, and level of confidence, suppose that n were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

61. Using the same $\bar{x}$, s_x, and n = 39, how would the error bound change if the confidence level were reduced to 90 percent? Why?

8.3 A Population Proportion

Use the following information to answer the next two exercises: Marketing companies are interested in knowing the population percentage of women who make the majority of household purchasing decisions.

62. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90 percent confident that the population proportion is estimated to within 0.05?

63. If it were later determined that it was important to be more than 90 percent confident and a new survey were commissioned, how would it affect the minimum number you need to survey? Why?

Use the following information to answer the next five exercises: Suppose a marketing company conducted a survey. It randomly surveyed 200 households and found that in 120 of them, the women made the majority of the purchasing decisions. We are interested in the population proportion of households where women make the majority of the purchasing decisions.

64. Identify the following:

a. x = _____
b. n = _____
c. p' = _____

65. Define the random variables X and P' in words.

66. Which distribution should you use for this problem?

67. Construct a 95 percent confidence interval for the population proportion of households where the women make the majority of the purchasing decisions. State the confidence interval, sketch the graph, and calculate the error bound.

68. List two difficulties the company might have in obtaining random results if this survey were done by email.

Use the following information to answer the next five exercises: Of 1,050 randomly selected adults, 360 identified themselves as manual laborers, 280 identified themselves as non-manual wage earners, 250 identified themselves as mid-level managers, and 160 identified themselves as executives. In the survey, 82 percent of manual laborers preferred trucks, 62 percent of non-manual wage earners preferred trucks, 54 percent of mid-level managers preferred trucks, and 26 percent of executives preferred trucks.

69. We are interested in finding the 95 percent confidence interval for the percentage of executives who prefer trucks. Define random variables X and P' in words.

70. Which distribution should you use for this problem?

71. Construct a 95 percent confidence interval. State the confidence interval, sketch the graph, and calculate the error bound.

72. Suppose we want to lower the sampling error. What is one way to accomplish that?

73. The sampling error given in the survey is ±2 percent. Explain what the ±2 percent means.

Use the following information to answer the next five exercises: A poll of 1,200 voters asked what the most significant issue was in the upcoming election. Sixty-five percent answered "the economy." We are interested in the population proportion of voters who believe the economy is the most important.

74. Define the random variable X in words.

75. Define the random variable P' in words.

76. Which distribution should you use for this problem?

77. Construct a 90 percent confidence interval, and state the confidence interval and the error bound.

78. What would happen to the confidence interval if the level of confidence were 95 percent?

Use the following information to answer the next 16 exercises: The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 p.m., Monday night, ages 8 to 12, beginning ice-skating class is picked. In that class are 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 to 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class are a random sample of the population.

79. What is being counted?

80. In words, define the random variable X.

81. Calculate the following:

a. $x =$ _____
b. $n =$ _____
c. $p' =$ _____

82. State the estimated distribution of X. $X \sim$ _____

83. Define a new random variable P'. What is p' estimating?

84. In words, define the random variable P'.

85. State the estimated distribution of P'. Construct a 92 percent confidence interval for the true proportion of girls in the ages 8 to 12 beginning ice-skating classes at the Ice Chalet.

86. How much area is in both tails (combined)?

87. How much area is in each tail?

88. Calculate the following:

a. lower limit
b. upper limit
c. error bound

89. The 92 percent confidence interval is _____.

90. Fill in the blanks on the graph with the areas, upper and lower limits of the confidence interval, and the sample proportion.

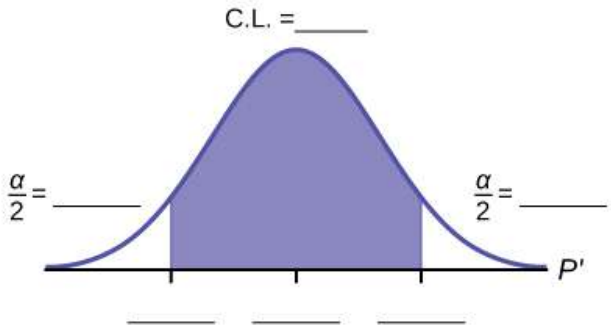


Figure 8.11

91. In one complete sentence, explain what the interval means.

92. Using the same p' and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

93. Using the same p' and $n = 80$, how would the error bound change if the confidence level were increased to 98 percent? Why?

94. If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

HOMEWORK

8.1 A Single Population Mean Using the Normal Distribution

95. Among various ethnic groups, the standard deviation of heights is known to be approximately three inches. We wish to construct a 95 percent confidence interval for the mean height of male Swedes. 48 male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 in.

a. i. $\bar{x}$ =_____
   ii. σ =_____
   iii. n =_____

b. In words, define the random variables X and $\bar{X}$.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95 percent confidence interval for the population mean height of male Swedes.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. What will happen to the level of confidence obtained if 1,000 male Swedes are surveyed instead of 48? Why?

96. Announcements for 84 upcoming engineering conferences were randomly picked from a stack of *IEEE Spectrum* magazines. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume the underlying population is normal.

a. In words, define the random variables X and $\bar{X}$.

b. Which distribution should you use for this problem? Explain your choice.

c. Construct a 95 percent confidence interval for the population mean length of engineering conferences.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

97. Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

a. i. $\bar{x}$ =_____
   ii. σ =_____
   iii. n =_____

b. In words, define the random variables X and $\bar{X}$.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 90 percent confidence interval for the population mean time to complete the tax forms.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, which changes should it make?

f. If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?

g. Suppose that the firm decided that it needed to be at least 96 percent confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

98. A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was two ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

a. i. $\bar{x}$ =_____
   ii. σ =_____
   iii. s_x =_____

b. In words, define the random variable X.

c. In words, define the random variable $\bar{X}$.

d. Which distribution should you use for this problem? Explain your choice.

e. Construct a 90 percent confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

f. Construct a 98 percent confidence interval for the population mean weight of the candies.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

g. In complete sentences, explain why the confidence interval in Part f is larger than the confidence interval in Part e.

h. In complete sentences, give an interpretation of what the interval in Part f means.

99. A camp director is interested in the mean number of letters each child sends during his or her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9, with a sample standard deviation of 2.8.

a. i. $\bar{x}$ =_____
   ii. σ =_____
   iii. n =_____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 90 percent confidence interval for the population mean number of letters campers send home.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?

100. What is meant by the term *90 percent confident* when constructing a confidence interval for a mean?

a. If we took repeated samples, approximately 90 percent of the samples would produce the same confidence interval.

b. If we took repeated samples, approximately 90 percent of the confidence intervals calculated from those samples would contain the sample mean.

c. If we took repeated samples, approximately 90 percent of the confidence intervals calculated from those samples would contain the true value of the population mean.

d. If we took repeated samples, the sample mean would equal the population mean in approximately 90 percent of the samples.

101. The Federal Election Commission collects information about campaign contributions and disbursements for candidates and political committees during each election cycle. During the 2012 campaign season, there were 1,619 candidates for the House of Representatives across the United States who received contributions from individuals. **Table 8.11** shows the total receipts from individuals for a random selection of 40 House candidates rounded to the nearest $100. The standard deviation for this data to the nearest hundred is σ = $909,200.

$3,600	$1,243,900	$10,900	$385,200	$581,500
$7,400	$2,900	$400	$3,714,500	$632,500
$391,000	$467,400	$56,800	$5,800	$405,200
$733,200	$8,000	$468,700	$75,200	$41,000
$13,300	$9,500	$953,800	$1,113,500	$1,109,300
$353,900	$986,100	$88,600	$378,200	$13,200
$3,800	$745,100	$5,800	$3,072,100	$1,626,700
$512,900	$2,309,200	$6,600	$202,400	$15,800

Table 8.11

a. Find the point estimate for the population mean.
b. Using 95 percent confidence, calculate the error bound.
c. Create a 95 percent confidence interval for the mean total individual contributions.
d. Interpret the confidence interval in the context of the problem.

102. The American Community Survey (ACS), part of the U.S. Census Bureau, conducts a yearly census similar to the one taken every 10 years, but with a smaller percentage of participants. The most recent survey estimates with 90 percent confidence that the mean household income in the United States falls between $69,720 and $69,922. Find the point estimate for mean U.S. household income and the error bound for mean U.S. household income.

103. The average height of young adult males has a normal distribution with standard deviation of 2.5 in. You want to estimate the mean height of students at your college or university to within 1 in. with 93 percent confidence. How many male students must you measure?

8.2 A Single Population Mean Using the Student's t-Distribution

104. In six packages of multicolored fruit snacks, there were five red snack pieces. The total number of snack pieces in the six bags was 68. We wish to calculate a 96 percent confidence interval for the population proportion of red snack pieces.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Calculate p'.
d. Construct a 96 percent confidence interval for the population proportion of red snack pieces per bag.
    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.
e. Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

105. A random survey of enrollment at 35 community colleges across the United States yielded the following figures: 6,414, 1,550, 2,109, 9,350, 21,828, 4,300, 5,944, 5,722, 2,825, 2,044, 5,481, 5,200, 5,853, 2,750, 10,012, 6,357, 27,000, 9,414, 7,681, 3,200, 17,500, 9,200, 7,380, 18,314, 6,557, 13,713, 17,768, 7,493, 2,771, 2,861, 1,263, 7,285, 28,165, 5,080, 11,622. Assume the underlying population is normal.

a. i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95 percent confidence interval for the population mean enrollment at community colleges in the United States.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. What will happen to the error bound and confidence interval if 500 community colleges are surveyed? Why?

106. Suppose that a committee is studying whether there is wasted time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for jury duty. The committee randomly surveyed 81 people who recently served as jurors. The sample mean wait time was 8 hr, with a sample standard deviation of 4 hr.

a. i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 95 percent confidence interval for the population mean time wasted.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

e. Explain in a complete sentence what the confidence interval means.

107. A pharmaceutical company makes a drug used during surgery. It is assumed that the distribution for the length of time the drug lasts is approximately normal. Researchers in a hospital used the drug on a random sample of nine patients. The effective period of the antibiotic drug for each patient (in hours) was as follows: 2.7, 2.8, 3.0, 2.3, 2.3, 2.2, 2.8, 2.1, and 2.4.

a. i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variable X in words.

c. Define the random variable $\bar{X}$ in words.

d. Which distribution should you use for this problem? Explain your choice.

e. Construct a 95 percent confidence interval for the population mean length of time.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

f. What does it mean to be *95 percent confident* in this problem?

108. Suppose that 14 children who were learning to ride two-wheel bikes were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of six months, with a sample standard deviation of three months. Assume that the underlying population distribution is normal.

a. i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____
b. Define the random variable X in words.
c. Define the random variable $\bar{X}$ in words.
d. Which distribution should you use for this problem? Explain your choice.
e. Construct a 99 percent confidence interval for the population mean length of time using training wheels.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
f. Why would the error bound change if the confidence level were lowered to 90 percent?

109. The Federal Election Commission (FEC) collects information about campaign contributions and disbursements for candidates and political committees during each election cycle. A political action committee (PAC) is a committee formed to raise money for candidates and campaigns. A Leadership PAC is a PAC formed by a federal politician (senator or representative) to raise money to help other candidates' campaigns.

The FEC has reported financial information for 556 Leadership PACs that operated during the 2011–2012 election cycle. The following table shows the total receipts during this cycle for a random selection of 20 Leadership PACs.

$46,500.00	$0	$40,966.50	$105,887.20	$5,175.00
$29,050.00	$19,500.00	$181,557.20	$31,500.00	$149,970.80
$2,555,363.20	$12,025.00	$409,000.00	$60,521.70	$18,000.00
$61,810.20	$76,530.80	$119,459.20	$0	$63,520.00
$6,500.00	$502,578.00	$705,061.10	$708,258.90	$135,810.00
$2,000.00	$2,000.00	$0	$1,287,933.80	$219,148.30

Table 8.12

$\bar{x} = \$251,\ 854.23$

$s = \$521,\ 130.41$

Use the sample data to construct a 96 percent confidence interval for the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle. Use the Student's *t*-distribution.

110. A major business magazine published data on the best small firms in 2012. These were firms that have been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. **Table 8.13** shows the ages of the corporate CEOs for a random sample of these firms.

48	58	51	61	56
59	74	63	53	50
59	60	60	57	46
55	63	57	47	55
57	43	61	62	49
67	67	55	55	49

Table 8.13

Use the sample data to construct a 90 percent confidence interval for the mean age of CEOs for these top small firms. Use the Student's *t*-distribution.

111. Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected, and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats, and the sample standard deviation is 4.1 seats.

a. i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variables X and $\bar{X}$ in words.

c. Which distribution should you use for this problem? Explain your choice.

d. Construct a 92 percent confidence interval for the population mean number of unoccupied seats per flight.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

112. In a recent sample of 84 used car sales costs, the sample mean was $6,425, with a standard deviation of $3,156. Assume the underlying distribution is approximately normal.

a. Which distribution should you use for this problem? Explain your choice.

b. Define the random variable $\bar{X}$ in words.

c. Construct a 95 percent confidence interval for the population mean cost of a used car.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

d. Explain what a *95 percent confidence interval* means for this study.

113. Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8, 8, 10, 7, 9, 9. Assume the underlying distribution is approximately normal.

a. Construct a 90 percent confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.

b. If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?

c. Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.

d. Calculate the mean.

e. Is the mean within the interval you calculated in Part a? Did you expect it to be? Why or why not?

114. A survey of the mean number of cents off given by coupons was conducted by randomly surveying one coupon per page from the coupons section of a local newspaper. The following data were collected: 20¢, 75¢, 50¢, 65¢, 30¢, 55¢, 40¢, 40¢, 30¢, 55¢, $1.50, 40¢, 65¢, 40¢. Assume the underlying distribution is approximately normal.

a. i. $\bar{x}$ = _____
   ii. s_x = _____
   iii. n = _____
   iv. $n - 1$ = _____

b. Define the random variables X and $\bar{X}$ in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95 percent confidence interval for the population mean worth of coupons.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If many random samples were collected with 14 samples as the size, which percentage of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

Use the following information to answer the next two exercises: A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16-oz serving size. The sample mean is 13.30, with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

115. Find the 95 percent confidence interval for the true population mean for the amount of soda served.
a. (12.42, 14.18)
b. (12.32, 14.29)
c. (12.50, 14.10)
d. Impossible to determine

116. Which of the following is the error bound?
a. 0.87
b. 1.98
c. 0.99
d. 1.74

8.3 A Population Proportion

117. Insurance companies are interested in knowing the population percentage of drivers who always buckle up before riding in a car.
a. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95 percent confident that the population proportion is estimated to within 0.03?
b. If it were later determined that it was important to be more than 95 percent confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

118. Suppose that the insurance companies did conduct a survey. They randomly surveyed 400 drivers and found that 320 claimed they always buckle up. We are interested in the population proportion of drivers who claim they always buckle up.

a. i. x = _____
   ii. n = _____
   iii. p' = _____
b. Define the random variables X and P' in words.
c. Which distribution should you use for this problem? Explain your choice.
d. Construct a 95 percent confidence interval for the population proportion who claim they always buckle up.
   i. State the confidence interval.
   ii. Sketch the graph.
   iii. Calculate the error bound.
e. If this survey were done by telephone, list three difficulties the companies might have in obtaining random results.

119. According to a recent survey of 1,200 people, 61 percent believe that the president is doing an acceptable job. We are interested in the population proportion of people who believe the president is doing an acceptable job.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 90 percent confidence interval for the population proportion of people who believe the president is doing an acceptable job.
    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.

120. An article regarding dating and marriage recently appeared in a major newspaper. Of the 1,709 randomly selected adults, 315 identified themselves as ethnicity A, 323 identified themselves as ethnicity B, 254 identified themselves as ethnicity C, and 779 identified themselves as ethnicity D. In this survey, 86 percent of ethnicity B said that they would welcome a person of ethnicity A into their families. Among ethnicity C, 77 percent would welcome a person of ethnicity D into their families, 71 percent would welcome a person of ethnicity A, and 66 percent would welcome a person of ethnicity B.

a. We are interested in finding the 95 percent confidence interval for the percent of all ethnicity B adults who would welcome a person of ethnicity D into their families. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95 percent confidence interval.
    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.

121. Refer to the information in **Exercise 8.120**.

a. Construct three 95 percent confidence intervals:
    i. percentage of all ethnicity C who would welcome a person of ethnicity D into their families
    ii. percentage of all ethnicity C who would welcome a person of ethnicity A into their families
    iii. percentage of all ethnicity C who would welcome a person of ethnicity B into their families
b. Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
c. For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
d. For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

122. Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5 percent of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight year period.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 97 percent confidence interval for the population proportion of people over 50 who ran and died in the same 8-year period.
    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.
d. Explain what a *97 percent confidence interval* means for this study.

123. A telephone poll of 1,000 adult Americans was reported in an issue of a national magazine. One of the questions asked, "What is the main problem facing the country?" Twenty percent responded "crime". We are interested in the population proportion of adult Americans who believe that crime is the main problem.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95 percent confidence interval for the population proportion of adult Americans who believe that crime is the main problem.
    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.
d. Suppose we want to lower the sampling error. What is one way to accomplish that?
e. The sampling error given by the group of researchers who conducted the poll is ±3 percent. In one to three complete sentences, explain what the ±3 percent represents.

124. Refer to **Exercise 8.123**. Another question in the poll asked, "[How much are] you worried about the quality of education in our schools?" Sixty-three percent responded "a lot". We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

a. Define the random variables X and P' in words.
b. Which distribution should you use for this problem? Explain your choice.
c. Construct a 95 percent confidence interval for the population proportion of adult Americans who are worried a lot about the quality of education in our schools.
    i. State the confidence interval.
    ii. Sketch the graph.
    iii. Calculate the error bound.
d. The sampling error given by the group of researchers who conducted the poll is ±3 percent. In one to three complete sentences, explain what the ±3 percent represents.

Use the following information to answer the next three exercises: According to a Field Poll, 79 percent of California adults (actual results are 400 out of 506 surveyed) believe that *education and our schools* is one of the top issues facing California. We wish to construct a 90 percent confidence interval for the true proportion of California adults who believe that education and the schools is one of the top issues facing California.

125. A point estimate for the true population proportion is _____.

a. 0.90
b. 1.27
c. 0.79
d. 400

126. A 90 percent confidence interval for the population proportion is _____.

a. (0.761, 0.820)
b. (0.125, 0.188)
c. (0.755, 0.826)
d. (0.130, 0.183)

127. The error bound is approximately _____.

a. 1.581
b. 0.791
c. 0.059
d. 0.030

Use the following information to answer the next two exercises: Five hundred eleven (511) homes in a certain southern California community are randomly surveyed to determine whether they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed meet the minimum recommendations for earthquake preparedness, and 338 do not.

128. Find the confidence interval at the 90 percent confidence level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.

a. (0.2975, 0.3796)
b. (0.6270, 0.6959)
c. (0.3041, 0.3730)
d. (0.6204, 0.7025)

129. The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is _____.

a. 0.6614
b. 0.3386
c. 173
d. 338

130. On May 23, 2013, a polling group reported that of the 1,005 people surveyed, 76 percent of U.S. workers believe that they will continue working past retirement age. The confidence level for this study was reported at 95 percent with a ±3 percent margin of error.

a. Determine the estimated proportion from the sample.
b. Determine the sample size.
c. Identify *CL* and α.
d. Calculate the error bound based on the information provided.
e. Compare the error bound in Part d to the margin of error reported by the polling group. Explain any differences between the values.
f. Create a confidence interval for the results of this study.
g. A reporter is covering the release of this study for a local news station. How should she explain the confidence interval to her audience?

131. A national survey of 1,000 adults was conducted on May 13, 2013, by a group of researchers. It concluded with 95 percent confidence that 49 percent to 55 percent of Americans believe that big-time college sports programs corrupt the process of higher education.

a. Find the point estimate and the error bound for this confidence interval.
b. Can we (with 95 percent confidence) conclude that more than half of all American adults believe this?
c. Use the point estimate from Part a and $n = 1{,}000$ to calculate a 75 percent confidence interval for the proportion of American adults who believe that major college sports programs corrupt higher education.
d. Can we (with 75 percent confidence) conclude that at least half of all American adults believe this?

132. A polling group recently conducted a survey asking adults across the United States about music preferences. When asked, 80 of the 571 participants download music weekly.

a. Create a 99 percent confidence interval for the true proportion of American adults who download music weekly.
b. This survey was conducted through automated telephone interviews on May 6 and 7, 2013. The error bound of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the survey's outcome that are not covered by the margin of error.
c. Without performing any calculations, describe how the confidence interval would change if the confidence level decreased from 99 percent to 90 percent.

133. You plan to conduct a survey on your college campus to learn about the political awareness of students. You want to estimate the true proportion of college students on your campus who voted in the 2012 presidential election with 95 percent confidence and a margin of error no greater than 5 percent. How many students must you interview?

134. In a recent poll, 9 of 48 respondents rated the likelihood of a certain event occurring in their community as *likely* or *very likely*. Use the plus-four method to create a 97 percent confidence interval for the proportion of American adults who believe that the event is likely or very likely. Explain what this confidence interval means in the context of the problem. A local poll in a New England town found that nine of 48 households think winter-proofing their cars is *very important*. Use the plus-four method to create a 97 percent confidence interval for the proportion of town residents who think winter-proofing their cars is *very important*. Explain what this confidence interval means in the context of this scenario.

REFERENCES

8.1 A Single Population Mean Using the Normal Distribution

Centers for Disease Control and Prevention. (n.d.). *National health and nutrition examination survey*. Retrieved from http://www.cdc.gov/nchs/nhanes.htm

Foothill De Anza Community College District. (n.d.). *Headcount enrollment trends by student demographics ten-year fall trends to most recently completed fall*. Retrieved from http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm

Kuczmarski, R. J., et al. (2002). *2000 CDC growth charts for the United States: Methods and development*. Retrieved from http://www.cdc.gov/growthcharts/2000growthchart-us.pdf

La, L., and; German, K. (n.d.). Cell phones with the highest radiation levels. *CNET*. Retrieved from http://reviews.cnet.com/cell-phone-radiation-levels/

U.S. Census Bureau. (n.d.). *American FactFinder*. Retrieved from http://factfinder2.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

U.S. Census Bureau. (2011). *Mean income in the past 12 months (in 2011 inflation-adjusted dollars): 2011 American Community Survey 1-year estimates*. Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1902&prodType=table

U.S. Federal Election Commission. (n.d.). *Disclosure data catalog: Candidate summary report 2012*. Retrieved from http://www.fec.gov/data/CandidateSummary.do?format=html&election_yr=2012

U.S. Federal Election Commission. (n.d.). *Metadata description of candidate summary file*. Retrieved from http://www.fec.gov/finance/disclosure/metadata/metadataforcandidatesummary.shtml

8.2 A Single Population Mean Using the Student's t-Distribution

Bloomberg Businessweek. (n.d.). Retrieved from http://www.businessweek.com/

Environmental Working Group. (n.d.). *Human toxome project: Mapping the pollution in people*. Retrieved from http://www.ewg.org/sites/humantoxome/participants/participant-group.php?group=in+utero%2Fnewborn

Federal Election Commission. (n.d.). *Disclosure data catalog: 2012 leadership PACs and sponsors*. Retrieved from http://www.fec.gov/data/index.jsp

Federal Election Commission. (n.d.). *Metadata description of leadership PAC list*. Retrieved from http://www.fec.gov/finance/disclosure/metadata/metadataLeadershipPacList.shtml

Forbes. (2013). America’s best small companies. Retrieved from http://www.forbes.com/best-small-companies/list/

Forbes. (n.d.). Retrieved from http://www.forbes.com/

Microsoft Bookshelf. (n.d.).

8.3 A Population Proportion

Jensen, T. (2013). *Democrats, Republicans divided on opinion of music icons*. Retrieved from http://www.publicpolicypolling.com/Day2MusicPoll.pdf

Madden, M., et al. (2013). *Teens, social media, and privacy*. Retrieved from http://www.pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy.aspx

Princeton Survey Research Associates International. (2012). *2012 teens and privacy management survey*. Retrieved from http://www.pewinternet.org/~/media//Files/Questionnaire/2013/Methods%20and%20Questions_Teens%20and%20Social%20Media.pdf

Rasmussen Reports. (2013). *52% say big-time college athletics corrupt education process*. Retrieved from http://www.rasmussenreports.com/public_content/lifestyle/sports/may_2013/52_say_big_time_college_athletics_corrupt_education_process

Saad, L. (2013). *Three in four U.S. workers plan to work past retirement age*. Retrieved from http://www.gallup.com/poll/162758/three-fourworkers-plan-work-past-retirement-age.aspx

The Field Poll. (n.d.). Retrieved from http://field.com/fieldpollonline/subscribers/

Zogby Analytics. (2013). *New SUNYIT/Zogby analytics poll: Few Americans worry about emergency situations occurring in their community; Only one in three have an emergency plan; 70% support infrastructure "investment" for national security*. Retrieved from http://www.zogbyanalytics.com/news/299-americans-neither-worried-norprepared-in-case-of-a-disaster-sunyit-zogby-analytics-poll

SOLUTIONS

1

a. 244

b. 15

c. 50

3 $N\left(244, \frac{15}{\sqrt{50}}\right)$

5 As the sample size increases, there will be less variability in the mean, so the interval size decreases.

7 X is the time in minutes it takes to complete the U.S. Census short form. $\bar{X}$ is the mean time it took a sample of 200 people to complete the U.S. Census short form.

9 CI: (7.9441, 8.4559)

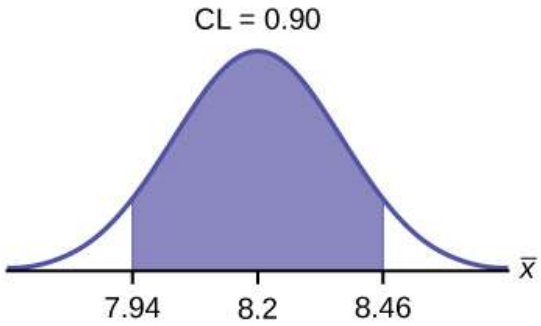


Figure 8.12

EBM = 0.26

11 The level of confidence would decrease, because decreasing n makes the confidence interval wider, so at the same error bound, the confidence level decreases.

13

a. $\bar{x}$ = 2.2

b. σ = 0.2

c. n = 20

15 $\bar{X}$ is the mean weight of a sample of 20 heads of lettuce.

17 *EBM* = 0.07
CI: (2.1264, 2.2736)

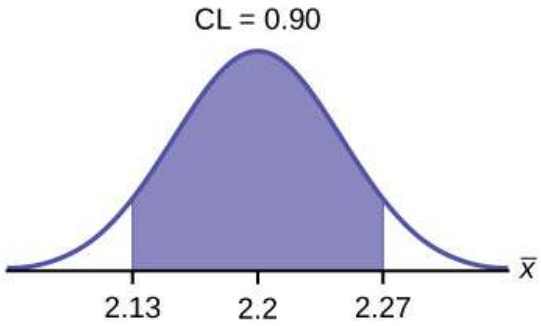


Figure 8.13

19 The interval is greater, because the level of confidence increased. If the only change made in the analysis is a change in confidence level, then all we are doing is changing how much area is being calculated for the normal distribution. Therefore, a larger confidence level results in larger areas and larger intervals.

21 The confidence level would increase.

23 30.4

25 σ

27 μ

29 normal

31 0.025

33 (24.52,36.28)

35 We are 95 percent confident that the true mean age for winter Foothill College students is between 24.52 and 36.28.

37 The error bound for the mean would decrease, because as the CL decreases, you need less area under the normal curve (which translates into a smaller interval) to capture the true population mean.

39 X is the number of hours a patient waits in the emergency room before being called back to be examined. $\bar{X}$ is the mean wait time of 70 patients in the emergency room.

41 CI: (1.3808, 1.6192)

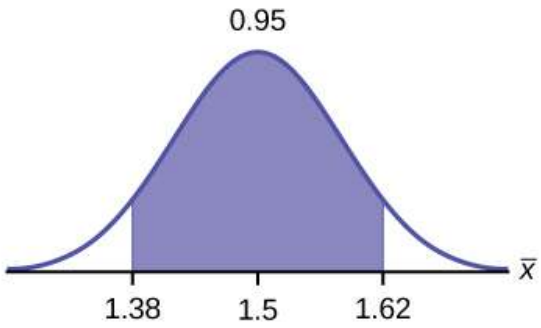


Figure 8.14

$EBM = 0.12$

43

a. $\bar{x} = 151$

b. $s_x = 32$

c. $n = 108$

d. $n - 1 = 107$

45 $\bar{X}$ is the mean number of hours spent watching television per month from a sample of 108 Americans.

47 CI: (142.92, 159.08)

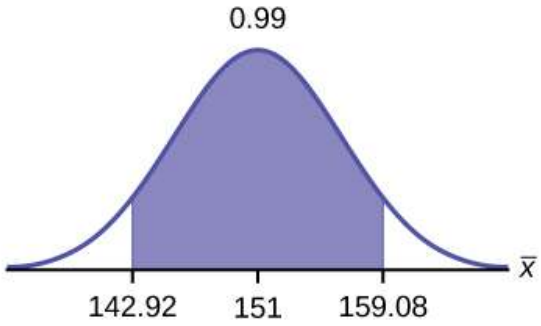


Figure 8.15

EBM = 8.08

49

a. 3.26

b. 1.02

c. 39

51 μ

53 t_{38}

55 0.025

57 (2.93, 3.59)

59 We are 95 percent confident that the true mean number of colors for national flags is between 2.93 colors and 3.59 colors.

60 The error bound would become EBM = 0.245. This error bound decreases, because as sample sizes increase, variability decreases, and we need less interval length to capture the true mean.

63 It would decrease, because the *z*-score would decrease, which would reduce the numerator and lower the number.

65 X is the number of *successes* where the woman makes the majority of the purchasing decisions for the household. P' is the percentage of households sampled where the woman makes the majority of the purchasing decisions for the household.

67 CI: (0.5321, 0.6679)

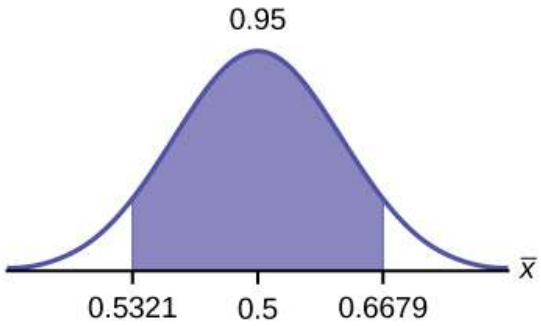


Figure 8.16

EBM: 0.0679

69 X is the number of *successes* where an executive prefers a truck. P' is the percentage of executives sampled who prefer a truck.

71 CI: (0.19432, 0.33068)

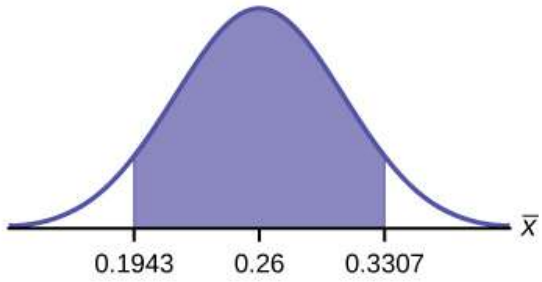


Figure 8.17

EBM: 0.0707

73 The sampling error means that the true mean can be 2 percent above or below the sample mean.

75 P' is the proportion of voters sampled who said the economy is the most important issue in the upcoming election.

77 CI: (0.62735, 0.67265); EBM: 0.02265

79 the number of girls, ages 8 to 12, in the 5 p.m. Monday night beginning ice-skating class

81

a. $x = 64$

b. $n = 80$

c. $p' = 0.8$

83 p

85 $P' \sim N\left(0.8, \sqrt{\frac{(0.8)(0.2)}{80}}\right)$ CI = (0.72171, 0.87829).

87 0.04

89 (0.72; 0.88)

91 With 92 percent confidence, we estimate the proportion of girls, ages 8 to 12, in a beginning ice-skating class at the Ice Chalet to be between 72 percent and 88 percent.

93 The error bound would increase. Assuming all other variables are kept constant, as the confidence level increases, the area under the curve corresponding to the confidence level becomes larger, which creates a wider interval and thus a larger

error.

95

a. i. 71

ii. 3

iii. 48

b. X is the height of a Swedish male, and is the mean height from a sample of 48 Swedish males.

c. Normal. We know the standard deviation for the population, and the sample size is greater than 30.

d. i. CI: (70.151, 71.49)

ii.

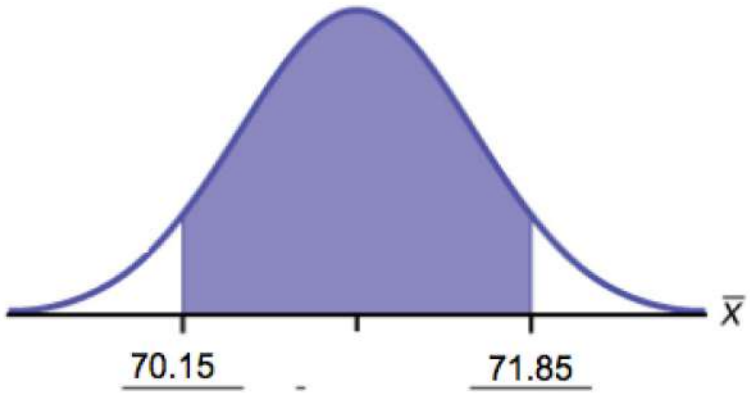


Figure 8.18

iii. EBM = 0.849

e. The confidence interval will decrease in size, because the sample size increased. Recall, when all factors remain unchanged, an increase in sample size decreases variability. Thus, we do not need as large an interval to capture the true population mean.

97

a. i. $\bar{x}$ = 23.6

ii. σ = 7

iii. n = 100

b. X is the time needed to complete an individual tax form. $\bar{X}$ is the mean time to complete tax forms from a sample of 100 customers.

c. $N\left(23.6, \frac{7}{\sqrt{100}}\right)$ because we know sigma.

d. ii. (22.228, 24.972)

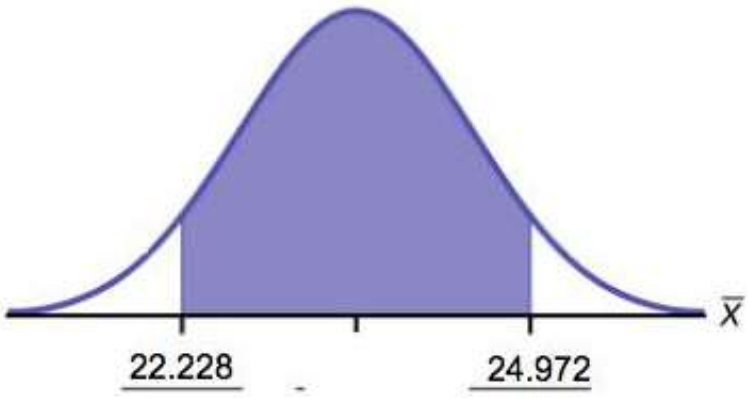


Figure 8.19

iii. *EBM* = 1.372

e. It will need to change the sample size. The firm needs to determine what the confidence level should be and then apply the error bound formula to determine the necessary sample size.

f. The confidence level would increase as a result of a larger interval. Smaller sample sizes result in more variability. To capture the true population mean, we need to have a larger interval.

g. According to the error bound formula, the firm needs to survey 206 people. Because we increase the confidence level, we need to increase either our error bound or the sample size.

99

a. i. 7.9

ii. 2.5

iii. 20

b. X is the number of letters a single camper will send home. $\bar{X}$ is the mean number of letters sent home from a sample of 20 campers.

c. $N\ 7.9\left(\frac{2.5}{\sqrt{20}}\right)$

d. i. CI: (6.98, 8.82)

ii.

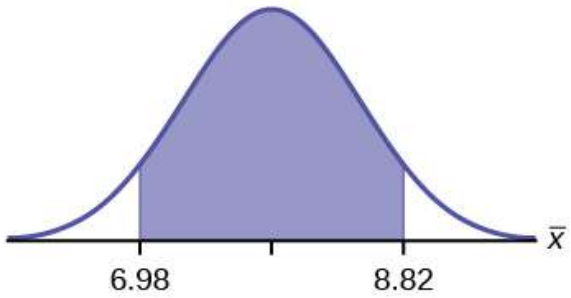


Figure 8.20

iii. *EBM*: 0.92

e. The error bound and confidence interval will decrease.

101

a. $\bar{x}$ = \$568,873

b. $CL = 0.95$, $\alpha = 1 - 0.95 = 0.05$, $z_{\frac{\alpha}{2}} = 1.96$

$EBM = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.96\ \frac{909200}{\sqrt{40}} = \$281{,}764$

c. $\bar{x} - EBM = 568{,}873 - 281{,}764 = 287{,}109$

$\bar{x} + EBM = 568{,}873 + 281{,}764 = 850{,}637$

Alternate solution:

Using the TI-83, 83+, 84, 84+ Calculator

1. Press STAT and arrow over to TESTS.
2. Arrow down to 7:ZInterval.
3. Press ENTER.
4. Arrow to Stats and press ENTER.
5. Arrow down and enter the following values:

   σ: 909,200

   $\bar{x}$: 568,873

   n: 40

   CL: 0.95
6. Arrow down to Calculate and press ENTER.
7. The confidence interval is (\$287,114, \$850,632).
8. Notice the small difference between the two solutions—these differences are simply due to rounding error in the hand calculations.

d. We estimate with 95 percent confidence that the mean amount of contributions received from all individuals by House candidates is between \$287,109 and \$850,637.

103 Use the formula for *EBM*, solved for n:

$n = \frac{z^2\sigma^2}{EBM^2}$ From the statement of the problem, you know that $\sigma = 2.5$, and you need $EBM = 1$. $z = z_{0.035} = 1.812$. (This is the value of z for which the area under the density curve to the *right* of z is 0.035.)

$n = \frac{z^2\sigma^2}{EBM^2} = \frac{1.812^2 2.5^2}{1^2} \approx 20.52$. You need to measure at least 21 male students to achieve your goal.

105

a. i. 8,629

ii. 6,944

iii. 35

iv. 34

b. t_{34}

c. i. CI: (6244, 11,014)

ii.

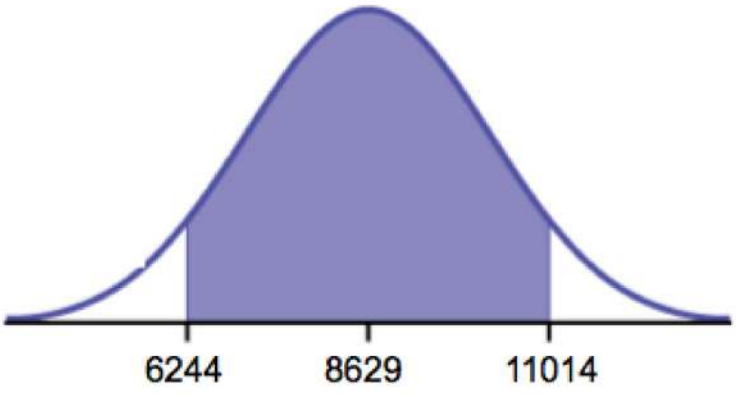


Figure 8.21

iii. EB = 2385

d. It will become smaller.

107

a. i. $\bar{x}$ = 2.51

ii. s_x = 0.318

iii. n = 9

iv. n - 1 = 8

b. The effective length of time for a tranquilizer

c. The mean effective length of time of tranquilizers from a sample of nine patients

d. We need to use a Student's *t*-distribution, because we do not know the population standard deviation.

e. i. CI: (2.27, 2.76)

ii. Check student's solution.

iii. *EBM*: 0.25

f. If we were to sample many groups of nine patients, 95 percent of the samples would contain the true population mean length of time.

109 $\bar{x} = \$251,854.23$; $s = \$521,130.41$. Note that we are not given the population standard deviation, only the standard deviation of the sample. There are 30 measures in the sample, so n = 30, and df = 30 - 1 = 29. CL = 0.96, so α = 1 - CL = 1 - 0.96 = 0.04. $\frac{\alpha}{2} = 0.02 t_{\frac{\alpha}{2}} = t_{0.02}$ = 2.150. $EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = 2.150\left(\frac{521,130.41}{\sqrt{30}}\right) \sim \$204,561.66$. $\bar{x}$ - *EBM* = \$251,854.23 - \$204,561.66 = \$47,292.57. $\bar{x}$ + *EBM* = \$251,854.23 + \$204,561.66 = \$456,415.89. We estimate with 96 percent confidence that the mean amount of money raised by all Leadership PACs during the 2011–2012 election cycle lies between \$47,292.57 and \$456,415.89.

Alternate Solution

Using the TI-83, 83+, 84, 84+ Calculator

STATTESTS8:TIntervalENTERENTERFreqC-LevelCalculateEnter

The difference between solutions arises from rounding differences.

111

i. $\bar{x}$ =

ii. s_x =

iii. n =

iv. n - 1 =

b. X is the number of unoccupied seats on a single flight. $\bar{X}$ is the mean number of unoccupied seats from a sample of 225 flights.

c. We will use a Student's t-distribution, because we do not know the population standard deviation.

d. i. CI: (11.12 , 12.08)

ii. Check student's solution.

iii. *EBM*: 0.48

113

a. i. CI: (7.64, 9.36)

ii.

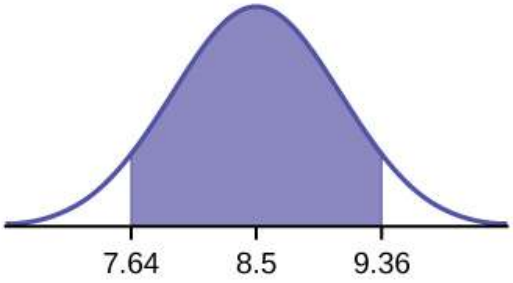


Figure 8.22

iii. *EBM*: 0.86

b. The sample should have been increased.

c. Answers will vary.

d. Answers will vary.

e. Answers will vary.

115 b

117

a. 1,068

b. The sample size would need to be increased, because the critical value increases as the confidence level increases.

119

a. X = the number of people who believe that the president is doing an acceptable job;

P' = the proportion of people in a sample who believe that the president is doing an acceptable job.

b. $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$

c. i. CI: (0.59, 0.63)

ii. Check student's solution.

iii. *EBM*: 0.02

121

a. i. (0.72, 0.82)

ii. (0.65, 0.76)

iii. (0.60, 0.72)

b. Yes, the intervals (0.72, 0.82) and (0.65, 0.76) overlap, and the intervals (0.65, 0.76) and (0.60, 0.72) overlap.

c. We can say that there does not appear to be a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a Latino person into their families.

d. We can say that there is a significant difference between the proportion of Asian adults who say that their families would welcome a white person into their families and the proportion of Asian adults who say that their families would welcome a black person into their families.

123

a. X = the number of adult Americans who believe that crime is the main problem; P' = the proportion of adult Americans who believe that crime is the main problem.

b. Because we are estimating a proportion, that $P' = 0.2$ and $n = 1{,}000$, the distribution we should use is $N\left(0.2, \sqrt{\frac{(0.2)(0.8)}{1000}}\right)$.

c. i. CI: (0.18, 0.22)

ii. Check student's solution.

iii. *EBM*: 0.02

d. One way to lower the sampling error is to increase the sample size.

e. The stated ± *3 percent* represents the maximum error bound. This means that those doing the study are reporting a maximum error of 3 percent. Thus, they estimate the percentage of adult Americans who the percentage of adult Americans who that crime is the main problem to be between 18 percent and 22 percent.

125 c

127 d

129 a

131

a. $p' = \frac{(0.55 + 0.49)}{2} = 0.52$; $EBP = 0.55 - 0.52 = 0.03$

b. No, the confidence interval includes values less than or equal to 0.50. It is possible that less than half of the population believe this.

c. $CL = 0.75$, so $\alpha = 1 - 0.75 = 0.25$ and $\frac{\alpha}{2} = 0.125$. $z_{\frac{\alpha}{2}} = 1.150$. (The area to the right of this z is 0.125, so the area to the left is $1 - 0.125 = 0.875$.)

$EBP = (1.150)\sqrt{\frac{0.52(0.48)}{1,\ 000}} \approx 0.018$

$(p' - EBP, p' + EBP) = (0.52 - 0.018, 0.52 + 0.018) = (0.502, 0.538)$

Alternate Solution

Using the TI-83, 83+, 84, 84+ Calculator

STAT TESTS A: 1-PropZinterval with $x = (0.52)(1{,}000)$, $n = 1{,}000$, CL = 0.75.

Answer is (0.502, 0.538).

d. Yes, this interval does not fall below 0.50, so we can conclude that at least half of all American adults believe that major sports programs corrupt education – but we do so with only 75 percent confidence.

133 $CL = 0.95$; $\alpha = 1 - 0.95 = 0.05$; $\frac{\alpha}{2} = 0.025$; $z_{\frac{\alpha}{2}} = 1.96$. Use $p' = q' = 0.5$.

$n = \frac{z_{\frac{\alpha}{2}}^2 p' q'}{EBP^2} = \frac{1.96^2(0.5)(0.5)}{0.05^2} = 384.16$. You need to interview at least 385 students to estimate the proportion to within 5 percent at 95 percent confidence.

9 | HYPOTHESIS TESTING WITH ONE SAMPLE

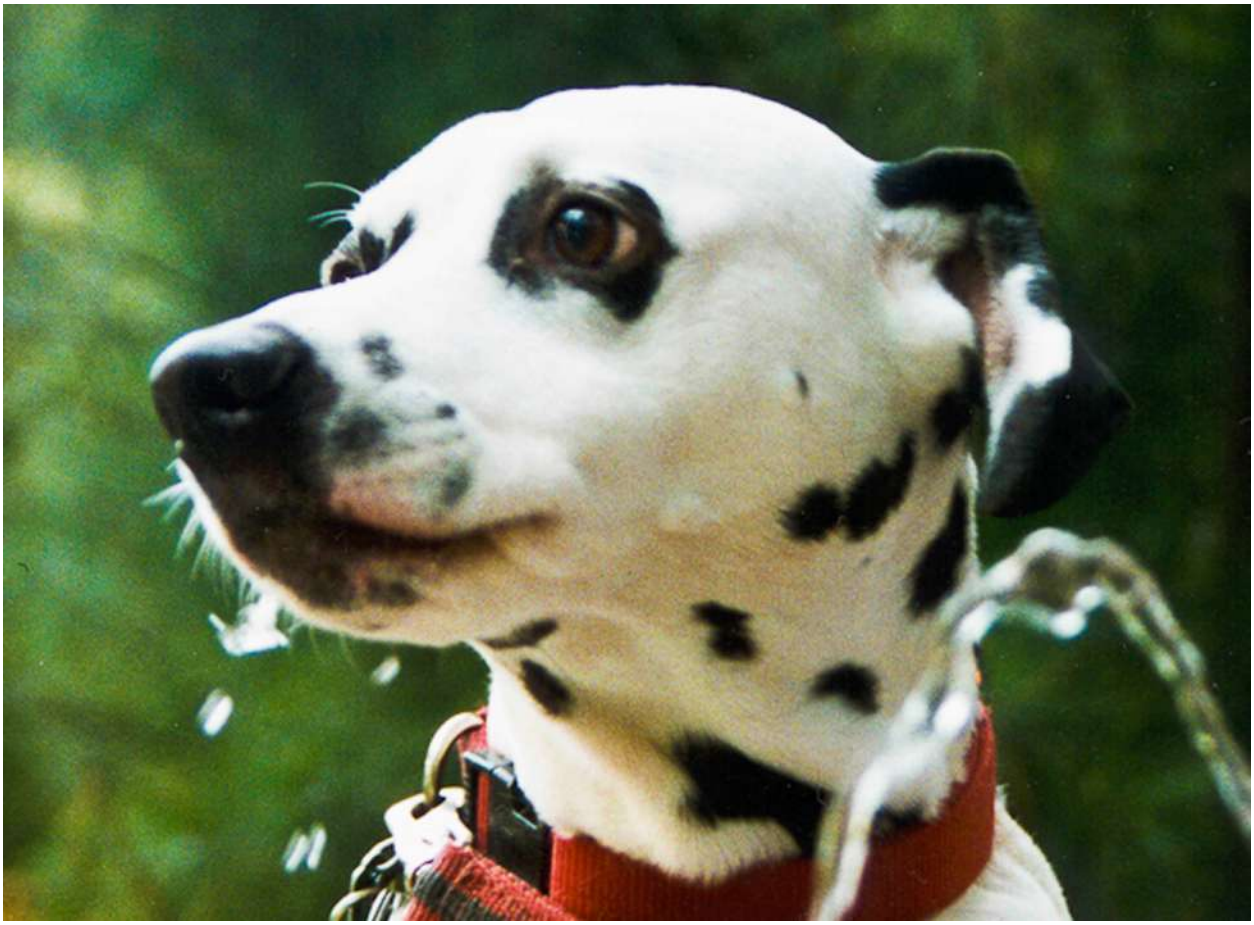


Figure 9.1 You can use a hypothesis test to decide if a dog breeder's claim that every Dalmatian has 35 spots is statistically sound. (credit: Robert Neff)

Introduction

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Differentiate between Type I and Type II errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown
- Conduct and interpret hypothesis tests for a single population proportion

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. **Confidence intervals** are one way to estimate a population parameter. Another way to make a statistical inference is to

make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on average. A tutoring service claims that its method of tutoring helps 90 percent of its students get an A or a B. A company says that women managers in their company earn an average of $60,000 per year.

A statistician will make a decision about these claims. This process is called **hypothesis testing.** A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence, based upon analyses of the data, to reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a statistician will do the following:

1. Set up two contradictory hypotheses.
2. Collect sample data. In homework problems, the data or summary statistics will be given to you.
3. Determine the correct distribution to perform the hypothesis test.
4. Analyze sample data by performing the calculations that ultimately will allow you to reject or decline to reject the null hypothesis.
5. Make a decision and write a meaningful conclusion.

NOTE

To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See **Appendix E**.

9.1 | Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

H_0, the —**null hypothesis:** a statement of no difference between sample means or proportions or no difference between a sample mean or proportion and a population mean or proportion. In other words, the difference equals 0.

H_a—, the **alternative hypothesis:** a claim about the population that is contradictory to H_0 and what we conclude when we reject H_0.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision.** There are two options for a decision. They are *reject* H_0 if the sample information favors the alternative hypothesis or *do not reject* H_0 or *decline to reject* H_0 if the sample information is insufficient to reject the null hypothesis.

Mathematical Symbols Used in H_0 and H_a:

H_0	H_a
equal (=)	not equal (≠) **or** greater than (>) **or** less than (<)
greater than or equal to (≥)	less than (<)
less than or equal to (≤)	more than (>)

Table 9.1

NOTE

H_0 always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers use = in the null hypothesis, even with

> or < as the symbol in the alternative hypothesis. This practice is acceptable because we only make the decision to reject or not reject the null hypothesis.

Example 9.1

H_0: No more than 30 percent of the registered voters in Santa Clara County voted in the primary election. $p \leq 30$
H_a: More than 30 percent of the registered voters in Santa Clara County voted in the primary election. $p > 30$

Try It

9.1 A medical trial is conducted to test whether or not a new medicine reduces cholesterol by 25 percent. State the null and alternative hypotheses.

Example 9.2

We want to test whether the mean GPA of students in American colleges is different from 2.0 (out of 4.0). The null and alternative hypotheses are the following:
H_0: $\mu = 2.0$
H_a: $\mu \neq 2.0$

Try It

9.2 We want to test whether the mean height of eighth graders is 66 inches. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: μ __ 66
b. H_a: μ __ 66

Example 9.3

We want to test if college students take fewer than five years to graduate from college, on the average. The null and alternative hypotheses are the following:
H_0: $\mu \geq 5$
H_a: $\mu < 5$

Try It

9.3 We want to test if it takes fewer than 45 minutes to teach a lesson plan. State the null and alternative hypotheses. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: μ __ 45
b. H_a: μ __ 45

Example 9.4

An article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third of the students pass. The same article stated that 6.6 percent of U.S. students take advanced placement exams and 4.4 percent pass. Test if the percentage of U.S. students who take advanced placement exams is more than 6.6 percent. State the null and alternative hypotheses.
H_0: $p \leq 0.066$
H_a: $p > 0.066$

Try It

9.4 On a state driver's test, about 40 percent pass the test on the first try. We want to test if more than 40 percent pass on the first try. Fill in the correct symbol (=, ≠, ≥, <, ≤, >) for the null and alternative hypotheses.

a. H_0: p __ 0.40

b. H_a: p __ 0.40

Collaborative Exercise

Bring to class a newspaper, some news magazines, and some internet articles. In groups, find articles from which your group can write null and alternative hypotheses. Discuss your hypotheses with the rest of the class.

9.2 | Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth, or falseness, of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	***H_0* IS ACTUALLY**	**...**
	True	False
Do not reject *H_0*	Correct outcome	Type II error
Reject *H_0*	Type I error	Correct outcome

Table 9.2

The four possible outcomes in the table are as follows:

1. The decision is *not to reject* H_0 when ***H_0* is true** (correct decision).
2. The decision is to *reject* H_0 when, in fact, ***H_0* is true** (incorrect decision known as a **Type I error**).
3. The decision is *not to reject* H_0 when, in fact, ***H_0* is false** (incorrect decision known as a **Type II error**).
4. The decision is to *reject* H_0 when ***H_0* is false** (correct decision whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a Type I error = ***P*(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = ***P*(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely zero.

The Power of the Test is $1 - \beta$. Ideally, we want a high power that is as close to one as possible. Increasing the sample size can increase the Power of the Test.

The following are examples of Type I and Type II errors.

Example 9.5

Suppose the null hypothesis, H_0, is: Frank's rock climbing equipment is safe.

Type I error: Frank does not go rock climbing because he considers that the equipment is not safe, when in fact, the equipment is really safe. Frank is making the mistake of rejecting the null hypothesis, when the equipment is actually safe!

Type II error: Frank goes climbing, thinking that his equipment is safe, but this is a mistake, and he painfully realizes that his equipment is not as safe as it should have been. Frank assumed that the null hypothesis was true, when it was not.

α = *probability* that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe. β = *probability* that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Try It

9.5 Suppose the null hypothesis, H_0, is: the blood cultures contain no traces of pathogen *X*. State the Type I and Type II errors.

Example 9.6

Suppose the null hypothesis, H_0, is: a tomato plant is alive when a class visits the school garden.

Type I error: The null hypothesis claims that the tomato plant is alive, and it is true, but the students make the mistake of thinking that the plant is already dead.

Type II error: The tomato plant is already dead (the null hypothesis is false), but the students do not notice it, and believe that the tomato plant is alive.

α = *probability* that the class thinks the tomato plant is dead when, in fact, it is alive = *P*(Type I error). β = *probability* that the class thinks the tomato plant is alive when, in fact, it is dead = *P*(Type II error).

The error with the greater consequence is the Type I error. (If the class thinks the plant is dead, they will not water it.)

Try It

9.6 Suppose the null hypothesis, H_0, is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

Example 9.7

It's a Boy Genetic Labs, a genetics company, claims to be able to increase the likelihood that a pregnancy will result in a boy being born. Statisticians want to test the claim. Suppose that the null hypothesis, H_0, is: It's a Boy

Genetic Labs has no effect on gender outcome.

Type I error: This error results when a true null hypothesis is rejected. In the context of this scenario, we would state that we believe that It's a Boy Genetic Labs influences the gender outcome, when in fact it has no effect. The probability of this error occurring is denoted by the Greek letter alpha, α.

Type II error: This error results when we fail to reject a false null hypothesis. In context, we would state that It's a Boy Genetic Labs does not influence the gender outcome of a pregnancy when, in fact, it does. The probability of this error occurring is denoted by the Greek letter beta, β.

The error with the greater consequence would be the Type I error since couples would use the It's a Boy Genetic Labs product in hopes of increasing the chances of having a boy.

Try It

9.7 *Red tide* is a bloom of poison-producing algae—a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries montors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kilogram of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

Example 9.8

A certain experimental drug claims a cure rate of at least 75 percent for males with a disease. Describe both the Type I and Type II errors in context. Which error is the more serious?

Type I: A patient believes the cure rate for the drug is less than 75 percent when it actually is at least 75 percent.

Type II: A patient believes the experimental drug has at least a 75 percent cure rate when it has a cure rate that is less than 75 percent.

In this scenario, the Type II error contains the more severe consequence. If a patient believes the drug works at least 75 percent of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

Try It

9.8 Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88 percent.

Identify the Type I and Type II errors from these four possible choices.

a. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88 percent when that percentage is actually less than 88 percent
b. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88 percent when the percentage is actually at least 88 percent
c. Reject the null hypothesis that the percentage of adults who have jobs is at least 88 percent when the percentage is actually at least 88 percent
d. Reject the null hypothesis that the percentage of adults who have jobs is at least 88 percent when that percentage is actually less than 88 percent

9.3 | Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. *Particular distributions are associated with hypothesis testing.* Perform tests of a population mean using a **normal distribution** or a **Student's *t*-distribution**. (Remember, use a Student's *t*-distribution when the population **standard deviation** is unknown and the distribution of the sample mean is approximately normal.) We perform tests of a population proportion using a normal distribution (usually n is large).

Assumptions

When you perform *a hypothesis test of a single population mean* μ using a **Student's *t*-distribution** (often called a *t*-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a **simple random sample** that comes from a population that is approximately **normally distributed**. You use the sample **standard deviation** to approximate the population standard deviation. Note that if the sample size is sufficiently large, a *t*-test will work even if the population is not approximately normally distributed.

When you perform a *hypothesis test of a single population mean* μ using a normal distribution (often called a *z*-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation which, in reality, is rarely known.

When you perform a *hypothesis test of a single population proportion* p, you take a simple random sample from the population. You must meet the conditions for a **binomial distribution**, which are the following: there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$. Remember that $q = 1 - p$.

9.4 | Rare Events, the Sample, and the Decision and Conclusion

Establishing the type of distribution, sample size, and known or unknown standard deviation can help you figure out how to go about a hypothesis test. However, there are several other factors you should consider when working out a hypothesis test.

Rare Events

The thinking process in hypothesis testing can be summarized as follows: You want to test whether or not a particular property of the population is true. You make an assumption about the true population mean for numerical data or the true population proportion for categorical data. This assumption is the null hypothesis. Then you gather sample data that is representative of the population. From this sample data you compute the sample mean (or the sample proportion). If the value that you observe is very unlikely to occur (a rare event) if the null hypothesis is true, then you wonder why this is happening. A plausible explanation is that the null hypothesis is false.

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket, and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely,

Ali is hoping that what the two of them were told is wrong and there are more $100 bills in the basket. A *rare event* has occurred (Didi getting the $100 bill) so Ali doubts the assumption about only one $100 bill being in the basket.

Using the Sample to Test the Null Hypothesis

After you collect data and obtain the test statistic (the sample mean, sample proportion, or other test statistic), you can determine the probability of obtaining that test statistic when the null hypothesis is true. This probability is called the ***p*-value**.

When the *p*-value is very small, it means that the observed test statistic is very unlikely to happen if the null hypothesis is true. This gives significant evidence to suggest that the null hypothesis is false, and to reject it in favor of the alternative hypothesis. In practice, to reject the null hypothesis we want the *p*-value to be smaller than 0.05 (5 percent) or sometimes even smaller than 0.01 (1 percent).

Example 9.9

Suppose a baker claims that his bread height is more than 15 cm, on average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the **standard deviation** for the height is 0.5 cm and the distribution of heights is normal.

The null hypothesis could be H_0: $\mu \leq 15$. The alternate hypothesis is H_a: $\mu > 15$.

The words *is more than* translates as a ">" so "$\mu > 15$" goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since σ *is known* ($\sigma = 0.5$ cm), the distribution for the population is known to be normal with mean $\mu = 15$ and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{10}} = 0.16$.

Suppose the null hypothesis is true (which is that the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the sample unexpectedly large? The hypothesis test works by asking the question how *unlikely* the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The *p*-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The p-value, then, is the probability that a sample mean is the same or greater than 17 cm when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means. In **Figure 9.2**, the *p*-value is the area under the normal curve to the right of 17. Using a normal distribution table or a calculator, we can compute that this probability is practically zero.

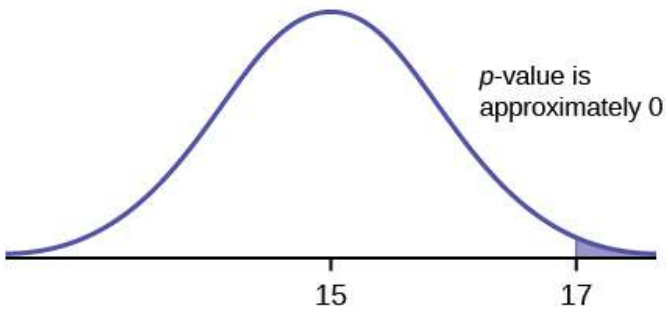


Figure 9.2

p-value = $P(\bar{x} > 17)$, which is approximately zero.

Because the *p*-value is almost 0, we conclude that obtaining a sample height of 17 cm or higher from 10 loaves of bread is very unlikely if the true mean height is 15 cm. We reject the null hypothesis and conclude that there is sufficient evidence to claim that the true population mean height of the baker's loaves of bread is higher than 15 cm.

Try It

9.9 A normal distribution has a standard deviation of 1. We want to verify a claim that the mean is greater than 12. A sample of 36 is taken with a sample mean of 12.5.

H_0: $\mu \leq 12$
H_a: $\mu > 12$
The *p*-value is 0.0013.
Draw a graph that shows the *p*-value.

Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the **null hypothesis** is to compare the *p*-value and a preset or preconceived α, also called the **level of significance of the test**. A preset α is the probability of a **Type I error** (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a *decision* to reject or not reject H_0, do as follows:

- If *p*-value $< \alpha$, reject H_0. The results of the sample data are significant. There is sufficient evidence to conclude that H_0 is an incorrect belief and that the **alternative hypothesis**, H_a, may be correct.
- If *p*-value $\geq \alpha$, do not reject H_0. The results of the sample data are not significant.There is not sufficient evidence to conclude that the alternative hypothesis, H_a, may be correct.
- When you *do not reject* H_0, it does not mean that you should believe that H_0 is true. It simply means that the sample data have *failed* to provide sufficient evidence to cast serious doubt about the truthfulness of H_0.

Conclusion: After you make your decision, write a thoughtful *conclusion* about the hypotheses in terms of the given problem.

Example 9.10

When using the *p*-value to evaluate a hypothesis test, you might find it useful to use the following mnemonic device:

If the *p*-value is low, the null must go.

If the *p*-value is high, the null must fly.

This memory aid relates a *p*-value less than the established alpha (the *p* is low) as rejecting the null hypothesis and, likewise, relates a *p*-value higher than the established alpha (the *p* is high) as not rejecting the null hypothesis.

Fill in the blanks.

Reject the null hypothesis when _____.

The results of the sample data _____.

Do not reject the null hypothesis when _____.

The results of the sample data _____.

Solution 9.10

Reject the null hypothesis when *the p-value is less than the established alpha value*. The results of the sample data **support the alternative hypothesis**.

Do not reject the null hypothesis when *the p-value is greater or equal to the established alpha value*. The results of the sample data **do not support the alternative hypothesis**.

Try It

9.10 It's a Boy Genetics Labs, a genetics company, claims their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

H_0: $p = 0.50$, H_a: $p > 0.50$

$\alpha = 0.01$

p-value = 0.025

Interpret the results and state a conclusion in simple, nontechnical terms.

9.5 | Additional Information and Full Hypothesis Test Examples

- In a **hypothesis test** problem, you may see words such as "the level of significance is 1 percent". The "1 percent" is the preconceived or preset α.
- The statistician setting up the hypothesis test selects the value of α to use *before* collecting the sample data.
- *If no level of significance is given, a common standard to use is $\alpha = 0.05$.*
- When you calculate the *p*-value and draw the picture, the *p*-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The **alternative hypothesis**, H_a, tells you if the test is left, right, or two-tailed. It is the *key* to conducting the appropriate test.
- H_a never has a symbol that contains an equal sign.
- *Thinking about the meaning of the* ***p*-value**: A data analyst should have more confidence that he made the correct decision to reject the null hypothesis with a smaller *p*-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large *p*-value such as 0.4, as opposed to a *p*-value of 0.056 (alpha = 0.05 is less than either number), a data analyst should have more confidence that she made the correct decision in not rejecting the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left-, right-, and two-tailed test.

Example 9.11

H_0: $\mu = 5$ H_a: $\mu < 5$

Test of a single population mean. H_a tells you the test is left-tailed. The picture of the *p*-value is as follows:

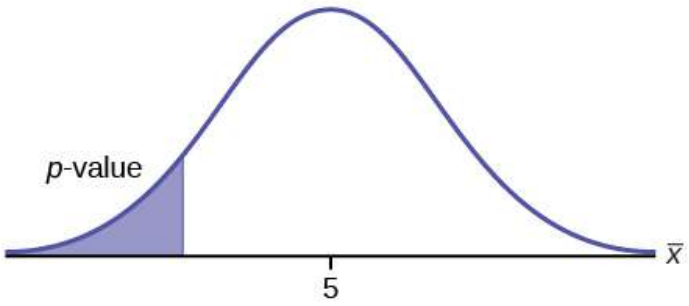


Figure 9.3

Try It

9.11 H_0: $\mu = 10$ H_a: $\mu < 10$

Assume the *p*-value is 0.0935. What type of test is this? Draw the picture of the *p*-value.

Example 9.12

H_0: $p \leq 0.2$ H_a: $p > 0.2$

This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the *p*-value is as follows:

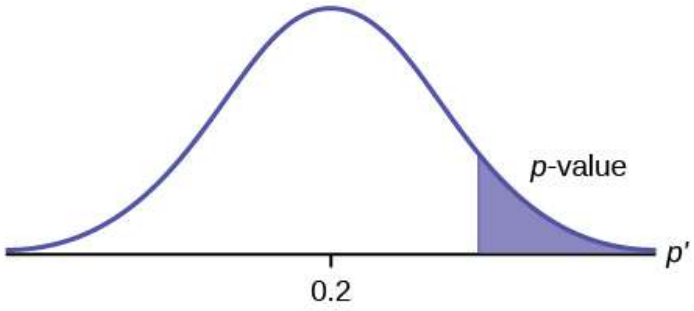


Figure 9.4

Try It

9.12 H_0: $\mu \leq 1$ H_a: $\mu > 1$

Assume the *p*-value is 0.1243. What type of test is this? Draw the picture of the *p*-value.

Example 9.13

H_0: $p = 50$ H_a: $p \neq 50$

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the *p*-value is as follows.

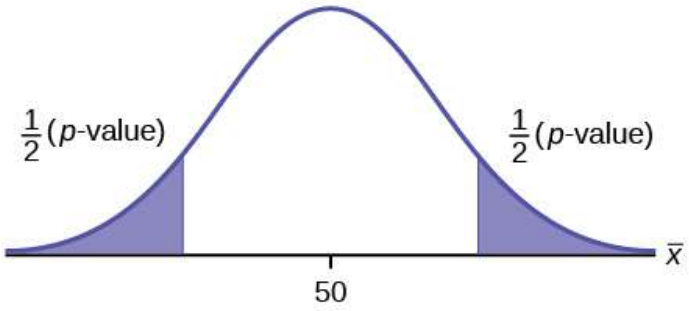


Figure 9.5

Try It

9.13 H_0: $p = 0.5$ H_a: $p \neq 0.5$

Assume the *p*-value is 0.2564. What type of test is this? Draw the picture of the *p*-value.

Full Hypothesis Test Examples

Example 9.14

Jeffrey, as an eight-year-old, established a mean time of 16.43 seconds for swimming the 25-yard freestyle, with a standard deviation of 0.8 seconds. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for 15 25-yard freestyle swims. For the 15 swims, Jeffrey's mean time was 16 seconds. Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Solution 9.14

Set up the hypothesis test:

Since the problem is about a mean, this is a *test of a single population mean*.

H_0: $\mu = 16.43$ H_a: $\mu < 16.43$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: $\bar{X}$ = the mean time to swim the 25-yard freestyle.

Distribution for the test: $\bar{X}$ is normal (population standard deviation is known: $\sigma = 0.8$)

with mean $\mu = 16.43$ and standard error of $\frac{0.8}{\sqrt{15}}$;

$\mu = 16.43$ comes from H_0 and not the data. $\sigma = 0.8$, and $n = 15$.

Using a table or a calculator, we can calculate the *p*-value as the area to the left of 16 under the normal curve:

p-value = $P(\bar{x} < 16) = 0.0187$ where the sample mean in the problem is given as 16.

p-value = 0.0187. The *p*-value is the area to the left of the sample mean given as 16.

Graph:

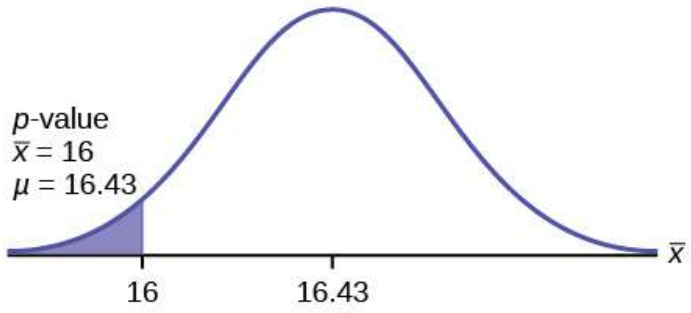


Figure 9.6

$\mu = 16.43$ comes from H_0. Our assumption is $\mu = 16.43$.

Interpretation of the *p*-value: *If H_0 is true*, there is a 0.0187 probability (1.87 percent), that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87 percent chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare α and the *p*-value:

$\alpha = 0.05$ *p*-value = 0.0187 $\alpha >$ *p*-value

Make a decision: Since $\alpha > p$-value, reject H_0.

An alternative approach is to find the z-test corresponding to the sample mean $\bar{x} = 16$. This is $z\text{-test} = \frac{\bar{x} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} = \frac{16 - 16.43}{\frac{0.8}{\sqrt{15}}} = -2.081729$.

The critical z-value = –1.645 for this test has probability 0.05 to its left tail, according to the Normal Table (see Appendices). Because the z-test is *to the left* of the critical z-value, we reject the null hypothesis.

This means that you reject $\mu = 16.43$. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but instead that he swims faster with the new goggles.

Conclusion: At the 5 percent significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The p-value can easily be calculated.

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS. Press 1:z-Test. Arrow over to Stats and press ENTER. Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ, 16 for the sample mean, and 15 for n. Arrow down to μ : (alternate hypothesis) and arrow over to < μ_0. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the p-value (p = 0.0187) but it also calculates the test statistic (z-score) for the sample mean. $\mu < 16.43$ is the alternative hypothesis. Do this set of instructions again except arrow to Draw(instead of Calculate). Press ENTER. A shaded graph appears with z = -2.08 (test statistic) and p = 0.0187 (p-value). Make sure when you use Draw that no other equations are highlighted in Y = and the plots are turned off.

When the calculator does a z-Test, the z-Test function finds the p-value by doing a normal probability calculation:

$P(\bar{x} < 16) =$ 2nd DISTR normcdf $(-10\text{^}99, 16, 16.43, 0.8/\sqrt{15})$.

The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually swims the 25-yard freestyle, on average, in 16.43 seconds. (Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is not evidence to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he actually does swim the 25-yard freestyle, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

HISTORICAL NOTE (EXAMPLE 9.11)

The traditional way to compare the two probabilities, α and the p-value, is to compare the critical value (z-score from α) to the test statistic (z-score from data). The calculated test statistic for the p-value is –2.08. (From the central limit theorem, the test statistic formula is $z = \frac{\bar{x} - \mu_X}{(\frac{\sigma_X}{\sqrt{n}})}$. For this problem, $\bar{x}$ = 16, μ_X = 16.43 from the null hypothesis, σ_X = 0.8, and n = 15.) You can find the critical value for α = 0.05 in the normal table (see **Appendix H: Tables**). The z-score for an area to the left equal to 0.05 is midway between –1.65 and –1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is –1.645. Since –1.645 > –2.08 (which demonstrates that $\alpha > p$-value), reject H_0. Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is

very common. For this problem, the *p*-value, 0.0187, is considerably smaller than α, 0.05. You can be confident about your decision to reject. The graph shows α, the *p*-value, and the test statistic and the critical value.

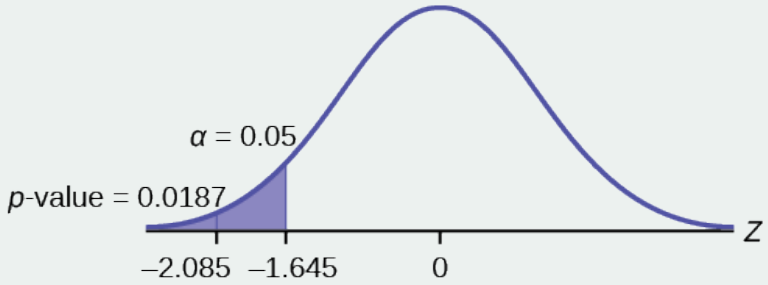


Figure 9.7

Try It

9.14 The mean throwing distance of a football by Marco, a high school freshman quarterback, is 40 yards, with a standard deviation of two yards. The team coach tells Marco to adjust his grip to get more distance. The coach records the distances for 20 throws. For the 20 throws, Marco's mean distance was 45 yards. The coach thought the different grip helped Marco throw farther than 40 yards. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume the throw distances for footballs are normal.

First, determine what type of test this is, set up the hypothesis test, find the *p*-value, sketch the graph, and state your conclusion.

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS. Press 1: *z*-Test. Arrow over to Stats and press ENTER. Arrow down and enter 40 for μ_0 (null hypothesis), 2 for σ, 45 for the sample mean, and 20 for *n*. Arrow down to μ: (alternative hypothesis) and set it either as <, ≠, or >. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the *p*-value but it also calculates the test statistic (*z*-score) for the sample mean. Select <, ≠, or > for the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with test statistic and *p*-value. Make sure when you use Draw that no other equations are highlighted in *Y* = and the plots are turned off.

Example 9.15

A college football coach records the mean weight that his players can bench press as 275 pounds, with a standard deviation of 55 pounds. Three of his players thought that the mean weight was more than that amount. They asked 30 of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3); 215(3); 225(1); 241(2); 252(2); 265(2); 275(2); 313(2); 316(5); 338(2); 341(1); 345(2); 368(2); 385(1).

Conduct a hypothesis test using a 2.5 percent level of significance to determine if the bench press mean is more than 275 pounds.

Solution 9.15

Set up the hypothesis test:

Since the problem is about a mean weight, this is a *test of a single population mean*.

H_0: $\mu = 275$ H_a: $\mu > 275$ This is a right-tailed test.

Calculating the distribution needed:

Random variable: $\bar{X}$ = the mean weight, in pounds, lifted by the football players.

Distribution for the test: It is normal because σ is known.

$$\bar{X} \sim N\left(275, \frac{55}{\sqrt{30}}\right)$$

$\bar{x} = 286.2$ pounds (from the data).

$\sigma = 55$ pounds. *Always use σ if you know it.* We assume $\mu = 275$ pounds unless our data shows us otherwise.

First, we compute the sample mean:

$$\bar{x} = \frac{205 + 205 + 205 + 215 + \cdots + 385}{30} = 286.2.$$

Next, we compute the *z*-test:

$$z\text{-test} = \frac{286.2 - 275}{\frac{55}{\sqrt{30}}} = 1.115362$$

Finally, the *p*-value is the probability to the right tail of the *z*-test, which we can compute from the table of *z*-scores as 0.5 –- 0.36650 = 0.1335.

() p-value $= P(\bar{x} > 286.2) = 0.1323$

Interpretation of the *p*-value: If H_0 is true, then there is a 0.1331 probability, 13.23 percent, that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23 percent chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.

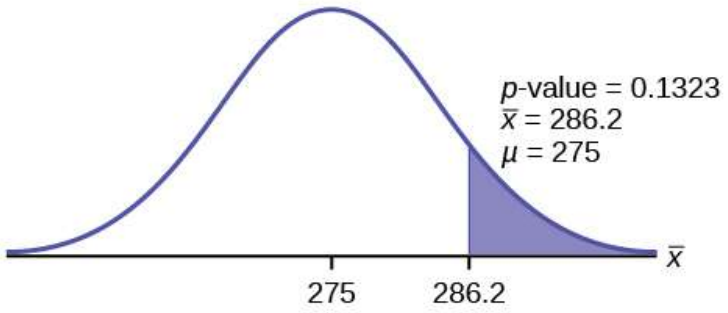


Figure 9.8

Compare α and the *p*-value:

$$\alpha = 0.025$$
$$p\text{-value} = 0.1323$$

Make a decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: At the 2.5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The *p*-value can easily be calculated.

Using the TI-83, 83+, 84, 84+ Calculator

Put the data and frequencies into lists. Press STAT and arrow over to TESTS. Press 1:*z*-Test. Arrow over to Data and press ENTER. Arrow down and enter 275 for μ_0, 55 for σ, the name of the list where you put the data, and the name of the list where you put the frequencies. Arrow down to μ: and arrow over to $> \mu_0$. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the *p*-value ($p = 0.1331$, a little different from the previous calculation—in it we used the sample mean rounded to one decimal place instead of the data), but also the test statistic (*z*-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 275$ is the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ (*p*-value). Make sure when you use Draw that no other equations are highlighted in *Y* = and the plots are turned off.

Example 9.16

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples 10 statistics students and obtains the scores 65; 65; 70; 67; 66; 63; 63; 68; 72; 71. He performs a hypothesis test using a 5 percent level of significance. The data are assumed to be from a normal distribution.

Solution 9.16

Set up the hypothesis test:

A 5 percent level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

H_0: $\mu = 65$ H_a: $\mu > 65$

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Determine the distribution needed:

Random variable: $\bar{X}$ = average score on the first statistics test.

Distribution for the test: If you read the problem carefully, you will notice that there is *no population standard deviation given*. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a Student's *t*-distribution.

Use *t*-distribution. Therefore, the distribution for the test is *t* with nine degrees of freedom.

Calculate the *p*-value using the Student's *t*-distribution:

First, we compute the sample mean as

$$\bar{x} = \frac{65 + 65 + \cdots + 71}{10} = 67.$$

Next, we compute the *t*-test as

$$t\text{-test} = \frac{\bar{x} - \mu_X}{\frac{s_X}{\sqrt{n}}} = \frac{67 - 65}{\frac{3.12}{\sqrt{10}}} \approx 1.98.$$

The *p*-value is the probability to the right tail of 1.98 in a *t*-distribution with nine degrees of freedom.

p-value = $P(\bar{x} > 67) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the *p*-value: If the null hypothesis is true, then there is a 0.0396 probability— (3.96 percent—)

that the sample mean is 65 or more.

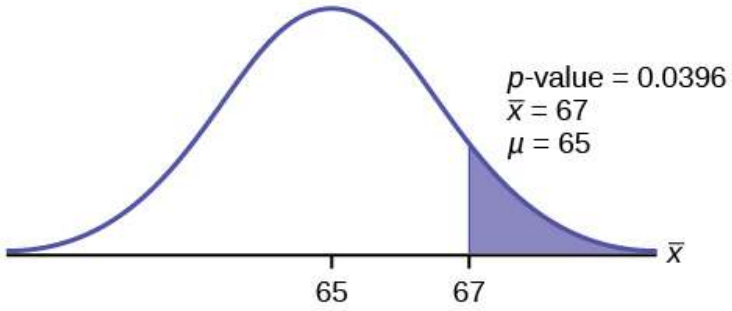


Figure 9.9

Compare α and the p-value:

Since $\alpha = 0.05$ and p-value = 0.0396, $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0.

Alternatively, according to a Student's t-distribution table (see Appendices), the critical t-value is 1.833. Since the t-test (1.98) is to the right of the critical t-value 1.833, we reject the null hypothesis.

This decision means we reject $\mu = 65$. In other words, we believe the average test score is more than 65.

Conclusion: At a 5 percent level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p-value can easily be calculated.

Using the TI-83, 83+, 84, 84+ Calculator

Put the data into a list. Press STAT and arrow over to TESTS. Press 2:T-Test. Arrow over to Data and press ENTER. Arrow down and enter 65 for μ_0, the name of the list where you put the data, and 1 for Freq:. Arrow down to μ: and arrow over to > μ_0. Press ENTER. Arrow down to Calculate and press ENTER. The calculator not only calculates the p-value ($p = 0.0396$) but it also calculates the test statistic (t-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternative hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with $t = 1.9781$ (test statistic) and $p = 0.0396$ (p-value). Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

Try It

9.16 It is believed that a stock price for a particular company will grow at a rate of $5 per week with a standard deviation of $1. An investor believes the stock won't grow as quickly. The changes in stock price are recorded for 10 weeks and are as follows: $4, $3, $2, $3, $1, $7, $2, $1, $1, $2. Perform a hypothesis test using a 5 percent level of significance. State the null and alternative hypotheses, find the p-value, state your conclusion, and identify the Type I and Type II errors.

Example 9.17

Joon believes that 50 percent of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is the same or different from 50 percent. Joon samples 100 first-time brides and 53 reply that they are younger than their grooms. For the hypothesis test, she uses a 1 percent level of significance.

Solution 9.17

Set up the hypothesis test:

The 1 percent level of significance means that $\alpha = 0.01$. This is a *test of a single population proportion*.

H_0: $p = 0.50$ H_a: $p \neq 0.50$

The words ***is the same or different from*** tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: P' = the percentage of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for P', the estimated proportion.

P' follows a normal distribution with mean value $\mu = p$, and standard error $\sigma = \sqrt{\frac{p \cdot q}{n}}$.

In our example, $p = q = 0.5$, and $n = 100$,
where $p = 0.50$, $q = 1 - p = 0.50$, and $n = 100$.

Calculate the p-value using the normal distribution for proportions:

First, we compute the sample proportion as $\hat{p} = \frac{53}{100} = 0.53$.

Next, the z-test is given by

$$z\text{-test} = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{0.53 - 0.50}{\sqrt{\frac{0.50 \times 0.50}{100}}} = 0.6.$$

Since the z-test is positive, we compute the area to the right tail of 0.6 in a normal distribution, $P(Z > 0.6) = 0.2742531$. Finally, because this is a two-sided test of significance, we multiply this probability times two to account for the left tail, and obtain

() p-value $= 2 \times 0.2742531 = 0.5485062$

where $x = 53$, $p' = \frac{x}{n} = \frac{53}{100} = 0.53$.

Interpretation of the *p*-value: If the null hypothesis is true, there is 0.5485 probability, (54.85 percent) that the sample (estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph in **Figure 9.9**).

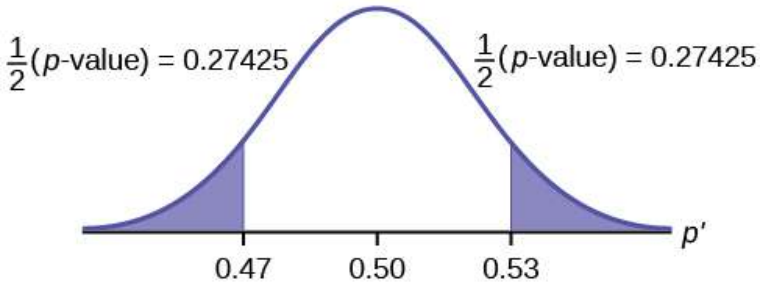


Figure 9.10

$\mu = p = 0.50$ comes from H_0, the null hypothesis.

$p' = 0.53$. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to $0.50 - 0.03 = 0.47$ where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p-value:

Since $\alpha = 0.01$ and p-value $= 0.5485$, $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, you cannot reject H_0.

Conclusion: At the 1 percent level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides who are younger than their grooms is different from 50 percent.

The p-value can easily be calculated.

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS. Press 5:1-PropZTest. Enter .5 for p_0, 53 for x and 100 for n. Arrow down to Prop and arrow to not equals p_0. Press ENTER. Arrow down to Calculate and press ENTER. The calculator calculates the p-value ($p = 0.5485$) and the test statistic (z-score). Prop not equals .5 is the alternate hypothesis. Do this set of instructions again except arrow to Draw (instead of Calculate). Press ENTER. A shaded graph appears with $z = 0.6$ (test statistic) and $p = 0.5485$ (p-value). Make sure when you use Draw that no other equations are highlighted in $Y =$ and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides who are younger than their grooms is different from 50 percent when, in fact, the proportion is actually 50 percent. Reject the null hypothesis when the null hypothesis is true.

The Type II error is there is not enough evidence to conclude that the proportion of first-time brides who are younger than their grooms differs from 50 percent when, in fact, the proportion does differ from 50 percent. Do not reject the null hypothesis when the null hypothesis is false.

Try It Σ

9.17 A teacher believes that 85 percent of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85 percent. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For the hypothesis test, use a 1 percent level of significance.

First, determine what type of test this is, set up the hypothesis test, find the p-value, sketch the graph, and state your conclusion.

Example 9.18

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30 percent. A cell phone company has reason to believe that the proportion is not 30 percent. Before the cell phone company starts a big advertising campaign, it conducts a hypothesis test. The company's marketing people survey 150 households with the result that 43 of the households have three cell phones.

Solution 9.18

Set up the hypothesis test:

H_0: $p = 0.30$ H_a: $p \neq 0.30$

Determine the distribution needed:

The **random variable** is P' = proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P' \sim N\left(0.30, \sqrt{\frac{(0.30) \cdot (0.70)}{150}}\right)$.

a. The value that helps determine the *p*-value is p'. Calculate p'.

Solution 9.18

a. $p' = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$$x = 43, \; n = 150$$

$$p' = \frac{43}{150.}$$

b. What is a *success* for this problem?

Solution 9.18

b. A success is having three cell phones in a household.

c. What is the level of significance?

Solution 9.18

c. The level of significance is the preset α. Since α is not given, assume that $\alpha = 0.05$.

d. Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.
Calculate the *p*-value.

Solution 9.18

d. First we compute the sample proportion $\hat{p} = \frac{43}{150} = 0.287$.

Next, the *z*-test is given by

$$z\text{-test} = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{0.287 - 0.30}{\sqrt{\frac{0.30 \times 0.70}{150}}} \approx -0.36.$$

Since the *z*-test is negative, we compute the area to the left tail of –0.36 in a normal distribution, $P(Z < -0.36) \approx 0.3607902$. Finally, because this is a two-sided test of significance, we multiply this probability times two to account for the right tail, and obtain $p\text{-value} = 2 \times 0.3607902 = 0.7215804$.

e. Make a decision. _____(Reject/Do not reject) H_0 because_____.

Solution 9.18

e. Assuming that $\alpha = 0.05$, $\alpha < p$-value. The decision is *do not reject* H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30 percent.

Try It

9.18 Marketers believe that 92 percent of adults in the United States own a cell phone. A cell phone manufacturer believes that number is actually lower. Two hundred American adults are surveyed, of which 174 report having cell phones. Use a 5 percent level of significance. State the null and alternative hypotheses, find the *p*-value, state your conclusion, and identify the Type I and Type II errors.

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter p. The distribution for the test is normal. The estimated proportion p' is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95 percent confidence interval computation. The poem is clever and humorous, so please enjoy it!

Example 9.19

My dog has so many fleas,
They do not come off with ease.
As for shampoo, I have tried many types
Even one called Bubble Hype,
Which only killed 25 percent of the fleas,
Unfortunately I was not pleased.

I've used all kinds of soap,
Until I had given up hope
Until one day I saw
An ad that put me in awe.

A shampoo used for dogs
Called GOOD ENOUGH to Clean a Hog
Guaranteed to kill more fleas.

I gave Fido a bath
And after doing the math
His number of fleas
Started dropping by 3's!

Before his shampoo
I counted 42.
At the end of his bath,
I redid the math
And the new shampoo had killed 17 fleas.
So now I was pleased.

Now it is time for you to have some fun
With the level of significance being .01,
You must help me figure out
Use the new shampoo or go without?

Solution 9.19

Set up the hypothesis test:

H_0: $p \leq 0.25$ H_a: $p > 0.25$

Determine the distribution needed:

In words, *clearly* state what your random variable $\bar{X}$ or P' represents.

P' = The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal:

$$N\left(0.25, \sqrt{\frac{(0.25)(1-0.25)}{42}}\right)$$

The *z*-test is given by

$$z\text{-test} = \frac{\bar{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{0.4048 - 0.25}{\sqrt{42}} \approx 2.316834.$$

Because this is a hypothesis test one-sided to the right, we compute the *p*-value as the area to the right tail of the *z*-test in a standard normal distribution, $P(Z > 3.32) \approx 0.0103$.

In one to two complete sentences, explain what the *p*-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 $\left(\frac{17}{42}\right)$ or more.

Use the previous information to sketch a picture of this situation. *Clearly* label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

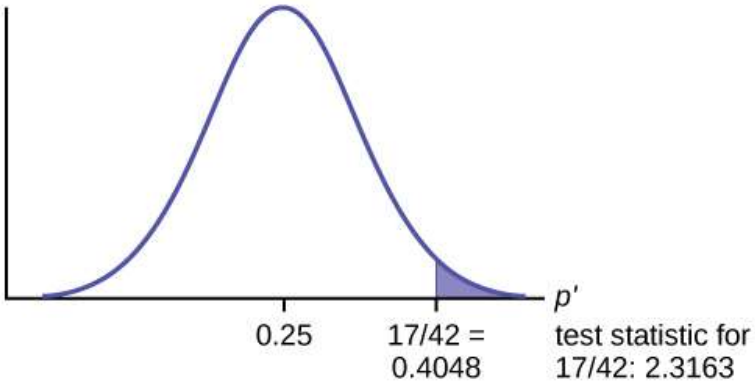


Figure 9.11

Compare α and the *p*-value:

Indicate the correct decision (*reject* or *do not reject* the null hypothesis) and the reason for it, and write an appropriate conclusion, using complete sentences.

Alpha	Decision	Reason for Decision
0.01	Do not reject H_0	$\alpha <$ *p*-value

Table 9.3

Conclusion: At the 1 percent level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25 percent.

Construct a 95 percent confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

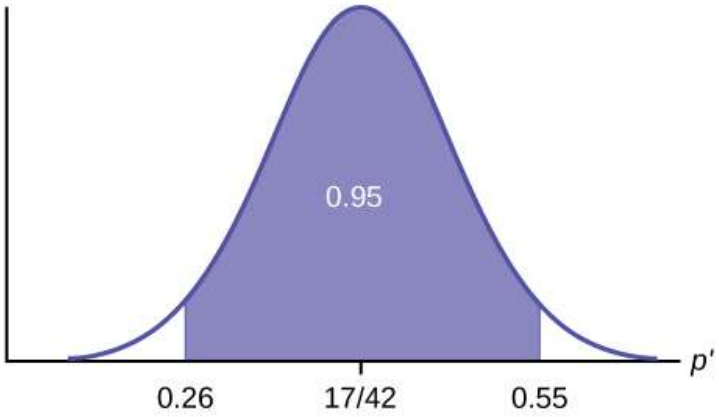


Figure 9.12

Confidence Interval: (0.26, 0.55). We are 95 percent confident that the true population proportion p of fleas that are killed by the new shampoo is between 26 percent and 55 percent.

NOTE

This test result is not very definitive since the p-value is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.

Example 9.20

The National Institute of Standards and Technology provides exact data on conductivity properties of materials. Following are conductivity measurements for 11 randomly selected pieces of a particular type of glass:

1.11, 1.07, 1.11, 1.07, 1.12, 1.08, 0.98, 0.98, 1.02, 0.95, 0.95

Is there convincing evidence that the average conductivity of this type of glass is greater than one? Use a significance level of 0.05. Assume the population is normal.

Solution 9.20

Let's follow a four-step process to answer this statistical question.

1. **State the question**: We need to determine if, at a 0.05 significance level, the average conductivity of the selected glass is greater than one. Our hypotheses will be as follows:
   a. H_0: $\mu \leq 1$
   b. H_a: $\mu > 1$
2. **Plan**: We are testing a sample mean without a known population standard deviation. Therefore, we need to use a Student's t-distribution. Assume the underlying population is normal.
3. **Do the calculations**: We will input the sample data into the TI-83 as follows.

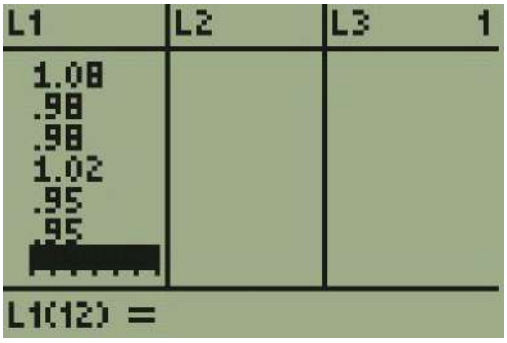


Figure 9.13

Figure 9.14

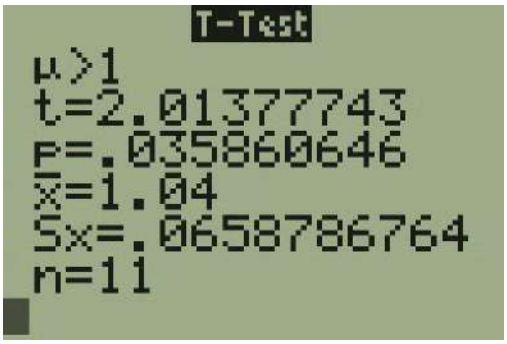


Figure 9.15

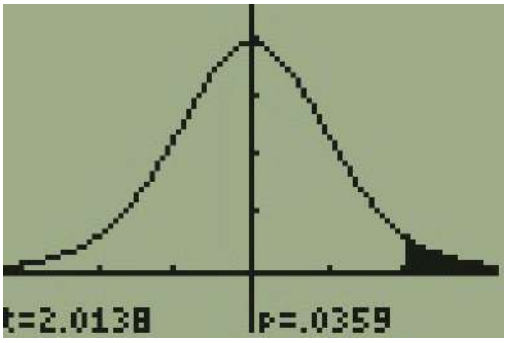


Figure 9.16

4. **State the conclusions**: Since the *p*-value ($p = 0.036$) is less than our alpha value, we will reject the null hypothesis. It is reasonable to state that the data support the claim that the average conductivity level is greater than one.

Example 9.21

In a study of 420,019 cell phone users, 172 of the subjects developed brain cancer. Test the claim that cell phone users developed brain cancer at a greater rate than that for non-cell phone users. The rate of brain cancer for non-cell phone users is 0.0340 percent. Since this is a critical issue, use a 0.005 significance level. Explain why the significance level should be so low in terms of a Type I error.

Solution 9.21

We will follow the four-step process.

1. We need to conduct a hypothesis test on the claimed cancer rate. Our hypotheses will be as follows:
   a. H_0: $p \leq 0.00034$
   b. H_a: $p > 0.00034$

   If we commit a Type I error, we are essentially accepting a false claim. Since the claim describes cancer-causing environments, we want to minimize the chances of incorrectly identifying causes of cancer.
2. We will be testing a sample proportion with $x = 172$ and $n = 420{,}019$. The sample is sufficiently large because we have $np = 420{,}019(0.00034) = 142.8$, $nq = 420{,}019(0.99966) = 419{,}876.2$, two independent outcomes, and a fixed probability of success $p = 0.00034$. Thus we will be able to generalize our results to the population.
3. The associated TI results are shown in the following figures.

Figure 9.17

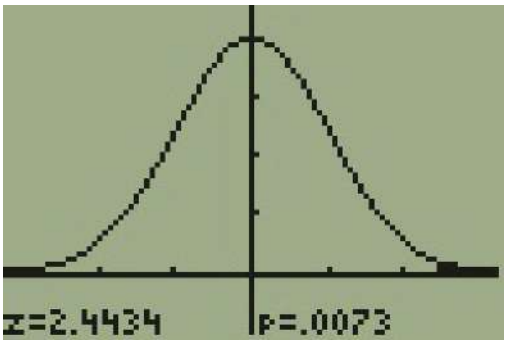


Figure 9.18

4. Since the *p*-value = 0.0073 is greater than our alpha value = 0.005, we cannot reject the null. Therefore, we conclude that there is not enough evidence to support the claim of higher brain cancer rates for the cell phone users.

9.6 | Hypothesis Testing of a Single Mean and Single Proportion

Stats Lab

9.1 Hypothesis Testing of a Single Mean and Single Proportion

Student Learning Outcomes

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

Television Survey

In a recent survey, it was stated that Americans watch television on average four hours per day. Assume that $\sigma = 2$. Using your class as the sample, conduct a hypothesis test to determine if the average for students at your school is lower.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
   a. Graph:

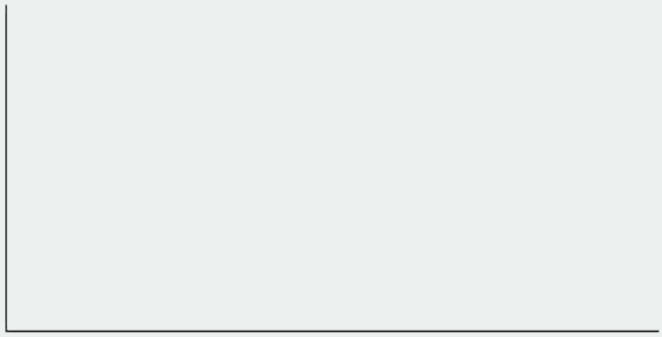


Figure 9.19

   b. Determine the *p*-value.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Language Survey

About 42.3 percent of Californians and 19.6 percent of all Americans over age five speak a language other than English at home. Using your class as the sample, conduct a hypothesis test to determine if the percentage of the students at your school who speak a language other than English at home is different from 42.3 percent.

1. H_0: _____
2. H_a: _____

3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    a. Graph:

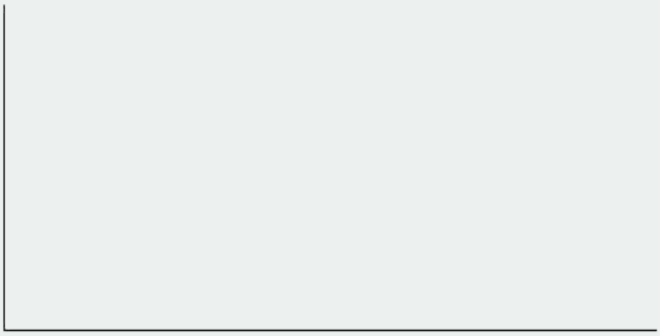


Figure 9.20

    b. Determine the *p*-value.
7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Jeans Survey

You've read in an article that young adults own an average of three pairs of jeans. Survey eight people from your class to determine if the average is higher than three. Assume the population is normal.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable. _____ = _____
4. The distribution to use for the test is _____.
5. Determine the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    a. Graph:

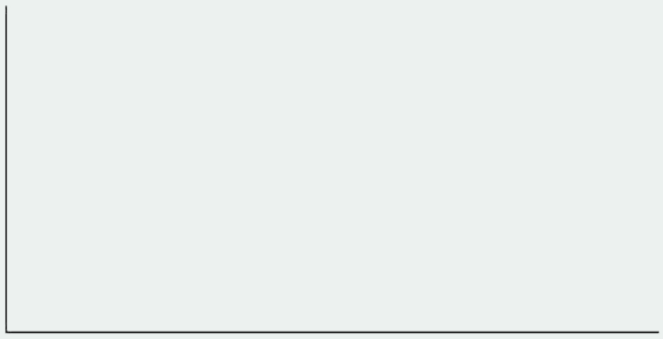


Figure 9.21

b. Determine the *p*-value.

7. Do you or do you not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

KEY TERMS

binomial distribution a discrete random variable (RV) that arises from Bernoulli trials; there are a fixed number, *n*, of independent trials
Independent means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in *n* trials. The notation is: $X \sim B(n, p)$ $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$.
The probability of exactly *x* successes in *n* trials is $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

confidence interval (*CI*) an interval estimate for an unknown population parameter
This depends on the following:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

hypothesis a statement about the value of a population parameter; in the case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a)

hypothesis testing based on sample evidence, a procedure for determining whether the hypothesis stated is a reasonable statement and should not be rejected, or is unreasonable and should be rejected

level of significance of the test probability of a Type I error (reject the null hypothesis when it is true)
Notation: α. In hypothesis testing, the level of significance is called the preconceived α or the preset α.

normal distribution a bell-shaped continuous random variable *X*, with center at the mean value (μ) and distance from the center to the inflection points of the bell curve given by the standard deviation (σ)
We write $X \sim N(\mu, \sigma)$. If the mean value is 0 and the standard deviation is 1, the random variable is called the standard normal distribution, and it is denoted with the letter *Z*.

***p*-value** the probability that an event will happen purely by chance assuming the null hypothesis is true; the smaller the *p*-value, the stronger the evidence is against the null hypothesis

standard deviation a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation

Student's *t*-distribution investigated and reported by William S. Gosset in 1908 and published under the pseudonym Student
The major characteristics of the random variable (RV) are as follows

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as *n* gets larger.
- There is a *family* of *t*-distributions: every representative of the family is completely defined by the number of degrees of freedom, which is one less than the number of data items.

Type 1 error the decision is to reject the null hypothesis when, in fact, the null hypothesis is true

Type 2 error the decision is not to reject the null hypothesis when, in fact, the null hypothesis is false

CHAPTER REVIEW

9.1 Null and Alternative Hypotheses

In a **hypothesis test**, sample data are evaluated in order to arrive at a decision about some type of claim. If certain conditions

about the sample are satisfied, then the claim can be evaluated for a population. In a hypothesis test, we do the following:

1. Evaluate the **null hypothesis**, typically denoted with H_0. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality (=, ≤, or ≥).
2. Always write the **alternative hypothesis**, typically denoted with H_a or H_1, using less than, greater than, or not equals symbols, i.e., (≠, >, or <).
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

9.2 Outcomes and the Type I and Type II Errors

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected.

The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively. The power of the test, $1 - \beta$, quantifies the likelihood that a test will yield the correct result of a true alternative hypothesis being accepted. A high power is desirable.

9.3 Distribution Needed for Hypothesis Testing

In order for a hypothesis test's results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's *t*-test should be used if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with an unknown standard deviation.
2. The normal test will work if the data come from a simple, random sample and the population is approximately normally distributed, or the sample size is large, with a known standard deviation.

When testing a single population proportion use a normal test for a single population proportion if the data come from a simple, random sample, fill the requirements for a binomial distribution, and the mean number of success and the mean number of failures satisfy the conditions: $np > 5$ and $nq > n$ where n is the sample size, p is the probability of a success, and q is the probability of a failure.

9.4 Rare Events, the Sample, and the Decision and Conclusion

When the probability of an event occurring is low, and it happens, it is called a rare event. Rare events are important to consider in hypothesis testing because they can inform your willingness not to reject or to reject a null hypothesis. To test a null hypothesis, find the *p*-value for the sample data and graph the results. When deciding whether or not to reject the null the hypothesis, keep these two parameters in mind:

1. $\alpha > p$-value, reject the null hypothesis.
2. $\alpha \leq p$-value, do not reject the null hypothesis.

9.5 Additional Information and Full Hypothesis Test Examples

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_0 and H_a. Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the *p*-value. (A *z*-score and a *t*-score are examples of test statistics.)
5. Compare the preconceived α with the *p*-value, make a decision (reject or do not reject H_0), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β. β is needed to help determine the sample size of the data that are used in calculating the *p*-value. Remember that the quantity $1 - \beta$ is called the **Power of the Test**. A high power is

desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same. If the power is low, the null hypothesis might not be rejected when it should be.

FORMULA REVIEW

9.1 Null and Alternative Hypotheses

H_0 and H_a are contradictory.

If H_0 has:	equal (=)	greater than or equal to (≥)	less than or equal to (≤)
then H_a has:	not equal (≠) **or** greater than (>) **or** less than (<)	less than (<)	greater than (>)

Table 9.4

If $\alpha \leq p$-value, then do not reject H_0.

If $\alpha > p$-value, then reject H_0.

α is preconceived. Its value is set before the hypothesis test starts. The p-value is calculated from the data.

9.2 Outcomes and the Type I and Type II Errors

α = probability of a Type I error = P(Type I error) = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = P(Type II error) = probability of not rejecting the null hypothesis when the null hypothesis is false.

9.3 Distribution Needed for Hypothesis Testing

If there is no given preconceived α, then use $\alpha = 0.05$.

Types of Hypothesis Tests

- Single population mean, *known* population variance (or standard deviation): **Normal test**.
- Single population mean, *unknown* population variance (or standard deviation): **Student's *t*-test**.
- Single population proportion: **Normal test**.
- For a *single population mean*, we may use a normal distribution with the following mean and standard deviation. Means: $\mu = \mu_{\bar{x}}$ and $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$.
- For a *single population proportion*, we may use a normal distribution with the following mean and standard deviation. Proportions: $\mu = \boldsymbol{p}$ and $\sigma = \sqrt{\frac{pq}{n}}$.

PRACTICE

9.1 Null and Alternative Hypotheses

1. You are testing that the mean speed of your cable internet connection is more than three megabits per second. What is the random variable? Describe it in words.

2. You are testing that the mean speed of your cable internet connection is more than three megabits per second. State the null and alternative hypotheses.

3. The American family has an average of two children. What is the random variable? Describe in words.

4. The mean entry level salary of an employee at a company is $58,000. You believe it is higher for IT professionals in the company. State the null and alternative hypotheses.

5. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.

6. A sociologist claims the probability that a person picked at random in Times Square in New York City is visiting the area is 0.83. You want to test to see if the claim is correct. State the null and alternative hypotheses.

7. In a population of fish, approximately 42 percent are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

8. Suppose that a recent article stated that the mean time students spend doing homework each week is 2.5 hours. A study was then done to see if the mean time has increased in the new century. A random sample of 26 students. The mean length of time the students spent on homework was 3 hours with a standard deviation of 1.8 hours. Suppose that it is somehow known that the population standard deviation is 1.5. If you were conducting a hypothesis test to determine if the mean length of homework has increased, what would the null and alternative hypotheses be? The distribution of the population is normal.

a. H_0: _____
b. H_a: _____

9. A random survey of 75 long-term marathon runners revealed that the mean length of time they've been running is 17.4 years with a standard deviation of 6.3 years. If you were conducting a hypothesis test to determine if the population mean time for these runners could likely be 15 years, what would the null and alternative hypotheses be?

a. H_0: _____
b. H_a: _____

10. Researchers published an article stating that in any one-year period, approximately 9.5 percent of American adults suffer from a particular type of disease. Suppose that in a survey of 100 people in a certain town, seven of them suffered from this disease. If you were conducting a hypothesis test to determine if the true proportion of people in that town suffering from this disease is lower than the percentage in the general adult American population, what would the null and alternative hypotheses be?

a. H_0: _____
b. H_a: _____

9.2 Outcomes and the Type I and Type II Errors

11. The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

12. A sleeping bag is tested to withstand temperatures of –15 °F. You think the bag cannot stand temperatures that low. State the Type I and Type II errors in complete sentences.

13. For **Exercise 9.12**, what are α and β in words?

14. In words, describe $1 - \beta$ for **Exercise 9.12**.

15. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.

16. A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, H_0, is: the surgical procedure will go well. Which is the error with the greater consequence?

17. The power of a test is 0.981. What is the probability of a Type II error?

18. A group of divers is exploring an old sunken ship. Suppose the null hypothesis, H_0, is the sunken ship does not contain buried treasure. State the Type I and Type II errors in complete sentences.

19. A microbiologist is testing a water sample for E. coli. Suppose the null hypothesis, H_0, is the sample does not contain E. coli. The probability that the sample does not contain E. coli, but the microbiologist thinks it does is 0.012. The probability that the sample does contain E. coli, but the microbiologist thinks it does not is 0.002. What is the power of this test?

20. A microbiologist is testing a water sample for E. coli. Suppose the null hypothesis, H_0, is the sample contains E-coli. Which is the error with the greater consequence?

9.3 Distribution Needed for Hypothesis Testing

21. Which two distributions can you use for hypothesis testing for this chapter?

22. Which distribution do you use when the standard deviation is not known? Assume sample size is large.

23. Which distribution do you use when the standard deviation is not known and you are testing one population mean? Assume sample size is large.

24. A population mean is 13. The sample mean is 12.8, and the sample standard deviation is two. The sample size is 20. What distribution should you use to perform a hypothesis test? Assume the underlying population is normal.

25. A population has a mean of 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

26. It is thought that 42 percent of respondents in a taste test would prefer Brand *A*. In a particular test of 100 people, 39 percent preferred Brand *A*. What distribution should you use to perform a hypothesis test?

27. You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. What must you assume about the distribution of the data?

28. You are performing a hypothesis test of a single population mean using a Student's *t*-distribution. The data are not from a simple random sample. Can you accurately perform the hypothesis test?

29. You are performing a hypothesis test of a single population proportion. What must be true about the quantities of np and nq?

30. You are performing a hypothesis test of a single population proportion. You find out that np is less than five. What must you do to be able to perform a valid hypothesis test?

31. You are performing a hypothesis test of a single population proportion. The data come from which distribution?

9.4 Rare Events, the Sample, and the Decision and Conclusion

32. When do you reject the null hypothesis?

33. The probability of winning the grand prize at a particular carnival game is 0.005. Is the outcome of winning very likely or very unlikely?

34. The probability of winning the grand prize at a particular carnival game is 0.005. Michele wins the grand prize. Is this considered a rare or common event? Why?

35. It is believed that the mean height of high school students who play basketball on the school team is 73 inches with a standard deviation of 1.8 inches. A random sample of 40 players is chosen. The sample mean was 71 inches, and the sample standard deviation was 1.5 inches. Do the data support the claim that the mean height is less than 73 inches? The *p*-value is almost zero. State the null and alternative hypotheses and interpret the *p*-value.

36. The mean age of graduate students at a university is at most 31 years with a standard deviation of two years. A random sample of 15 graduate students is taken. The sample mean is 32 years and the sample standard deviation is three years. Are the data significant at the 1 percent level? The *p*-value is 0.0264. State the null and alternative hypotheses and interpret the *p*-value.

37. Does the shaded region represent a low or a high *p*-value compared to a level of significance of 1 percent?

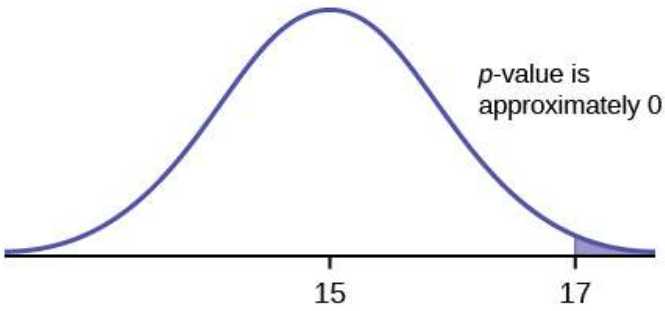


Figure 9.22

38. What should you do when $\alpha > p$-value?

39. What should you do if $\alpha = p$-value?

40. If you do not reject the null hypothesis, then it must be true. Is that statement correct? State why or why not in complete sentences.

Use the following information to answer the next seven exercises: Suppose that a recent article stated that the mean time students spend doing homework each week is 2.5 hours. A study was then done to see if the mean time has increased in the new century. A random sample of 26 students was taken. The mean length of time they did homework each week was three hours with a standard deviation of 1.8 hours. Suppose that it is somehow known that the population standard deviation is 1.5. Conduct a hypothesis test to determine if the mean length of time doing homework each week has increased. Assume

the distribution of homework times is approximately normal.

41. Is this a test of means or proportions?

42. What symbol represents the random variable for this test?

43. In words, define the random variable for this test.

44. Is σ known and, if so, what is it?

45. Calculate the following:

a. $\bar{x}$ _____
b. σ _____
c. s_x _____
d. n _____

46. Since both σ and s_x are given, which should be used? In one to two complete sentences, explain why.

47. State the distribution to use for the hypothesis test.

48. A random survey of 75 long-term marathon runners revealed that the mean length of time they have been running is 17.4 years with a standard deviation of 6.3 years. Conduct a hypothesis test to determine if the population mean time is likely to be 15 years.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0: _____ H_a : _____
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Is the population standard deviation known and, if so, what is it?
g. Calculate the following:
   i. $\bar{x}$ = _____
   ii. s = _____
   iii. n = _____
h. Which test should be used?
i. State the distribution to use for the hypothesis test.
j. Find the p-value.
k. At a pre-conceived $\alpha = 0.05$, give your answer for each of the following:
   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

9.5 Additional Information and Full Hypothesis Test Examples

49. Assume H_0: $\mu = 9$ and H_a: $\mu < 9$. Is this a left-tailed, right-tailed, or two-tailed test?

50. Assume H_0: $\mu \leq 6$ and H_a: $\mu > 6$. Is this a left-tailed, right-tailed, or two-tailed test?

51. Assume H_0: $p = 0.25$ and H_a: $p \neq 0.25$. Is this a left-tailed, right-tailed, or two-tailed test?

52. Draw the general graph of a left-tailed test.

53. Draw the graph of a two-tailed test.

54. A bottle of water is labeled as containing 16 fluid ounces of water. You believe it is less than that. What type of test would you use?

55. Your friend claims that his mean golf score is 63. You want to show that it is higher than that. What type of test would you use?

56. A bathroom scale claims to be able to identify correctly any weight within a pound. You think that it cannot be that accurate. What type of test would you use?

57. You flip a coin and record whether it shows heads or tails. You know the probability of getting heads is 50 percent, but you think it is less for this particular coin. What type of test would you use?

58. If the alternative hypothesis has a not equals ($\neq$) symbol, you know to use which type of test?

59. Assume the null hypothesis states that the mean is at least 18. Is this a left-tailed, right-tailed, or two-tailed test?

60. Assume the null hypothesis states that the mean is at most 12. Is this a left-tailed, right-tailed, or two-tailed test?

61. Assume the null hypothesis states that the mean is equal to 88. The alternative hypothesis states that the mean is not equal to 88. Is this a left-tailed, right-tailed, or two-tailed test?

HOMEWORK

9.1 Null and Alternative Hypotheses

62. Some of the following statements refer to the null hypothesis, some to the alternate hypothesis.

State the null hypothesis, H_0, and the alternative hypothesis. H_a, in terms of the appropriate parameter (μ or p).

a. The mean number of years Americans work before retiring is 34.
b. At most 60 percent of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. Twenty-nine percent of high school students take physical education daily.
e. Less than 5 percent of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in her lifetime is not more than 10.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11 percent for women.
j. Private universities' mean tuition cost is more than $20,000 per year.

63. A recent survey of 273 randomly selected teens living in Massachusetts asked about social media. Sixty-three said that they routinely use a certain app to share pictures. The researchers want to determine if there is good evidence that more than 30 percent of teens use this app. The alternative hypothesis is as follows:

a. $p < 0.30$
b. $p \leq 0.30$
c. $p \geq 0.30$
d. $p > 0.30$

64. A statistics instructor believes that fewer than 20 percent of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is as follows:

a. $p = 0.20$
b. $p > 0.20$
c. $p < 0.20$
d. $p \leq 0.20$

65. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The null and alternative hypotheses are as follows:

a. H_o: $\bar{x} = 4.5$, H_a : $\bar{x} > 4.5$
b. H_o: $\mu \geq 4.5$, H_a: $\mu < 4.5$
c. H_o: $\mu = 4.75$, H_a: $\mu > 4.75$
d. H_o: $\mu = 4.5$, H_a: $\mu > 4.5$

9.2 Outcomes and the Type I and Type II Errors

66. State the Type I and Type II errors in complete sentences given the following statements.

a. The mean number of years Americans work before retiring is 34.
b. At most 60 percent of Americans vote in presidential elections.
c. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
d. 29 percent of high school students take physical education every day.
e. Less than 5 percent of adults ride the bus to work in Los Angeles.
f. The mean number of cars a person owns in his or her lifetime is not more than 10.
g. About half of Americans prefer to live away from cities, given the choice.
h. Europeans have a mean paid vacation each year of six weeks.
i. The chance of developing breast cancer is under 11 percent for women.
j. Private universitie' mean tuition cost is more than $20,000 per year.

67. For Statements A–J in **Exercise 9.66**, answer the following in complete sentences.

a. State a consequence of committing a Type I error.
b. State a consequence of committing a Type II error.

68. When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the U.S. Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is *the drug is unsafe*. What is the Type II error?

a. To conclude the drug is safe when, in fact, it is unsafe.
b. Not to conclude the drug is safe when, in fact, it is safe.
c. To conclude the drug is safe when, in fact, it is safe.
d. Not to conclude the drug is unsafe when, in fact, it is unsafe.

69. A statistics instructor believes that fewer than 20 percent of Evergreen Valley College (EVC) students attended the opening midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing. The Type I error is to conclude that the percent of EVC students who attended is _____.

a. at least 20 percent, when, in fact, it is less than 20 percent.
b. 20 percent, when, in fact, it is 20 percent.
c. less than 20 percent, when, in fact, it is at least 20 percent.
d. less than 20 percent, when, in fact, it is less than 20 percent.

70. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5 percent, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The Type II error is not to reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours

a. is more than seven hours.
b. is at most seven hours.
c. is at least seven hours.
d. is less than seven hours.

71. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test. The Type I error is

a. to conclude that the current mean hours per week is higher than 4.5, when, in fact, it is higher.
b. to conclude that the current mean hours per week is higher than 4.5, when, in fact, it is the same.
c. to conclude that the mean hours per week currently is 4.5, when, in fact, it is higher.
d. to conclude that the mean hours per week currently is no higher than 4.5, when, in fact, it is not higher.

9.3 Distribution Needed for Hypothesis Testing

72. It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5 percent, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average? The distribution to be used for this test is $\bar{X} \sim$ _____

a. $N(7.24, \frac{1.93}{\sqrt{22}})$
b. $N(7.24, 1.93)$
c. t_{22}
d. t_{21}

9.4 Rare Events, the Sample, and the Decision and Conclusion

73. The National Institute of Mental Health published an article stating that in any one-year period approximately 9.5 percent of American adults suffer from depression or a depressive illness. Suppose that in a survey of 100 people in a certain town, seven of them suffered from depression or a depressive illness. Conduct a hypothesis test to determine if the true proportion of people in that town suffering from depression or a depressive illness is lower than the percent in the general adult American population.

a. Is this a test of one mean or proportion?
b. State the null and alternative hypotheses.
   H_0: _____ H_a: _____
c. Is this a right-tailed, left-tailed, or two-tailed test?
d. What symbol represents the random variable for this test?
e. In words, define the random variable for this test.
f. Calculate the following:
   i. x = _____
   ii. n = _____
   iii. p' = _____
g. Calculate σ_x = _____. Show the formula setup.
h. State the distribution to use for the hypothesis test.
i. Find the p-value.
j. At a pre-conceived $\alpha = 0.05$, give your answer for each of the following:
   i. Decision:
   ii. Reason for the decision:
   iii. Conclusion (write out in a complete sentence):

9.5 Additional Information and Full Hypothesis Test Examples

For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in ***Appendix E, Solution Sheets****. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.*

NOTE

If you are using a Student's-*t*-distribution for one of the following homework problems, you may assume that the underlying population is normally distributed. In general, you must first prove that assumption, however.

74. A particular brand of tires claims that its deluxe tire averages at least 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles with a standard deviation of 9,800 miles. Using alpha = 0.05, are the data highly inconsistent with the claim?

75. In 2009, President Barack Obama announced a new national fuel economy and emissions policy for cars and light trucks. It stated that the combined fleet fuel economy for an auto manufacturer of cars and light trucks will have to average 35.5 mpg or better by 2016. From past studies on fuel economy, it is known that the standard deviation of a typical fleet is 7.6 mpg. An auto manufacturer selects a random sample of 55 cars and light trucks and finds the sample mean fuel economy to be 34.6 mpg with a standard deviation of 10.3 mpg. Can the manufacturer claim that their fleet meets the fuel economy standard in the 2016 policy at the 5 percent level?

76. The cost of a daily newspaper varies from city to city. However, the variation among prices remains steady with a standard deviation of 20¢. A study was done to test the claim that the mean cost of a daily newspaper is $1.00. Twelve costs yield a mean cost of 95¢ with a standard deviation of 18¢. Do the data support the claim at the 1 percent level?

77. An article in the *San Jose Mercury News* stated that students in the California state university system take 4.5 years, on average, to finish their undergraduate degrees. Suppose you believe that the mean time is longer. You conduct a survey of 49 students and obtain a sample mean of 5.1 with a sample standard deviation of 1.2. Do the data support your claim at the 1 percent level?

78. The mean number of sick days an employee takes per year is believed to be about 10. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let x = the number of sick days they took for the past year. Should the personnel team believe that the mean number is 10?

79. In 1955, *Life Magazine* reported that the 25-year-old mother of three worked, on average, an 80-hour week. Recently, many groups have been studying whether or not the women's movement has, in fact, resulted in an increase in the average work week for women (combining employment and at-home work). Suppose a study was done to determine if the mean work week has increased. Eighty-one women were surveyed with the following results. The sample mean was 83; the sample standard deviation was 10. Does it appear that the mean work week has increased for women at the 5 percent level?

80. Your statistics instructor claims that 60 percent of the students who take her Elementary Statistics class go through life feeling more enriched. For some reason that she can't quite figure out, most people don't believe her. You decide to check this out on your own. You randomly survey 64 of her past Elementary Statistics students and find that 34 feel more enriched as a result of her class. Now, what do you think?

81. A Nissan Motor Corporation advertisement read, “The average man’s I.Q. is 107. The average brown trout’s I.Q. is 4. So why can’t man catch brown trout?” Suppose you believe that the brown trout’s mean I.Q. is greater than four. You catch 12 brown trout. A fish psychologist determines the I.Q.s as follows: 5, 4, 7, 3, 6, 4, 5, 3, 6, 3, 8, 5. Conduct a hypothesis test of your belief.

82. Refer to **Exercise 9.81**. Conduct a hypothesis test to see if your decision and conclusion would change if your belief were that the brown trout’s mean I.Q. is **not** four.

83. According to an article in *Newsweek*, the natural ratio of girls to boys is 100:105. In China, the birth ratio is 100: 114 (46.7 percent girls). Suppose you don’t believe the reported figures of the percentage of girls born in China. You conduct a study. In this study, you count the number of girls and boys born in 150 randomly chosen recent births. There are 60 girls and 90 boys born of the 150. Based on your study, do you believe that the percentage of girls born in China is 46.7?

84. A group of researchers research a common contagious disease. A newspaper found that 13 percent of Americans have been diagnosed with the disease in the last year. The researchers doubt that the percentage is really that high. It conducts its own survey. Out of 76 Americans surveyed, only two had been diagnosed with the disease. Would you agree with the newspaper's poll? In complete sentences, give three reasons why polls might give different results.

85. The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it’s shorter. She asks 10 engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours?

Data (length of mean work week): 70, 45, 55, 60, 65, 55, 55, 60, 50, 55.

86. Use the *Lap time* data for Lap 4 (see **Appendix C: Data Sets**) to test the claim that Terri finishes Lap 4, on average, in less than 129 seconds. Use all 20 races given.

87. Use the *Initial Public Offering* data (see **Appendix C: Data Sets**) to test the claim that the mean offer price was $18 per share. Do not use all the data. Use your random number generator to randomly survey 15 prices.

NOTE

The following questions were written by past students. They are excellent problems!

88. "Asian Family Reunion," by Chau Nguyen

Every two years it comes around.
We all get together from different towns.
In my honest opinion,
It's not a typical family reunion.
Not forty, or fifty, or sixty,
But how about seventy companions!
The kids would play, scream, and shout
One minute they're happy, another they'll pout.
The teenagers would look, stare, and compare
From how they look to what they wear.
The men would chat about their business
That they make more, but never less.
Money is always their subject
And there's always talk of more new projects.
The women get tired from all of the chats
They head to the kitchen to set out the mats.
Some would sit and some would stand
Eating and talking with plates in their hands.
Then come the games and the songs
And suddenly, everyone gets along!
With all that laughter, it's sad to say
That it always ends in the same old way.
They hug and kiss and say "good-bye"
And then they all begin to cry!
I say that 60 percent shed their tears
But my mom counted 35 people this year.
She said that boys and men will always have their pride,
So we won't ever see them cry.
I myself don't think she's correct,
So could you please try this problem to see if you object?

89. "Blowing Bubbles," by Sondra Prull

Studying stats just made me tense,
I had to find some sane defense.
Some light and lifting simple play
To float my math anxiety away.
Blowing bubbles lifts me high
Takes my troubles to the sky.
POIK! They're gone, with all my stress
Bubble therapy is the best.
The label said each time I blew
The average number of bubbles would be at least 22.
I blew and blew and this I found
From 64 blows, they all are round!
But the number of bubbles in 64 blows
Varied widely, this I know.
20 per blow became the mean
They deviated by 6, and not 16.
From counting bubbles, I sure did relax
But now I give to you your task.
Was 22 a reasonable guess?
Find the answer and pass this test!

90. "Dalmatian Darnation," by Kathy Sparling

A greedy dog breeder named Spreckles
Bred puppies with numerous freckles
The Dalmatians he sought
Possessed spot upon spot
The more spots, he thought, the more shekels.
His competitors did not agree
That freckles would increase the fee.
They said, “Spots are quite nice
But they don't affect price;
One should breed for improved pedigree.”
The breeders decided to prove
This strategy was a wrong move.
Breeding only for spots
Would wreak havoc, they thought.
His theory they want to disprove.
They proposed a contest to Spreckles
Comparing dog prices to freckles.
In records they looked up
One hundred one pups:
Dalmatians that fetched the most shekels.
They asked Mr. Spreckles to name
An average spot count he'd claim
To bring in big bucks.
Said Spreckles, “Well, shucks,
It's for one hundred one that I aim.”
Said an amateur statistician
Who wanted to help with this mission.
“Twenty-one for the sample
Standard deviation's ample.”
They examined one hundred and one
Dalmatians that fetched a good sum.
They counted each spot,
Mark, freckle, and dot
And tallied up every one.
Instead of one hundred one spots
They averaged ninety-six dots
Can they muzzle Spreckles’
Obsession with freckles
Based on all the dog data they've got?

91. *Macaroni and Cheese, please!!* by Nedda Misherghi and Rachelle Hall

As a poor starving student I don't have much money to spend for even the bare necessities. So my favorite and main staple food is macaroni and cheese. It's high in taste and low in cost and nutritional value.

One day, as I sat down to determine the meaning of life, I got a serious craving for this, oh, so important, food of my life. So I went down the street to Greatway to get a box of macaroni and cheese, but it was SO expensive! $2.02 !!! Can you believe it? It made me stop and think. The world is changing fast. I had thought that the mean cost of a box (the normal size, not some super-gigantic-family-value-pack) was at most $1, but now I wasn't so sure. However, I was determined to find out. I went to 53 of the closest grocery stores and surveyed the prices of macaroni and cheese. Here are the data I wrote in my notebook:

Price per box of Mac and Cheese

- 5 stores @ $2.02
- 15 stores @ $0.25
- 3 stores @ $1.29
- 6 stores @ $0.35
- 4 stores @ $2.27
- 7 stores @ $1.50
- 5 stores @ $1.89
- 8 stores @ $0.75

I could see that the cost varied but I had to sit down to figure out whether or not I was right. If it does turn out that this mouth-watering dish is at most $1, then I'll throw a big cheesy party in our next statistics lab, with enough macaroni and cheese for just me. After all, as a poor starving student I can't be expected to feed our class of animals!

92. "William Shakespeare: The Tragedy of Hamlet, Prince of Denmark," by Jacqueline Ghodsi THE CHARACTERS (in order of appearance):

- HAMLET, Prince of Denmark and student of statistics
- POLONIUS, Hamlet's tutor
- HORATIO, friend to Hamlet and fellow student

Scene: The great library of the castle, in which Hamlet does his lessons

Act I

The day is fair, but the face of Hamlet is clouded. He paces the large room. His tutor, Polonius, is reprimanding Hamlet regarding the latter's recent experience. Horatio is seated at the large table at right stage.

POLONIUS: My Lord, how cans't thou admit that thou hast seen a ghost! It is but a figment of your imagination!

HAMLET: I beg to differ; I know of a certainty that five-and-seventy in one hundred of us, condemned to the whips and scorns of time as we are, have gazed upon a spirit of health, or goblin damn'd, be their intents wicked or charitable.

POLONIUS: If thou dost insist upon thy wretched vision then let me invest your time; be true to thy work and speak to me through the reason of the null and alternate hypotheses. (He turns to Horatio.) Did not Hamlet himself say, "What a piece of work is man, how noble in reason, how infinite in faculties"? Then let not this foolishness persist. Go, Horatio, make a survey of three-and-sixty and discover what the true proportion be. For my part, I will never succumb to this fantasy, but deem man to be devoid of all reason should thy proposal of at least five-and-seventy in one hundred hold true.

HORATIO (to Hamlet): What should we do, my Lord?

HAMLET: Go to thy purpose, Horatio.

HORATIO: To what end, my Lord?

HAMLET: That you must teach me. But let me conjure you by the rights of our fellowship, by the consonance of our youth, but the obligation of our ever-preserved love, be even and direct with me, whether I am right or no.

Horatio exits, followed by Polonius, leaving Hamlet to ponder alone.

Act II

The next day, Hamlet awaits anxiously the presence of his friend, Horatio. Polonius enters and places some books upon the table just a moment before Horatio enters.

POLONIUS: So, Horatio, what is it thou didst reveal through thy deliberations?

HORATIO: In a random survey, for which purpose thou thyself sent me forth, I did discover that one-and-forty believe fervently that the spirits of the dead walk with us. Before my God, I might not this believe, without the sensible and true avouch of mine own eyes.

POLONIUS: Give thine own thoughts no tongue, Horatio. (Polonius turns to Hamlet.) But look to't I charge you, my Lord. Come Horatio, let us go together, for this is not our test. (Horatio and Polonius leave together.)

HAMLET: To reject, or not reject, that is the question: whether 'tis nobler in the mind to suffer the slings and arrows of outrageous statistics, or to take arms against a sea of data, and, by opposing, end them. (Hamlet resignedly attends to his task.)

(Curtain falls)

93. "Untitled," by Stephen Chen

I've often wondered how software is released and sold to the public. Ironically, I work for a company that sells products with known problems. Unfortunately, most of the problems are difficult to create, which makes them difficult to fix. I usually use the test program X, which tests the product, to try to create a specific problem. When the test program is run to make an error occur, the likelihood of generating an error is 1 percent.

So, armed with this knowledge, I wrote a new test program Y that will generate the same error that test program X creates, but more often. To find out if my test program is better than the original, so that I can convince the management that I'm right, I ran my test program to find out how often I can generate the same error. When I ran my test program 50 times, I generated the error twice. While this may not seem much better, I think that I can convince the management to use my test program instead of the original test program. Am I right?

94. "Japanese Girls' Names"

by Kumi Furuichi

It used to be very typical for Japanese girls' names to end with "ko." The trend might have started around my grandmothers' generation and its peak might have been around my mother's generation. "Ko" means "child" in Chinese characters. Parents would name their daughters with "ko" attaching to other Chinese characters that have meanings that they want their daughters to become, such as Sachiko—happy child, Yoshiko—a good child, Yasuko—a healthy child, and so on.

However, I noticed recently that only two out of nine of my Japanese girlfriends at this school have names that end with "ko." More and more, parents seem to have become creative, modernized, and, sometimes, westernized in naming their children.

I have a feeling that, while 70 percent or more of my mother's generation would have names with "ko" at the end, the proportion has dropped among my peers. I wrote down all my Japanese friends', ex-classmates', coworkers', and acquaintances' names that I could remember. Following are the names. Some are repeats. Test to see if the proportion has dropped for this generation.

Ai, Akemi, Akiko, Ayumi, Chiaki, Chie, Eiko, Eri, Eriko, Fumiko, Harumi, Hitomi, Hiroko, Hiroko, Hidemi, Hisako, Hinako, Izumi, Izumi, Junko, Junko, Kana, Kanako, Kanayo, Kayo, Kayoko, Kazumi, Keiko, Keiko, Kei, Kumi, Kumiko, Kyoko, Kyoko, Madoka, Maho, Mai, Maiko, Maki, Miki, Miki, Mikiko, Mina, Minako, Miyako, Momoko, Nana, Naoko, Naoko, Naoko, Noriko, Rieko, Rika, Rika, Rumiko, Rei, Reiko, Reiko, Sachiko, Sachiko, Sachiyo, Saki, Sayaka, Sayoko, Sayuri, Seiko, Shiho, Shizuka, Sumiko, Takako, Takako, Tomoe, Tomoe, Tomoko, Touko, Yasuko, Yasuko, Yasuyo, Yoko, Yoko, Yoko, Yoshiko, Yoshiko, Yoshiko, Yuka, Yuki, Yuki, Yukiko, Yuko, Yuko.

95. "Phillip’s Wish," by Suzanne Osorio

My nephew likes to play
Chasing the girls makes his day.
He asked his mother
If it is okay
To get his ear pierced.
She said, “No way!”
To poke a hole through your ear,
Is not what I want for you, dear.
He argued his point quite well,
Says even my macho pal, Mel,
Has gotten this done.
It’s all just for fun.
C’mon please, mom, please, what the hell.
Again Phillip complained to his mother,
Saying half his friends (including their brothers)
Are piercing their ears
And they have no fears
He wants to be like the others.
She said, “I think it’s much less.
We must do a hypothesis test.
And if you are right,
I won’t put up a fight.
But, if not, then my case will rest.”
We proceeded to call fifty guys
To see whose prediction would fly.
Nineteen of the fifty
Said piercing was nifty
And earrings they’d occasionally buy.
Then there’s the other thirty-one,
Who said they’d never have this done.
So now this poem’s finished.
Will his hopes be diminished,
Or will my nephew have his fun?

96. "The Craven," by Mark Salangsang

Once upon a morning dreary
In stats class I was weak and weary.
Pondering over last night’s homework
Whose answers were now on the board
This I did and nothing more.
While I nodded nearly napping
Suddenly, there came a tapping.
As someone gently rapping,
Rapping my head as I snore.
Quoth the teacher, “Sleep no more.”
“In every class you fall asleep,”
The teacher said, his voice was deep.
“So a tally I’ve begun to keep
Of every class you nap and snore.
The percentage being forty-four.”
“My dear teacher I must confess,
While sleeping is what I do best.
The percentage, I think, must be less,
A percentage less than forty-four.”
This I said and nothing more.
“We’ll see,” he said and walked away,
And fifty classes from that day
He counted till the month of May
The classes in which I napped and snored.
The number he found was twenty-four.
At a significance level of 0.05,
Please tell me am I still alive?
Or did my grade just take a dive
Plunging down beneath the floor?
Upon thee I hereby implore.

97. Toastmasters International cites a report by Gallup Poll that 40 percent of Americans fear public speaking. A student believes that less than 40 percent of students at her school fear public speaking. She randomly surveys 361 schoolmates and finds that 135 report they fear public speaking. Conduct a hypothesis test to determine if the percentage at her school is less than 40.

98. Sixty-eight percent of online courses taught at community colleges nationwide were taught by full-time faculty. To test if 68 percent also represents California’s percent for full-time faculty teaching the online classes, Long Beach City College (LBCC) in California was randomly selected for comparison. In the same year, 34 of the 44 online courses LBCC offered were taught by full-time faculty. Conduct a hypothesis test to determine if 68 percent represents California. Note: For more accurate results, use more California community colleges and this past year's data.

99. According to an article in a local poll, a city found that 14 percent of its residents walk for exercise. Suppose that a survey is conducted to determine this year’s rate. Nine out of 70 randomly chosen city residents replied that they walk for exercise. Conduct a hypothesis test to determine if the rate is still 14 percent or if it has decreased.

100. The mean age of De Anza College students in a previous term was 26.6 years old. An instructor thinks the mean age for online students is older than 26.6. She randomly surveys 56 online students and finds that the sample mean is 29.4 with a standard deviation of 2.1. Conduct a hypothesis test.

101. Registered nurses earned an average annual salary of $69,110. For that same year, a survey was conducted of 41 California registered nurses to determine if the annual salary is higher than $69,110 for California nurses. The sample average was $71,121 with a sample standard deviation of $7,489. Conduct a hypothesis test.

102. La Leche League International reports that the mean age of weaning a child from breastfeeding is age four to five worldwide. In America, most nursing mothers wean their children much earlier. Suppose a random survey is conducted of 21 U.S. mothers who recently weaned their children. The mean weaning age was nine months (3/4 year) with a standard deviation of 4 months. Conduct a hypothesis test to determine if the mean weaning age in the United States is less than four years old.

103. Harley Davidson motorcycles are the largest selling motorcycle in the United States, with 14 percent of all motorcycles sold in 2012. Interestingly, a random sample of 1,945 stolen motorcycles was selected, and it was found that just 8 percent of them were Harleys. Is there good evidence that the proportion of Harleys among stolen motorcycles is significantly less than their share of all motorcycles? After conducting the test, what decision and conclusion would you make?

a. Reject H_0: There is sufficient evidence to conclude that the proportion of Harleys stolen is significantly less than their share of all motorcycles
b. Do not reject H_0: There is not sufficient evidence to conclude that the proportion of Harleys stolen is significantly less than their share of all motorcycles
c. Do not reject H_0: There is sufficient evidence to conclude that the proportion of Harleys stolen is significantly more than their share of all motorcycles
d. Reject H_0: There is sufficient evidence to conclude that the proportion of Harleys stolen is significantly more than their share of all motorcycles

104. A statistics instructor believes that fewer than 20 percent of Evergreen Valley College (EVC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 of them attended the midnight showing.

At a 1 percent level of significance, what is an appropriate conclusion?

a. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20 percent.
b. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is more than 20 percent.
c. There is sufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is less than 20 percent.
d. There is insufficient evidence to conclude that the percent of EVC students who attended the midnight showing of Harry Potter is at least 20 percent.

105. Previously, an organization reported that teenagers spent 4.5 hours per week, on average, on the phone. The organization thinks that, currently, the mean is higher. Fifteen randomly chosen teenagers were asked how many hours per week they spend on the phone. The sample mean was 4.75 hours with a sample standard deviation of 2.0. Conduct a hypothesis test.

At a significance level of $a = 0.05$, what is the correct conclusion?

a. There is enough evidence to conclude that the mean number of hours is more than 4.75.
b. There is enough evidence to conclude that the mean number of hours is more than 4.5.
c. There is not enough evidence to conclude that the mean number of hours is more than 4.5.
d. There is not enough evidence to conclude that the mean number of hours is more than 4.75.

Hypothesis testing: For the following 10 exercises, answer each question.

a. State the null and alternate hypotheses.
b. State the *p*-value.
c. State alpha.
d. What is your decision?
e. Write a conclusion.
f. Answer any other questions asked in the problem.

106. A research group is studying a particular infectious disease. In 2011 at least 18 percent of nursing home residents had the disease. An Introduction to Statistics class in Daviess County, KY, conducted a hypothesis test at the nursing home (approximately 1,200 residents) to determine if the local nursing home's incidence was lower. One hundred fifty residents were chosen at random and surveyed. Of the 150 residents surveyed, 82 have the disease. Use a significance level of 0.05 and, using appropriate statistical evidence, conduct a hypothesis test and state the conclusions.

107. A recent survey in the *New York Times Almanac* indicated that 48.8 percent of families own stock. A broker wanted to determine if this survey could be valid. He surveyed a random sample of 250 families and found that 142 owned some type of stock. At the 0.05 significance level, can the survey be considered to be accurate?

108. Driver error can be listed as the cause of approximately 54 percent of all fatal auto accidents, according to the American Automobile Association. Thirty randomly selected fatal accidents are examined, and it is determined that 14 were caused by driver error. Using $\alpha = 0.05$, is the AAA proportion accurate?

109. The U.S. Department of Energy reported that 51.7 percent of homes were heated by natural gas. A random sample of 221 homes in Kentucky found that 115 were heated by natural gas. Does the evidence support the claim for Kentucky at the $\alpha = 0.05$ level? Are the results applicable across the country? Why?

110. For Americans using library services, the American Library Association claims that at most 67 percent of patrons borrow books. The library director in Owensboro, KY, feels this is not true, so she asked a local college statistic class to conduct a survey. The class randomly selected 100 patrons and found that 82 borrowed books. Did the class demonstrate that the percentage was higher in Owensboro, KY? Use $\alpha = 0.01$ level of significance. What is the possible proportion of patrons who do borrow books from the Owensboro Library?

111. The Weather Underground reported that the mean amount of summer rainfall for the northeastern United States is at least 11.52 inches. Ten cities in the northeast are randomly selected and the mean rainfall amount is calculated to be 7.42 inches with a standard deviation of 1.3 inches. At the $\alpha = 0.05$ level, can it be concluded that the mean rainfall was below the reported average? What if $\alpha = 0.01$? Assume the amount of summer rainfall follows a normal distribution.

112. A survey in the *New York Times Almanac* finds the mean commute time (one way) is 25.4 minutes for the 15 largest US cities. The Austin, TX, chamber of commerce feels that Austin's commute time is less and wants to publicize this fact. The mean for 25 randomly selected commuters is 22.1 minutes with a standard deviation of 5.3 minutes. At the $\alpha = 0.10$ level, is the Austin, TX, commute significantly less than the mean commute time for the 15 largest U.S. cities?

113. A report by the Gallup Poll found that a woman visits her doctor, on average, at most 5.8 times each year. A random sample of 20 women results in these yearly visit totals:
3; 2; 1; 3; 7; 2; 9; 4; 6; 6; 8; 0; 5; 6; 4; 2; 1; 3; 4; 1
At the $\alpha = 0.05$ level, can it be concluded that the sample mean is higher than 5.8 visits per year?

114. According to the *New York Times Almanac* the mean family size in the United States is 3.18. A sample of a college math class resulted in the following family sizes:
5; 4; 5; 4; 4; 3; 6; 4; 3; 3; 5; 5; 6; 3; 3; 2; 7; 4; 5; 2; 2; 2; 3; 2
At $\alpha = 0.05$, is the class's mean family size greater than the national average? Does the *Almanac* result remain valid? Why?

115. The student academic group on a college campus claims that freshman students study at least 2.5 hours per day, on average. One Introduction to Statistics class was skeptical. The class took a random sample of 30 freshman students and found a mean study time of 137 minutes with a standard deviation of 45 minutes. At $\alpha = 0.01$ level, is the student academic group's claim correct?

REFERENCES

9.1 Null and Alternative Hypotheses

Centers for Disease Control and Prevention. (n.d.). *Physical activity facts*. Retrieved from http://www.cdc.gov/healthyschools/physicalactivity/facts.htm

National Institute of Mental Health. (n.d.). *Publications about depression*. Retrieved from http://www.nimh.nih.gov/publicat/depression.cfm

9.5 Additional Information and Full Hypothesis Test Examples

American Automobile Association. (n.d.). Retrieved from www.aaa.com

American Library Association. (n.d.). Retrieved from www.ala.org

Amit Schitai. (n.d.). *Data*.

Bureau of Labor Statistics. (n.d.). *Occupational employment statistics*. Retrieved from http://www.bls.gov/oes/current/oes291111.htm

Centers for Disease Control and Prevention. (n.d.). Retrieved from www.cdc.gov

De Anza College. (2006). *Foothill-De Anza Community College District*. Retrieved from http://research.fhda.edu/factbook/DAdemofs/Fact_sheet_da_2006w.pdf

Federal Bureau of Investigation. (n.d.). *Uniform crime reports and index of crime in Daviess in the state of Kentucky enforced by Daviess County from 1985 to 2005*. Retrieved from http://www.disastercenter.com/kentucky/crime/3868.htm

Gallup. (n.d.). Retrieved from www.gallup.com

Johansen, C., Boice, Jr. J., McLaughlin, J., & Olsen, J. (2001, Feb. 7). Cellular telephones and cancer—a nationwide cohort study in Denmark. *Journal of National Cancer Institute*, 93(3), 203–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11158188

La Leche League International. (n.d.). Retrieved from http://www.lalecheleague.org/Law/BAFeb01.html

Online Learning Consortium. (2005 Nov.). *Growing by degrees: Online education in the United States, 2005*. Newburyport, MA: Allen, I. E., & Seaman, J. Available at http://files.eric.ed.gov/fulltext/ED530062.pdf

Toastmasters International. (n.d.). Retrieved from http://toastmasters.org/artisan/detail.asp?CategoryID=1&SubCategoryID=10&ArticleID=429&Page=1

U.S. Census Bureau. (n.d.). *Language use*. Retrieved from https://www.census.gov/topics/population/language-use.html

U.S. Census Bureau. (n.d.). *QuickFacts*. Retrieved from https://www.census.gov/quickfacts/table/PST045216/00

U.S. Department of Energy. (n.d.). Retrieved from http://energy.gov

Weather Underground. (n.d.). Retrieved from www.wunderground.com

SOLUTIONS

1 The random variable is the mean Internet speed in megabits per second.

3 The random variable is the mean number of children an American family has.

5 The random variable is the proportion of people picked at random in Times Square visiting the city.

7

a. H_0: $p = 0.42$

b. H_a: $p < 0.42$

9

a. H_0: $\mu = 15$

b. H_a: $\mu \neq 15$

11 Type I: The mean price of mid-sized cars is $32,000, but we conclude that it is not $32,000. Type II: The mean price of mid-sized cars is not $32,000, but we conclude that it is $32,000.

13 α = the probability that you think the bag cannot withstand –15 degrees F, when, in fact, it can. β = the probability that you think the bag can withstand –15 degrees F, when, in fact, it cannot.

15 Type I: The procedure will go well, but the doctors think it will not. Type II: The procedure will not go well, but the doctors think it will.

17 0.019

19 0.998

21 A normal distribution or a Student's *t*-distribution

23 Use a Student's *t*-distribution

25 a normal distribution for a single population mean

27 It must be approximately normally distributed.

29 They must both be greater than five.

31 binomial distribution

33 The outcome of winning is very unlikely.

35 H_0: $\mu >= 73$
H_a: $\mu < 73$
The *p*-value is almost zero, which means there is sufficient data to conclude that the mean height of high school students who play basketball on the school team is less than 73 inches at the 5 percent level. The data do support the claim.

37 The shaded region shows a low *p*-value.

39 Do not reject H_0.

41 means

43 the mean time spent on homework for 26 students

45

a. 3

b. 1.5

c. 1.8

d. 26

47 $\bar{X} \sim N\left(2.5, \frac{1.5}{\sqrt{26}}\right)$

49 This is a left-tailed test.

51 This is a two-tailed test.

53

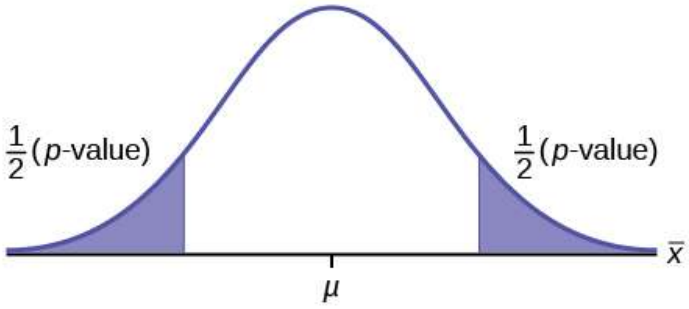


Figure 9.23

55 a right-tailed test

57 a left-tailed test

59 This is a left-tailed test.

61 This is a two-tailed test.

62

a. H_0: $\mu = 34$; H_a: $\mu \neq 34$

b. H_0: $p \leq 0.60$; H_a: $p > 0.60$

c. H_0: $\mu \geq 100{,}000$; H_a: $\mu < 100{,}000$

d. H_0: $p = 0.29$; H_a: $p \neq 0.29$

e. H_0: $p = 0.05$; H_a: $p < 0.05$

f. H_0: $\mu \leq 10$; H_a: $\mu > 10$

g. H_0: $p = 0.50$; H_a: $p \neq 0.50$

h. H_0: $\mu = 6$; H_a: $\mu \neq 6$

i. H_0: $p \geq 0.11$; H_a: $p < 0.11$

j. H_0: $\mu \leq 20{,}000$; H_a: $\mu > 20{,}000$

64 c

66

a. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.

b. Type I error: We conclude that more than 60 percent of Americans vote in presidential elections, when the actual percentage is at most 60 percent.Type II error: We conclude that at most 60 percent of Americans vote in presidential elections when, in fact, more than 60 percent do.

c. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is at least $100,000. Type II error: We conclude that the mean starting salary is at least $100,000 when, in fact, it is less than $100,000.

d. Type I error: We conclude that the proportion of high school seniors who take physical education daily is not 29%, when it really is 29%. Type II error: We conclude that the proportion of high school seniors who take physical education daily is 29% when, in fact, it is not 29%.

e. Type I error: We conclude that fewer than 5 percent of adults ride the bus to work in Los Angeles, when the percentage that do is really 29%. Type II error: We conclude that 29%. or more adults ride the bus to work in Los Angeles when, in fact, fewer that 29% do.

f. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10. Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.

g. Type I error: We conclude that the proportion of Americans who prefer to live away from cities is not about half, though the actual proportion is about half. Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.

h. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We conclude that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.

i. Type I error: We conclude that the proportion is less than 11 percent, when it is really at least 11 percent. Type II error: We conclude that the proportion of women who develop breast cancer is at least 11 percent, when in fact it is less than 11 percent.

j. Type I error: We conclude that the average tuition cost at private universities is more than $20,000, though in reality it is at most $20,000. Type II error: We conclude that the average tuition cost at private universities is at most $20,000 when, in fact, it is more than $20,000.

68 b

70 d

72 d

74

a. H_0: $\mu \geq 50{,}000$

b. H_a: $\mu < 50{,}000$

c. Let $\bar{X}$ = the average lifespan of a brand of tires.

d. normal distribution

e. $z = -2.315$

f. p-value = 0.0103

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: There is sufficient evidence to conclude that the mean lifespan of the tires is less than 50,000 miles.

i. (43,537, 49,463)

75

a. H_0: $\mu \geq 35.5$

b. H_a: $\mu < 35.5$

c. Let $\bar{x}$ = the average mpg for the sample of cars and trucks in the fleet

d. normal distribution

e. z = -0.648

f. p-value = 0.2578

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. Conclusion: There is sufficient evidence to support the claim that the manufacturer's fleet meets the fuel economy standards in the 2016 policy.

i. (31.88 mpg, 37.32 mpg)

76

a. H_0: $\mu = \$1.00$

b. H_a: $\mu \neq \$1.00$

c. Let $\bar{x}$ = the average cost of a daily newspaper.

d. normal distribution

e. $z = –0.866$

f. p-value = 0.3865

g. Check student's solution.

h.
   i. Alpha: 0.01
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.01.
   iv. Conclusion: There is sufficient evidence to support the claim that the mean cost of daily papers is $1. The mean cost could be $1.

i. ($0.84, $1.06)

78

a. H_0: $\mu = 10$

b. H_a: $\mu \neq 10$

c. Let $\bar{X}$ = the mean number of sick days an employee takes per year.

d. Student's *t*-distribution

e. $t = -1.12$

f. *p*-value = 0.300

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The *p*-value is greater than 0.05.
   iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that the mean number of sick days is not 10.

i. (4.9443, 11.806)

80

a. H_0: $p \geq 0.6$

b. H_a: $p < 0.6$

c. Let P' = the proportion of students who feel more enriched as a result of taking elementary statistics.

d. normal for a single proportion

e. 1.12

f. *p*-value = 0.1308

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The *p*-value is greater than 0.05.
   iv. Conclusion: There is insufficient evidence to conclude that less than 60 percent of her students feel more enriched.

i. Confidence interval: (0.409, 0.654)
The "plus-4s" confidence interval is (0.411, 0.648)

82

a. H_0: $\mu = 4$

b. H_a: $\mu \neq 4$

c. Let $\bar{X}$ the average I.Q. of a set of brown trout.

d. two-tailed Student's *t*-test

e. $t = 1.95$

f. *p*-value = 0.076

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The *p*-value is greater than 0.05

iv. Conclusion: There is insufficient evidence to conclude that the average IQ of brown trout is not four.

i. (3.8865, 5.9468)

84

a. H_0: $p \geq 0.13$

b. H_a: $p < 0.13$

c. Let P' = the proportion of Americans who have the disease

d. normal for a single proportion

e. –2.688

f. p-value = 0.0036

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: The p-value is less than 0.05.

iv. Conclusion: There is sufficient evidence to conclude that the percentage of Americans who have been diagnosed with the disease is less than 13 percent.

i. (0, 0.0623).
The *plus-4s* confidence interval is (0.0022, 0.0978)

86

a. H_0: $\mu \geq 129$

b. H_a: $\mu < 129$

c. Let $\bar{X}$ = the average time in seconds that Terri finishes Lap 4.

d. Student's t-distribution

e. $t = 1.209$

f. 0.8792

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Do not reject the null hypothesis.

iii. Reason for decision: The p-value is greater than 0.05.

iv. Conclusion: There is insufficient evidence to conclude that Terri's mean lap time is less than 129 seconds.

i. (128.63, 130.37)

88

a. H_0: $p = 0.60$

b. H_a: $p < 0.60$

c. Let P' = the proportion of family members who shed tears at a reunion.

d. normal for a single proportion

e. –1.71

f. 0.0438

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: p-value < alpha

iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the proportion of family members who shed tears at a reunion is less than 0.60. However, the test is weak because the p-value and alpha are quite close, so other tests should be done.

i. We are 95 percent confident that between 38.29 percent and 61.71 percent of family members will shed tears at a family reunion. (0.3829, 0.6171). The *plus-4s* confidence interval (see chapter 8) is (0.3861, 0.6139)

Note that here the *large-sample* 1 – PropZTest provides the approximate p-value of 0.0438. Whenever a p-value based on a normal approximation is close to the level of significance, the exact p-value based on binomial probabilities should be calculated whenever possible. This is beyond the scope of this course.

89

a. H_0: $\mu \geq 22$

b. H_a: $\mu < 22$

c. Let $\bar{X}$ = the mean number of bubbles per blow.

d. Student's t-distribution

e. –2.667

f. p-value = 0.00486

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: The p-value is less than 0.05.

iv. Conclusion: There is sufficient evidence to conclude that the mean number of bubbles per blow is less than 22.

i. (18.501, 21.499)

91

a. H_0: $\mu \leq 1$

b. H_a: $\mu > 1$

c. Let $\bar{X}$ = the mean cost in dollars of macaroni and cheese in a certain town.

d. Student's t-distribution

e. $t = 0.340$

f. p-value = 0.36756

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Do not reject the null hypothesis.

iii. Reason for decision: The p-value is greater than 0.05

iv. Conclusion: The mean cost could be $1, or less. At the 5 percent significance level, there is insufficient evidence to conclude that the mean price of a box of macaroni and cheese is more than $1.

i. (0.8291, 1.241)

93

a. H_0: $p = 0.01$

b. H_a: $p > 0.01$

c. Let P' = the proportion of errors generated

d. Normal for a single proportion

e. 2.13

f. 0.0165

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: The *p*-value is less than 0.05.

iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the proportion of errors generated is more than 0.01.

i. Confidence interval: (0, 0.094).
The *plus-4s* confidence interval is (0.004, 0.144).

95

a. H_0: $p = 0.50$

b. H_a: $p < 0.50$

c. Let P' = the proportion of friends that has a pierced ear.

d. normal for a single proportion

e. –1.70

f. *p*-value = 0.0448

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: The *p*-value is less than 0.05. (However, they are very close.)

iv. Conclusion: There is sufficient evidence to support the claim that less than 50 percent of his friends have pierced ears.

i. Confidence interval: (0.245, 0.515): The *plus-4s* confidence interval is (0.259, 0.519).

97

a. H_0: $p = 0.40$

b. H_a: $p < 0.40$

c. Let P' = the proportion of schoolmates who fear public speaking.

d. normal for a single proportion

e. –1.01

f. *p*-value = 0.1563

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Do not reject the null hypothesis.

iii. Reason for decision: The *p*-value is greater than 0.05.

iv. Conclusion: There is insufficient evidence to support the claim that less than 40 percent of students at the school fear public speaking.

i. Confidence interval: (0.3241, 0.4240): The *plus-4s* confidence interval is (0.3257, 0.4250).

99

a. H_0: $p = 0.14$

b. H_a: $p < 0.14$

c. Let P' = the proportion of nursing home residents that have the disease.

d. normal for a single proportion

e. –0.2756

f. p-value = 0.3914

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: The p-value is greater than 0.05.
   iv. At the 5 percent significance level, there is insufficient evidence to conclude that the proportion of nursing home residents that have the disease is less than 0.14.

i. Confidence interval: (0.0502, 0.2070): The *plus-4s* confidence interval (see chapter 8) is (0.0676, 0.2297).

101

a. H_0: μ = 69,110

b. H_a: μ > 69,110

c. Let $\bar{X}$ = the mean salary in dollars for California registered nurses.

d. Student's t-distribution

e. t = 1.719

f. p-value: 0.0466

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: The p-value is less than 0.05.
   iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean salary of California registered nurses exceeds $69,110.

i. ($68,757, $73,485)

103

a. H_0: p ≥ 0.14, H_a: p < 0.14

b. p-value < 0.0002

c. Alpha: 0.05

d. Reject the null hypothesis.

e. At the 5 percent significance level, there is sufficient evidence to conclude that the proportion of Harleys stolen is significantly less than their share of all motorcycles. (conclusion a)

105 c

107

a. H_0: p = 0.488 H_a: $p \neq 0.488$

b. p-value = 0.0114

c. alpha = 0.05

d. Reject the null hypothesis.

e. At the 5 percent level of significance, there is enough evidence to conclude that 48.8 percent of families own stocks.

f. The survey does not appear to be accurate.

109

a. H_0: $p = 0.517$ H_a: $p \neq 0.517$
b. p-value = 0.9203.
c. alpha = 0.05.
d. Do not reject the null hypothesis.
e. At the 5 percent significance level, there is not enough evidence to conclude that the proportion of homes in Kentucky that are heated by natural gas is 0.517.
f. However, we cannot generalize this result to the entire nation. First, the sample's population is only the state of Kentucky. Second, it is reasonable to assume that homes in the extreme north and south will have extreme high usage and low usage, respectively. We would need to expand our sample base to include these possibilities if we wanted to generalize this claim to the entire nation.

111

a. H_0: $\mu \geq 11.52$ H_a: $\mu < 11.52$
b. p-value = 0.000002 which is almost 0.
c. alpha = 0.05.
d. Reject the null hypothesis.
e. At the 5 percent significance level, there is enough evidence to conclude that the mean amount of summer rain in the northeaster US is less than 11.52 inches, on average.
f. We would make the same conclusion if alpha was 1 percent because the p-value is almost 0.

113

a. H_0: $\mu \leq 5.8$ H_a: $\mu > 5.8$
b. p-value = 0.9987
c. alpha = 0.05
d. Do not reject the null hypothesis.
e. At the 5 percent level of significance, there is not enough evidence to conclude that a woman visits her doctor, on average, more than 5.8 times a year.

115

a. H_0: $\mu \geq 150$ H_a: $\mu < 150$
b. p-value = 0.0622
c. alpha = 0.01
d. Do not reject the null hypothesis.
e. At the 1 percent significance level, there is not enough evidence to conclude that freshmen students study less than 2.5 hours per day, on average.
f. The student academic group's claim appears to be correct.

10 | HYPOTHESIS TESTING WITH TWO SAMPLES

Figure 10.1 If you want to test a claim that involves two groups (the types of breakfasts eaten east and west of the Mississippi River), you can use a slightly different technique when conducting a hypothesis test. (credit: Chloe Lim)

Introduction

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Classify hypothesis tests by type
- Conduct and interpret hypothesis tests for two population means, population standard deviations known
- Conduct and interpret hypothesis tests for two population means, population standard deviations unknown
- Conduct and interpret hypothesis tests for two population proportions
- Conduct and interpret hypothesis tests for matched or paired samples

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart

attacks. Over the last few years, newspapers and magazines have reported various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether the SAT or GRE preparatory courses really help raise their scores.

You have learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is the same, just expanded.

To compare two means or two proportions, you work with two groups. The groups are classified as *independent groups* or *matched pairs*. *Independent groups* consist of two samples that are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. *Matched pairs* consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

NOTE

This chapter relies on either a calculator or a computer to calculate the degrees of freedom, the test statistics, and *p* values. TI-83+ and TI-84 instructions are included, as well as the test statistic formulas. When using a TI-83+ or TI-84 calculator, we do not need to separate two population means, independent groups, or population variances unknown into large and small sample sizes. However, most statistical computer software has the ability to differentiate these tests.

This chapter deals with the following hypothesis tests:

- Independent groups (samples are independent)
  - Test of two population means
  - Test of two population proportions
- Matched or paired samples (samples are dependent)
  - Test of the two population proportions by testing one population mean of differences

10.1 | Two Population Means with Unknown Standard Deviations

1. The two independent samples are simple random samples from two distinct populations.
2. For the two distinct populations
   - if the sample sizes are small, the distributions are important (should be normal), and
   - if the sample sizes are large, the distributions are not important (need not be normal)

The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Aspin-Welch *t*-test. The degrees of freedom formula was developed by Aspin-Welch.

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. To account for the variation, we take the difference of the sample means, $\bar{X}_1 - \bar{X}_2$, and divide by the standard error to standardize the difference. The result is a *t*-score test statistic.

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated **standard deviation**, or **standard error**, of the difference in sample means, $\bar{X}_1 - \bar{X}_2$.

The standard error is calculated as follows:

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$$

The test statistic (*t*-score) is calculated as follows:

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

where

- s_1 and s_2, the sample standard deviations, are estimates of σ_1 and σ_2, respectively,
- σ_1 and σ_1 are the unknown population standard deviations,
- $\bar{x}_1$ and $\bar{x}_2$ are the sample means, and
- μ_1 and μ_2 are the population means.

The number of **degrees of freedom (*df*)** requires a somewhat complicated calculation. However, a computer or calculator calculates it easily. The *df* are not always a whole number. The test statistic calculated previously is approximated by the Student's *t*-distribution with *df* as follows:

Degrees of freedom

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

When both sample sizes n_1 and n_2 are five or larger, the Student's *t* approximation is very good. Notice that the sample variances $(s_1)^2$ and $(s_2)^2$ are not pooled. (If the question comes up, do not pool the variances.)

It is not necessary to compute this by hand. A calculator or computer easily computes it.

Example 10.1 Independent groups

The average amount of time boys and girls aged 7 to 11 spend playing sports each day is believed to be the same. A study is done and data are collected, resulting in the data in **Table 10.1**. Each populations has a normal distribution.

	Sample Size	Average Number of Hours Playing Sports per Day	Sample Standard Deviation
Girls	9	2	0.866
Boys	16	3.2	1.00

Table 10.1

Is there a difference in the mean amount of time boys and girls aged 7 to 11 play sports each day? Test at the 5 percent level of significance.

Solution 10.1

The population standard deviations are not known. Let *g* be the subscript for girls and *b* be the subscript for boys. Then, μ_g is the population mean for girls and μ_b is the population mean for boys. This is a test of two **independent groups**, two population **means**.

Random variable: $\bar{X}_g - \bar{X}_b$ = difference in the sample mean amount of time girls and boys play sports each day.

H_0: $\mu_g = \mu_b$ H_0: $\mu_g - \mu_b = 0$

H_a: $\mu_g \neq \mu_b$ H_a: $\mu_g - \mu_b \neq 0$

The words *the same* tell you H_0 has an "=". Since there are no other words to indicate H_a, assume it says *is different.* This is a two-tailed test.

Distribution for the test: Use t_{df} where *df* is calculated using the *df* formula for independent groups, two population means. Using a calculator, *df* is approximately 18.8462. *Do not pool the variances.*

Calculate the *p*-value using a Student's *t*-distribution: *p*-value = 0.0054

Graph:

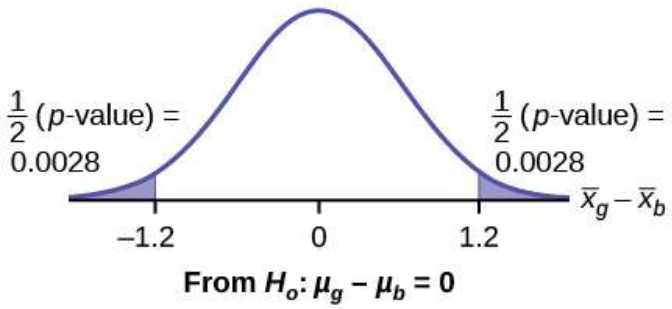


Figure 10.2

$s_g = 0.866$

$s_b = 1$

So, $\bar{x}_g - \bar{x}_b = 2 - 3.2 = -1.2$

Half the *p*-value is below –1.2, and half is above 1.2.

Make a decision: Since $\alpha >$ *p*-value, reject H_0. This means you reject $\mu_g = \mu_b$. The means are different.

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT. Arrow over to TESTS and press 4:2-SampTTest. Arrow over to Stats and press ENTER. Arrow down and enter 2 for the first sample mean, 0.866 for Sx1, 9 for n1, 3.2 for the second sample mean, 1 for Sx2, and 16 for n2. Arrow down to μ1: and arrow to does not equal μ2. Press ENTER. Arrow down to Pooled: and No. Press ENTER. Arrow down to Calculate and press ENTER. The *p*-value is $p = 0.0054$, the *df*s are approximately 18.8462, and the test statistic is –3.14. Do the procedure again, but instead of Calculate do Draw.

Conclusion—: At the 5 percent level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that girls and boys aged 7 to 11 play sports per day is different (mean number of hours boys aged 7 to 11 play sports per day is greater than the mean number of hours played by girls OR the mean number of hours girls aged 7 to 11 play sports per day is greater than the mean number of hours played by boys).

Try It Σ

10.1 Two samples are shown in **Table 10.2**. Both have normal distributions. The means for the two populations are thought to be the same. Is there a difference in the means? Test at the 5 percent level of significance.

	Sample Size	Sample Mean	Sample Standard Deviation
Population A	25	5	1
Population B	16	4.7	1.2

Table 10.2

NOTE

When the sum of the sample sizes is larger than 30 ($n_1 + n_2 > 30$), you can use the normal distribution to approximate the Student's *t*.

Example 10.2

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is 4 math classes with a standard deviation of 1.5 math classes. College B samples nine graduates. Their average is 3.5 math classes with a standard deviation of 1 math class. The community group believes that a student who graduates from College A *has taken more math classes*, on average. Both populations have a normal distribution. Test at a 1 percent significance level. Answer the following questions:

a. Is this a test of two means or two proportions?

Solution 10.2
a. two means

b. Are the populations standard deviations known or unknown?

Solution 10.2
b. unknown

c. Which distribution do you use to perform the test?

Solution 10.2
c. Student's *t*

d. What is the random variable?

Solution 10.2
d. $\bar{X}_A - \bar{X}_B$

e. What are the null and alternate hypotheses? Write the null and alternate hypotheses in symbols.

Solution 10.2

e. $H_o : \mu_A \leq \mu_B$

$H_a : \mu_A > \mu_B$

f. Is this test right-, left-, or two-tailed?

Solution 10.2

f.

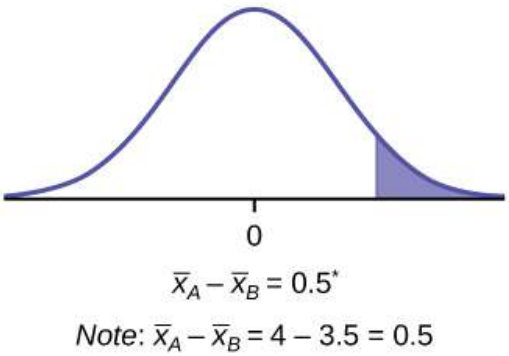


Figure 10.3

right

g. What is the *p*-value?

Solution 10.2

g. 0.1928

h. Do you reject or not reject the null hypothesis?

Solution 10.2

h. do not reject

i. **Conclusion:**

Solution 10.2

i. At the 1 percent level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from College A has taken more math classes, on average, than a student who graduates from College B.

Try It

10.2 A study is done to determine if Company A retains its workers longer than Company B. Company A samples 15 workers, and their average time with the company is 5 years with a standard deviation of 1.2. Company B samples 20 workers, and their average time with the company is 4.5 years with a standard deviation of 0.8. The populations are normally distributed.

a. Are the population standard deviations known?

b. Conduct an appropriate hypothesis test. At the 5 percent significance level, what is your conclusion?

Example 10.3

A professor at a large community college wanted to determine whether there is a difference in the means of final exam scores between students who took his statistics course online and the students who took his face-to-face statistics class. He believed that the mean of the final exam scores for the online class would be lower than that of the face-to-face class. Was the professor correct? The randomly selected 30 final exam scores from each group are listed in **Table 10.3** and **Table 10.4**.

67.6	41.2	85.3	55.9	82.4	91.2	73.5	94.1	64.7	64.7
70.6	38.2	61.8	88.2	70.6	58.8	91.2	73.5	82.4	35.5
94.1	88.2	64.7	55.9	88.2	97.1	85.3	61.8	79.4	79.4

Table 10.3 Online Class

77.9	95.3	81.2	74.1	98.8	88.2	85.9	92.9	87.1	88.2
69.4	57.6	69.4	67.1	97.6	85.9	88.2	91.8	78.8	71.8
98.8	61.2	92.9	90.6	97.6	100	95.3	83.5	92.9	89.4

Table 10.4 Face-to-Face Class

Is the mean of the final exam scores of the online class lower than the mean of the final exam scores of the face-to-face class? Test at a 5 percent significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
f. Is this test right-, left-, or two-tailed?
g. What is the *p*-value?
h. Do you reject or not reject the null hypothesis?
i. At the _____ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.

(See the conclusion in **Example 10.2**, and write yours in a similar fashion.)

Using the TI-83, 83+, 84, 84+ Calculator

First put the data for each group into two lists (such as L1 and L2). Press STAT. Arrow over to TESTS and press 4:2SampTTest. Make sure Data is highlighted and press ENTER. Arrow down and enter L1 for the first list and L2 for the second list. Arrow down to μ_1: and arrow to $\neq \mu_2$ (does not equal). Press ENTER. Arrow down to Pooled: No. Press ENTER. Arrow down to Calculate and press ENTER.

NOTE

Be careful not to mix up the information for Group 1 and Group 2!

Solution 10.3

a. two means

b. unknown

c. Student's *t*

d. $\bar{X}_1 - \bar{X}_2$

e. 1. H_0: $\mu_1 = \mu_2$ Null hypothesis: The means of the final exam scores are equal for the online and face-to-face statistics classes.

   2. H_a: $\mu_1 < \mu_2$ Alternative hypothesis: The mean of the final exam scores of the online class is less than the mean of the final exam scores of the face-to-face class.

f. left-tailed

g. *p*-value = 0.0011

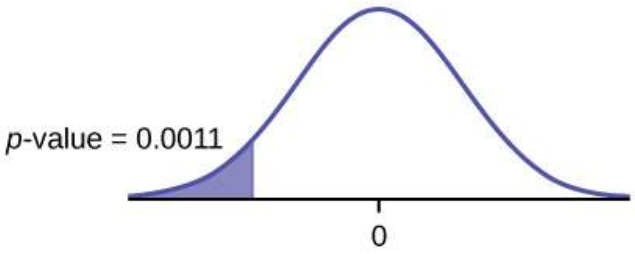


Figure 10.4

h. Reject the null hypothesis.

i. The professor was correct. The evidence shows that the mean of the final exam scores for the online class is lower than that of the face-to-face class.
At the 5 percent level of significance, from the sample data, there is (is/is not) sufficient evidence to conclude that the mean of the final exam scores for the online class is less than the mean of final exam scores of the face-to-face class.

Cohen's Standards for Small, Medium, and Large Effect Sizes

Cohen's *d* is a measure of effect size based on the differences between two means. Cohen's *d*, named for U.S. statistician Jacob Cohen, measures the relative strength of the differences between the means of two populations based on sample data. The calculated value of effect size is then compared to Cohen's standards of small, medium, and large effect sizes.

Size of Effect	*d*
Small	0.2
medium	0.5
Large	0.8

Table 10.5 Cohen's Standard Effect Sizes

Cohen's *d* is the measure of the difference between two means divided by the pooled standard deviation: $d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$

where $s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$.

Example 10.4

Calculate Cohen's *d* for **Example 10.2**. Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

Solution 10.4

$\mu_1 = 4$ $s_1 = 1.5$ $n_1 = 11$
$\mu_2 = 3.5$ $s_2 = 1$ $n_2 = 9$
$d = 0.384$
The effect is small because 0.384 is between Cohen's value of 0.2 for small effect size and 0.5 for medium effect size. The size of the differences of the means for the two colleges is small, indicating that there is not a significant difference between them.

Example 10.5

Calculate Cohen's *d* for **Example 10.3**. Is the size of the effect small, medium, or large? Explain what the size of the effect means for this problem.

Solution 10.5

$d = 0.834$; large, because 0.834 is greater than Cohen's 0.8 for a large effect size. The size of the differences between the means of the final exam scores of online students and students in a face-to-face class is large, indicating a significant difference.

Try It

10.5 Weighted alpha is a measure of risk-adjusted performance of stocks over a period of a year. A high positive weighted alpha signifies a stock whose price has risen, while a small positive weighted alpha indicates an unchanged stock price during the time period. Weighted alpha is used to identify companies with strong upward or downward trends. The weighted alpha for the top 30 stocks of banks in the Northeast and in the West as identified by Nasdaq on May 24, 2013 are listed in **Table 10.6** and **Table 10.7**, respectively.

94.2	75.2	69.6	52.0	48.0	41.9	36.4	33.4	31.5	27.6
77.3	71.9	67.5	50.6	46.2	38.4	35.2	33.0	28.7	26.5
76.3	71.7	56.3	48.7	43.2	37.6	33.7	31.8	28.5	26.0

Table 10.6 Northeast

126.0	70.6	65.2	51.4	45.5	37.0	33.0	29.6	23.7	22.6
116.1	70.6	58.2	51.2	43.2	36.0	31.4	28.7	23.5	21.6
78.2	68.2	55.6	50.3	39.0	34.1	31.0	25.3	23.4	21.5

Table 10.7 West

Is there a difference in the weighted alpha of the top 30 stocks of banks in the Northeast and in the West? Test at a 5 percent significance level. Answer the following questions:

a. Is this a test of two means or two proportions?
b. Are the population standard deviations known or unknown?
c. Which distribution do you use to perform the test?
d. What is the random variable?
e. What are the null and alternative hypotheses? Write the null and alternative hypotheses in words and in symbols.
f. Is this test right-, left-, or two-tailed?
g. What is the *p*-value?
h. Do you reject or not reject the null hypothesis?
i. At the _____ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.
j. Calculate Cohen's *d* and interpret it.

10.2 | Two Population Means with Known Standard Deviations

Even though this situation is not likely (knowing the population standard deviations), the following example illustrates hypothesis testing for independent means, known population standard deviations. The sampling distribution for the difference between the means is normal, and both populations must be normal. The random variable is $\bar{X}_1 - \bar{X}_2$. The normal distribution has the following format:
Normal distribution is

$$\bar{X}_1 - \bar{X}_2 \sim N\left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right].$$

The standard deviation is

$$\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}.$$

The test statistic (*z*-score) is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}}.$$

Example 10.6

Independent groups, population standard deviations known: The mean lasting time of two competing floor waxes is to be compared. Twenty floors are randomly assigned to test each wax. Both populations have a normal distribution. The data are recorded in **Table 10.8**.

Wax	Sample Mean Number of Months Floor Wax Lasts	Population Standard Deviation
1	3	0.33
2	2.9	0.36

Table 10.8

Does the data indicate that Wax 1 is more effective than Wax 2? Test at a 5 percent level of significance.

Solution 10.6

This is a test of two independent groups, two population means, population standard deviations known.

Random Variable: $\bar{X}_1 - \bar{X}_2$ = difference in the mean number of months the competing floor waxes last.

H_0: $\mu_1 \leq \mu_2$

H_a: $\mu_1 > \mu_2$

The words *is more effective* says that **Wax 1 lasts longer than Wax 2**, on average. *Longer* is a > symbol and goes into H_a. Therefore, this is a right-tailed test.

Distribution for the test: The population standard deviations are known, so the distribution is normal. Using the formula, the distribution is

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \sqrt{\frac{0.33^2}{20} + \frac{0.36^2}{20}}\right).$$

Since $\mu_1 \leq \mu_2$, then $\mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is zero.

Calculate the *p* value using the normal distribution: p value = 0.1799

Graph:

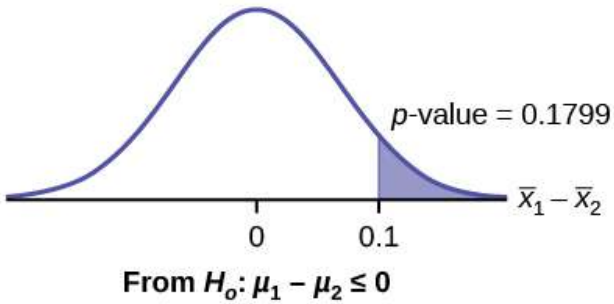


Figure 10.5

$\bar{X}_1 - \bar{X}_2 = 3 - 2.9 = 0.1$

Compare α and the *p* value: $\alpha = 0.05$ and p value = 0.1799. Therefore, $\alpha < p$ value.

Make a decision: Since $\alpha < p$ value, do not reject H_0.

Conclusion: At the 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the mean time Wax 1 lasts is longer (Wax 1 is more effective) than the mean time Wax 2 lasts.

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT. Arrow over to TESTS and press 3:2-SampZTest. Arrow over to Stats and press ENTER. Arrow down and enter .33 for sigma1, .36 for sigma2, 3 for the first sample mean, 20 for n1, 2.9 for the second sample mean, and 20 for n2. Arrow down to μ1: and arrow to > μ_2. Press ENTER. Arrow down to Calculate and press ENTER. The p value is p = 0.1799, and the test statistic is 0.9157. Do the procedure again, but instead of Calculate do Draw.

Try It

10.6 The means of the number of revolutions per minute of two competing engines are to be compared. Thirty engines are randomly assigned to be tested. Both populations have normal distributions. **Table 10.9** shows the result. Do the data indicate that Engine 2 has higher RPM than Engine 1? Test at a 5 percent level of significance.

Engine	Sample Mean Number of RPM	Population Standard Deviation
1	1,500	50
2	1,600	60

Table 10.9

Example 10.7

An interested citizen wanted to know if Democratic U.S. senators are older than Republican U.S. senators, on average. On May 26, 2013, the mean age of 30 randomly selected Republican senators was 61 years 247 days (61.675 years) with a standard deviation of 10.17 years. The mean age of 30 randomly selected Democratic senators was 61 years 257 days (61.704 years) with a standard deviation of 9.55 years.

Do the data indicate that Democratic senators are older than Republican senators, on average? Test at a 5 percent level of significance.

Solution 10.7

This is a test of two independent groups, two population means. The population standard deviations are unknown, but the sum of the sample sizes is 30 + 30 = 60, which is greater than 30, so we can use the normal approximation to the Student's-*t* distribution.
Subscripts: 1: Democratic senators; 2: Republican senators

Random variable: $\bar{X}_1 - \bar{X}_2$ = difference in the mean age of Democratic and Republican U.S. senators.

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

H_a: $\mu_1 > \mu_2$ H_a: $\mu_1 - \mu_2 > 0$

The words *older than* translates as a > symbol and goes into H_a. Therefore, this is a right-tailed test.

Distribution for the test: The distribution is the normal approximation to the Student's *t* for means, independent groups. Using the formula, the distribution is

$$() \bar{X}_1 - \bar{X}_2 \sim N[0, \sqrt{\frac{(9.55)^2}{30} + \frac{(10.17)^2}{30}}]$$

Since $\mu_1 \leq \mu_2$, $\mu_1 - \mu_2 \leq 0$ and the mean for the normal distribution is zero.

Calculating the *p* value using the normal distribution gives *p* value = 0.4040.

Graph:

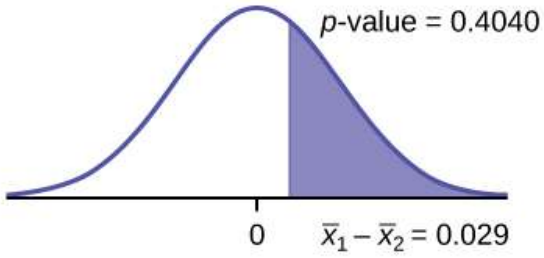


Figure 10.6

Compare α and the p value: $\alpha = 0.05$ and p value = 0.4040. Therefore, $\alpha < p$ value.

Make a decision: Since $\alpha < p$ value, do not reject H_0.

Conclusion: At the 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of Democratic senators is greater than the mean age of the Republican senators.

10.3 | Comparing Two Independent Population Proportions

When conducting a hypothesis test that compares two independent population proportions, the following characteristics should be present:

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least five, and the number of failures is at least five, for each of the samples.
3. Growing literature states that the population must be at least 10 or 20 times the size of the sample. This keeps each population from being over-sampled and causing incorrect results.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, H_0: $p_A = p_B$. To conduct the test, we use a **pooled proportion**, p_c.

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B}.$$

The distribution for the differences is

$$P'_A - P'_B \sim N[0, \sqrt{p_c(1 - p_c)(\frac{1}{n_A} + \frac{1}{n_B})}].$$

The test statistic (z-score) is

$$z = \frac{(p'_A - p'_B) - (p_A - p_B)}{\sqrt{p_c(1 - p_c)(\frac{1}{n_A} + \frac{1}{n_B})}}.$$

Example 10.8

Two types of medication for hives are being tested to determine if there is a difference in the proportions of adult patient reactions. Twenty out of a random sample of 200 adults given Medication A still had hives 30 minutes

after taking the medication. Twelve out of another random sample of 200 adults given Medication B still had hives 30 minutes after taking the medication. Test at a 1 percent level of significance.

Solution 10.8

The problem asks for a difference in proportions, making it a test of two proportions.

Let A and B be the subscripts for Medication A and Medication B, respectively. Then, p_A and p_B are the desired population proportions.

Random Variable:

$P'_A - P'_B$ = difference in the proportions of adult patients who did not react after 30 minutes to Medication A and to Medication B.

H_0: $p_A = p_B$

$p_A - p_B = 0$

H_a: $p_A \neq p_B$

$p_A - p_B \neq 0$

The words *is a difference* tell you the test is two-tailed.

Distribution for the test: Since this is a test of two binomial population proportions, the distribution is normal:

$$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20 + 12}{200 + 200} = 0.08 \quad 1 - p_c = 0.92$$

$$P'_A - P'_B \sim N\left[0, \sqrt{(0.08)(0.92)\left(\frac{1}{200} + \frac{1}{200}\right)}\right]$$

$P'_A - P'_B$ follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution: p-value = 0.1404.

Estimated proportion for group A: $p'_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$

Estimated proportion for group B: $p'_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$

Graph:

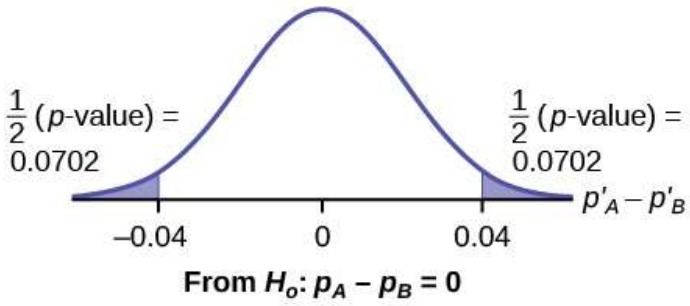


Figure 10.7

$P'_A - P'_B = 0.1 - 0.06 = 0.04$.

Half the p-value is below –0.04, and half is above 0.04.

Compare α and the p-value: $\alpha = 0.01$ and the p-value = 0.1404. $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: At a 1 percent level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to Medication A and Medication B.

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 20 for x1, 200 for n1, 12 for x2, and 200 for n2. Arrow down to p1: and arrow to not equal p2. Press ENTER. Arrow down to Calculate and press ENTER. The p-value is $p = 0.1404$, and the test statistic is 1.47. Do the procedure again, but instead of Calculate do Draw.

Try It

10.8 Two types of valves are being tested to determine if there is a difference in pressure tolerances. Fifteen out of a random sample of 100 of Valve A cracked under 4,500 psi. Six out of a random sample of 100 of Valve B cracked under 4,500 psi. Test at a 5 percent level of significance.

Example 10.9

A research study was conducted about gender differences in texting. The researcher believed that the proportion of girls involved in texting is less than the proportion of boys involved. The data collected in spring 2010 among a random sample of middle and high school students in a large school district in the southern United States is summarized in **Table 10.9**. Is the proportion of girls sending texts less than the proportion of boys texting? Test at a 1 percent level of significance.

	Males	**Females**
Sent texts	183	156
Total number surveyed	2231	2169

Table 10.10

Solution 10.9

This is a test of two population proportions. Let M and F be the subscripts for males and females. Then, p_M and p_F are the desired population proportions.

Random variable:

$p'_F - p'_M$ = difference in the proportions of males and females who sent texts.

H_0: $p_F = p_M$ H_0: $p_F - p_M = 0$

H_a: $p_F < p_M$ H_a: $p_F - p_M < 0$

The words *less than* tell you the test is left-tailed.

Distribution for the test: Since this is a test of two population proportions, the distribution is normal:

$$p_c = \frac{x_F + x_M}{n_F + n_M} = \frac{156 + 183}{2169 + 2231} = 0.077$$

$$1 - p_c = 0.923$$

Therefore,

$$p'_F - p'_M \sim N\left(0, \sqrt{(0.077)(0.923)\left(\frac{1}{2169} + \frac{1}{2231}\right)}\right)$$

$p'_F - p'_M$ follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution:
p-value = 0.1045
Estimated proportion for females: 0.0719
Estimated proportion for males: 0.082

Graph:

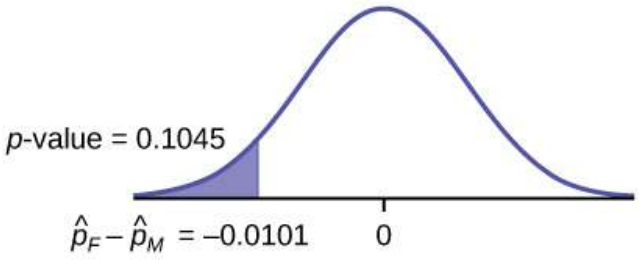


Figure 10.8

Decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: At the 1 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the proportion of girls sending texts is less than the proportion of boys sending texts.

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 156 for x1, 2169 for n1, 183 for x2, and 2231 for n2. Arrow down to p1: and arrow to less than p2. Press ENTER. Arrow down to Calculate and press ENTER. The *p*-value is $p = 0.1045$ and the test statistic is $z = -1.256$.

Example 10.10

Researchers conducted a study of smartphone use (Phone A versus Phone B) among adults. A cell phone company claimed that Phone B smartphones are more popular with whites (non-Hispanic) than with African Americans. The results of the survey indicate that of the 232 African American cell phone owners randomly sampled, 5 percent own Phone B. Of the 1,343 white cell phone owners randomly sampled, 10 percent own Phone B. Test at the 5 percent level of significance. Is the proportion of white Phone B owners greater than the proportion of African American Phone B owners?

Solution 10.10

This is a test of two population proportions. Let *W* and *A* be the subscripts for the whites and African Americans. Then, p_W and p_A are the desired population proportions.

Random variable:

$p'_W - p'_A$ = difference in the proportions of Phone A and Phone B users.

H_0: $p_W = p_A$ H_0: $p_W - p_A = 0$

H_a: $p_W > p_A$ H_a: $p_W - p_A > 0$

The words *more popular* indicate that the test is right-tailed.

Distribution for the test: The distribution is approximately normal.

$$p_c = \frac{x_W + x_A}{n_W + n_A} = \frac{134 + 12}{1343 + 232} = 0.0927$$

$$1 - p_c = 0.9073$$

Therefore,

$$p'_W - p'_A \backsim N\left(0, \sqrt{(0.0927)(0.9073)\left(\frac{1}{1343} + \frac{1}{232}\right)}\right)$$

$p'_W - p'_A$ follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution:

p-value = 0.0077
Estimated proportion for group A: 0.10
Estimated proportion for group B: 0.05

Graph:

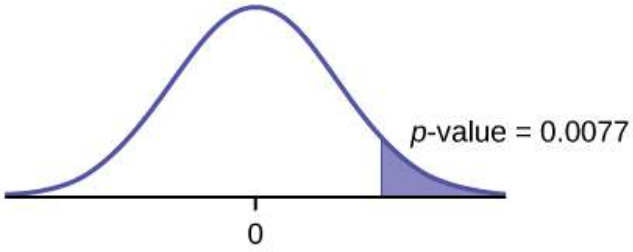


Figure 10.9

Decision: Since $\alpha > p$-value, reject the H_0.

Conclusion: At the 5 percent level of significance, from the sample data, there is sufficient evidence to conclude that a larger proportion of white cell phone owners use Phone B than African Americans.

Using the TI-83, 83+, 84, 84+ Calculator

TI-83+ and TI-84: Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 135 for x1, 1343 for n1, 12 for x2, and 232 for n2. Arrow down to p1: and arrow to greater than p2. Press ENTER. Arrow down to Calculate and press ENTER. The *p*-value is $p = 0.0092$, and the test statistic is $z = 2.33$.

Try It Σ

10.10 A group of citizens wanted to know if the proportion of homeowners in their small city was different in 2011 than in 2010. Their research showed that of the 113,231 available homes in their city in 2010, 7,622 of them were owned by the families who live there. In 2011, 7,439 of the 104,873 of the available homes were owned by city residents. Test at a 5 percent significance level. Answer the following questions:

a. Is this a test of two means or two proportions?

b. Which distribution do you use to perform the test?

c. What is the random variable?

d. What are the null and alternative hypotheses? Write the null and alternative hypotheses in symbols.

e. Is this test right-, left-, or two-tailed?

f. What is the *p*-value?

g. Do you reject or not reject the null hypothesis?

h. At the _____ level of significance, from the sample data, there _____ (is/is not) sufficient evidence to conclude that _____.

10.4 | Matched or Paired Samples (Optional)

When using a hypothesis test for matched or paired samples, the following characteristics should be present:

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.
6. Either the matched pairs have differences that come from a population that is normal or the number of differences is sufficiently large so that distribution of the sample mean of differences is approximately normal.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, μ_d, is then tested using a Student's-*t* test for a single population mean with $n - 1$ degrees of freedom, where n is the number of differences.

The test statistic (*t*-score) is

$$t = \frac{\bar{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}.$$

Example 10.11

A study was conducted to investigate the effectiveness of pain-reducing medication. Results for randomly selected subjects are shown in **Table 10.10**. A lower score indicates less pain. The *before* value is matched to an *after* value, and the differences are calculated. The differences have a normal distribution. Are the sensory measurements, on average, lower after the medication? Test at a 5 percent significance level.

Subject:	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Table 10.11

Solution 10.11

Corresponding *before* and *after* values form matched pairs. (Calculate *after* – *before*.)

After Data	Before Data	Difference
6.8	6.6	0.2
2.4	6.5	–4.1
7.4	9	–1.6
8.5	10.3	–1.8
8.1	11.3	–3.2
6.1	8.1	–2
3.4	6.3	–2.9
2	11.6	–9.6

Table 10.12

The data for the test are the differences: {0.2, –4.1, –1.6, –1.8, –3.2, –2, –2.9, –9.6}

The sample mean and sample standard deviation of the differences are: $\overline{x_d} = -3.13$ and $s_d = 2.91$

Verify these values.

Let μ_d be the population mean for the differences. We use the subscript d to denote *differences*.

Random variable: $\bar{X}_d$ = the mean difference of the sensory measurements.

H_0: $\mu_d \geq 0$

The null hypothesis is zero or positive, meaning that there is the same or more pain felt after taking the medication. That means the subject shows no improvement. μ_d is the population mean of the differences.

H_a: $\mu_d < 0$

The alternative hypothesis is negative, meaning there is less pain felt after taking the medication. That means the subject shows improvement. The score should be lower after taking the medication, so the difference ought to be negative to indicate improvement.

Distribution for the test: The distribution is a Student's t with $df = n - 1 = 8 - 1 = 7$. Use t_7. Note —that the test is for a single population mean.

Calculate the *p*-value using the Student's-*t* distribution: p-value = 0.0095

Graph:

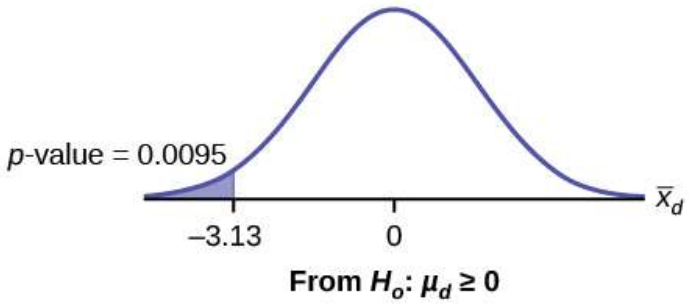


Figure 10.10

$\bar{X}_d$ is the random variable for the differences.

The sample mean and sample standard deviation of the differences are as follows:

$\bar{x}_d = -3.13$

$\bar{s}_d = 2.91$

Compare α and the *p*-value: $\alpha = 0.05$ and *p*-value = 0.0095. $\alpha >$ *p*-value.

Make a decision: Since $\alpha >$ *p*-value, reject H_0. This means that $\mu_d < 0$ and there is improvement.

Conclusion: At a 5 percent level of significance, from the sample data, there is sufficient evidence to conclude that the sensory measurements, on average, are lower after taking the medication. The medication appears to be effective in reducing pain.

NOTE

For the TI-83+ and TI-84 calculators, you can either calculate the differences ahead of time (*after* - *before*) and put the differences into a list or you can put the *after* data into a first list and the *before* data into a second list. Then, go to a third list and arrow up to the name. Enter `1st list name – 2nd list name`. The calculator will do the subtraction, and you will have the differences in the third list.

Using the TI-83, 83+, 84, 84+ Calculator

Use your list of differences as the data. Press `STAT` and arrow over to `TESTS`. Press `2:T-Test`. Arrow over to `Data` and press `ENTER`. Arrow down and enter `0` for μ_0, the name of the list where you put the data, and `1` for Freq:. Arrow down to μ: and arrow over to `< μ0`. Press `ENTER`. Arrow down to `Calculate` and press `ENTER`. The *p*-value is 0.0094, and the test statistic is –3.04. Do these instructions again except, arrow to `Draw` instead of `Calculate`. Press `ENTER`.

Try It

10.11 A study was conducted to investigate how effective a new diet was in lowering cholesterol. Results for the randomly selected subjects are shown in the table. The differences have a normal distribution. Are the subjects' cholesterol levels lower on average after the diet? Test at the 5 percent level.

Subject	A	B	C	D	E	F	G	H	I
Before	209	210	205	198	216	217	238	240	222
After	199	207	189	209	217	202	211	223	201

Table 10.13

Example 10.12

A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked four of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

Weight (in pounds)	**Player 1**	**Player 2**	**Player 3**	**Player 4**
Amount of weight lifted prior to the class	205	241	338	368
Amount of weight lifted after the class	295	252	330	360

Table 10.14

The coach wants to know if the strength development class makes his players stronger, on average.
Record the *differences* data. Calculate the differences by subtracting the amount of weight lifted prior to the class from the weight lifted after completing the class. The data for the differences are: {90, 11, -8, -8}. Assume the differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation.

$\bar{x}_d = 21.3$, $s_d = 46.7$

NOTE

The data given here would indicate that the distribution is right-skewed. The difference 90 may be an extreme outlier. It is pulling the sample mean to be 21.3 (positive). The means of the other three data values are negative.

Using the difference data, this becomes a test of a single _____.

Define the random variable: $\bar{X}_d$ is the mean difference in the maximum lift per player.

The distribution for the hypothesis test is t_3.

H_0: $\mu_d \leq 0$, H_a: $\mu_d > 0$

Graph:

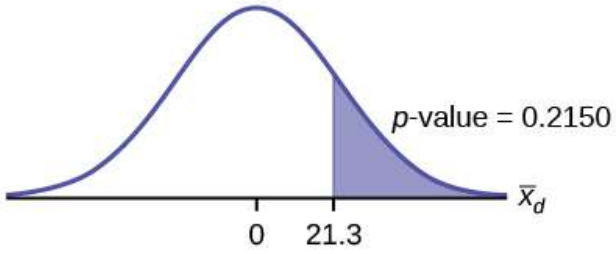


Figure 10.11

Calculate the *p*-value: The *p*-value is 0.2150.

Decision: If the level of significance is 5 percent, the decision is not to reject the null hypothesis, because $\alpha <$ *p*-value.

What is the conclusion?

At a 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the strength development class helped make the players stronger, on average.

Try It

10.12 A new prep class was designed to improve SAT test scores. Five students were selected at random. Their scores on two practice exams were recorded, one before the class and one after. The data are recorded in **Table 10.15**. Are the scores, on average, higher after the class? Test at a 5 percent level.

SAT Scores	Student 1	Student 2	Student 3	Student 4
Score before class	1840	1960	1920	2150
Score after class	1920	2160	2200	2100

Table 10.15

Example 10.13

Seven eighth-graders at Kennedy Middle School measured how far they could push the shot put with their dominant (writing) hand and their weaker (nonwriting) hand. They thought that they could push equal distances with both hands. The data are collected and recorded in **Table 10.16**.

Distance (in feet) using	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Dominant Hand	30	26	34	17	19	26	20
Weaker Hand	28	14	27	18	17	26	16

Table 10.16

Conduct a hypothesis test to determine whether the mean difference in distances between the children's dominant versus weaker hands is significant.

Record the *differences* data. Calculate the differences by subtracting the distances with the weaker hand from the distances with the dominant hand. The data for the differences are: {2, 12, 7, –1, 2, 0, 4}. The differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation. $\bar{x}_d = 3.71$, $s_d = 4.5$.

Random variable: $\bar{X}_d$ = mean difference in the distances between the hands.

Distribution for the hypothesis test: t_6

H_0: $\mu_d = 0$ H_a: $\mu_d \neq 0$

Graph:

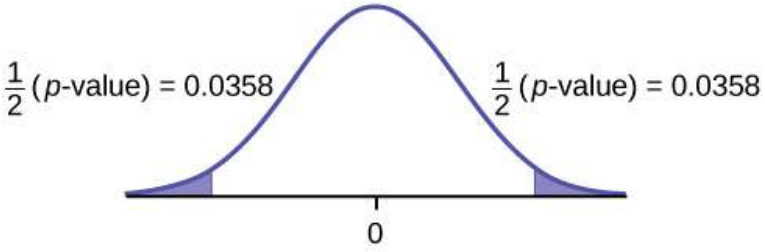


Figure 10.12

Calculate the *p*-value: The *p*-value is 0.0716 (using the data directly).

Test statistic = 2.18. *p*-value = 0.0719 using $\left(\bar{x}_d = 3.71,\ s_d = 4.5\right)$.

Decision: Assume $\alpha = 0.05$. Since $\alpha < p$-value, do not reject H_0.

Conclusion: At the 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the children's weaker and dominant hands to push the shot put.

Try It Σ

10.13 Five ball players think they can throw the same distance with their dominant hand (throwing) and off-hand (catching hand). The data were collected and recorded in **Table 10.17**. Conduct a hypothesis test to determine whether the mean 5 difference in distances between the dominant and off-hand is significant. Test at the 5 percent level.

	Player 1	Player 2	Player 3	Player 4	Player 5
Dominant Hand	120	111	135	140	125
Off-Hand	105	109	98	111	99

Table 10.17

10.5 | Hypothesis Testing for Two Means and Two Proportions

Stats Lab

10.1 Hypothesis Testing for Two Means and Two Proportions

Student Learning Outcomes

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

Supplies:

- The business section from two consecutive days' newspapers
- Three small packages of multicolored chocolates
- Five small packages of peanut butter candies

Increasing Stocks Survey

Look at yesterday's newspaper business section. Conduct a hypothesis test to determine if the proportion of New York Stock Exchange (NYSE) stocks that increased is greater than the proportion of NASDAQ stocks that increased. As randomly as possible, choose 40 NYSE stocks and 32 NASDAQ stocks and complete the following statements.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable.
4. The distribution to use for the test is _____.
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    a. Graph:

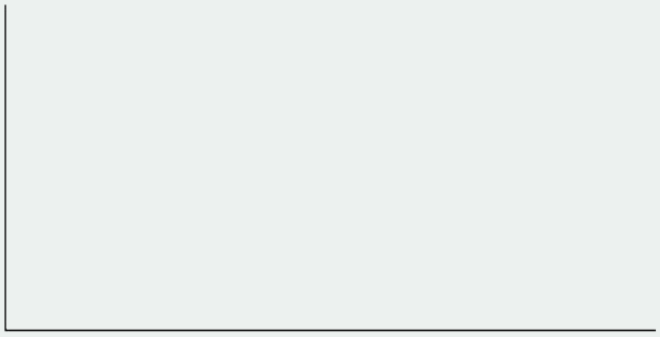


Figure 10.13

    b. Calculate the *p* value.
7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Decreasing Stocks Survey

Randomly pick eight stocks from the newspaper. Using two consecutive days' business sections, test whether the stocks went down, on average, for the second day.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable.
4. The distribution to use for the test is _____.
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    a. Graph

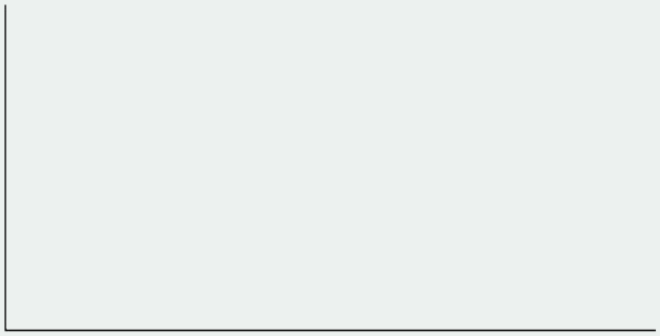


Figure 10.14

    b. Calculate the *p* value:
7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Candy Survey

Buy three small packages of multicolored chocolates and five small packages of peanut butter candies (same net weight as the multicolored chocolates). Test whether the mean number of candy pieces per package is the same for the two brands.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable.
4. What distribution should be used for this test?
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
    a. Graph

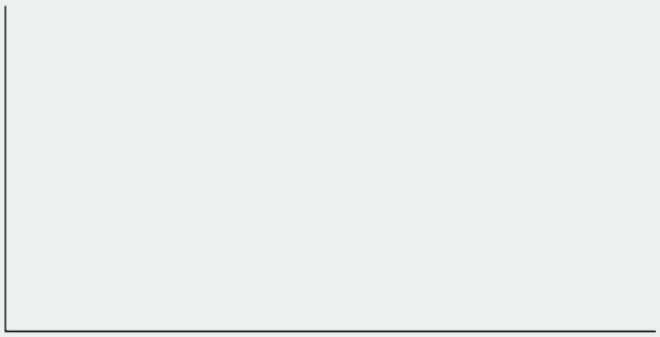


Figure 10.15

b. Calculate the *p* value.

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Shoe Survey

Test whether women have, on average, more pairs of shoes than men. Include all forms of sneakers, shoes, sandals, and boots. Use your class as the sample.

1. H_0: _____
2. H_a: _____
3. In words, define the random variable.
4. The distribution to use for the test is _____.
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
   a. Graph

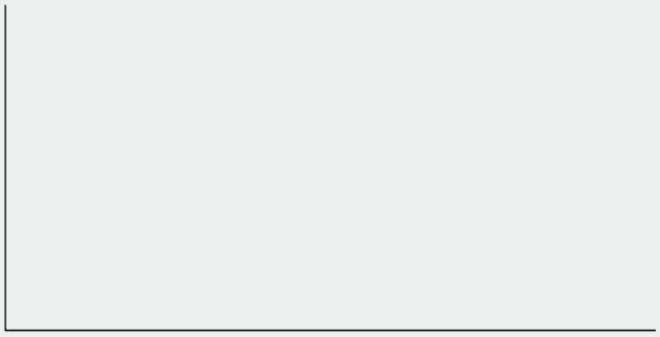


Figure 10.16

b. Calculate the *p* value.

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

KEY TERMS

degrees of freedom (*df*) the number of objects in a sample that are free to vary

pooled proportion estimate of the common value of p_1 and p_2

standard deviation a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation

variable (random variable) a characteristic of interest in a population being studied.
Common notation for variables are uppercase Latin letters X, Y, Z,... Common notation for a specific value from the domain (set of all possible values of a variable) are lowercase Latin letters x, y, z,.... For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, ... Variables in statistics differ from variables in intermediate algebra in two ways:

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color, then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x of the random variable X takes only after performing the experiment.

CHAPTER REVIEW

10.1 Two Population Means with Unknown Standard Deviations

Two population means from independent samples where the population standard deviations are not known

- Random variable: $\bar{X}_1 - \bar{X}_2$ = the difference of the sampling means
- Distribution: Student's t-distribution with degrees of freedom (variances not pooled)

10.2 Two Population Means with Known Standard Deviations

A hypothesis test of two population means from independent samples where the population standard deviations are known (typically approximated with the sample standard deviations) will have these characteristics:

- Random variable: $\bar{X}_1 - \bar{X}_2$ = the difference of the means
- Distribution: normal distribution

10.3 Comparing Two Independent Population Proportions

Test of two population proportions from independent samples

- Random variable: $\hat{p}_A - \hat{p}_B =$ difference between the two estimated proportions
- Distribution: normal distribution

10.4 Matched or Paired Samples (Optional)

A hypothesis test for matched or paired samples (t-test) has these characteristics:

- Test the differences by subtracting one measurement from the other measurement
- Random variable: $\bar{x}_d$ = mean of the differences.
- Distribution: Student's t distribution with $n - 1$ degrees of freedom.
- If the number of differences is small (less than 30), the differences must follow a normal distribution.
- Two samples are drawn from the same set of objects.
- Samples are dependent.

FORMULA REVIEW

10.1 Two Population Means with Unknown Standard Deviations

Standard error: $SE = \sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$

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

Degrees of freedom:

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

where:

s_1 and s_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.

$\bar{x}_1$ and $\bar{x}_2$ are the sample means.

Cohen's *d* is the measure of effect size:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$$

where $s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$.

10.2 Two Population Means with Known Standard Deviations

Normal distribution:

$$\bar{X}_1 - \bar{X}_2 \sim N\left[\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}\right].$$

Generally, $\boldsymbol{\mu_1 - \mu_2 = 0}$.

Test statistic (*z*-score):

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

Generally, $\boldsymbol{\mu_1 - \mu_2 = 0}$.

where

σ_1 and σ_2 are the known population standard deviations, n_1 and n_2 are the sample sizes, $\bar{x}_1$ and $\bar{x}_2$ are the sample means, and μ_1 and μ_2 are the population means.

10.3 Comparing Two Independent Population Proportions

Pooled proportion: $p_c = \frac{x_F + x_M}{n_F + n_M}$

Distribution for the differences:

$$p'_A - p'_B \sim N\left[0, \sqrt{p_c(1 - p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}\right]$$

where the null hypothesis is H_0: $p_A = p_B$ or H_0: $p_A - p_B = 0$

Test statistic (*z*-score): $z = \frac{(p'_A - p'_B)}{\sqrt{p_c(1 - p_c)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$

where the null hypothesis is H_0: $p_A = p_B$ or H_0: $p_A - p_B = 0$

and where

p'_A and p'_B are the sample proportions, p_A and p_B are the population proportions,

P_c is the pooled proportion, and n_A and n_B are the sample sizes.

10.4 Matched or Paired Samples (Optional)

Test statistic (*t*-score): $t = \frac{\bar{x}_d - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$

where:

$\bar{x}_d$ is the mean of the sample differences, μ_d is the mean of the population differences, s_d is the sample standard deviation of the differences, and *n* is the sample size.

PRACTICE

10.1 Two Population Means with Unknown Standard Deviations

Use the following information to answer the next 15 exercises. Indicate if the hypothesis test is for

a. independent group means, population standard deviations, and/or variances known,

b. independent group means, population standard deviations, and/or variances unknown,

c. matched or paired samples,

d. single mean,

e. two proportions, or

f. single proportion.

1. It is believed that 70 percent of males pass their drivers test in the first attempt, while 65 percent of females pass the test in the first attempt. Of interest is whether the proportions are equal.

2. A new laundry detergent is tested on consumers. Of interest is the proportion of consumers who prefer the new brand over the leading competitor. A study is done to test this.

3. A new windshield treatment claims to repel water more effectively. Ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. A hypothesis test is conducted.

4. The known standard deviation in salary for all mid-level professionals in the financial industry is $11,000. Company A and Company B are in the financial industry. Suppose samples are taken of mid-level professionals from Company A and from Company B. The sample mean salary for mid-level professionals in Company A is $80,000. The sample mean salary for mid-level professionals in Company B is $96,000. Company A and Company B management want to know if their mid-level professionals are paid differently, on average.

5. The average worker in Germany gets eight weeks of paid vacation.

6. According to a television commercial, 80% of dentists agree that a brand of fluoridated toothpaste is the best on the market.

7. It is believed that the average grade on an English essay in a particular school system is higher for females than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of 3, and a random sample of 25 males had a mean score of 76 with a standard deviation of 4.

8. The league mean batting average is 0.280 with a known standard deviation of 0.06. The Rattlers and the Vikings belong to the league. The mean batting average for a sample of eight Rattlers is 0.210, and the mean batting average for a sample of eight Vikings is 0.260. There are 24 players on the Rattlers and 19 players on the Vikings. Are the batting averages of the Rattlers and Vikings statistically different?

9. In a random sample of 100 forests in the United States, 56 were coniferous or contained conifers. In a random sample of 80 forests in Mexico, 40 were coniferous or contained conifers. Is the proportion of conifers in the United States statistically more than the proportion of conifers in Mexico?

10. A new medicine is said to help improve sleep. Eight subjects are picked at random and given the medicine. The mean hours slept for each person were recorded before starting the medication and after.

11. It is thought that teenagers sleep more than adults on average. A study is done to verify this. A sample of 16 teenagers has a mean of 8.9 hours slept and a standard deviation of 1.2. A sample of 12 adults has a mean of 6.9 hours slept and a standard deviation of 0.6.

12. Varsity athletes practice five times a week, on average.

13. A sample of 12 in-state graduate school programs at School A has a mean tuition of $64,000 with a standard deviation of $8,000. At School B, a sample of 16 in-state graduate programs has a mean tuition of $80,000 with a standard deviation of $6,000. On average, are the mean tuitions different?

14. A new WiFi range booster is being offered to consumers. A researcher tests the native range of 12 different routers under the same conditions. The ranges are recorded. Then, the researcher uses the new WiFi range booster and records the new ranges. Does the new WiFi range booster do a better job?

15. A high school principal claims that 30 percent of student athletes drive themselves to school, while 4 percent of nonathletes drive themselves to school. In a sample of 20 student athletes, 45 percent drive themselves to school. In a sample of 35 nonathlete students, 6 percent drive themselves to school. Is the percent of student athletes who drive themselves to school more than the percent of nonathletes?

Use the following information to answer the next three exercises: A study is done to determine which of two soft drinks has more sugar. There are 13 cans of Beverage A in a sample and six cans of Beverage B. The mean amount of sugar in Beverage A is 36 grams with a standard deviation of 0.6 grams. The mean amount of sugar in Beverage B is 38 grams with

a standard deviation of 0.8 grams. The researchers believe that Beverage B has more sugar than Beverage A, on average. Both populations have normal distributions.

16. Are standard deviations known or unknown?

17. What is the random variable?

18. Is this a one-tailed or two-tailed test?

Use the following information to answer the next 12 exercises. The U.S. Centers for Disease Control reports that the mean life expectancy was 47.6 years for whites born in 1900 and 33.0 years for nonwhites. Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 whites, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 nonwhites, the mean life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the mean life spans in the county were the same for whites and nonwhites.

19. Is this a test of means or proportions?

20. State the null and alternative hypotheses.
a. H_0: _____
b. H_a: _____

21. Is this a right-tailed, left-tailed, or two-tailed test?

22. In symbols, what is the random variable of interest for this test?

23. In words, define the random variable of interest for this test.

24. Which distribution (normal or Student's *t*) would you use for this hypothesis test?

25. Explain why you chose the distribution you did for **Exercise 10.24**.

26. Calculate the test statistic and *p*-value.

27. Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the *p*-value.

28. Find the *p*-value.

29. At a preconceived $\alpha = 0.05$, write the following:
a. Your decision:
b. The reason for your decision:
c. Your conclusion (write out in a complete sentence):

30. Does it appear that the means are the same? Why or why not?

10.2 Two Population Means with Known Standard Deviations

Use the following information to answer the next five exercises. The mean speeds of fastball pitches from two different baseball pitchers are to be compared. A sample of 14 fastball pitches is measured from each pitcher. The populations have normal distributions. **Table 10.18** shows the result. Scouters believe that Rodriguez pitches a speedier fastball.

Pitcher	Sample Mean Speed of Pitches (mph)	Population Standard Deviation
Wesley	86	3
Rodriguez	91	7

Table 10.18

31. What is the random variable?

32. State the null and alternative hypotheses.

33. What is the test statistic?

34. What is the *p* value?

35. At the 1 percent significance level, what is your conclusion?

Use the following information to answer the next five exercises. A researcher is testing the effects of plant food on plant growth. Nine plants have been given the plant food. Another nine plants have not been given the plant food. The heights of

the plants are recorded after eight weeks. The populations have normal distributions. The following table is the result. The researcher thinks the food makes the plants grow taller.

Plant Group	Sample Mean Height of Plants (inches)	Population Standard Deviation
Food	16	2.5
No food	14	1.5

Table 10.19

36. Is the population standard deviation known or unknown?

37. State the null and alternative hypotheses.

38. What is the *p* value?

39. Draw the graph of the *p* value.

40. At the 1 percent significance level, what is your conclusion?

Use the following information to answer the next five exercises. Two metal alloys are being considered as material for ball bearings. The mean melting point of the two alloys is to be compared. Fifteen pieces of each metal are being tested. Both populations have normal distributions. The following table is the result. It is believed that Alloy Zeta has a different melting point.

	Sample Mean Melting Temperatures (°F)	Population Standard Deviation
Alloy Gamma	800	95
Alloy Zeta	900	105

Table 10.20

41. State the null and alternative hypotheses.

42. Is this a right-, left-, or two-tailed test?

43. What is the *p* value?

44. Draw the graph of the *p* value.

45. At the 1 percent significance level, what is your conclusion?

10.3 Comparing Two Independent Population Proportions

Use the following information for the next five exercises. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS_1 had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS_2 had system failures within the first eight hours of operation. OS_2 is believed to be more stable (have fewer crashes) than OS_1.

46. Is this a test of means or proportions?

47. What is the random variable?

48. State the null and alternative hypotheses.

49. What is the *p*-value?

50. What can you conclude about the two operating systems?

Use the following information to answer the next 12 exercises. In the recent U.S. Census, 3 percent of the U.S. population reported being of two or more races. However, the percent varies tremendously from state to state. Suppose that two random surveys are conducted. In the first random survey, out of 1,000 North Dakotans, only 9 people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races. Conduct

a hypothesis test to determine if the population percents are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

51. Is this a test of means or proportions?

52. State the null and alternative hypotheses.

a. H_0: _____
b. H_a: _____

53. Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

54. What is the random variable of interest for this test?

55. In words, define the random variable for this test.

56. Which distribution (normal or Student's *t*) would you use for this hypothesis test?

57. Explain why you chose the distribution you did for the **Exercise 10.56**.

58. Calculate the test statistic.

59. Sketch a graph of the situation. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the *p*-value.

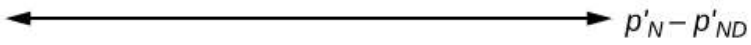


Figure 10.17

60. Find the *p*-value.

61. At a preconceived $\alpha = 0.05$, write the following:

a. Your decision:
b. The reason for your decision:
c. Your conclusion (write out in a complete sentence):

62. Does it appear that the proportion of Nevadans who are two or more races is higher than the proportion of North Dakotans? Why or why not?

10.4 Matched or Paired Samples (Optional)

Use the following information to answer the next five exercises. A study was conducted to test the effectiveness of a software patch in reducing system failures over a six-month period. Results for randomly selected installations are shown in **Table 10.21**. The *before* value is matched to an *after* value, and the differences are calculated. The differences have a normal distribution. Test at the 1 percent significance level.

Installation	A	B	C	D	E	F	G	H
Before	3	6	4	2	5	8	2	6
After	1	5	2	0	1	0	2	2

Table 10.21

63. What is the random variable?

64. State the null and alternative hypotheses.

65. What is the *p*-value?

66. Draw the graph of the *p*-value.

67. What conclusion can you draw about the software patch?

Use the following information to answer next five exercises. A study was conducted to test the effectiveness of a juggling class. Before the class started, six subjects juggled as many balls as they could at once. After the class, the same six subjects juggled as many balls as they could. The differences in the number of balls are calculated. The differences have a normal

distribution. Test at the 1 percent significance level.

Subject	A	B	C	D	E	F
Before	3	4	3	2	4	5
After	4	5	6	4	5	7

Table 10.22

68. State the null and alternative hypotheses.

69. What is the *p*-value?

70. What is the sample mean difference?

71. Draw the graph of the *p*-value.

72. What conclusion can you draw about the juggling class?

Use the following information to answer the next five exercises. A doctor wants to know if a blood pressure medication is effective. Six subjects have their blood pressures recorded. After twelve weeks on the medication, the same six subjects have their blood pressure recorded again. For this test, only systolic pressure is of concern. Test at the 1 percent significance level.

Patient	A	B	C	D	E	F
Before	161	162	165	162	166	171
After	158	159	166	160	167	169

Table 10.23

73. State the null and alternative hypotheses.

74. What is the test statistic?

75. What is the *p*-value?

76. What is the sample mean difference?

77. What is the conclusion?

HOMEWORK

10.1 Two Population Means with Unknown Standard Deviations

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in ***Appendix E****. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.*

NOTE

If you are using a Student's *t*-distribution for a homework problem in what follows, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption.)

78. The mean number of English courses taken in a two-year period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of 3 English courses with a standard deviation of 0.8. The females took an average of 4 English courses with a standard deviation of 1.0. Are the means statistically the same?

79. A student at a four-year college claims that mean enrollment at four-year colleges is higher than at two-year colleges in the United States. Two surveys are conducted. Of the 35 two-year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191.

80. At Rachel's eleventh birthday party, eight girls were timed to see how long (in seconds) they could sit perfectly still in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be zero. Test their hypothesis.

Relaxed time (seconds)	Jumping time (seconds)
26	21
47	40
30	28
22	21
23	25
45	43
37	35
29	32

Table 10.24

81. Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean mechanical engineering salary is lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry-level mechanical engineers and 60 entry-level electrical engineers. Their mean salaries were $46,100 and $46,700, respectively. Their standard deviations were $3,450 and $4,210, respectively. Conduct a hypothesis test to determine if you agree that the mean entry-level mechanical engineering salary is lower than the mean entry-level electrical engineering salary.

82. Marketing companies have collected data implying that teenage girls use more ringtones on their smartphones than teenage boys do. In one study of 40 randomly chosen teenage girls and boys (20 of each) with smartphones, the mean number of ringtones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls' mean is higher than the boys' mean.

*Use the information from **Appendix C** to answer the next four exercises.*

83. Using the data from Lap 1 only, conduct a hypothesis test to determine if the mean time for completing a lap in races is the same as it is in practices.

84. Repeat the test in **Exercise 10.83**, but use Lap 5 data this time.

85. Repeat the test in **Exercise 10.83**, but this time combine the data from Laps 1 and 5.

86. In two to three complete sentences, explain in detail how you might use Terri Vogel's data to answer the following question: Does Terri Vogel drive faster in races than she does in practices?

Use the following information to answer the next two exercises. The Eastern and Western Major League Soccer conferences have a new Reserve Division that allows new players to develop their skills. Data for a randomly picked date showed the following annual goals.

Western	Eastern
Los Angeles 9	D United 9
FC Dallas 3	Chicago 8
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

Table 10.25

Conduct a hypothesis test to answer the next two exercises.

87. The **exact** distribution for the hypothesis test is
a. the normal distribution
b. the Student's *t*-distribution
c. the uniform distribution
d. the exponential distribution

88. If the level of significance is 0.05, the conclusion is:
a. There is sufficient evidence to conclude that the **W** Division teams score fewer goals, on average, than the **E** teams.
b. There is insufficient evidence to conclude that the **W** Division teams score more goals, on average, than the **E** teams.
c. There is insufficient evidence to conclude that the **W** teams score fewer goals, on average, than the **E** teams.
d. There is not sufficient evidence to determine a conclusion.

89. Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The *day* subscript refers to the statistics day students. The *night* subscript refers to the statistics night students. Which of the following is a concluding statement:
a. There is sufficient evidence to conclude that statistics night students' mean on Exam 2 is better than the statistics day students' mean on Exam 2.
b. There is insufficient evidence to conclude that the statistics day students' mean on Exam 2 is better than the statistics night students' mean on Exam 2.
c. There is insufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.
d. There is sufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.

90. Researchers interviewed people in a certain industry in Canada and the United States. The mean age of the 100 Canadians upon entering this industry was 18 with a standard deviation of 6. The mean age of the 130 Americans upon entering this industry was 20 with a standard deviation of 8. Is the mean age of entering this industry in Canada lower than the mean age in the United States? Test at a 1 percent significance level.

91. A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet. The powder diet group had a mean weight loss of 42 pounds with a standard deviation of 12 pounds. The liquid diet group had a mean weight loss of 45 pounds with a standard deviation of 14 pounds.

92. Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91, respectively. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The *day* subscript refers to the statistics day students. The *night* subscript refers to the statistics night students. An appropriate alternative hypothesis for the hypothesis test is

a. $\mu_{day} > \mu_{night}$
b. $\mu_{day} < \mu_{night}$
c. $\mu_{day} = \mu_{night}$
d. $\mu_{day} \neq \mu_{night}$

10.2 Two Population Means with Known Standard Deviations

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in ***Appendix E****. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.*

NOTE

If you are using a Student's *t*-distribution for one of the following homework problems, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption.)

93. A study is done to determine if students in the California state university system take longer to graduate, on average, than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. Suppose that from years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively. The following data are collected. The California state university system students took on average 4.5 years with a standard deviation of 0.8. The private university students took on average 4.1 years with a standard deviation of 0.3.

94. Parents of teenage boys often complain that auto insurance costs more, on average, for teenage boys than for teenage girls. A group of concerned parents examines a random sample of insurance bills. The mean annual cost for 36 teenage boys was $679. For 23 teenage girls, it was $559. From past years, it is known that the population standard deviation for each group is $180. Determine whether you believe that the mean cost for auto insurance for teenage boys is greater than that for teenage girls.

95. A group of transfer-bound students wondered if they will spend the same mean amount on texts and supplies each year at their four-year university as they have at their community college. They conducted a random survey of 54 students at their community college and 66 students at their local four-year university. The sample means were $947 and $1,011, respectively. The population standard deviations are known to be $254 and $87, respectively. Conduct a hypothesis test to determine if the means are statistically the same.

96. Some manufacturers claim that nonhybrid sedan cars have a lower mean miles per gallon (mpg) than hybrid ones. Suppose that consumers test 21 hybrid sedans and get a mean of 31 mpg with a standard deviation of 7 mpg. Thirty-one nonhybrid sedans get a mean of 22 mpg with a standard deviation of 4 mpg. Suppose that the population standard deviations are known to be 6 and 3, respectively. Conduct a hypothesis test to evaluate the manufacturers' claim.

97. A baseball fan wanted to know if there is a difference between the number of games played in a World Series when the American League won the series versus when the National League won the series. From 1922 to 2012, the population standard deviation of games won by the American League was 1.14, and the population standard deviation of games won by the National League was 1.11. Of 19 randomly selected World Series games won by the American League, the mean number of games won was 5.76. The mean number of 17 randomly selected games won by the National League was 5.42. Conduct a hypothesis test.

98. One of the questions in a study of marital satisfaction of dual-career couples was to rate the statement "I'm pleased with the way we divide the responsibilities for childcare." The ratings went from 1 (strongly agree) to 5 (strongly disagree). **Table 10.26** contains 10 of the paired responses for husbands and wives. Conduct a hypothesis test to see if the mean difference in the husband's versus the wife's satisfaction level is negative (meaning that, within the partnership, the husband is happier than the wife).

Wife's Score	2	2	3	3	4	2	1	1	2	4
Husband's Score	2	2	1	3	2	1	1	1	2	4

Table 10.26

10.3 Comparing Two Independent Population Proportions

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in ***Appendix E****. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.*

NOTE

If you are using a Student's *t*-distribution for one of the following homework problems, including for paired data, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption.)

99. A recent drug survey showed an increase in the use of prescription medication among local senior citizens as compared to the national percent. Suppose that a survey of 100 local seniors and 100 national seniors is conducted to see if the proportion of prescription medication use is higher locally or nationally. Locally, 65 senior citizens reported taking prescription medication within the past month, while 60 national seniors reported using them.

100. Elizabeth Mjelde, an art history professor, was interested in whether the value from the Golden Ratio formula, $\left(\frac{\text{larger} + \text{smaller dimension}}{\text{larger dimension}}\right)$, was the same in the Whitney Exhibit for works from 1900 to 1919 as for works from 1920 to 1942. Thirty-seven early works were sampled, averaging 1.74 with a standard deviation of 0.11. Sixty-five of the later works were sampled, averaging 1.746 with a standard deviation of 0.1064. Do you think that there is a significant difference in the Golden Ratio calculation?

101. A year was randomly picked from 1985 to the present. In that year, there were 2,051 Hispanic students at Cabrillo College out of a total of 12,328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2,441 students. In general, do you think that the percent of Hispanic students at the two colleges is basically the same or different?

Use the following information to answer the next three exercises. Neuroinvasive West Nile virus is a severe disease that affects a person's nervous system. It is spread by the *Culex* species of mosquito. In the United States in 2010, there were 629 reported cases of neuroinvasive West Nile virus out of a total of 1,021 reported cases, and there were 486 neuroinvasive reported cases out of a total of 712 cases reported in 2011. Is the 2011 proportion of neuroinvasive West Nile virus cases more than the 2010 proportion of neuroinvasive West Nile virus cases? Using a 1 percent level of significance, conduct an appropriate hypothesis test.

- *2011* subscript: 2011 group.
- *2010* subscript: 2010 group

102. This is

a. a test of two proportions
b. a test of two independent means
c. a test of a single mean
d. a test of matched pairs.

103. An appropriate null hypothesis is

a. $p_{2011} \leq p_{2010}$
b. $p_{2011} \geq p_{2010}$
c. $\mu_{2011} \leq \mu_{2010}$
d. $p_{2011} > p_{2010}$

104. The *p*-value is 0.0022. At a 1 percent level of significance, what is the appropriate conclusion?

a. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile virus is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile virus.
b. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile virus is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile virus.
c. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile virus is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile virus.
d. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile virus is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile virus.

105. Researchers conducted a study to find out if there is a difference in the use of e-readers by different age groups. Randomly selected participants were divided into two age groups. In the 16- to 29-year-old group, 7 percent of the 628 surveyed use e-readers, while 11 percent of the 2,309 participants 30 years old and older use e-readers.

106. Adults aged 18 years and older were randomly selected for a survey about a specific disease. The researchers wanted to determine if the proportion of women who have the disease is less than the proportion of southern men who do. The results are shown in **Table 10.27**. Test at the 1 percent level of significance.

	Number diagnosed with disease	Sample size
Men	42,769	155,525
Women	67,169	248,775

Table 10.27

107. Two computer users were discussing tablet computers. A higher proportion of people ages 16 to 29 use tablets than of people age 30 and older. **Table 10.28** details the number of tablet owners for each age group. Test at the 1 percent level of significance.

	16–29 year olds	30 years and older
Own a Tablet	69	231
Sample Size	628	2,309

Table 10.28

108. A group of friends debated whether more men use smartphones than women. They consulted a research study of smartphone use among adults. The results of the survey indicate that of the 973 men randomly sampled, 379 use smartphones. For women, 404 of the 1,304 who were randomly sampled use smartphones. Test at the 5 percent level of significance.

109. While her husband spent 2.5 hours picking out new speakers, a statistician decided to determine whether the percent of men who enjoy shopping for electronic equipment is higher than the percent of women who do. The population was Saturday afternoon shoppers. Out of 67 men, 24 said they enjoyed the activity. Eight of the 24 women surveyed claimed to enjoy the activity. Interpret the results of the survey.

110. We are interested in whether children's educational computer software costs less, on average, than children's entertainment software. Thirty-six educational software titles were randomly picked from a catalog. The mean cost was $31.14 with a standard deviation of $4.69. Thirty-five entertainment software titles were randomly picked from the same catalog. The mean cost was $33.86 with a standard deviation of $10.87. Decide whether children's educational software costs less, on average, than children's entertainment software.

111. A researcher recently claimed that the proportion of college-age males who wear at least one piece of jewelery is as high as the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 wear at least one piece of jewelery. Out of 92 females, 47 wear at least one piece of jewelery. Do you believe that the proportion of males has reached the proportion of females?

112. Use the data sets found in **Appendix C** to answer this exercise. Is the proportion of race laps Terri completes slower than 130 seconds less than the proportion of practice laps she completes slower than 135 seconds?

113. *To Breakfast or Not to Breakfast?* by Richard Ayore

In the American society, birthdays are one of those days that everyone looks forward to. People of different ages and peer groups gather to mark the 18th, 20th, …, birthdays. During this time, one looks back to see what he or she has achieved for the past year and also focuses ahead for more to come.

If, by any chance, I am invited to one of these parties, my experience is always different. Instead of dancing around with my friends while the music is booming, I get carried away by memories of my family back home in Kenya. I remember the good times I had with my brothers and sister while we did our daily routine.

Every morning, I remember we went to the shamba (garden) to weed our crops. I remember one day arguing with my brother as to why he always remained behind just to join us an hour later. In his defense, he said that he preferred waiting for breakfast before he came to weed. He said, "This is why I always work more hours than you guys!"

And so, to prove him wrong or right, we decided to give it a try. One day we went to work as usual without breakfast, and recorded the time we could work before getting tired and stopping. On the next day, we all ate breakfast before going to work. We recorded how long we worked again before getting tired and stopping. Of interest was our mean increase in work time. Though not sure, my brother insisted that it was more than two hours. Using the data in **Table 10.29**, solve our problem.

Work hours with breakfast	Work hours without breakfast
8	6
7	5
9	5
5	4
9	7
8	7
10	7
7	5
6	6
9	5

Table 10.29

10.4 Matched or Paired Samples (Optional)

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in ***Appendix E****. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.*

NOTE

If you are using a Student's *t*-distribution for the homework problems, including for paired data, you may assume that the underlying population is normally distributed. (When using these tests in a real situation, you must first prove that assumption.)

114. Ten individuals went on a low-fat diet for 12 weeks to lower their cholesterol. The data are recorded in **Table 10.30**. Do you think that their cholesterol levels were significantly lowered?

Starting cholesterol level	Ending cholesterol level
140	140
220	230
110	120
240	220
200	190
180	150
190	200
360	300
280	300
260	240

Table 10.30

Use the following information to answer the next two exercises. A new preventative medication was tried on a group of 224 patients who had the same risk factors for a disease. 45 patients developed the disease after four years. In a control group of 224 patients, 68 developed the disease after four years. We want to test whether the method of treatment reduces the proportion of patients who develop the disease after four years.

Let the subscript t = treated patient and ut = untreated patient.

115. The appropriate hypotheses are

a. H_0: $p_t < p_{ut}$ and H_a: $p_t \geq p_{ut}$
b. H_0: $p_t \leq p_{ut}$ and H_a: $p_t > p_{ut}$
c. H_0: $p_t = p_{ut}$ and H_a: $p_t \neq p_{ut}$
d. H_0: $p_t = p_{ut}$ and H_a: $p_t < p_{ut}$

116. If the p-value is 0.0062, what is the conclusion? Use $\alpha = 0.05$.

a. The method has no effect.
b. There is sufficient evidence to conclude that the method reduces the proportion of patients who develop the disease after four years.
c. There is sufficient evidence to conclude that the method increases the proportion of patients who develop the disease after four years.
d. There is insufficient evidence to conclude that the method reduces the proportion of patients who develop the disease after four years.

Use the following information to answer the next two exercises. An experiment is conducted to show that blood pressure can be consciously reduced in people trained in a biofeedback exercise program. Six subjects were randomly selected, and blood pressure measurements were recorded before and after the training. The difference between blood pressures was calculated (after – before), producing the following results: $\bar{x}_d = -10.2$ $s_d = 8.4$. Using the data, test the hypothesis that the blood pressure has decreased after the training.

117. The distribution for the test is

a. t_5
b. t_6
c. $N(-10.2, 8.4)$
d. $N(-10.2, \frac{8.4}{\sqrt{6}})$

118. If $\alpha = 0.05$, the p-value and the conclusion are

a. 0.0014; There is sufficient evidence to conclude that the blood pressure decreased after the training.
b. 0.0014; There is sufficient evidence to conclude that the blood pressure increased after the training.
c. 0.0155; There is sufficient evidence to conclude that the blood pressure decreased after the training.
d. 0.0155; There is sufficient evidence to conclude that the blood pressure increased after the training.

119. A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four new students. She records their 18-hole scores before learning the technique and then after having taken her class. She conducts a hypothesis test. The data are as follows.

	Player 1	Player 2	Player 3	Player 4
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

Table 10.31

The correct decision is

a. reject H_0.
b. do not reject H_0.

120. A local research group is studying a chronic disease. They believe the number of cases of the disease is higher in 2013 than in 2012 in the southern United States. The group compared the estimates of new cases by southern state in 2012 and 2013. The results are in **Table 10.32**.

Southern States	2012	2013
Alabama	3,450	3,720
Arkansas	2,150	2,280
Florida	15,540	15,710
Georgia	6,970	7,310
Kentucky	3,160	3,300
Louisiana	3,320	3,630
Mississippi	1,990	2,080
North Carolina	7,090	7,430
Oklahoma	2,630	2,690
South Carolina	3,570	3,580
Tennessee	4,680	5,070
Texas	15,050	14,980
Virginia	6,190	6,280

Table 10.32

121. A traveler wanted to know if the prices of hotels are different in the 10 cities that he visits the most often. The list of the cities with the corresponding prices for his two favorite hotel chains is in **Table 10.33**. Test at the 1 percent level of significance.

Cities	Hyatt Regency prices in dollars	Hilton prices in dollars
Atlanta	107	169
Boston	358	289
Chicago	209	299
Dallas	209	198
Denver	167	169
Indianapolis	179	214
Los Angeles	179	169
New York City	625	459
Philadelphia	179	159
Washington, DC	245	239

Table 10.33

122. A politician asked his staff to determine whether the underemployment rate in the Northeast decreased from 2011 to 2012. The results are in **Table 10.34**.

Northeastern States	2011	2012
Connecticut	17.3	16.4
Delaware	17.4	13.7
Maine	19.3	16.1
Maryland	16.0	15.5
Massachusetts	17.6	18.2
New Hampshire	15.4	13.5
New Jersey	19.2	18.7
New York	18.5	18.7
Ohio	18.2	18.8
Pennsylvania	16.5	16.9
Rhode Island	20.7	22.4
Vermont	14.7	12.3
West Virginia	15.5	17.3

Table 10.34

BRINGING IT TOGETHER: HOMEWORK

Use the following information to answer the next 10 exercises. Indicate which of the following choices best identifies the hypothesis test.

A. Independent group means, population standard deviations and/or variances known

B. Independent group means, population standard deviations and/or variances unknown

C. Matched or paired samples

D. Single mean

E. Two proportions

F. Single proportion

123. A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. The population standard deviations are two pounds and three pounds, respectively. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet.

124. A new chocolate bar is taste-tested on consumers. Of interest is whether the proportion of children who like the new chocolate bar is greater than the proportion of adults who like it.

125. The mean number of English courses taken in a two-year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 9 males and 16 females.

126. A football league reported that the mean number of touchdowns per game was five. A study is done to determine if the mean number of touchdowns has decreased.

127. A study is done to determine if students in the California state university system take longer to graduate than students enrolled in private universities. One hundred students from both the California state university system and private universities are surveyed. From years of research, it is known that the population standard deviations are 1.5811 years and 1 year, respectively.

128. According to a doctor's magazine, 75 percent of senior citizens think that yearly checkups are very important. A study is done to verify this.

129. According to a recent study, U.S. companies have a mean maternity leave of six weeks.

130. A recent survey showed an increase in use of prescription medication among local senior citizens as compared to the national percent. Suppose that a survey of 100 local senior citizens and 100 national senior citizens is conducted to see if the proportion of prescription medication use is higher locally than nationally.

131. A new SAT study course is tested on 12 individuals. Pre-course and post-course scores are recorded. Of interest is the mean increase in SAT scores. The following data are collected:

Pre-course score	Post-course score
1	300
960	920
1010	1100
840	880
1100	1070
1250	1320
860	860
1330	1370
790	770
990	1040
1110	1200
740	850

Table 10.35

132. According to a statistics college professor, 68 percent of his students pass the final exam. A graduate researcher designs a study to determine if this claim is true.

133. Lesley E. Tan investigated the relationship between left-handedness versus right-handedness and motor competence in preschool children. Random samples of 41 left-handed preschool children and 41 right-handed preschool children were given several tests of motor skills to determine if there is evidence of a difference between the children based on this experiment. The experiment produced the means and standard deviations shown in **Table 10.36**. Determine the appropriate test and best distribution to use for that test.

	Left-handed	Right-handed
Sample size	41	41
Sample mean	97.5	98.1
Sample standard deviation	17.5	19.2

Table 10.36

a. Two independent means, normal distribution
b. Two independent means, Student's *t*-distribution
c. Matched or paired samples, Student's *t*-distribution
d. Two population proportions, normal distribution

134. A golf instructor is interested in determining if her new technique for improving players' golf scores is effective. She takes four new students. She records their 18-hole scores before learning the technique and after having taken her class. She conducts a hypothesis test. The data are shown in **Table 10.37**.

	Player 1	**Player 2**	**Player 3**	**Player 4**
Mean score before class	83	78	93	87
Mean score after class	80	80	86	86

Table 10.37

This is

a. a test of two independent means.
b. a test of two proportions.
c. a test of a single mean.
d. a test of a single proportion.

REFERENCES

10.1 Two Population Means with Unknown Standard Deviations

Baseball-Almanac. (2013). *World series history*. Retrieved from http://www.baseball-almanac.com/ws/wsmenu.shtml

Graduating Engineer + Computer Careers. (n.d.). Retrieved from http://www.graduatingengineer.com

Microsoft Bookshelf. (n.d.).

Nasdaq. (n.d.). *Sectoring by industry groups*. Retrieved from http://www.nasdaq.com/markets/barchart-sectors.aspx

Prostitution Research and Education. (2013). *Strip clubs: Where prostitution and trafficking happen*. Retrieved from www.prostitutionresearch.com/ProsViolPosttrauStress.html

U.S. Senate. (n.d.). Retrieved from www.senate.gov

Wikipedia. (n.d.). *List of current United States Senators by age*. Retrieved from http://en.wikipedia.org/wiki/List_of_current_United_States_Senators_by_age

10.2 Two Population Means with Known Standard Deviations

Centers for Disease Control and Prevention. (2008, July 18). State-specific prevalence of obesity among adults—United States, 2007. *MMWR*, 57(28), 765–768. Retrieved from http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5728a1.htm

Federal Bureau of Investigation. (n.d.). *Texas Crime Rates 1960–1012*. Available at http://www.disastercenter.com/crime/txcrime.htm

Hinduja, S. (2013). Sexting Research and Gender Differences. *Cyberbulling Research Center*. Retrieved from http://cyberbullying.us/blog/sexting-research-and-gender-differences/

Humes, K. R., Jones, N. A., & Ramirez, R. R. (2011 March). *Overview of race and Hispanic origin: 2010* (2010 Census Briefs). Washington, DC: U.S. Census Bureau. Available online at http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf

Smith, A. (2011, July 11). 35% of American adults own a smartphone. Pew Internet. Available online at http://www.pewinternet.org/~/media/Files/Reports/2011/PIP_Smartphones.pdf

Visually. (2013). *Smart phone users, by the numbers*. Retrieved from http://visual.ly/smart-phone-users-numbers

10.3 Comparing Two Independent Population Proportions

American Cancer Society. (n.d.). Retrieved from http://www.cancer.org/index

Centers for Disease Control and Prevention. (n.d.). *West Nile virus*. Retrieved from http://www.cdc.gov/ncidod/dvbid/westnile/index.htm

Chancellor's Office, California Community Colleges. (1994, Nov.).

Educational Resources. (n.d.).

Gallup. (2013). *State of the states*. Retrieved from http://www.gallup.com/poll/125066/State-States.aspx?ref=interactive

Hilton Hotels. (n.d.). Retrieved from http://www.hilton.com

Hyatt Hotels. Retrieved from http://hyatt.com

San Jose Museum of Art. (n.d.). Whitney exhibit (on loan).

U.S. Department of Health and Human Services. (n.d). *Statistics*. Retrieved from https://www.hhs.gov/

SOLUTIONS

1 two proportions

3 matched or paired samples

5 single mean

7 independent group means, population standard deviations and/or variances unknown

9 two proportions

11 independent group means, population standard deviations and/or variances unknown

13 independent group means, population standard deviations and/or variances unknown

15 two proportions

17 The random variable is the difference between the mean amounts of sugar in the two soft drinks.

19 means

21 two-tailed

23 the difference between the mean life spans of whites and nonwhites

25 This is a comparison of two population means with unknown population standard deviations.

27 Check student's solution.

29

a. Reject the null hypothesis.

b. p-value < 0.05

c. There is not enough evidence at the 5 percent level of significance to support the claim that life expectancy in the 1900s is different between whites and nonwhites.

31 the difference in mean speeds of the fastball pitches of the two pitchers

33 –2.46

35 At the 1 percent significance level, we can reject the null hypothesis. There is sufficient data to conclude that the mean speed of Rodriguez's fastball is faster than Wesley's.

37 Subscripts: 1 = Food, 2 = No Food
H_0: $\mu_1 \leq \mu_2$
H_a: $\mu_1 > \mu_2$

39

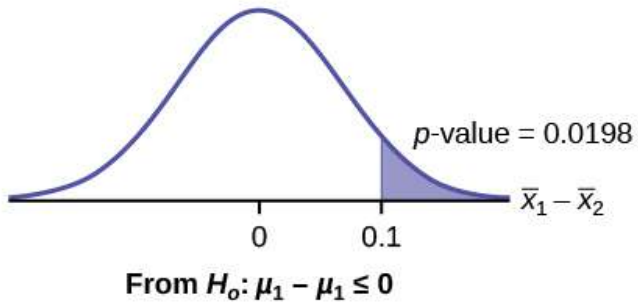


Figure 10.18

41 Subscripts: 1 = Gamma, 2 = Zeta
H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

43 0.0062

45 There is sufficient evidence to reject the null hypothesis. The data support that the melting point for Alloy Zeta is different from the melting point of Alloy Gamma.

47 $P'_{OS1} - P'_{OS2}$ = difference in the proportions of phones that had system failures within the first eight hours of operation with OS_1 and OS_2.

49 0.1018

51 proportions

53 right-tailed

55 The random variable is the difference in proportions (percents) of the populations that are of two or more races in Nevada and North Dakota.

57 Our sample sizes are much greater than five each, so we use the normal for two proportions distribution for this hypothesis test.

59 Check student's solution.

61

a. Reject the null hypothesis.

b. p-value < alpha

c. At the 5 percent significance level, there is sufficient evidence to conclude that the proportion (percent) of the population that is of two or more races in Nevada is statistically higher than that in North Dakota.

63 the mean difference of the system failures

65 0.0067

67 With a *p*-value 0.0067, we can reject the null hypothesis. There is enough evidence to support that the software patch is effective in reducing the number of system failures.

69 0.0021

71

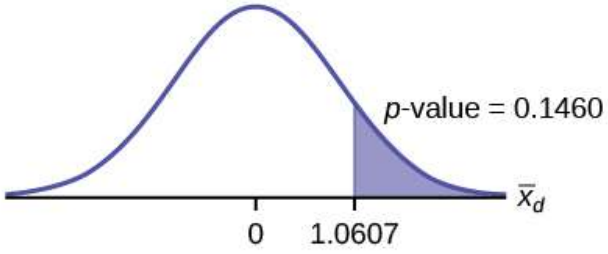


Figure 10.19

73 H_0: $\mu_d \geq 0$ H_a: $\mu_d < 0$

75 0.0699

77 We decline to reject the null hypothesis. There is not sufficient evidence to support that the medication is effective.

79 Subscripts: 1: two-year colleges, 2: four-year colleges

a. H_0: $\mu_1 \geq \mu_2$

b. H_a: $\mu_1 < \mu_2$

c. $\bar{X}_1 - \bar{X}_2$ is the difference between the mean enrollments of the two-year colleges and the four-year colleges.

d. Student's *t*

e. test statistic: -0.2480

f. *p*-value: 0.4019

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject.
   iii. Reason for Decision: *p*-value > alpha
   iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean enrollment at four-year colleges is higher than at two-year colleges.

81 Subscripts: 1: mechanical engineering, 2: electrical engineering

a. H_0: $\mu_1 \geq \mu_2$

b. H_a: $\mu_1 < \mu_2$

c. $\bar{X}_1 - \bar{X}_2$ is the difference between the mean entry-level salaries of mechanical engineers and electrical engineers.

d. t_{108}

e. test statistic: $t = -0.82$

f. *p*-value: 0.2061

g. Check student's solution.

h.
   i. Alpha: 0.05

ii. Decision: Do not reject the null hypothesis.

iii. Reason for Decision: *p*-value > alpha

iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that the mean entry-level salaries of mechanical engineers is lower than that of electrical engineers.

83

a. H_0: $\mu_1 = \mu_2$

b. H_a: $\mu_1 \neq \mu_2$

c. $\bar{X}_1 - \bar{X}_2$ is the difference between the mean times for completing a lap in races and in practices.

d. $t_{20.32}$

e. test statistic: –4.70

f. *p*-value: 0.0001

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for Decision: *p*-value < alpha

iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

85

a. H_0: $\mu_1 = \mu_2$

b. H_a: $\mu_1 \neq \mu_2$

c. is the difference between the mean times for completing a lap in races and in practices.

d. $t_{40.94}$

e. test statistic: –5.08

f. *p*-value: zero

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for Decision: *p*-value < alpha

iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different from that in practices.

88 c

90 Test: two independent sample means, population standard deviations unknown. Random variable: $\bar{X}_1 - \bar{X}_2$
Distribution: H_0: $\mu_1 = \mu_2$, H_a: $\mu_1 < \mu_2$
The mean age of entering the industry in Canada is lower than the mean age in the United States.

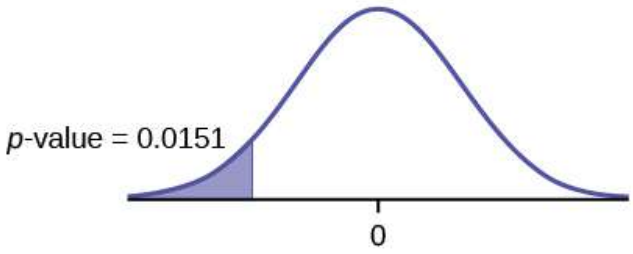


Figure 10.20

Graph: left-tailed p-value : 0.0151 Decision: Do not reject H_0. Conclusion: At the 1 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of entering the industry in Canada is lower than the mean age in the United States.

92 d

94 Subscripts: 1 = boys, 2 = girls

a. H_0: $\mu_1 \leq \mu_2$

b. H_a: $\mu_1 > \mu_2$

c. The random variable is the difference in the mean auto insurance costs for boys and girls.

d. normal

e. test statistic: $z = 2.50$

f. p value: 0.0062

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for Decision: p value < alpha
   iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean cost of auto insurance for teenage boys is greater than that for girls.

96 Subscripts: 1 = non-hybrid sedans, 2 = hybrid sedans

a. H_0: $\mu_1 \geq \mu_2$

b. H_a: $\mu_1 < \mu_2$

c. The random variable is the difference in the mean miles per gallon of nonhybrid sedans and hybrid sedans.

d. normal

e. test statistic: 6.36

f. p-value: 0

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p value < alpha
   iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the mean miles per gallon of non-hybrid sedans is less than that of hybrid sedans.

98

a. H_0: $\mu_d = 0$

b. H_a: $\mu_d < 0$

c. The random variable X_d is the average difference between husband's and wife's satisfaction level.

d. t_9

e. test statistic: $t = -1.86$

f. p value: 0.0479

g. Check student's solution

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis, but run another test.

iii. Reason for Decision: p value < alpha

iv. Conclusion: This is a weak test because alpha and the p value are close. However, there is insufficient evidence to conclude that the mean difference is negative.

101 Subscripts: 1 = Cabrillo College, 2 = Lake Tahoe College

a. H_0: $p_1 = p_2$

b. H_a: $p_1 \neq p_2$

c. The random variable is the difference between the proportions of Hispanic students at Cabrillo College and Lake Tahoe College.

d. normal for two proportions

e. test statistic: 4.29

f. p-value: 0.00002

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: p-value < alpha

iv. Conclusion: There is sufficient evidence to conclude that the proportions of Hispanic students at Cabrillo College and Lake Tahoe College are different.

103 a

105 Test: two independent sample proportions. Random variable: $p'_1 - p'_2$ Distribution:
H_0: $p_1 = p_2$
H_a: $p_1 \neq p_2$ The proportion of e-reader users is different for the 16- to 29-year-old users from that of the 30 and older users. Graph: two-tailed

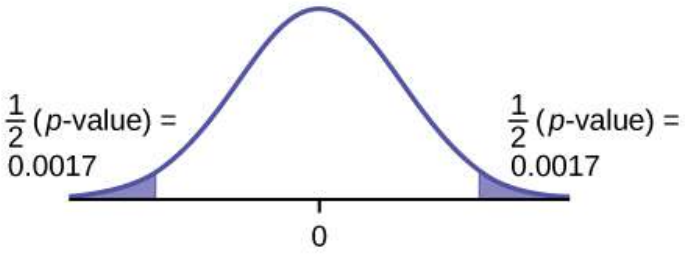


Figure 10.21

p-value : 0.0033 Decision: Reject the null hypothesis. Conclusion: At the 5 percent level of significance, from the sample data, there is sufficient evidence to conclude that the proportion of e-reader users 16 to 29 years old is different from the proportion of e-reader users 30 and older.

107 Test: two independent sample proportions Random variable: $p'_1 - p'_2$ Distribution: H_0: $p_1 = p_2$
H_a: $p_1 > p_2$ A higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older. Graph: right-tailed

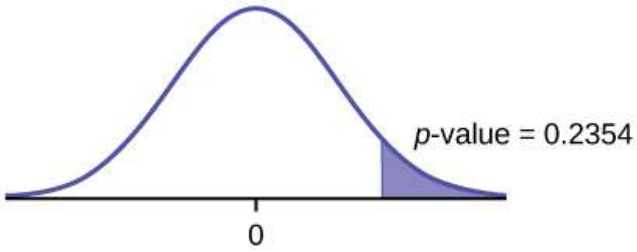


Figure 10.22

p-value: 0.2354 Decision: Do not reject the H_0. Conclusion: At the 1 percent level of significance, from the sample data, there is not sufficient evidence to conclude that a higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

109 Subscripts: 1: men; 2: women

a. H_0: $p_1 \leq p_2$

b. H_a: $p_1 > p_2$

c. $P'_1 - P'_2$ is the difference between the proportions of men and women who enjoy shopping for electronic equipment.

d. normal for two proportions

e. test statistic: 0.22

f. *p*-value: 0.4133

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Do not reject the null hypothesis.

iii. Reason for Decision: *p*-value > alpha

iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that the proportion of men who enjoy shopping for electronic equipment is more than the proportion of women.

111

a. H_0: $p_1 = p_2$

b. H_a: $p_1 \neq p_2$

c. $P'_1 - P'_2$ is the difference between the proportions of men and women that have at least one pierced ear.

d. normal for two proportions

e. test statistic: –4.82

f. *p*-value: zero

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for Decision: *p*-value < alpha

iv. Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that the proportions of males and females with at least one pierced ear is different.

113

a. H_0: $\mu_d = 0$

b. H_a: $\mu_d > 0$

c. The random variable X_d is the mean difference in work times on days when eating breakfast and on days when not eating breakfast.

d. t_9

e. test statistic: 4.8963

f. p-value: 0.0004

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for Decision: p-value < alpha

iv. Conclusion: At the 5 percent level of significance, there is sufficient evidence to conclude that the mean difference in work times on days when eating breakfast and on days when not eating breakfast has increased.

114 p-value = 0.1494 At the 5 percent significance level, there is insufficient evidence to conclude that the medication lowered cholesterol levels after 12 weeks.

116 b

118 c

120 Test: two matched pairs or paired samples (t-test) Random variable: $\bar{X}_d$ Distribution: t_{12} H_0: $\mu_d = 0$ H_a: $\mu_d > 0$ The mean of the differences of new female breast cancer cases in the south between 2013 and 2012 is greater than zero. The estimate for new female breast cancer cases in the south is higher in 2013 than in 2012. Graph: right-tailed p-value: 0.0004

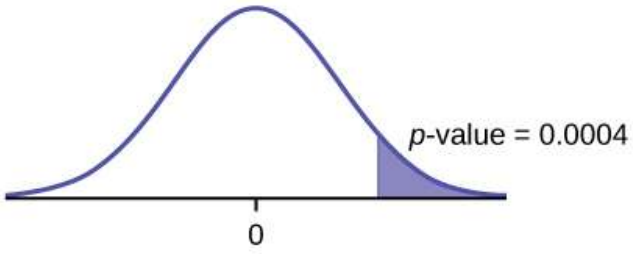


Figure 10.23

Decision: Reject H_0. Conclusion: At the 5 percent level of significance, from the sample data, there is sufficient evidence to conclude that there was a higher estimate of new female breast cancer cases in 2013 than in 2012.

122 Test: matched or paired samples (t-test) Difference data: {–0.9, –3.7, –3.2, –0.5, 0.6, –1.9, –0.5, 0.2, 0.6, 0.4, 1.7, –2.4, 1.8} Random Variable: $\bar{X}_d$ Distribution: H_0: $\mu_d = 0$ H_a: $\mu_d < 0$ The mean of the differences of the rate of underemployment in the northeastern states between 2012 and 2011 is less than zero. The underemployment rate went down from 2011 to 2012. Graph: left-tailed.

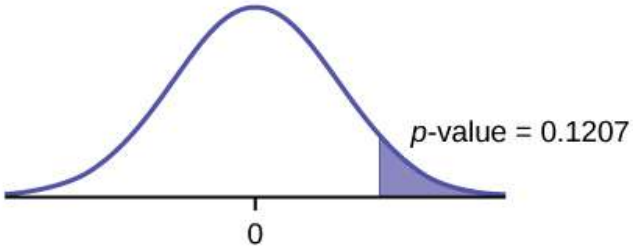


Figure 10.24

p-value: 0.1207 Decision: Do not reject H_0. Conclusion: At the 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that there was a decrease in the underemployment rates of the northeastern states from 2011 to 2012.

124 e

126 d

128 f

130 e

132 f The graduate researcher will be comparing a sample proportion to a population proportion or claim. Thus, the study includes the hypothesis test of a single proportion. A two proportion hypothesis test compares two sample proportions.

134 a

11 | THE CHI-SQUARE DISTRIBUTION

Figure 11.1 The chi-square distribution can be used to find relationships between two things, like grocery prices at different stores. (credit: Pete/flickr)

Introduction

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Interpret the chi-square probability distribution as the sample size changes
- Conduct and interpret chi-square goodness-of-fit hypothesis tests
- Conduct and interpret chi-square test of independence hypothesis tests
- Conduct and interpret chi-square homogeneity hypothesis tests
- Conduct and interpret chi-square single variance hypothesis tests

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to such questions. This distribution is called the chi-square distribution.

In this chapter, you will learn the three major applications of the chi-square distribution:

- The goodness-of-fit test, which determines if data fit a particular distribution, such as in the lottery example
- The test of independence, which determines if events are independent, such as in the movie example
- The test of a single variance, which tests variability, such as in the coffee example

NOTE

Though the chi-square distribution depends on calculators or computers for most of the calculations, there is a table available (see **Appendix G**). TI-83+ and TI-84 calculator instructions are included in the text.

Collaborative Exercise

Look in the sports section of a newspaper or on the internet for some sports data: baseball averages, basketball scores, golf tournament scores, football odds, swimming times, and the like. Plot a histogram and a boxplot using your data. See if you can determine a probability distribution that your data fits. Have a discussion with the class about your choice.

11.1 | Facts About the Chi-Square Distribution

The notation for the **chi-square distribution** is

$$\chi \sim \chi^2_{df}$$

where df = degrees of freedom, which depends on how chi-square is being used. If you want to practice calculating chi-square probabilities then use $df = n - -1$. The degrees of freedom for the three major uses are calculated differently.

For the χ^2 distribution, the population mean is $\mu = df$, and the population standard deviation is $\sigma = \sqrt{2(df)}$.

The random variable is shown as χ^2, but it may be any uppercase letter.

The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables is

$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots + (Z_k)^2$, where the following are true:

- The curve is nonsymmetrical and skewed to the right.
- There is a different chi-square curve for each df.

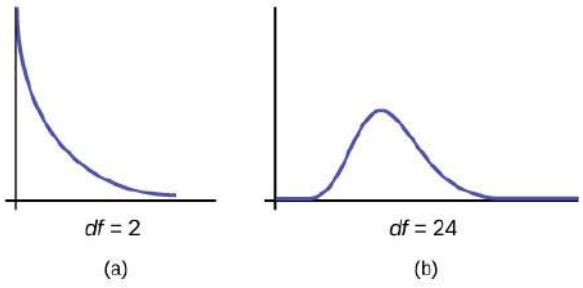


Figure 11.2

- The test statistic for any test is always greater than or equal to zero.

- When $df > 90$, the chi-square curve approximates the normal distribution. For $X \sim \chi^2_{1,000}$, the mean, $\mu = df = 1{,}000$ and the standard deviation, $\sigma = \sqrt{2(1{,}000)} = 44.7$. Therefore, $X \sim N(1{,}000, 44.7)$, approximately.
- The mean, μ, is located just to the right of the peak.

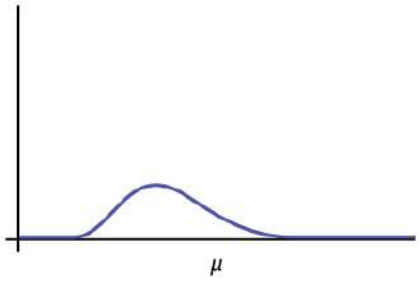


Figure 11.3

11.2 | Goodness-of-Fit Test

In this type of hypothesis test, you determine whether the data fit a particular distribution. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test, meaning the distribution for the hypothesis test is chi-square, to determine if there is a fit. The null and the alternative hypotheses for this test may be written in sentences or may be stated as equations or inequalities.

The test statistic for a goodness-of-fit test is:

$$\sum_k \frac{(O-E)^2}{E}$$

where

- O = **observed values** (data),
- E = **expected values** (from theory), and
- k = the number of different data cells or categories.

The observed values are the data values, and the expected values are the values you would expect to get if the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The number of degrees of freedom is df = (number of categories – 1).

The goodness-of-fit test is almost always right-tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

NOTE

The expected value for each cell needs to be at least five for you to use this test.

Example 11.1

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism rate follows faculty perception. The faculty expected that a group of 100 students would miss class according to **Table 11.1**.

Number of Absences per Term	Expected Number of Students
0–2	50
3–5	30
6–8	12
9–11	6
12+	2

Table 11.1

A random survey across all mathematics courses was then done to determine the number of observed absences in a course. **Table 11.2** displays the results of that survey.

Number of Absences per Term	Actual Number of Students
0–2	35
3–5	40
6–8	20
9–11	1
12+	4

Table 11.2

Determine the null and alternative hypotheses needed to conduct a goodness-of-fit test.

H_0: Student absenteeism *fits* faculty perception.

The alternative hypothesis is the opposite of the null hypothesis.

H_a: Student absenteeism *does not fit* faculty perception.

a. Can you use the information as it appears in the charts to conduct the goodness-of-fit test?

Solution 11.1

a. **No.** Notice that the expected number of absences for the 12+ entry is less than five; it is two. Combine that group with the 9–11 group to create new tables where the number of students for each entry is at least five. The new results are in **Table 11.2** and **Table 11.3**.

Number of Absences per Term	Expected Number of Students
0–2	50
3–5	30
6–8	12
9+	8

Table 11.3

Number of Absences per Term	Actual Number of Students
0–2	35
3–5	40
6–8	20
9+	5

Table 11.4

b. What is the number of degrees of freedom (*df*)?

Solution 11.1

b. There are four cells or categories in each of the new tables.

df = number of cells – 1 = 4 – 1 = 3.

Try It

11.1 A factory manager needs to understand how many products are defective versus how many are produced. The number of expected defects is listed in **Table 11.5**.

Number Produced	Number Defective
0–100	5
101–200	6
201–300	7
301–400	8
401–500	10

Table 11.5

A random sample was taken to determine the actual number of defects. **Table 11.6** shows the results of the survey.

Number Produced	Number Defective
0–100	5
101–200	7
201–300	8
301–400	9
401–500	11

Table 11.6

State the null and alternative hypotheses needed to conduct a goodness-of-fit test, and state the degrees of freedom.

Example 11.2

Employers want to know which days of the week employees are absent in a five-day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week they had the highest number of employee absences. The results were distributed as in **Table 11.6**. For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five-day work week? Test at a 5 percent significance level.

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Absences	15	12	9	9	15

Table 11.7 Day of the Week Employees Were Most Absent

Solution 11.2

The null and alternative hypotheses are as follows:

- H_0: The absent days occur with equal frequencies; that is, they fit a uniform distribution.
- H_a: The absent days occur with unequal frequencies; that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then, out of 60 absent days (the total in the sample: 15 + 12 + 9 + 9 + 15 = 60) there would be 12 absences on Monday, 12 on Tuesday, 12 on Wednesday, 12 on Thursday, and 12 on Friday. These numbers are the *expected* (E) values. The values in the table are the *observed* (O) values or data.

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

- Expected (E) values (12, 12, 12, 12, 12)
- Observed (O) values (15, 12, 9, 9, 15)
- $(O - E)$
- $(O - E)^2$
- $\frac{(O - E)^2}{E}$

Now add (sum) the last column. The sum is three. This is the χ^2 test statistic.

To find the p-value, calculate $P(\chi^2 > 3)$. This test is right-tailed. Use a computer or calculator to find the p-value. You should get p-value = 0.5578.

The dfs are the number of cells – 1 = 5 – 1 = 4.

Using the TI-83, 83+, 84, 84+ Calculator

Press `2nd DISTR`. Arrow down to `χ²cdf`. Press `ENTER`. Enter `(3,10^99,4)`. Rounded to four decimal places, you should see .5578, which is the p-value.

Next, complete a graph like the following one with the proper labeling and shading. You should shade the right tail.

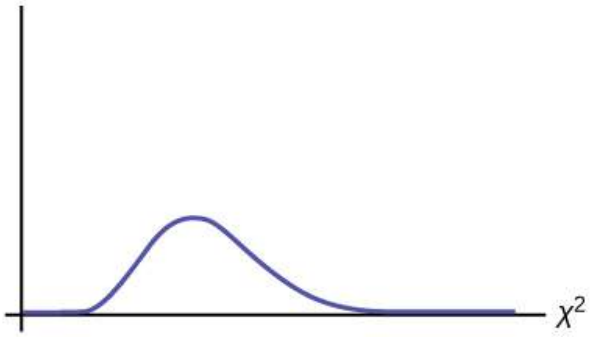


Figure 11.4

The decision is not to reject the null hypothesis.

Conclusion: At a 5 percent level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

Using the TI-83, 83+, 84, 84+ Calculator

TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example, **Example 11.3**, has the calculator instructions. The newer TI-84 calculators have in `STAT TESTS` the test `Chi2 GOF`. To run the test, put the observed values—the data—into a first list and the expected values—the values you expect if the null hypothesis is true—into a second list. Press `STAT TESTS` and `Chi2 GOF`. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press `Calculate` or `Draw`. Make sure you clear any lists before you start. **To Clear Lists in the calculators:** Go into `STAT EDIT` and arrow up to the list name area of the particular list. Press `CLEAR` and then arrow down. The list will be cleared. Alternatively, you can press `STAT` and press 4 for `ClrList`. Enter the list name and press `ENTER`.

Try It

11.2 Teachers want to know which night each week their students are doing most of their homework. Most teachers think that students do homework equally throughout the week. Suppose a random sample of 56 students were asked on which night of the week they did the most homework. The results were distributed as in **Table 11.8**.

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Number of Students	11	8	10	7	10	5	5

Table 11.8

From the population of students, do the nights for the highest number of students doing the majority of their homework occur with equal frequencies during a week? What type of hypothesis test should you use?

Example 11.3

One study indicates that the number of televisions that American families have is distributed (this is the *given* distribution for the American population) as in **Table 11.9**.

Number of Televisions	Percent
0	10
1	16
2	55
3	11
4+	8

Table 11.9

The table contains expected (*E*) percents.

A random sample of 600 families in the far western U.S. resulted in the data in **Table 11.10**.

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
4+	15
	Total = 600

Table 11.10

The table contains observed (*O*) frequency values.

At the 1 percent significance level, does it appear that the distribution *number of televisions* of far western U.S. families is different from the distribution for the American population as a whole?

Solution 11.3

This problem asks you to test whether the far western U.S. families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (*E*) frequencies, multiply the percentage by 600. The expected frequencies are shown in **Table 11.10**.

Number of Televisions	Percent	Expected Frequency
0	10	(0.10)(600) = 60
1	16	(0.16)(600) = 96
2	55	(0.55)(600) = 330
3	11	(0.11)(600) = 66
more than 3	8	(0.08)(600) = 48

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter 0.10 * 600.

H_0: The *number of televisions* distribution of far western U.S. families is the same as the *number of televisions* distribution of the American population.

H_a: The *number of televisions* distribution of far western U.S. families is different from the *number of televisions* distribution of the American population.

Distribution for the test: χ^2_4 where df = (the number of cells) – 1 = 5 – 1 = 4.

NOTE

$df \neq 600 - 1$

Calculate the test statistic: $\chi^2 = 29.65$

Graph

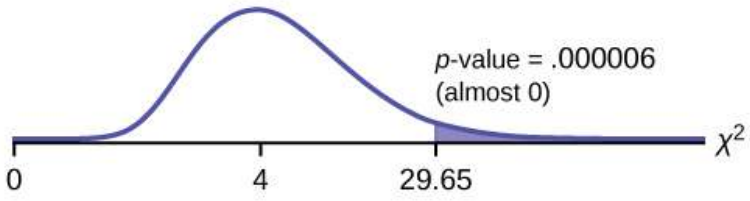


Figure 11.5

Probability statement: p-value = $P(\chi^2 > 29.65) = .000006$

Compare α and the p-value:

- $\alpha = .01$
- p-value = 0.000006

So, $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_o.

This means you reject the hypothesis that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1 percent significance level, from the data, there is sufficient evidence to conclude that the *number of televisions* distribution for the far western United States is different from the *number of televisions* distribution for the American population as a whole.

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and ENTER. Make sure to clear lists L1, L2, and L3 if they have data in them—see the note at the end of **Example 11.2**. Into L1, put the observed frequencies 66, 119, 349, 60, 15. Into L2, put the expected frequencies .10*600, .16*600, .55*600, .11*600, .08*600. Arrow over to list L3 and up to the name area L3. Enter (L1-L2)^2/L2 and ENTER. Press 2nd QUIT. Press 2nd LIST and arrow over to MATH. Press 5. You should see sum (Enter L3). Rounded to two decimal places, you should see 29.65. Press 2nd DISTR. Press 7 or Arrow down to 7:χ2cdf and press ENTER. Enter (29.65,1E99,4). Rounded to four places, you should see 5.77E-6 = .000006 (rounded to six decimal places), which is the *p*-value.

The newer TI-84 calculators have in STAT TESTS the test Chi2 GOF. To run the test, put the observed values (the data) into a first list and the expected values—the values you expect if the null hypothesis is true—into a second list. Press STAT TESTS and Chi2 GOF. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press Calculate or Draw. Make sure you clear any lists before you start.

Try It Σ

11.3 The expected percentage of the number of pets students have in their homes is distributed (this is the given distribution for the student population of the United States) as in **Table 11.12**.

Number of Pets	Percent
0	18
1	25
2	30
3	18
4+	9

Table 11.12

A random sample of 1,000 students from the eastern United States resulted in the data in **Table 11.13**.

Number of Pets	Frequency
0	210
1	240
2	320
3	140
4+	90

Table 11.13

At the 1 percent significance level, does it appear that the distribution *number of pets* of students in the eastern United

States is different from the distribution for the United States student population as a whole? What is the *p*-value?

Example 11.4

Suppose you flip two coins 100 times. The results are 20 *HH*, 27 *HT*, 30 *TH*, and 23 *TT*. Are the coins fair? Test at a 5 percent significance level.

Solution 11.4

This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is {*HH*, *HT*, *TH*, *TT*}. Out of 100 flips, you would expect 25 *HH*, 25 *HT*, 25 *TH*, and 25 *TT*. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins (20 *HH*, 27 *HT*, 30 *TH*, 23 *TT*) fit the expected distribution?"

Random variable: Let X = the number of heads in one flip of the two coins. X takes on the values 0, 1, 2. There are 0, 1, or 2 heads in the flip of two coins. Therefore, the number of cells is three. Since X = the number of heads, the observed frequencies are 20 for two heads, 57 for one head, and 23 for zero heads or both tails. The expected frequencies are 25 for two heads, 50 for one head, and 25 for zero heads or both tails. This test is right-tailed.

H_0: The coins are fair.

H_a: The coins are not fair.

Distribution for the test: χ^2_2 where $df = 3 - 1 = 2$.

Calculate the test statistic: $\chi^2 = 2.14$.

Graph

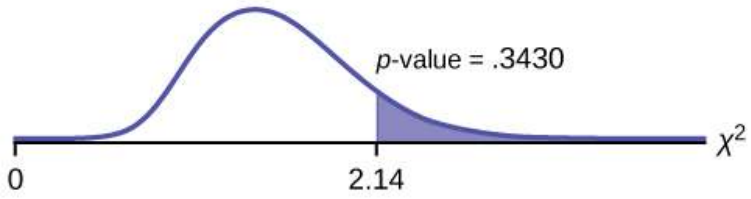


Figure 11.6

Probability statement: p-value = $P(\chi^2 > 2.14) = 0.3430$.

Compare α and the p-value:

- $\alpha = .05$
- p-value = 0.3430

$\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, do not reject H_0.

Conclusion: There is insufficient evidence to conclude that the coins are not fair.

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and ENTER. Make sure you clear lists L1, L2, and L3 if they have data in them. Into L1, put the observed frequencies 20, 57, 23. Into L2, put the expected frequencies 25, 50, 25. Arrow over to list L3 and up to the name area L3. Enter (L1-L2)^2/L2 and ENTER. Press 2nd QUIT. Press 2nd LIST and

arrow over to MATH. Press 5. You should see sum. Enter L3. Rounded to two decimal places, you should see 2.14. Press 2nd DISTR. Arrow down to 7:χ2cdf—or press 7. Press ENTER. Enter 2.14,1E99,2). Rounded to four places, you should see .3430, which is the *p*-value.

The newer TI-84 calculators have in STAT TESTS the test Chi2 GOF. To run the test, put the observed values—the data—into a first list and the expected values—the values you expect if the null hypothesis is true—into a second list. Press STAT TESTS and Chi2 GOF. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press Calculate or Draw. Make sure you clear any lists before you start.

Try It Σ

11.4 Students in a social studies class hypothesize that the literacy rates around the world for every region are 82 percent. **Table 11.14** shows the actual literacy rates around the world broken down by region. What are the test statistic and the degrees of freedom?

MDG Region	Adult Literacy Rate (%)
Developed regions	99
Commonwealth of Independent States	99.5
Northern Africa	67.3
Sub-Saharan Africa	62.5
Latin America and the Caribbean	91
Eastern Asia	93.8
Southern Asia	61.9
Southeastern Asia	91.9
Western Asia	84.5
Oceania	66.4

Table 11.14

11.3 | Test of Independence

Tests of independence involve using a **contingency table** of observed (data) values.

The test statistic for a **test of independence** is similar to that of a goodness-of-fit test

$$\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$$

where

- O = observed values,
- E = expected values,
- i = the number of rows in the table, and

- j = the number of columns in the table.

There are $i \cdot j$ terms of the form $\frac{(O-E)^2}{E}$.

A test of independence determines whether two factors are independent. You first encountered the term *independence* in **Probability Topics**. As a review, consider the following example.

NOTE

The expected value for each cell needs to be at least five for you to use this test.

Example 11.5

Suppose A = a speeding violation in the last year and B = a cell phone user while driving. If A and B are independent, then $P(A \text{ AND } B) = P(A)P(B)$. A AND B is the event that a driver received a speeding violation last year and also used a cell phone while driving. Suppose, in a study of drivers who received speeding violations in the last year, and who used cell phones while driving, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 used cell phones while driving and 450 did not.

Let y = expected number of drivers who used a cell phone while driving and received speeding violations.

If A and B are independent, then $P(A \text{ AND } B) = P(A)P(B)$. By substitution,

$$\frac{y}{755} = \left(\frac{70}{755}\right)\left(\frac{305}{755}\right).$$

Solve for y: $y = \frac{(70)(305)}{755} = 28.3$.

About 28 people from the sample are expected to use cell phones while driving and to receive speeding violations.

In a test of independence, we state the null and alternative hypotheses in words. Since the contingency table consists of two factors, the null hypothesis states that the factors are independent and the alternative hypothesis states that they are not independent (dependent). If we do a test of independence using the example, then the null hypothesis is the following:

H_0: Being a cell phone user while driving and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to use cell phones while driving and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, as it is in a goodness-of-fit.

The number of degrees of freedom for the test of independence is

df = (number of columns – 1)(number of rows – 1).

The following formula calculates the **expected number** (E):

$$E = \frac{\text{(row total)(column total)}}{\text{total number surveyed}}$$

Try It

11.5 A sample of 300 students is taken. Of the students surveyed, 50 were music students, while 250 were not. 97 were on the honor roll, while 203 were not. If we assume being a music student and being on the honor roll are independent events, what is the expected number of music students who are also on the honor roll?

Example 11.6

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and non-students. In **Table 11.15** is a sample of the adult volunteers and the number of hours they volunteer per week.

Type of Volunteer	1–3 Hours	4–6 Hours	7–9 Hours	Row Total
Community College Students	111	96	48	255
Four-year College Students	96	133	61	290
Non-students	91	150	53	294
Column Total	298	379	162	839

Table 11.15 Number of Hours Worked per Week by Volunteer Type (Observed) The table contains **observed (O)** values (data).

Is the number of hours volunteered **independent** of the type of volunteer?

Solution 11.6

The **observed values** and the question at the end of the problem, "Is the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are *number of hours volunteered* and *type of volunteer*. This test is always right-tailed.

H_0: The number of hours volunteered is *independent* of the type of volunteer.

H_a: The number of hours volunteered is *dependent* on the type of volunteer.

The expected result are in **Table 11.15**.

Type of Volunteer	1-3 Hours	4–6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103	131	56
Nonstudents	104.42	132.81	56.77

Table 11.16 Number of Hours Worked per Week by Volunteer Type (Expected) The table contains **expected** (*E*) values (data).

For example, the calculation for the expected frequency for the top-left cell is

$$E = \frac{(\text{row total})(\text{column total})}{\text{total number surveyed}} = \frac{(255)(298)}{839} = 90.57.$$

Calculate the test statistic: $\chi^2 = 12.99$ (calculator or computer)

Distribution for the test: χ^2_4

$df = (3 \text{ columns} - 1)(3 \text{ rows} - 1) = (2)(2) = 4$

Graph

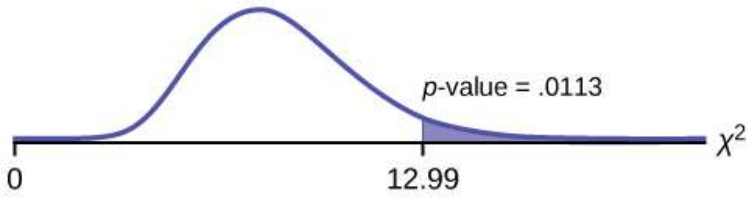


Figure 11.7

Probability statement: p-value = $P(\chi^2 > 12.99) = 0.0113$

Compare α and the p-value: Since no α is given, assume $\alpha = 0.05$. p-value = 0.0113. $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0. This means that the factors are not independent.

Conclusion: At a 5 percent level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on each other.

For the example in **Table 11.15**, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

Using the TI-83, 83+, 84, 84+ Calculator

Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 3 ENTER 3 ENTER. Enter the table values by row from **Table 11.15**. Press ENTER after each. Press 2nd QUIT. Press STAT and arrow over to TESTS. Arrow down to C:χ2-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 12.9909 and the p-value = .0113. Do the procedure a second time, but arrow down to Draw instead of Calculate.

Try It

11.6 The Bureau of Labor Statistics gathers data about employment in the United States. A sample is taken to calculate the number of U.S. citizens working in one of several industry sectors over time. **Table 11.17** shows the results:

Industry Sector	2000	2010	2020	Total
Non-agriculture Wage and Salary	13,243	13,044	15,018	41,305
Goods-producing, Excluding Agriculture	2,457	1,771	1,950	6,178
Services-providing	10,786	11,273	13,068	35,127
Agriculture, Forestry, Fishing, and Hunting	240	214	201	655
Non-agriculture Self-employed and Unpaid Family Worker	931	894	972	2,797
Secondary Wage and Salary Jobs in Agriculture and Private Household Industries	14	11	11	36
Secondary Jobs as a Self-employed or Unpaid Family Worker	196	144	152	492

Table 11.17

Industry Sector	2000	2010	2020	Total
Total	27,867	27,351	31,372	86,590

Table 11.17

We want to know if the change in the number of jobs is independent of the change in years. State the null and alternative hypotheses and the degrees of freedom.

Example 11.7

De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and need to succeed in school. **Table 11.18** shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Need to Succeed in School	High Anxiety	Med-High Anxiety	Medium Anxiety	Med-Low Anxiety	Low Anxiety	Row Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

Table 11.18 Need to Succeed in School vs. Anxiety Level

a. How many high anxiety level students are expected to have a high need to succeed in school?

Solution 11.7

a. The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

$$E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = \frac{155 \cdot 57}{400} = 22.09$$

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

b. If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?

Solution 11.7

b. The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} =$ _____

Solution 11.7

c. $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}} = 8.19$

d. The expected number of students who have a med-low anxiety level and a low need to succeed in school is about _____.

Solution 11.7

d. 8

Try It

11.7 Refer back to the information in **Try It**. How many services-providing jobs are there expected to be in 2020? How many nonagriculture wage and salary jobs are there expected to be in 2020?

11.4 | Test for Homogeneity

The goodness-of-fit test can be used to decide whether a population fits a given distribution, but it will not suffice to decide whether two populations follow the same unknown distribution. A different test, called the **test for homogeneity**, can be used to draw a conclusion about whether two populations have the same distribution. To calculate the test statistic for a test for homogeneity, follow the same procedure as with the test of independence.

NOTE

The expected value for each cell needs to be at least five for you to use this test.

Hypotheses

H_0: The distributions of the two populations are the same.

H_a: The distributions of the two populations are not the same.

Test Statistic

Use a χ^2 test statistic. It is computed in the same way as the test for independence.

Degrees of freedom (*df*)

df = number of columns – 1

Requirements

All values in the table must be greater than or equal to five.

Common Uses

Comparing two populations. For example: men vs. women, before vs. after, east vs. west. The variable is categorical with more than two possible response values.

Example 11.8

Do male and female college students have the same distribution of living arrangements? Use a level of significance of 0.05. Suppose that 250 randomly selected male college students and 300 randomly selected female college students were asked about their living arrangements: dormitory, apartment, with parents, other.

The results are shown in **Table 11.18**. Do male and female college students have the same distribution of living arrangements?

	Dormitory	Apartment	With Parents	Other
Males	72	84	49	45
Females	91	86	88	35

Table 11.19 Distribution of Living Arragements for College Males and College Females

Solution 11.8

H_0: The distribution of living arrangements for male college students is the same as the distribution of living arrangements for female college students.

H_a: The distribution of living arrangements for male college students is not the same as the distribution of living arrangements for female college students.

Degrees of freedom (*df*):
df = number of columns – 1 = 4 – 1 = 3

Distribution for the test: χ^2_3

Calculate the test statistic: χ^2 = 10.1287 (calculator or computer)

Probability statement: *p*-value = $P(\chi^2 > 10.1287)$ = 0.0175

Using the TI-83, 83+, 84, 84+ Calculator

Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 2 ENTER 4 ENTER. Enter the table values by row. Press ENTER after each. Press 2nd QUIT. Press STAT and arrow over to TESTS. Arrow down to C:χ2-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 10.1287 and the *p*-value = 0.0175. Do the procedure a second time but arrow down to Draw instead of Calculate.

Compare α and the *p*-value: Since no α is given, assume α = 0.05. *p*-value = 0.0175. α > *p*-value.

Make a decision: Since α > *p*-value, reject H_0. This means that the distributions are not the same.

Conclusion: At a 5 percent level of significance, from the data, there is sufficient evidence to conclude that the distributions of living arrangements for male and female college students are not the same.

Notice that the conclusion is only that the distributions are not the same. We cannot use the test for homogeneity to draw any conclusions about how they differ.

Try It

11.8 Do families and singles have the same distribution of cars? Suppose that 100 randomly selected families and 200

randomly selected singles were asked what type of car they drove: sport, sedan, hatchback, truck, van/SUV. The results are shown in **Table 11.20**. Do families and singles have the same distribution of cars? Test at a level of significance of 0.05.

	Sport	Sedan	Hatchback	Truck	Van/SUV
Family	5	15	35	17	28
Single	45	65	37	46	7

Table 11.20

Example 11.9

Both before and after a recent earthquake, surveys were conducted asking voters which of the three candidates they planned on voting for in the upcoming city council election. Has there been a change since the earthquake? Use a level of significance of 0.05. **Table 11.20** shows the results of the survey. Has there been a change in the distribution of voter preferences since the earthquake?

	Perez	Chung	Stevens
Before	167	128	135
After	214	197	225

Table 11.21

Solution 11.9

H_0: The distribution of voter preferences was the same before and after the earthquake.

H_a: The distribution of voter preferences was not the same before and after the earthquake.

Degrees of freedom (*df*):
df = number of columns – 1 = 3 – 1 = 2

Distribution for the test: χ^2_2

Calculate the test statistic: $\chi^2 = 3.2603$ (calculator or computer)

Probability statement: p-value=$P(\chi^2 > 3.2603) = 0.1959$

Using the TI-83, 83+, 84, 84+ Calculator

Press the MATRX key and arrow over to EDIT. Press 1:[A]. Press 2 ENTER 3 ENTER. Enter the table values by row. Press ENTER after each. Press 2nd QUIT. Press STAT and arrow over to TESTS. Arrow down to C:χ2-TEST. Press ENTER. You should see Observed:[A] and Expected:[B]. Arrow down to Calculate. Press ENTER. The test statistic is 3.2603 and the p-value = 0.1959. Do the procedure a second time but arrow down to Draw instead of Calculate.

Compare α and the p-value: $\alpha = 0.05$ and the p-value = 0.1959. $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, do not reject H_o.

Conclusion: At a 5 percent level of significance, from the data, there is insufficient evidence to conclude that the distribution of voter preferences was not the same before and after the earthquake.

Try It

11.9 Ivy League schools receive many applications, but only some can be accepted. At the schools listed in Table 11.22, two types of applications are accepted: regular and early decision.

Application Type Accepted	Brown	Columbia	Cornell	Dartmouth	Penn	Yale
Regular	2,115	1,792	5,306	1,734	2,685	1,245
Early Decision	577	627	1,228	444	1,195	761

Table 11.22

We want to know if the number of regular applications accepted follows the same distribution as the number of early applications accepted. State the null and alternative hypotheses, the degrees of freedom and the test statistic, sketch the graph of the p-value, and draw a conclusion about the test of homogeneity.

11.5 | Comparison of the Chi-Square Tests

You have seen the χ^2 test statistic used in three different circumstances. The following bulleted list is a summary that will help you decide which χ^2 test is the appropriate one to use.

- **Goodness-of-Fit:** Use the goodness-of-fit test to decide whether a population with an unknown distribution *fits* a known distribution. In this case there will be a single qualitative survey question or a single outcome of an experiment from a single population. Goodness-of-fit is typically used to see if the population is uniform (all outcomes occur with equal frequency), the population is normal, or the population is the same as another population with a known distribution. The null and alternative hypotheses are as follows:
  H_0: The population fits the given distribution.
  H_a: The population does not fit the given distribution.
- **Independence:** Use the test for independence to decide whether two variables (factors) are independent or dependent. In this case there will be two qualitative survey questions or experiments and a contingency table will be constructed. The goal is to see if the two variables are unrelated/independent or related/dependent. The null and alternative hypotheses are as follows:
  H_0: The two variables (factors) are independent.
  H_a: The two variables (factors) are dependent.
- **Homogeneity:** Use the test for homogeneity to decide if two populations with unknown distributions have the same distribution. In this case there will be a single qualitative survey question or experiment given to two different populations. The null and alternative hypotheses are as follows:
  H_0: The two populations follow the same distribution.
  H_a: The two populations have different distributions.

11.6 | Test of a Single Variance

A **test of a single variance** assumes that the underlying distribution is normal. The null and alternative hypotheses are stated in terms of the **population variance** or population standard deviation. The test statistic is

$$\frac{(n-1)s^2}{\sigma^2}$$

where

- n = the total number of data,
- s^2 = sample variance, and
- σ^2 = population variance.

You may think of s as the random variable in this test. The number of degrees of freedom is $df = n - 1$. A test of a single variance may be right-tailed, left-tailed, or two-tailed. **Example 11.10** will show you how to set up the null and alternative hypotheses. The null and alternative hypotheses contain statements about the population variance.

Example 11.10

Math instructors are not only interested in how their students do on exams, on average, but how the exam scores vary. To many instructors, the variance, or standard deviation, may be more important than the average.

Suppose a math instructor believes that the standard deviation for his final exam is five points. One of his best students thinks otherwise. The student claims that the standard deviation is more than five points. If the student were to conduct a hypothesis test, what would the null and alternative hypotheses be?

Solution 11.10

Even though we are given the population standard deviation, we can set up the test using the population variance as follows:

- H_0: $\sigma^2 = 5^2$
- H_a: $\sigma^2 > 5^2$

Try It

11.10 A scuba instructor wants to record the collective depths each of his students dives during their checkout. He is interested in how the depths vary, even though everyone should have been at the same depth. He believes the standard deviation is three feet. His assistant thinks the standard deviation is less than three feet. If the instructor were to conduct a test, what would the null and alternative hypotheses be?

Example 11.11

With individual lines at its various windows, a post office finds that the standard deviation for normally distributed waiting times for customers on Friday afternoon is 7.2 minutes. The post office experiments with a single, main waiting line and finds that for a random sample of 25 customers, the waiting times for customers have a standard deviation of 3.5 minutes.

With a significance level of 5 percent, test the claim that a single line causes lower variation among waiting times (shorter waiting times) for customers.

Solution 11.11

Since the claim is that a single line causes less variation, this is a test of a single variance. The parameter is the population variance, σ^2, or the population standard deviation, σ.

Random variable: The sample standard deviation, s, is the random variable. Let s = standard deviation for the waiting times.

- H_0: $\sigma^2 = 7.2^2$
- H_a: $\sigma^2 < 7.2^2$

The word *less* tells you this is a left-tailed test.

Distribution for the test: χ^2_{24}, where

- n = the number of customers sampled, and
- $df = n - 1 = 25 - 1 = 24$.

Calculate the test statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(3.5)^2}{7.2^2} = 5.67$$

where $n = 25$, $s = 3.5$, and $\sigma = 7.2$.

Graph

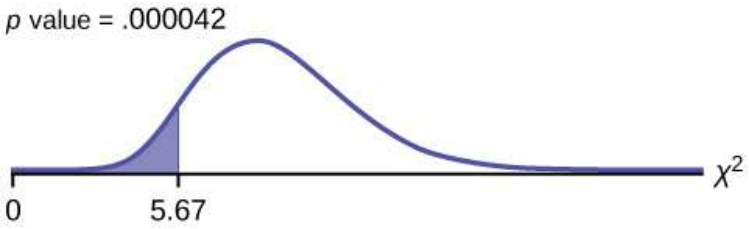


Figure 11.8

Probability statement: p-value = $P(\chi^2 < 5.67) = 0.000042$

Compare α and the p-value:
$\alpha = 0.05$
p-value = 0.000042
$\alpha > p$-value

Make a decision: Since $\alpha > p$-value, reject H_0. This means that you reject $\sigma^2 = 7.2^2$. In other words, you do not think the variation in waiting times is 7.2 minutes; you think the variation in waiting times is less.

Conclusion: At a 5 percent level of significance, from the data, there is sufficient evidence to conclude that a single line causes a lower variation among the waiting times *or* with a single line, the customer waiting times vary less than 7.2 minutes.

Using the TI-83, 83+, 84, 84+ Calculator

In `2nd DISTR`, use `7:χ2cdf`. The syntax is `(lower, upper, df)` for the parameter list. For **Example 11.11**, `χ2cdf(-1E99,5.67,24)`. The p-value = 0.000042.

Try It

11.11 The FCC conducts broadband speed tests to measure how much data per second passes between a consumer's computer and the internet. As of August 2012, the standard deviation of internet speeds across internet service providers (ISPs) was 12.2 percent. Suppose a sample of 15 ISPs is taken, and the standard deviation is 13.2. An analyst

claims that the standard deviation of speeds is more than what was reported. State the null and alternative hypotheses, compute the degrees of freedom, calculate the test statistic, sketch the graph of the *p*-value, and draw a conclusion. Test at the 1 percent significance level.

11.7 | Lab 1: Chi-Square Goodness-of-Fit

Stats Lab

11.1 Lab 1: Chi-Square Goodness-of-Fit

Student Learning Outcome

- The student will evaluate data collected to determine if they fit either the uniform or exponential distributions.

Collect the Data

Go to your local supermarket. Ask 30 people as they leave for the total amount on their grocery receipts. Or, ask 3 cashiers for the last 10 amounts. Be sure to include the express lane, if it is open.

NOTE

You may need to combine two categories so that each cell has an expected value of at least five.

1. Record the values.

Table 11.23

2. Construct a histogram of the data. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.

Figure 11.9

3. Calculate the following:
   a. $\bar{x}$ = _____
   b. s = _____
   c. s^2 = _____

Uniform Distribution

Test to see if grocery receipts follow the uniform distribution.

1. Using your lowest and highest values, $X \sim U$ (_____, _____).
2. Divide the distribution into fifths.
3. Calculate the following:
   a. lowest value = _____
   b. 20^{th} percentile = _____
   c. 40^{th} percentile = _____
   d. 60^{th} percentile = _____
   e. 80^{th} percentile = _____
   f. highest value = _____
4. For each fifth, count the observed number of receipts and record it. Then determine the expected number of receipts and record that.

Fifth	Observed	Expected
1^{st}		
2^{nd}		
3^{rd}		
4^{th}		
5^{th}		

Table 11.24

5. H_0: _____
6. H_a: _____
7. What distribution should you use for a hypothesis test?
8. Why did you choose this distribution?
9. Calculate the test statistic.
10. Find the *p*-value.
11. Sketch a graph of the situation. Label and scale the *x*-axis. Shade the area corresponding to the *p*-value.

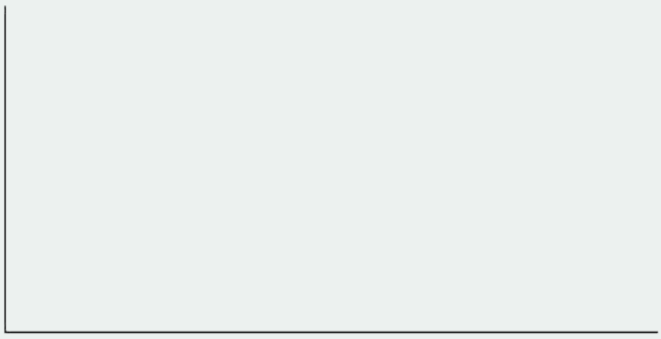


Figure 11.10

12. State your decision.
13. State your conclusion in a complete sentence.

Exponential Distribution

Test to see if grocery receipts follow the exponential distribution with decay parameter $\frac{1}{\bar{x}}$.

1. Using $\frac{1}{\bar{x}}$ as the decay parameter, $X \sim Exp$(_____).
2. Calculate the following:
   a. lowest value = _____
   b. first quartile = _____
   c. 37th percentile = _____
   d. median = _____
   e. 63rd percentile = _____
   f. 3rd quartile = _____
   g. highest value = _____
3. For each cell, count the observed number of receipts and record it. Then determine the expected number of receipts and record that.

Cell	Observed	Expected
1st		
2nd		
3rd		
4th		
5th		
6th		

Table 11.25

4. H_0: _____

5. H_a: _____
6. What distribution should you use for a hypothesis test?
7. Why did you choose this distribution?
8. Calculate the test statistic.
9. Find the *p*-value.
10. Sketch a graph of the situation. Label and scale the *x*-axis. Shade the area corresponding to the *p*-value.

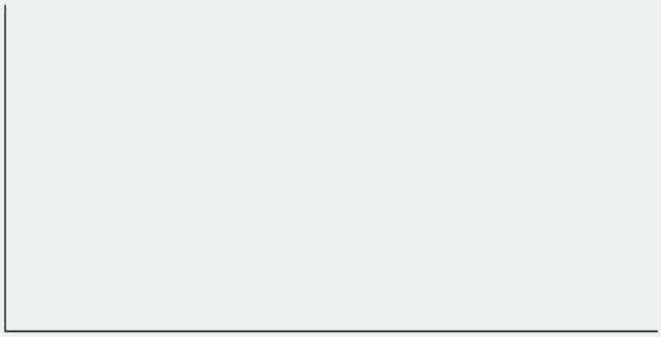


Figure 11.11

11. State your decision.
12. State your conclusion in a complete sentence.

Discussion Questions

1. Did your data fit either distribution? If so, which?
2. In general, do you think it's likely that data could fit more than one distribution? In complete sentences, explain why or why not.

11.8 | Lab 2: Chi-Square Test of Independence

Stats Lab

11.2 Lab 2: Chi-Square Test of Independence

Student Learning Outcome

- The student will evaluate if there is a significant relationship between favorite type of snack and gender.

Collect the Data

1. Using your class as a sample, complete the following chart. Ask one another what your favorite snack is, then total the results.

NOTE

You may need to combine two food categories so that each cell has an expected value of at least five.

	Sweets (candy & baked goods)	Ice Cream	Chips & Pretzels	Fruits & Vegetables	Total
Male					
Female					
Total					

Table 11.26 Favorite Type of Snack

2. Looking at **Table 11.26**, does it appear to you that there is a dependence between gender and favorite type of snack food? Why or why not?

Hypothesis Test

Conduct a hypothesis test to determine if the factors are independent:

1. H_0: _____
2. H_a: _____
3. What distribution should you use for a hypothesis test?
4. Why did you choose this distribution?
5. Calculate the test statistic.
6. Find the p value.
7. Sketch a graph of the situation. Label and scale the x axis. Shade the area corresponding to the p value.

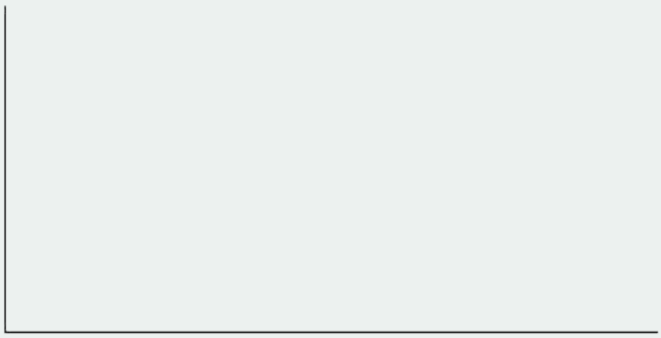


Figure 11.12

8. State your decision.
9. State your conclusion in a complete sentence.

Discussion Questions

1. Is the conclusion of your study the same as or different from your answer to answer to Question 2 under **Collect the Data**?
2. Why do you think that occurred?

KEY TERMS

contingency table a table that displays sample values for two different factors that may be dependent or contingent on each other; facilitates determining conditional probabilities

CHAPTER REVIEW

11.1 Facts About the Chi-Square Distribution

The chi-square distribution is a useful tool for assessment in a series of problem categories. These problem categories include primarily (i) whether a data set fits a particular distribution, (ii) whether the distributions of two populations are the same, (iii) whether two events might be independent, and (iv) whether there is a different variability than expected within a population.

An important parameter in a chi-square distribution is the degrees of freedom *df* in a given problem. The random variable in the chi-square distribution is the sum of squares of *df* standard normal variables, which must be independent. The key characteristics of the chi-square distribution also depend directly on the degrees of freedom.

The chi-square distribution curve is skewed to the right, and its shape depends on the degrees of freedom *df*. For $df > 90$, the curve approximates the normal distribution. Test statistics based on the chi-square distribution are always greater than or equal to zero. Such application tests are almost always right-tailed tests.

11.2 Goodness-of-Fit Test

To assess whether a data set fits a specific distribution, you can apply the goodness-of-fit hypothesis test that uses the chi-square distribution. The null hypothesis for this test states that the data come from the assumed distribution. The test compares observed values against the values you would expect to have if your data followed the assumed distribution. The test is almost always right-tailed. Each observation or cell category must have an expected value of at least five.

11.3 Test of Independence

To assess whether two factors are independent, you can apply the test of independence that uses the chi-square distribution. The null hypothesis for this test states that the two factors are independent. The test compares observed values to expected values. The test is right-tailed. Each observation or cell category must have an expected value of at least five.

11.4 Test for Homogeneity

To assess whether two data sets are derived from the same distribution, which need not be known, you can apply the test for homogeneity that uses the chi-square distribution. The null hypothesis for this test states that the populations of the two data sets come from the same distribution. The test compares the observed values against the expected values if the two populations followed the same distribution. The test is right-tailed. Each observation or cell category must have an expected value of at least five.

11.5 Comparison of the Chi-Square Tests

The goodness-of-fit test is typically used to determine if data fits a particular distribution. The test of independence makes use of a contingency table to determine the independence of two factors. The test for homogeneity determines whether two populations come from the same distribution, even if this distribution is unknown.

11.6 Test of a Single Variance

To test variability, use the chi-square test of a single variance. The test may be left-, right-, or two-tailed, and its hypotheses are always expressed in terms of the variance or standard deviation.

FORMULA REVIEW

11.1 Facts About the Chi-Square Distribution

$\chi^2 = (Z_1)^2 + (Z_2)^2 + \ldots (Z_{df})^2$ chi-square distribution random variable

$\mu_{\chi^2} = df$ chi-square distribution population mean

$\sigma_{\chi^2} = \sqrt{2(df)}$ chi-square distribution population standard deviation

11.2 Goodness-of-Fit Test

$\sum_{k} \frac{(O-E)^2}{E}$ goodness-of-fit test statistic where

O: observed values
E: expected values

k: number of different data cells or categories

$df = k - 1$ degrees of freedom

11.3 Test of Independence

Test of Independence

- The number of degrees of freedom is equal to (number of columns–1)(number of rows–1).
- The test statistic is $\sum_{(i \cdot j)} \frac{(O-E)^2}{E}$ where O = observed values, E = expected values, i = the number of rows in the table, and j = the number of columns in the table.
- If the null hypothesis is true, the expected number $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}}$.

11.4 Test for Homogeneity

$\sum_{i \cdot j} \frac{(O-E)^2}{E}$ Homogeneity test statistic where O = observed values
E = expected values
i = number of rows in data contingency table
j = number of columns in data contingency table

$df = (i - 1)(j - 1)$ degrees of freedom

11.6 Test of a Single Variance

$\chi^2 = \frac{(n-1) \cdot s^2}{\sigma^2}$ Test of a single variance statistic where
n: sample size
s: sample standard deviation
σ: population standard deviation

$df = n - 1$ degrees of freedom

Test of a Single Variance

- Use the test to determine variation.
- The degrees of freedom is the number of samples – 1.
- The test statistic is $\frac{(n-1) \cdot s^2}{\sigma^2}$, where n = the total number of data, s^2 = sample variance, and σ^2 = population variance.
- The test may be left-, right-, or two-tailed.

PRACTICE

11.1 Facts About the Chi-Square Distribution

1. If the number of degrees of freedom for a chi-square distribution is 25, what is the population mean and standard deviation?

2. If $df > 90$, the distribution is _____. If $df = 15$, the distribution is _____.

3. When does the chi-square curve approximate a normal distribution?

4. Where is μ located on a chi-square curve?

5. Is it more likely the *df* is 90, 20, or 2 in the graph?

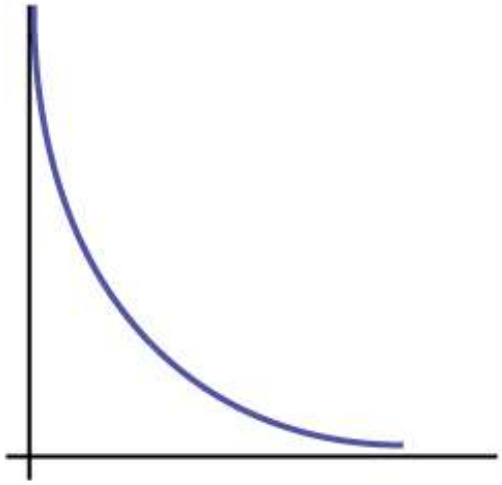


Figure 11.13

11.2 Goodness-of-Fit Test

Determine the appropriate test to be used in the next three exercises.

6. An archeologist is calculating the distribution of the frequency of the number of artifacts she finds in a dig site. Based on previous digs, the archeologist creates an expected distribution broken down by grid sections in the dig site. Once the site has been fully excavated, she compares the actual number of artifacts found in each grid section to see if her expectation was accurate.

7. An economist is deriving a model to predict outcomes on the stock market. He creates a list of expected points on the stock market index for the next two weeks. At the close of each day's trading, he records the actual points on the index. He wants to see how well his model matched what actually happened.

8. A personal trainer is putting together a weight-lifting program for her clients. For a 90-day program, she expects each client to lift a specific maximum weight each week. As she goes along, she records the actual maximum weights her clients lifted. She wants to know how well her expectations met with what was observed.

Use the following information to answer the next five exercises. A teacher predicts the distribution of grades on the final exam. The predictions are shown in **Table 11.27**.

Grade	Proportion
A	0.25
B	0.30
C	0.35
D	0.10

Table 11.27

The actual distribution for a class of 20 is in **Table 11.28**.

Grade	Frequency
A	7

Table 11.28

Grade	Frequency
B	7
C	5
D	1

Table 11.28

9. $df =$ _____

10. State the null and alternative hypotheses.

11. χ^2 test statistic = _____

12. p-value = _____

13. At the 5 percent significance level, what can you conclude?

Use the following information to answer the next nine exercises. The cumulative number of cases of a chronic disease reported for Santa Clara County is broken down by ethnicity as in **Table 11.29**.

Ethnicity	Number of Cases
White	2,229
Hispanic	1,157
Black/African American	457
Asian, Pacific Islander	232
	Total = 4,075

Table 11.29

The percentage of each ethnic group in Santa Clara County is as in **Table 11.30**.

Ethnicity	% of Total County Population	Number Expected (round to two decimal places)
White	42.9%	1,748.18
Hispanic	26.7%	
Black/African American	2.6%	
Asian, Pacific Islander	27.8%	
	Total = 100%	

Table 11.30

14. If the ethnicities of patients followed the ethnicities of the total county population, fill in the expected number of cases per ethnic group.
Perform a goodness-of-fit test to determine whether the occurrence of disease cases follows the ethnicities of the general population of Santa Clara County.

15. H_0: _____

16. H_a: _____

17. Is this a right-tailed, left-tailed, or two-tailed test?

18. degrees of freedom = _____

19. χ^2 test statistic = _____

20. p-value = _____

21. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

Figure 11.14

Let $\alpha = 0.05$.

Decision: _____

Reason for the decision: _____

Conclusion (write out in complete sentences): _____

22. Does it appear that the pattern of disease cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

11.3 Test of Independence

Determine the appropriate test to be used in the next three exercises.

23. A pharmaceutical company is interested in the relationship between age and presentation of symptoms for a common viral infection. A random sample is taken of 500 people with the infection across different age groups.

24. The owner of a baseball team is interested in the relationship between player salaries and team winning percentage. He takes a random sample of 100 players from different organizations.

25. A marathon runner is interested in the relationship between the brand of shoes runners wear and their run times. She takes a random sample of 50 runners and records their run times and the brand of shoes they were wearing.

Use the following information to answer the next seven exercises: Transit Railroads is interested in the relationship between travel distance and the ticket class purchased. A random sample of 200 passengers is taken. **Table 11.31** shows the results. The railroad wants to know if a passenger's choice in ticket class is independent of the distance the passenger must travel.

Traveling Distance	Third Class	Second Class	First Class	Total
1–100 miles	21	14	6	41
101–200 miles	18	16	8	42
201–300 miles	16	17	15	48

Table 11.31

Traveling Distance	Third Class	Second Class	First Class	Total
301–400 miles	12	14	21	47
401–500 miles	6	6	10	22
Total	73	67	60	200

Table 11.31

26. State the hypotheses.
H_0: _____
H_a: _____

27. df = _____

28. How many passengers are expected to travel between 201 and 300 miles and purchase second-class tickets?

29. How many passengers are expected to travel between 401 and 500 miles and purchase first-class tickets?

30. What is the test statistic?

31. What is the p-value?

32. What can you conclude at the 5 percent level of significance?

Use the following information to answer the next ten exercises. An article in the *New England Journal of Medicine* discussed a study on people who used a certain product in California and Hawaii. In one part of the report, the self-reported ethnicity and product-use levels per day were given. Of the people using the product at most 10 times per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 whites. Of the people using the product 11 to 20 times per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 whites. Of the people using the product 21 to 30 times per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese Americans, and 6,062 whites. Of the people using the product at least 31 times per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 whites.

33. Complete the table.

Product use Per Day	African American	Native Hawaiian	Latino	Japanese American	White	TOTALS
1–10						
11–20						
21–30						
31+						
TOTALS						

Table 11.32

34. State the hypotheses.
H_0: _____
H_a: _____

35. Enter expected values in **Table 11.32**. Round to two decimal places.

Calculate the following values.

36. df = _____

37. χ^2 test statistic = _____

38. p-value = _____

39. Is this a right-tailed, left-tailed, or two-tailed test? Explain why.

40. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the *p*-value.

Figure 11.15

State the decision and conclusion (in a complete sentence) for the following levels of α.

41. $\alpha = 0.05$

a. Decision: _____
b. Reason for the decision: _____
c. Conclusion (write out in a complete sentence): _____

42. $\alpha = 0.01$

a. Decision: _____
b. Reason for the decision: _____
c. Conclusion (write out in a complete sentence): _____

11.4 Test for Homogeneity

43. A math teacher wants to see if two of her classes have the same distribution of test scores. What test should she use?

44. What are the null and alternative hypotheses for **Exercise 11.43**?

45. A market researcher wants to see if two different stores have the same distribution of sales throughout the year. What type of test should he use?

46. A meteorologist wants to know if East and West Australia have the same distribution of storms. What type of test should she use?

47. What condition must be met to use the test for homogeneity?

Use the following information to answer the next five exercises. Do private practice doctors and hospital doctors have the same distribution of working hours? Suppose that a sample of 100 private practice doctors and 150 hospital doctors are selected at random and asked about the number of hours a week they work. The results are shown in **Table 11.33**.

	20–30	30–40	40–50	50–60
Private Practice	16	40	38	6
Hospital	8	44	59	39

Table 11.33

48. State the null and alternative hypotheses.

49. *df* = _____

50. What is the test statistic?

51. What is the *p*-value?

52. What can you conclude at the 5 percent significance level?

11.5 Comparison of the Chi-Square Tests

53. Which test do you use to decide whether an observed distribution is the same as an expected distribution?

54. What is the null hypothesis for the type of test from **Exercise 11.53**?

55. Which test would you use to decide whether two factors have a relationship?

56. Which test would you use to decide if two populations have the same distribution?

57. How are tests of independence similar to tests for homogeneity?

58. How are tests of independence different from tests for homogeneity?

11.6 Test of a Single Variance

Use the following information to answer the next three exercises. An archer's standard deviation for his hits is six, where the data are measured in distance from the center of the target. An observer claims the standard deviation is less than six.

59. What type of test should be used?

60. State the null and alternative hypotheses.

61. Is this a right-tailed, left-tailed, or two-tailed test?

Use the following information to answer the next three exercises. The standard deviation of heights for students in a school is 0.81. A random sample of 50 students is taken, and the standard deviation of heights of the sample is 0.96. A researcher in charge of the study believes the standard deviation of heights for the school is greater than 0.81.

62. What type of test should be used?

63. State the null and alternative hypotheses.

64. $df =$ _____

Use the following information to answer the next four exercises: The average waiting time in a doctor's office varies. The standard deviation of waiting times in a doctor's office is 3.4 minutes. A random sample of 30 patients in the doctor's office has a standard deviation of waiting times of 4.1 minutes. One doctor believes the variance of waiting times is greater than originally thought.

65. What type of test should be used?

66. What is the test statistic?

67. What is the *p*-value?

68. What can you conclude at the 5 percent significance level?

HOMEWORK

11.1 Facts About the Chi-Square Distribution

Decide whether the following statements are true or false.

69. As the number of degrees of freedom increases, the graph of the chi-square distribution looks more and more symmetrical.

70. The standard deviation of the chi-square distribution is twice the mean.

71. The mean and the median of the chi-square distribution are the same if $df = 24$.

11.2 Goodness-of-Fit Test

For each problem, use a solution sheet to solve the hypothesis test problem. Go to ***Appendix E*** *for the chi-square solution sheet. Round expected frequency to two decimal places.*

72. A six-sided die is rolled 120 times. Fill in the expected frequency column. Then, conduct a hypothesis test to determine if the die is fair. The data in **Table 11.34** are the result of the 120 rolls.

Face Value	Frequency	Expected Frequency
1	15	
2	29	
3	16	
4	15	
5	30	
6	15	

Table 11.34

73. The marital status distribution of the U.S. male population, ages 15 and older, is as shown in **Table 11.35**.

Marital Status	%	Expected Frequency
Never Married	31.3%	
Married	56.1%	
Widowed	2.5%	
Divorced/Separated	10.1%	

Table 11.35

Suppose that a random sample of 400 U.S. males, 18 to 24 years old, yielded the following frequency distribution. We are interested in whether this age group of males fits the distribution of the U.S. adult population. Calculate the frequency one would expect when surveying 400 people. Fill in **Table 11.35**, rounding to two decimal places.

Marital Status	Frequency
Never Married	140
Married	238
Widowed	2
Divorced/Separated	20

Table 11.36

Use the following information to answer the next two exercises. The columns in **Table 11.37** contain the Race/Ethnicity of U.S. Public Schools for a recent year, the percentages for the Advanced Placement Examinee Population for that class, and the Overall Student Population. Suppose the right column contains the results of a survey of 1,000 local students from that year who took an AP exam.

Race/Ethnicity	AP Examinee Population	Overall Student Population	Survey Frequency
Asian, Asian American, or Pacific Islander	10.2%	5.4%	113
Black or African American	8.2%	14.5%	94
Hispanic or Latino	15.5%	15.9%	136
American Indian or Alaska Native	0.6%	1.2%	10
White	59.4%	61.6%	604
Not Reported/Other	6.1%	1.4%	43

Table 11.37

74. Perform a goodness-of-fit test to determine whether the local results follow the distribution of the U.S. overall student population based on ethnicity.

75. Perform a goodness-of-fit test to determine whether the local results follow the distribution of U.S. AP examinee population, based on ethnicity.

76. The city of South Lake Tahoe, California, has an Asian population of 1,419 out of a total population of 23,609. Suppose that a survey of 1,419 self-reported Asians in the borough of Manhattan in the New York City area yielded the data in **Table 11.38**. Conduct a goodness-of-fit test to determine if the self-reported subgroups of Asians in Manhattan fit that of the South Lake Tahoe area.

Race	South Lake Tahoe Frequency	Manhattan Frequency
Asian Indian	131	174
Chinese	118	557
Filipino	1,045	518
Japanese	80	54
Korean	12	29
Vietnamese	9	21
Other	24	66

Table 11.38

Use the following information to answer the next two exercises. UCLA conducted a survey of more than 263,000 college freshmen from 385 colleges in fall 2005. The results of students' expected majors by gender were reported in *The Chronicle of Higher Education (2/2/2006)*. Suppose a survey of 5,000 graduating females and 5,000 graduating males was done as a follow-up last year to determine what their actual majors were. The results are shown in the tables for **Exercise 11.77** and **Exercise 11.78**. The second column in each table does not add to 100 percent because of rounding.

77. Conduct a goodness-of-fit test to determine if the actual college majors of graduating females fit the distribution of their expected majors.

Major	Females—Expected Major	Females—Actual Major
Arts & Humanities	14%	670
Biological Sciences	8.4%	410
Business	13.1%	685
Education	13%	650
Engineering	2.6%	145
Physical Sciences	2.6%	125
Professional	18.9%	975
Social Sciences	13%	605
Technical	0.4%	15
Other	5.8%	300
Undecided	8%	420

Table 11.39

78. Conduct a goodness-of-fit test to determine if the actual college majors of graduating males fit the distribution of their expected majors.

Major	Males—Expected Major	Males—Actual Major
Arts & Humanities	11%	600
Biological Sciences	6.7%	330
Business	22.7%	1,130
Education	5.8%	305
Engineering	15.6%	800
Physical Sciences	3.6%	175
Professional	9.3%	460
Social Sciences	7.6%	370
Technical	1.8%	90
Other	8.2%	400
Undecided	6.6%	340

Table 11.40

Read the statement and decide whether it is true or false.

79. In a goodness-of-fit test, the expected values are the values we would expect if the null hypothesis were true.

80. In general, if the observed values and expected values of a goodness-of-fit test are not close together, then the test statistic can get very large and on a graph will be way out in the right tail.

81. Use a goodness-of-fit test to determine if high school principals believe that students are absent equally during the week.

82. The test to use to determine if a six-sided die is fair is a goodness-of-fit test.

83. In a goodness-of-fit test, if the *p*-value is 0.0113, in general, do not reject the null hypothesis.

84. A sample of 212 commercial businesses was surveyed for recycling one commodity; a commodity here means any one type of recyclable material such as plastic or aluminum. **Table 11.41** shows the business categories in the survey, the sample size of each category, and the number of businesses in each category that recycle one commodity. Based on the study, on average half of the businesses were expected to be recycling one commodity. As a result, the last column shows the expected number of businesses in each category that recycle one commodity. At the 5 percent significance level, perform a hypothesis test to determine if the observed number of businesses that recycle one commodity follows the uniform distribution of the expected values.

Business Type	Number in Class	Observed Number that Recycle One Commodity	Expected Number that Recycle One Commodity
Office	35	19	17.5
Retail/ Wholesale	48	27	24
Food/ Restaurants	53	35	26.5
Manufacturing/ Medical	52	21	26
Hotel/Mixed	24	9	12

Table 11.41

85. **Table 11.42** contains information from a survey of 499 participants classified according to their age groups. The researchers making the survey wanted to find out how many people were diagnosed with a particular disease within the last year. The second column shows the percentage of people with the disease per age class among the study participants. The last column comes from a different study at the national level that shows the corresponding percentages of people with the disease in the same age classes in the United States. Perform a hypothesis test at the 5 percent significance level to determine whether the survey participants are a representative sample of the people with the disease nationwide.

Age Class (years)	% of People Diagnosed	% of Expected U.S. Average
20–30	75%	32.6%
31–40	26.5%	32.6%
41–50	13.6%	36.6%
51–60	21.9%	36.6%
61–70	21%	39.7%

Table 11.42

11.3 Test of Independence

*For each problem, use a solution sheet to solve the hypothesis test problem. Go to **Appendix E** for the chi-square solution sheet. Round expected frequency to two decimal places.*

86. A recent debate about where in the U.S. skiers believe the skiing is best prompted the following survey. Test to see if the best ski area is independent of the level of the skier.

U.S. Ski Area	Beginner	Intermediate	Advanced
Tahoe	20	30	40
Utah	10	30	60
Colorado	10	40	50

Table 11.43

87. Car manufacturers are interested in whether there is a relationship between the size of car an individual drives and the number of people in the driver's family—that is, whether car size and family size are independent. To test this, suppose that 800 car owners were randomly surveyed with the results in **Table 11.44**. Conduct a test of independence.

Family Size	Sub & Compact	Mid-Size	Full-Size	Van & Truck
1	20	35	40	35
2	20	50	70	80
3–4	20	50	100	90
5+	20	30	70	70

Table 11.44

88. College students may be interested in whether their majors have any effect on starting salaries after graduation. Suppose that 300 recent graduates were surveyed as to their majors in college and their starting salaries after graduation. **Table 11.45** shows the data. Conduct a test of independence.

Major	< $50,000	$50,000–$68,999	$69,000 +
English	5	20	5
Engineering	10	30	60
Nursing	10	15	15
Business	10	20	30
Psychology	20	30	20

Table 11.45

89. Some travel agents claim that honeymoon hotspots vary according to age of the bride. Suppose that 280 recent brides were interviewed as to where they spent their honeymoons. The information is given in **Table 11.46**. Conduct a test of independence.

Location	20–29	30–39	40–49	50+
Niagara Falls	15	25	25	20
Poconos	15	25	25	10
Europe	10	25	15	5
Virgin Islands	20	25	15	5

Table 11.46

90. A manager of a sports club keeps information concerning the main sport in which members participate and their ages. To test whether there is a relationship between the age of a member and his or her choice of sport, 643 members of the sports club are randomly selected. Conduct a test of independence.

Sport	18–25	26–30	31–40	41+
Racquetball	42	58	30	46
Tennis	58	76	38	65
Swimming	72	60	65	33

Table 11.47

91. A major food manufacturer is concerned that the sales for its skinny french fries have been decreasing. As a part of a feasibility study, the company conducts research into the types of fries sold across the country to determine if the type of fries sold is independent of the area of the country. The results of the study are shown in **Table 11.48**. Conduct a test of independence.

Type of Fries	Northeast	South	Central	West
Skinny Fries	70	50	20	25
Curly Fries	100	60	15	30
Steak Fries	20	40	10	10

Table 11.48

92. According to Dan Leonard, an independent insurance agent in the Buffalo, New York area, the following is a breakdown of the amount of life insurance purchased by males in the following age groups. He is interested in whether the age of the male and the amount of life insurance purchased are independent events. Conduct a test for independence.

Age of Males	None	< $200,000	$200,000–$400,000	$401,001–$1,000,000	$1,000,001+
20–29	40	15	40	0	5
30–39	35	5	20	20	10
40–49	20	0	30	0	30
50+	40	30	15	15	10

Table 11.49

93. Suppose that 600 thirty-year-olds were surveyed to determine whether there is a relationship between the level of education an individual has and salary. Conduct a test of independence.

Annual Salary	Not a High School Graduate	High School Graduate	College Graduate	Masters or Doctorate
< $30,000	15	25	10	5
$30,000–$40,000	20	40	70	30
$40,000–$50,000	10	20	40	55
$50,000–$60,000	5	10	20	60
$60,000+	0	5	10	150

Table 11.50

Read the statement and decide whether it is true or false.

94. The number of degrees of freedom for a test of independence is equal to the sample size minus one.

95. The test for independence uses tables of observed and expected data values.

96. The test to use when determining if the college or university a student chooses to attend is related to his or her socioeconomic status is a test for independence.

97. In a test of independence, the expected number is equal to the row total multiplied by the column total divided by the total surveyed.

98. An ice cream maker performs a nationwide survey about favorite flavors of ice cream in different geographic areas of the United States. Based on **Table 11.51**, do the numbers suggest that geographic location is independent of favorite ice cream flavors? Test at the 5 percent significance level.

U.S. Region/ Flavor	Strawberry	Chocolate	Vanilla	Rocky Road	Mint Chocolate Chip	Pistachio	Row Total
West	12	21	22	19	15	8	97
Midwest	10	32	22	11	15	6	96
East	8	31	27	8	15	7	96
South	15	28	30	8	15	6	102
Column Total	45	112	101	46	60	27	391

Table 11.51

99. **Table 11.52** provides results of a recent survey of the youngest online entrepreneurs whose net worth is estimated at one million dollars or more. Their ages range from 17 to 30. Each cell in the table illustrates the number of entrepreneurs who correspond to the specific age group and their net worth. Are the ages and net worth independent? Perform a test of independence at the 5 percent significance level.

Age Group/Net Worth Value (in millions of U.S. dollars)	1–5	6–24	≥25	Row Total
17–25	8	7	5	20
26–30	6	5	9	20
Column Total	14	12	14	40

Table 11.52

100. A 2013 poll in California surveyed people about a new tax. The results are presented in **Table 11.53** and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a test of independence at the 5 percent significance level.

Opinion/ Ethnicity	Asian American	White/Non-Hispanic	African American	Latino	Row Total
Against Tax	48	433	41	160	682
In Favor of Tax	54	234	24	147	459
No Opinion	16	43	16	19	94
Column Total	118	710	81	326	1,235

Table 11.53

11.4 Test for Homogeneity

For each word problem, use a solution sheet to solve the hypothesis test problem. Go to ***Appendix E*** *for the chi-square solution sheet. Round expected frequency to two decimal places.*

101. A psychologist is interested in testing whether there is a difference in the distribution of personality types for business majors and social science majors. The results of the study are shown in **Table 11.54**. Conduct a test of homogeneity. Test at a 5 percent level of significance.

	Open	Conscientious	Extrovert	Agreeable	Neurotic
Business	41	52	46	61	58
Social Science	72	75	63	80	65

Table 11.54

102. Do men and women select different breakfasts? The breakfasts ordered by randomly selected men and women at a popular breakfast place are shown in **Table 11.55**. Conduct a test for homogeneity at a 5 percent level of significance.

	French Toast	Pancakes	Waffles	Omelettes
Men	47	35	28	53
Women	65	59	55	60

Table 11.55

103. A fisherman is interested in whether the distribution of fish caught in Green Valley Lake is the same as the distribution of fish caught in Echo Lake. Of the 191 randomly selected fish caught in Green Valley Lake, 105 were rainbow trout, 27 were other trout, 35 were bass, and 24 were catfish. Of the 293 randomly selected fish caught in Echo Lake, 115 were rainbow trout, 58 were other trout, 67 were bass, and 53 were catfish. Perform a test for homogeneity at a 5 percent level of significance.

104. In 2007, the United States had 1.5 million homeschooled students, according to the U.S. National Center for Education Statistics. In **Table 11.56**, you can see that parents decide to homeschool their children for different reasons, and some reasons are ranked by parents as more important than others. According to the survey results shown in the table, is the distribution of applicable reasons the same as the distribution of the most important reason? Provide your assessment at the 5 percent significance level. Did you expect the result you obtained?

Reasons for Homeschooling	Applicable Reason (in thousands of respondents)	Most Important Reason (in thousands of respondents)	Row Total
Concern About the Environment of Other Schools	1,321	309	1,630
Dissatisfaction with Academic Instruction at Other Schools	1,096	258	1,354
To Provide Religious or Moral Instruction	1,257	540	1,797
Child Has Special Needs, Other Than Physical or Mental	315	55	370
Nontraditional Approach to Child's Education	984	99	1,083
Other Reasons (e.g., finances, travel, family time, etc.)	485	216	701
Column Total	5,458	1,477	6,935

Table 11.56

105. When looking at energy consumption, we are often interested in detecting trends over time and how they correlate among different countries. The information in **Table 11.57** shows the average energy use in units of kg of oil equivalent per capita in the United States and the joint European Union countries (EU) for the six-year period 2005 to 2010. Do the energy use values in these two areas come from the same distribution? Perform the analysis at the 5 percent significance level.

Year	European Union	United States	Row Total
2010	3,413	7,164	10,557
2009	3,302	7,057	10,359
2008	3,505	7,488	10,993
2007	3,537	7,758	11,295
2006	3,595	7,697	11,292
2005	3,613	7,847	11,460
Column Total	20,965	45,011	65,976

Table 11.57

106. The Insurance Institute for Highway Safety collects safety information about all types of cars every year and publishes a report of top safety picks among all cars, makes, and models. **Table 11.58** presents the number of top safety picks in six car categories for the two years 2009 and 2013. Analyze the table data to conclude whether the distribution of cars that earned the top safety picks safety award has remained the same between 2009 and 2013. Derive your results at the 5 percent significance level.

Year/Car Type	Small	Mid-Size	Large	Small SUV	Mid-Size SUV	Large SUV	Row Total
2009	12	22	10	10	27	6	87
2013	31	30	19	11	29	4	124
Column Total	43	52	29	21	56	10	211

Table 11.58

11.5 Comparison of the Chi-Square Tests

*For each word problem, use a solution sheet to solve the hypothesis test problem. Go to **Appendix E** for the chi-square solution sheet. Round expected frequency to two decimal places.*

107. Is there a difference between the distribution of community college statistics students and the distribution of university statistics students in what technology they use on their homework? Of some randomly selected community college students, 43 used a computer, 102 used a calculator with built-in statistics functions, and 65 used a table from the textbook. Of some randomly selected university students, 28 used a computer, 33 used a calculator with built-in statistics functions, and 40 used a table from the textbook. Conduct an appropriate hypothesis test using a 0.05 level of significance.

Read the statement and decide whether it is true or false.

108. If $df = 2$, the chi-square distribution has a shape that reminds us of the exponential.

11.6 Test of a Single Variance

Use the following information to answer the next 12 exercises. Suppose an airline claims that its flights are consistently on time with an average delay of at most 15 minutes. It claims that the average delay is so consistent that the variance is no more than 150 minutes. Doubting the consistency part of the claim, a disgruntled traveler calculates the delays for his next 25 flights. The average delay for those 25 flights is 22 minutes with a standard deviation of 15 minutes.

109. Is the traveler disputing the claim about the average or about the variance?

110. A sample standard deviation of 15 minutes is the same as a sample variance of _____ minutes.

111. Is this a right-tailed, left-tailed, or two-tailed test?

112. H_0: _____

113. df = _____

114. chi-square test statistic = _____

115. p-value = _____

116. Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade the p-value.

117. Let $\alpha = 0.05$
Decision: _____
Conclusion (write out in a complete sentence): _____

118. How did you know to test the variance instead of the mean?

119. If an additional test were done on the claim of the average delay, which distribution would you use?

120. If an additional test were done on the claim of the average delay, but 45 flights were surveyed, which distribution would you use?

*For each word problem, use a solution sheet to solve the hypothesis test problem. Go to **Appendix E** for the chi-square solution sheet. Round expected frequency to two decimal places.*

121. A plant manager is concerned her equipment may need recalibrating. It seems that the actual weight of the 15-ounce cereal boxes it fills has been fluctuating. The standard deviation should be at most 0.5 ounces. To determine if the machine needs to be recalibrated, 84 randomly selected boxes of cereal from the next day's production were weighed. The standard deviation of the 84 boxes was 0.54. Does the machine need to be recalibrated?

122. Consumers may be interested in whether the cost of a particular calculator varies from store to store. Based on surveying 43 stores, which yielded a sample mean of $84 and a sample standard deviation of $12, test the claim that the standard deviation is greater than $15.

123. Isabella, an accomplished Bay-to-Breakers runner, claims that the standard deviation for her time to run the 7.5 mile race is at most 3 minutes. To test her claim, Isabella looks up five of her race times. They are 55 minutes, 61 minutes, 58 minutes, 63 minutes, and 57 minutes.

124. Airline companies are interested in the consistency of the number of babies on each flight so that they have adequate safety equipment. They are also interested in the variation of the number of babies. Suppose that an airline executive believes the average number of babies on flights is six with a variance of nine at most. The airline conducts a survey. The results of the 18 flights surveyed give a sample average of 6.4 with a sample standard deviation of 3.9. Conduct a hypothesis test of the airline executive's belief.

125. The number of births per woman in China is 1.6, down from 5.91 in 1966. This fertility rate has been attributed to the law passed in 1979 restricting births to one per woman. Suppose that a group of students studied whether the standard deviation of births per woman was greater than 0.75. They asked 50 women across China the number of births they had. The results are shown in **Table 11.59**. Does the students' survey indicate that the standard deviation is greater than 0.75?

# of Births	Frequency
0	5
1	30
2	10
3	5

Table 11.59

126. According to an avid aquarist, the average number of fish in a 20-gallon tank is 10, with a standard deviation of two. His friend, also an aquarist, does not believe that the standard deviation is two. She counts the number of fish in 15 other 20-gallon tanks. Based on the results that follow, do you think that the standard deviation is different from two? Data: 11; 10; 9; 10; 10; 11; 11; 10; 12; 9; 7; 9; 11; 10; and 11.

127. The manager of Frenchies is concerned that patrons are not consistently receiving the same amount of French fries with each order. The chef claims that the standard deviation for a 10-ounce order of fries is at most 1.5 ounces, but the manager thinks that it may be higher. He randomly weighs 49 orders of fries, which yields a mean of 11 ounces and a standard deviation of 2 ounces.

128. You want to buy a specific computer. A sales representative of the manufacturer claims that retail stores sell this computer at an average price of $1,249 with a very narrow standard deviation of $25. You find a website that has a price comparison for the same computer at a series of stores as follows: $1,299; $1,229.99; $1,193.08; $1,279; $1,224.95; $1,229.99; $1,269.95; and $1,249. Can you argue that pricing has a larger standard deviation than claimed by the manufacturer? Use the 5 percent significance level. As a potential buyer, what would be the practical conclusion from your analysis?

129. A company packages apples by weight. One of the weight grades is Class A apples. Class A apples have a mean weight of 150 grams, and there is a maximum allowed weight tolerance of 5 percent above or below the mean for apples in the same consumer package. A batch of apples is selected to be included in a Class A apple package. Given the following apple weights of the batch, does the fruit comply with the Class A grade weight tolerance requirements? Conduct an appropriate hypothesis test.

(a) At the 5 percent significance level

(b) At the 1 percent significance level

Weights in selected apple batch (in grams): 158; 167; 149; 169; 164; 139; 154; 150; 157; 171; 152; 161; 141; 166; and 172.

BRINGING IT TOGETHER: HOMEWORK

130.

a. Explain why a goodness-of-fit test and a test of independence are generally right-tailed tests.
b. If you did a left-tailed test, what would you be testing?

REFERENCES

11.1 Facts About the Chi-Square Distribution

Parade Magazine. (n.d.). Retrieved from https://parade.com/

Santa Clara County Public Health Department. (2011, May). *HIV/AIDS epidemiology Santa Clara County*. Retrieved from http://sccgov.iqm2.com/Citizens/FileOpen.aspx?Type=4&ID=32762

11.2 Goodness-of-Fit Test

College Board. (n.d.). Retrieved from http://www.collegeboard.com

Ma, Y., et al. (2003). Association between eating patterns and obesity in a free-living US adult population. *American Journal of Epidemiology 158*(1), 85–92.

Ogden, C. L., et al. (2012, January). *Prevalence of obesity in the United States, 2009–2010* (NCHS Data Brief No. 82). Hyattsville, MD: National Center for Health Statistics. Retrieved from http://www.cdc.gov/nchs/data/databriefs/db82.pdf

Stevens, B. J. (n.d.). *Multi-family and commercial solid waste and recycling survey*. Arlington County, VA. Retrieved from http://www.arlingtonva.us/departments/EnvironmentalServices/SW/file84429.pdf

U.S. Census Bureau. (n.d.). *Current population reports*. Retrieved from https://www.census.gov/main/www/cprs.html

U.S. Census Bureau. (n.d). Retrieved from https://www.census.gov/

11.3 Test of Independence

Harris Interactive. (n.d.). Retrieved from http://www.statisticbrain.com/favorite-flavor-of-ice-cream/

Statistics Brain. (2016, June 29). *Youngest online entrepreneurs list*. Retrieved from http://www.statisticbrain.com/youngest-online-entrepreneur-list

11.4 Test for Homogeneity

Bielick, S. (2008, December). 1.5 million homeschooled students in the United States in 2007 (NCES 2009030).

Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009030

Bielick, S. (2008, December). 1.5 million homeschooled students in the United States in 2007—supplemental tables (NCES 2009030). Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubs2009/2009030_sup.pdf

Insurance Institute for Highway Safety. (n.d.). Ratings. Retrieved from www.iihs.org/iihs/ratings

World Bank Group. (2014). *Energy use (kg of oil equivalent per capita*). Retrieved from http://data.worldbank.org/indicator/EG.USE.PCAP.KG.OE/countries

11.6 Test of a Single Variance

Apple Insider. (n.d.). Retrieved from http://appleinsider.com/mac_price_guide

World Bank. (n.d.). Retrieved from http://www.worldbank.org/

SOLUTIONS

1 mean = 25 and standard deviation = 7.0711

3 when the number of degrees of freedom is greater than 90

5 $df = 2$

7 a goodness-of-fit test

9 3

11 2.04

13 We decline to reject the null hypothesis. There is not enough evidence to suggest that the observed test scores are significantly different from the expected test scores.

15 H_0: the distribution of disease cases follows the ethnicities of the general population of Santa Clara County.

17 right-tailed

19 2016.136

21 Graph: Check student's solution. Decision: Reject the null hypothesis. Reason for decision: p-value < alpha Conclusion: The make-up of cases does not fit the ethnicities of the general population of Santa Clara County.

23 a test of independence

25 a test of independence

27 8

29 6.6

31 0.0435

33

Product-use Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	Totals
1–10	9,886	2,745	12,831	8,378	7,650	41,490
11–20	6,514	3,062	4,932	10,680	9,877	35,065
21–30	1,671	1,419	1,406	4,715	6,062	15,273
31+	759	788	800	2,305	3,970	8,622
Totals	18,830	8,014	19,969	26,078	27,559	10,0450

Table 11.60

35

Product Use Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White
1-10	7,777.57	3,310.11	8,248.02	10,771.29	11,383.01
11-20	6,573.16	2797.52	6970.76	9,103.29	9,620.27

Table 11.61

Product Use Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White
21-30	2,863.02	1,218.49	3,036.20	3,965.05	4,190.23
31+	1,616.25	687.87	1,714.01	2,238.37	2,365.49

Table 11.61

37 10,301.8

39 right-tailed

41

a. Reject the null hypothesis.

b. *p*-value < alpha

c. There is sufficient evidence to conclude that product use is dependent on ethnic group.

43 test for homogeneity

45 test for homogeneity

47 All values in the table must be greater than or equal to five.

49 3

51 0.00005

53 a goodness-of-fit test

55 a test for independence

57 Answers will vary. Sample answer: Tests of independence and tests for homogeneity both calculate the test statistic the same way $\sum_{(ij)} \frac{(O-E)^2}{E}$. In addition, all values must be greater than or equal to five.

59 a test of a single variance

61 a left-tailed test

63 H_0: $\sigma^2 = 0.81^2$; H_a: $\sigma^2 > 0.81^2$

65 a test of a single variance

67 0.0542

69 true

71 false

73

Marital Status	%	Expected Frequency
Never Married	31.3%	125.2
Married	56.1%	224.4
Widowed	2.5%	10
Divorced/Separated	10.1%	40.4

Table 11.62

a. The data fit the distribution.

b. The data do not fit the distribution.
c. 3
d. chi-square distribution with $df = 3$
e. 19.27
f. 0.0002
g. Check student's solution.
h.
   i. Alpha = 0.05
   ii. Decision: Reject null hypothesis.
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: Data do not fit the distribution.

75

a. H_0: The local results follow the distribution of the U.S. AP examinee population.
b. H_a: The local results do not follow the distribution of the U.S. AP examinee population.
c. $df = 5$
d. chi-square distribution with $df = 5$
e. chi-square test statistic = 13.4
f. p-value = 0.0199
g. Check student's solution.
h.
   i. Alpha = 0.05
   ii. Decision: Reject null when $a = 0.05$.
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: Local data do not fit the AP examinee distribution.
   v. Decision: Do not reject null when $a = 0.01$
   vi. Conclusion: There is insufficient evidence to conclude that local data do not follow the distribution of the U.S. AP examinee distribution.

77

a. H_0: The actual college majors of graduating females fit the distribution of their expected majors.
b. H_a: The actual college majors of graduating females do not fit the distribution of their expected majors.
c. $df = 10$
d. chi-square distribution with $df = 10$
e. test statistic = 11.48
f. p-value = 0.3211
g. Check student's solution.
h.
   i. Alpha = 0.05
   ii. Decision: Do not reject null hypothesis when $a = 0.05$ and $a = 0.01$.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: There is insufficient evidence to conclude that the distribution of actual college majors of graduating females do not fit the distribution of their expected majors.

79 true

81 true

83 false

85

a. H_0: Surveyed individuals fit the distribution of expected patients.
b. H_a: The surveyed individuals do not fit the distribution of patients.
c. $df = 4$
d. chi-square distribution with $df = 4$
e. test statistic = 54.01
f. p-value = 0
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: At the 5 percent level of significance from the data, there is sufficient evidence to conclude that the surveyed patients with the disease do not fit the distribution of expected patients.

87

a. H_0: Car size is independent of family size.
b. H_a: Car size is dependent on family size.
c. $df = 9$
d. chi-square distribution with $df = 9$
e. test statistic = 15.8284
f. p-value = 0.0706
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that car size and family size are dependent.

89

a. H_0: Honeymoon locations are independent of bride's age.
b. H_a: Honeymoon locations are dependent on bride's age.
c. $df = 9$
d. chi-square distribution with $df = 9$
e. test statistic = 15.7027
f. p-value = 0.0734
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that honeymoon location and bride age are dependent.

91

a. H_0: The types of fries sold are independent of the location.

b. H_a: The types of fries sold are dependent on the location.

c. $df = 6$

d. chi-square distribution with $df = 6$

e. test statistic =18.8369

f. p-value = 0.0044

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: p-value < alpha

iv. Conclusion: At the 5 percent significance level, there is sufficient evidence that types of fries and location are dependent.

93

a. H_0: Salary is independent of level of education.

b. H_a: Salary is dependent on level of education.

c. $df = 12$

d. chi-square distribution with $df = 12$

e. test statistic = 255.7704

f. p-value = 0

g. Check student's solution.

h. Alpha: 0.05

Decision: Reject the null hypothesis.

Reason for decision: p-value < alpha

Conclusion: At the 5 percent significance level, there is sufficient evidence to conclude that salary and level of education are dependent.

95 true

97 true

99

a. H_0: Age is independent of the youngest online entrepreneurs' net worth.

b. H_a: Age is dependent on the net worth of the youngest online entrepreneurs.

c. $df = 2$

d. chi-square distribution with $df = 2$

e. test statistic = 1.76

f. p-value = 0.4144

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Do not reject the null hypothesis.

iii. Reason for decision: p-value > alpha

iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that age and net worth for the youngest online entrepreneurs are dependent.

101

a. H_0: The distribution for personality types is the same for both majors.

b. H_a: The distribution for personality types is not the same for both majors.

c. $df = 4$

d. chi-square with $df = 4$

e. test statistic = 3.01

f. p-value = 0.5568

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: There is insufficient evidence to conclude that the distribution of personality types is different for business and social science majors.

103

a. H_0: The distribution for fish caught is the same in Green Valley Lake and in Echo Lake.

b. H_a: The distribution for fish caught is not the same in Green Valley Lake and in Echo Lake.

c. 3

d. chi-square with $df = 3$

e. 11.75

f. p-value = 0.0083

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: There is evidence to conclude that the distribution of fish caught is different in Green Valley Lake and in Echo Lake.

105

a. H_0: The distribution of average energy use in the United States is the same as in Europe between 2005 and 2010.

b. H_a: The distribution of average energy use in the United States is not the same as in Europe between 2005 and 2010.

c. $df = 4$

d. chi-square with $df = 4$

e. test statistic = 2.7434

f. p-value = 0.7395

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: At the 5 percent significance level, there is insufficient evidence to conclude that the average energy use values in the United States and EU are not derived from different distributions for the period from 2005 to 2010.

107

a. H_0: The distribution for technology use is the same for community college students and university students.

b. H_a: The distribution for technology use is not the same for community college students and university students.

c. 2

d. chi-square with $df = 2$

e. 7.05

f. p value = 0.0294

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: p value < alpha

iv. Conclusion: There is sufficient evidence to conclude that the distribution of technology use for statistics homework is not the same for statistics students at community colleges and at universities.

110 225

112 H_0: $\sigma^2 \leq 150$

114 36

116 Check student's solution.

118 The claim is that the variance is no more than 150 minutes.

120 a student's t or normal distribution

122

a. H_0: $\sigma = 15$

b. H_a: $\sigma > 15$

c. $df = 42$

d. chi-square with $df = 42$

e. test statistic = 26.88

f. p-value = 0.9663

g. Check student's solution.

h. i. Alpha = 0.05

ii. Decision: Do not reject null hypothesis.

iii. Reason for decision: p-value > alpha

iv. Conclusion: There is insufficient evidence to conclude that the standard deviation is greater than 15.

124

a. H_0: $\sigma \leq 3$

b. H_a: $\sigma > 3$

c. $df = 17$

d. chi-square distribution with $df = 17$

e. test statistic = 28.73

f. p-value = 0.0371

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Reject the null hypothesis.

iii. Reason for decision: p-value < alpha

iv. Conclusion: There is sufficient evidence to conclude that the standard deviation is greater than three.

126

a. H_0: $\sigma = 2$

b. H_a: $\sigma \neq 2$

c. $df = 14$

d. chi-square distiribution with $df = 14$

e. chi-square test statistic = 5.2094

f. p-value = 0.0346

g. Check student's solution.

h.
   i. Alpha = 0.05
   ii. Decision: Reject the null hypothesis
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: There is sufficient evidence to conclude that the standard deviation is different than two.

128 The sample standard deviation is \$34.29. $H_0 : \sigma^2 = 25^2$
$H_a : \sigma^2 > 25^2$
$df = n - 1 = 7$

Test statistic: $x^2 = \quad x_7^2 = \quad \frac{(n-1)s^2}{25^2} = \quad \frac{(8-1)(34.29)^2}{25^2} = 13.169$;

p-value: $P\left(x_7^2 > 13.169\right) = 1 - P\left(x_7^2 \quad \leq 13.169\right) = .0681$

Alpha: 0.05
Decision: Do not reject the null hypothesis.
Reason for decision: p-value > alpha
Conclusion: At the 5 percent level, there is insufficient evidence to conclude that the variance is more than 625.

130

a. The test statistic is always positive and if the expected and observed values are not close together, the test statistic is large and the null hypothesis will be rejected.

b. Testing to see if the data fits the distribution too well or is too perfect.

12 | LINEAR REGRESSION AND CORRELATION

Figure 12.1 Linear regression and correlation can help you determine whether an auto mechanic's salary is related to his work experience. (credit: Joshua Rothhaas)

Introduction

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Discuss basic ideas of linear regression and correlation
- Create and interpret a line of best fit
- Calculate and interpret the correlation coefficient
- Calculate and interpret outliers

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship, and how strong is it?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

The type of data described in the examples is **bivariate** data—*bi*—for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will study the simplest form of regression—*linear regression*—with one independent variable (x). This involves data that fit a line in two dimensions. You will also study correlation, which measures the strength of a relationship.

12.1 | Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form

$$y = a + bx$$

where a and b are constant numbers.

The variable x is the *independent variable*; y is the *dependent variable*. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example 12.1

The following examples are linear equations.

$$y = 3 + 2x$$
$$y = -0.01 + 1.2x$$

Try It

12.1 Is the following an example of a linear equation?

$y = -0.125 - 3.5x$

The graph of a linear equation of the form $y = a + bx$ is a straight line. Any line that is not vertical can be described by this equation.

Example 12.2

Graph the equation $y = -1 + 2x$.

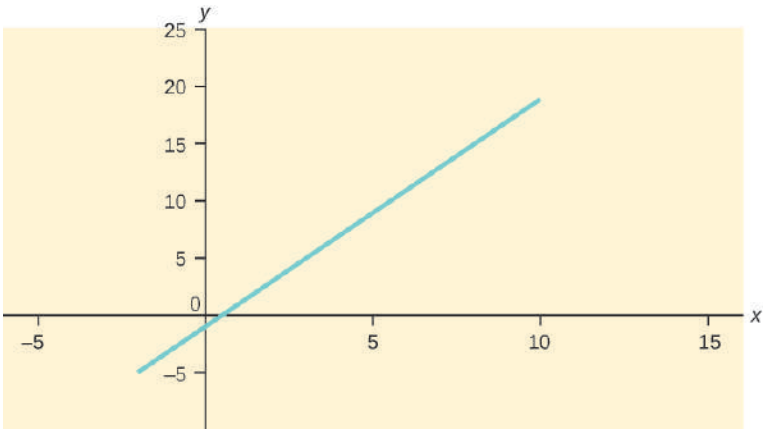


Figure 12.2

Try It

12.2 Is the following an example of a linear equation? Why or why not?

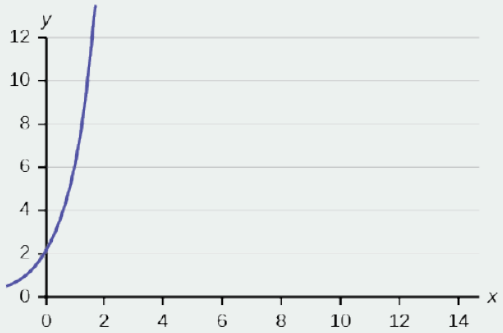


Figure 12.3

Example 12.3

Aaron's Word Processing Service does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the total cost in terms of the number of hours required to complete the job.

Solution 12.3

Let x = the number of hours it takes to get the job done.
Let y = the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then (32)(x) is the cost of the word processing only. The total cost is $y = 31.50 + 32x$.

Try It

12.3 Emma's Extreme Sports hires hang-gliding instructors and pays them a fee of $50 per class, as well as $20 per student in the class. The total cost Emma pays depends on the number of students in a class. Find the equation that expresses the total cost in terms of the number of students in a class.

Slope and *y*-interceptof a Linear Equation

For the linear equation $y = a + bx$, b = slope and a = y-inttercept. From algebra, recall that the slope is a number that describes the steepness of a line; the y-intercept is the y-coordinate of the point (0, a), where the line crosses the y-axis.

Please note that in previous courses you learned $y = mx + b$ was the slope-intercept form of the equation, where m represented the slope and b represented the y-intercept. In this text, the form $y = a + bx$ is used, where a is the y-intercept and b is the slope. The key is remembering the coefficient of x is the slope, and the constant number is the y-intercept.

Figure 12.4 Three possible graphs of $y = a + bx$. (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Example 12.4

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the independent and dependent variables? What is the y-intercept, and what is the slope? Interpret them using complete sentences.

Solution 12.4

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns $15 for each hour she tutors.

Try It

12.4 Ethan repairs household appliances such as dishwashers and refrigerators. For each visit, he charges $25 plus $20 per hour of work. A linear equation that expresses the total amount of money Ethan earns per visit is $y = 25 + 20x$.

What are the independent and dependent variables? What is the y-intercept, and what is the slope? Interpret them using complete sentences.

12.2 | The Regression Equation

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data with a scatter plot that appear to *fit* a straight line. This is called a *line of best fit or least-squares regression line*.

Collaborative Exercise

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height. For each set of data, plot the points on graph paper. Make your graph big enough and *use a ruler*. Then, *by eye*, draw a line that appears to *fit* the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your line so it crosses the y-axis. Using the slopes and the y-intercepts, write your equation of *best fit*. Do you think everyone will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches?

Example 12.5

A random sample of 11 statistics students produced the data in **Table 12.1**, where x is the third exam score out of 80 and y is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Table 12.1

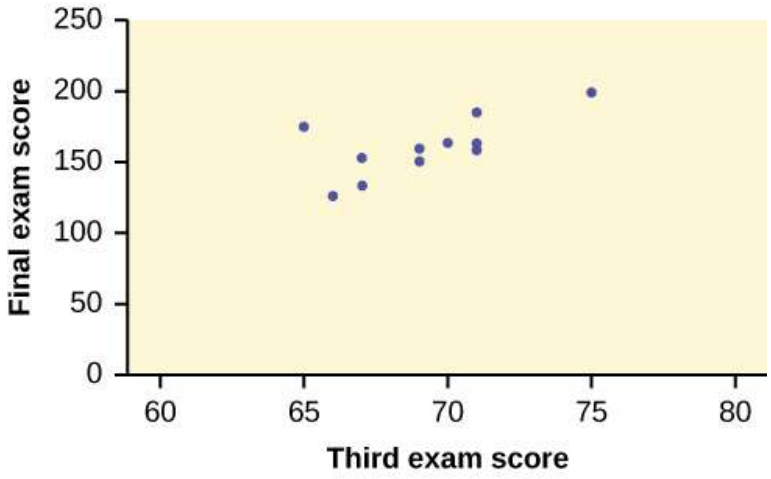


Figure 12.5 Using the x- and y-coordinates in the table, we plot the points on a graph to create the scatter plot showing the scores on the final exam based on scores from the third exam.

Try It

12.5 SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in **Table 12.2** show different depths in feet, with the maximum dive times in minutes. Use your calculator to find the least squares regression line and predict the maximum dive time for 110 feet.

x (depth)	*y* (maximum dive time)
50	80
60	55
70	45
80	35
90	25
100	22

Table 12.2

The third exam score, *x*, is the independent variable, and the final exam score, *y*, is the dependent variable. We will plot a regression line that best *fits* the data. If each of you were to fit a line *by eye*, you would draw different lines. We can obtain a line of best fit using either the median-–median line approach or by calculating the least-squares regression line.

Let''s first find the line of best fit for the relationship between the third exam score and the final exam score using the median-median line approach. Remember that this is the data from **Example 12.5** after the ordered pairs have been listed by ordering *x* values. If multiple data points have the same *y* values, then they are listed in order from least to greatest *y* (see data values where $x = 71$). We first divide our scores into three groups of approximately equal numbers of *x* values per group. The first and third groups have the same number of *x* values. We must remember first to put the *x* values in ascending order. The corresponding *y* values are then recorded. However, to find the median, we first must rearrange the *y* values in each group from the least value to the greatest value. **Table 12.3** shows the correct ordering of the *x* values but does not show a reordering of the *y* values.

x (third exam score)	*y* (final exam score)
65	175
66	126
67	133
67	153
69	151
69	159
70	163
71	159
71	163
71	185
75	198

Table 12.3

With this set of data, the first and last groups each have four *x* values and four corresponding *y* values. The second group has three *x* values and three corresponding *y* values. We need to organize the *x* and *y* values per group and find the median *x* and *y* values for each group. Let's now write out our *y* values for each group in ascending order. For group 1, the *y* values in order are 126, 133, 153, and 175. For group 2, the *y* values are already in order. For group 3, the *y* values are also already in order. We can represent these data as shown in **Table 12.4**, but notice that we have *broken* the ordered pairs; (65, 126) is not a data point in our original set:

Group	*x* (third exam score)	*y* (final exam score)	Median *x* value	Median *y* value
1	65 66 67 67	126 133 153 175	66.5	143
2	69 69 70	151 159 163	69	159
3	71 71 71 75	159 163 185 198	71	174

Table 12.4

When this is completed, we can write the ordered pairs for the median values. This allows us to find the slope and *y*-intercept of the –median-median line.

The ordered pairs are (66.5, 143), (69, 159), and (71, 174).

The slope can be calculated using the formula $m - \frac{y_2 - y_1}{x_2 - x_1}$. Substituting the median *x* and *y* values from the first and third groups gives $m = \frac{174 - 143}{71 - 66.5}$, which simplifies to $m \approx 6.9$.

The *y*-intercept may be found using the formula $b = \frac{\Sigma y - m\Sigma x}{3}$, which means the quantity of the sum of the median *y* values minus the slope times the sum of the median *x* values divided by three.

The sum of the median *x* values is 206.5, and the sum of the median *y* values is 476. Substituting these sums and the slope into the formula gives $b = \frac{476 - 6.9(206.5)}{3}$, which simplifies to $b \approx -316.3$.

The line of best fit is represented as $y = mx + b$.

Thus, the equation can be written as $y = 6.9x - 316.3$.

The median–median line may also be found using your graphing calculator. You can enter the *x* and *y* values into two separate lists; choose Stat, Calc, Med-Med, and press Enter. The slope, *a*, and *y*-intercept, *b*, will be provided. The calculator shows a slight deviation from the previous manual calculation as a result of rounding. Rounding to the nearest tenth, the calculator gives the –median-median line of $y = 6.9x - 315.5$. Each point of data is of the the form (x, y), and each point of the line of best fit using least-squares linear regression has the form $(x, \hat{y})$.

The $\hat{y}$ is read *y hat* and is the *estimated value of y*. It is the value of *y* obtained using the regression line. It is not generally equal to *y* from data, but it is still important because it can help make predictions for other values.

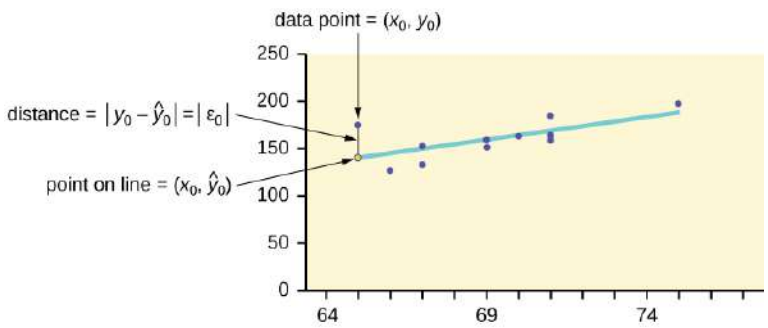


Figure 12.6

The term $y_0 - \hat{y}_0 = \varepsilon_0$ is called the *error* or *residual*. It is not an error in the sense of a mistake. The *absolute value of a residual* measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line, or it measures how far the estimate is from the actual data value.

If the observed data point lies above the line, the residual is positive and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative and the line overestimates that actual data value for y.

In **Figure 12.6**, $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

ε = the Greek letter epsilon

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, \ldots, 11$.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add them, you get the sum of ε squared from $i = 1$ to $i = 11$, as shown below.

$$(\varepsilon_1)^2 + (\varepsilon_2)^2 + \ldots + (\varepsilon_{11})^2 = \sum_{i=1}^{11} \varepsilon^2.$$

This is called the **sum of squared errors (SSE)**.

Using calculus, you can determine the values of a and b that make the SSE a minimum. When you make the SSE a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation

$$\hat{y} = a + bx$$

where

$$a = \bar{y} - b\bar{x}$$

and $b = \dfrac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$.

The sample means of the x values and the y values are $\bar{x}$ and $\bar{y}$, respectively. The best-fit line always passes through the point $(\bar{x}, \bar{y})$.

The slope (b) can be written as $b = r\left(\frac{s_y}{s_x}\right)$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient, which shows the relationship between the x and y values. This will be discussed in more detail in the next section.

Least-Squares Criteria for Best Fit

The process of fitting the best-fit line is called **linear regression**. We assume that the data are scattered about a straight line. To find that line, we minimize the sum of the squared errors (SSE), or make it as small as possible. Any other line you might choose would have a higher SSE than the best-fit line. This best-fit line is called the *least-squares regression line*.

NOTE

Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best-fit line and create a scatter plot are shown at the end of this section.

Third Exam vs. Final Exam Example

The graph of the line of best fit for the third exam/final exam example is as follows:

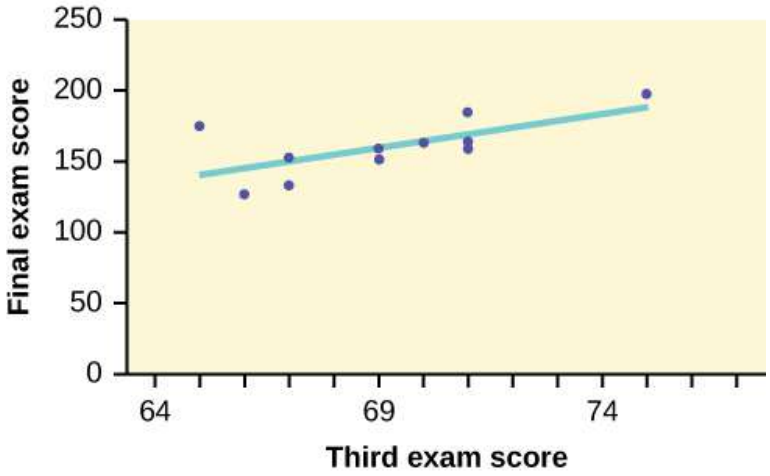


Figure 12.7

The least-squares regression line (best-fit line) for the third exam/final exam example has the equation

$$\hat{y} = -173.51 + 4.83x.$$

Understanding and Interpreting the *y*-intercept

The *y*-intercept, *a*, of the line describes where the plot line crosses the *y*-axis. The *y*-intercept of the best-fit line tells us the best value of the relationship when *x* is zero. In some cases, it does not make sense to figure out what *y* is when $x = 0$. For example, in the third exam vs. final exam example, the *y*-intercept occurs when the third exam score, or *x*, is zero. Since all the scores are grouped around a passing grade, there is no need to figure out what the final exam score, or *y*, would be when the third exam was zero.

However, the *y*-intercept is very useful in many cases. For many examples in science, the *y*-intercept gives the baseline reading when the experimental conditions aren''t applied to an experimental system. This baseline indicates how much the experimental condition affects the system. It could also be used to ensure that equipment and measurements are calibrated properly before starting the experiment.

In biology, the concentration of proteins in a sample can be measured using a chemical assay that changes color depending on how much protein is present. The more protein present, the darker the color. The amount of color can be measured by the absorbance reading. **Table 12.5** shows the expected absorbance readings at different protein concentrations. This is called a *standard curve* for the assay.

Concentration (mM)	Absorbance (mAU)
125	0.021
250	0.023
500	0.068
750	0.086
1,000	0.105
1,500	0.124
2,000	0.146

Table 12.5

The scatter plot **Figure 12.8** includes the line of best fit.

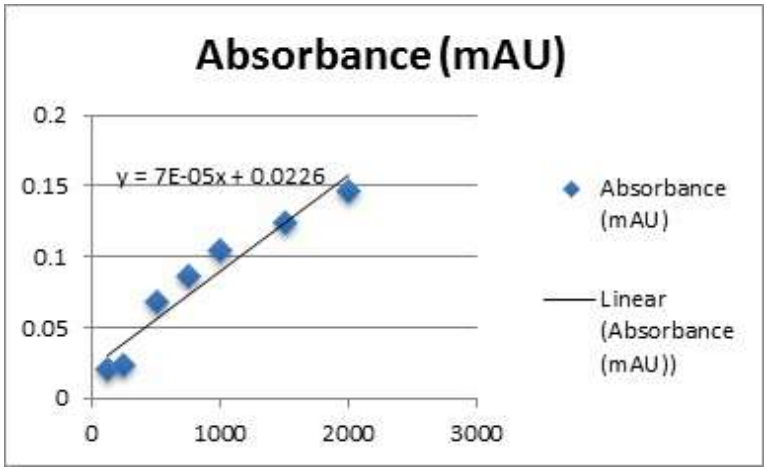


Figure 12.8

The *y*-intercept of this line occurs at 0.0226 mAU. This means the assay gives a reading of 0.0226 mAU when there is no protein present. That is, it is the baseline reading that can be attributed to something else, which, in this case, is some other non-protein chemicals that are absorbing light. We can tell that this line of best fit is reasonable because the *y*-intercept is small, close to zero. When there is no protein present in the sample, we expect the absorbance to be very small, or close to zero, as well.

Understanding Slope

The slope of the line, *b*, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

Interpretation of the Slope: The slope of the best-fit line tells us how the dependent variable (*y*) changes for every one unit increase in the independent (*x*) variable, on average.

Third Exam vs. Final Exam Example

Slope: The slope of the line is $b = 4.83$.
Interpretation: For a 1-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

Using the TI-83, 83+, 84, 84+ Calculator

Using the Linear Regression T Test: `LinRegTTest`

1. In the `STAT` list editor, enter the *x* data in list `L1` and the *y* data in list `L2`, paired so that the corresponding (*x*, *y*) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)
2. On the `STAT TESTS` menu, scroll down and select `LinRegTTest`. (Be careful to select `LinRegTTest`. Some calculators may also have a different item called `LinRegTInt`.)
3. On the `LinRegTTest` input screen, enter `Xlist: L1`, `Ylist: L2`, and `Freq: 1`.
4. On the next line, at the prompt β or ρ, highlight $\neq$ `0` and press `ENTER`.
5. Leave the line for `RegEQ:` blank.
6. Highlight `Calculate` and press `ENTER`.

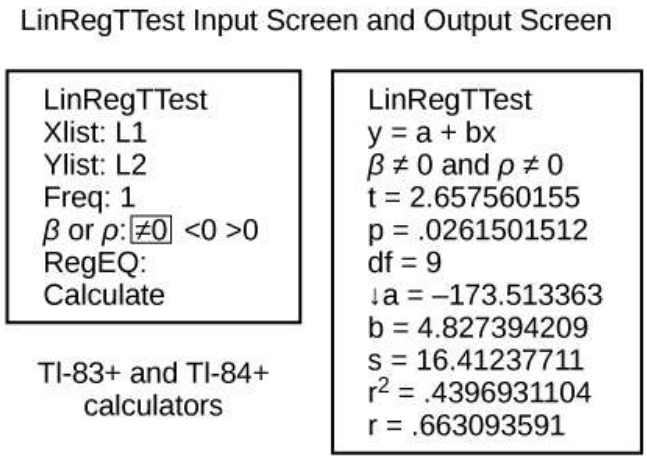


Figure 12.9

The output screen contains a lot of information. For now, let's focus on a few items from the output and return to the other items later.
The second line says $y = a + bx$. Scroll down to find the values $a = –173.513$ and $b = 4.8273$.

The equation of the best-fit line is $\hat{y} = –173.51 + 4.83x$.
The two items at the bottom are $r^2 = .43969$ and $r = .663$. For now, just note where to find these values; we examine them in the next two sections.

Graphing the Scatter Plot and Regression Line

1. We are assuming the x data are already entered in list `L1` and the y data are in list `L2`.
2. Press `2nd STATPLOT ENTER` to use `Plot 1`.
3. On the input screen for `PLOT 1`, highlight `On`, and press `ENTER`.
4. For `TYPE`, highlight the first icon, which is the scatter plot, and press `ENTER`.
5. Indicate `Xlist: L1` and `Ylist: L2`.
6. For `Mark`, it does not matter which symbol you highlight.
7. Press the `ZOOM` key and then the number `9` (for menu item `ZoomStat`); the calculator fits the window to the data.
8. To graph the best-fit line, press the `Y=` key and type the equation –173.5 + 4.83X into equation Y1. (The `X` key is immediately left of the `STAT` key.) Press `ZOOM 9` again to graph it.
9. Optional: If you want to change the viewing window, press the `WINDOW` key. Enter your desired window using `Xmin`, `Xmax`, `Ymin`, and `Ymax`.

NOTE

Another way to graph the line after you create a scatter plot is to use `LinRegTTest`.

1. Make sure you have done the scatter plot. Check it on your screen.
2. Go to `LinRegTTest` and enter the lists.
3. At `RegEq`, press `VARS` and arrow over to `Y-VARS`. Press 1 for `1:Function`. Press 1 for `1:Y1`. Then, arrow down to `Calculate` and do the calculation for the line of best fit.
4. Press `Y=` (you will see the regression equation).
5. Press `GRAPH`, and the line will be drawn.

The Correlation Coefficient *r*

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you determine whether the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatter plot) of the strength of the relationship between *x* and *y*.

The **correlation coefficient, *r*,** developed by Karl Pearson during the early 1900s, is numeric and provides a measure of the strength and direction of the linear association between the independent variable *x* and the dependent variable *y*.

If you suspect a linear relationship between *x* and *y*, then *r* can measure the strength of the linear relationship.

What the Value of *r* Tells Us

- The value of *r* is always between –1 and +1. In other words, $-1 \leq r \leq 1$.
- The size of the correlation *r* indicates the strength of the linear relationship between *x* and *y*. Values of *r* close to –1 or to +1 indicate a stronger linear relationship between *x* and *y*.
- If $r = 0$, there is absolutely no linear relationship between *x* and *y* (no linear correlation).
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all the original data points lie on a straight line. Of course, in the real world, this does not generally happen.

What the Sign of *r* Tells Us

- A positive value of *r* means that when *x* increases, *y* tends to increase and when *x* decreases, *y* tends to decrease (positive correlation).
- A negative value of *r* means that when *x* increases, *y* tends to decrease and when *x* decreases, *y* tends to increase (negative correlation).
- The sign of *r* is the same as the sign of the slope, *b*, of the best-fit line.

NOTE

A strong correlation does not suggest that *x* causes *y* or *y* causes *x*. We say correlation does not imply causation.

The correlation coefficient is calculated as the quantity of data points times the sum of the quantity of the *x*-coordinates times the *y*-coordinates, minus the quantity of the sum of the *x*-coordinates times the sum of the *y*-coordinates, all divided by the square root of the quantity of data points times the sum of the *x*-coordinates squared minus the square of the sum of the *x*-coordinates, times the number of data points times the sum of the *y*-coordinates squared minus the square of the sum of the *y*-coordinates. It can be summarized by the following equation:

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

where *n* is the number of data points.

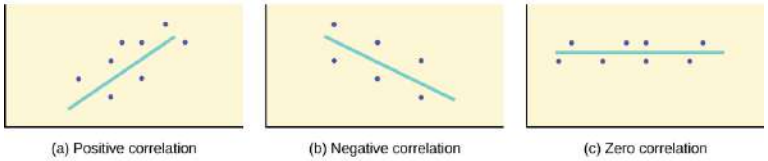


Figure 12.10 (a) A scatter plot showing data with a positive correlation: $0 < r < 1$. (b) A scatter plot showing data with a negative correlation: $-1 < r < 0$. (c) A scatter plot showing data with zero correlation: $r = 0$.

The formula for *r* looks formidable. However, computer spreadsheets, statistical software, and many calculators can calculate *r* quickly. The correlation coefficient, *r*, is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

The variable r^2 is called the **coefficient of determination** and it is the square of the correlation coefficient, but it is usually stated as a percentage, rather than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percentage of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.
- $1 - r^2$, when expressed as a percentage, represents the percentage of variation in y that is *not* explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the **third exam/final exam example** introduced in the previous section.

- The line of best fit is: $\hat{y} = -173.51 + 4.83x$.
- The correlation coefficient is $r = .6631$.
- The coefficient of determination is $r^2 = .6631^2 = .4397$.

Interpret r^2 in the context of this example.

- Approximately 44 percent of the variation (0.4397 is approximately 0.44) in the final exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore, the rest of the variation (1 – 0.44 = 0.56 or 56 percent) in the final exam grades cannot be explained by the variation of the grades on the third exam with the best-fit regression line. These are the variation of the points that are not as close to the regression line as others.

12.3 | Testing the Significance of the Correlation Coefficient (Optional)

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the correlation coefficient r and the sample size n, together.

We perform a hypothesis test of the *significance of the correlation coefficient* to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But, because we have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

The symbol for the population correlation coefficient is ρ, the Greek letter rho.
ρ = population correlation coefficient (unknown).
r = sample correlation coefficient (known; calculated from sample data).

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is *close to zero* or *significantly different from zero*. We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes the correlation coefficient is significantly different from zero, we say the correlation coefficient is *significant*.

- Conclusion: There is sufficient evidence to conclude there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes the correlation coefficient is not significantly different from zero (it is close to zero), we say the correlation coefficient is *not significant*.

- Conclusion: There is insufficient evidence to conclude there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero.
- What the conclusion means: There is not a significant linear relationship between x and y. Therefore, we *cannot* use the regression line to model a linear relationship between x and y in the population.

NOTE

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.

- If r is not significant *or* if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If r is significant and the scatter plot shows a linear trend, the line may *not* be appropriate or reliable for prediction *outside* the domain of observed x values in the data.

Performing the Hypothesis Test

- Null hypothesis: H_0: $\rho = 0$.
- Alternate hypothesis: H_a: $\rho \neq 0$.

What the Hypothesis Means in Words:

- **Null hypothesis H_0:** The population correlation coefficient *is not* significantly different from zero. There *is not* a significant linear relationship (correlation) between x and y in the population.
- **Alternate hypothesis H_a:** The population correlation coefficient *is* significantly different from zero. There *is* a significant linear relationship (correlation) between x and y in the population.

Drawing a Conclusion:

There are two methods to make a conclusion. The two methods are equivalent and give the same result.

- **Method 1: Use the *p*-value.**
- **Method 2: Use a table of critical values.**

In this chapter, we will always use a significance level of 5 percent, $\alpha = 0.05$.

NOTE

Using the *p*-value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$. But, the table of critical values provided in this textbook assumes we are using a significance level of 5 percent, $\alpha = 0.05$. If we wanted to use a significance level different from 5 percent with the critical value method, we would need different tables of critical values that are not provided in this textbook.

METHOD 1: Using a *p*-value to Make a Decision

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the *p*-value using LinRegTTEST:

1. Complete the same steps as the LinRegTTest performed previously in this chapter, making sure on the line prompt for β or σ, $\neq$ 0 is highlighted.
2. When looking at the output screen, the *p*-value is on the line that reads p =.

If the *p*-value is less than the significance level ($\alpha = 0.05$):

- Decision: Reject the null hypothesis.
- Conclusion: There is sufficient evidence to conclude there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.

If the *p*-value is *not* less than the significance level ($\alpha = 0.05$):

- Decision: Do not reject the null hypothesis.
- Conclusion: There is insufficient evidence to conclude there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero.

You will use technology to calculate the *p*-value, but it is useful to know that the *p*-value is calculated using a t distribution with $n - 2$ degrees of freedom and that the *p*-value is the combined area in both tails.

An alternative way to calculate the *p*-value (p) given by LinRegTTest is the command 2*tcdf(abs(t),10^99, n–2) in 2nd DISTR.

Third Exam vs. Final Exam Example: *p*-value Method

- Consider the **third exam/final exam example**.

- The line of best fit is $\hat{y} = -173.51 + 4.83x$, with $r = 0.6631$, and there are $n = 11$ data points.
- Can the regression line be used for prediction? Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?

H_0: $\rho = 0$

H_a: $\rho \neq 0$

$\alpha = 0.05$

- The p-value is 0.026 (from LinRegTTest on a calculator or from computer software).
- The p-value, 0.026, is less than the significance level of $\alpha = 0.05$.
- Decision: Reject the null hypothesis H_0.
- Conclusion: There is sufficient evidence to conclude there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

METHOD 2: Using a Table of Critical Values to Make a Decision

The **95 Percent Critical Values of the Sample Correlation Coefficient Table** (**Table 12.9**) can be used to give you a good idea of whether the computed value of r is significant. Use it to find the critical values using the degrees of freedom, $df = n - 2$. The table has already been calculated with $\alpha = 0.05$. The table tells you the positive critical value, but you should also make that number negative to have two critical values. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may use the line for prediction. If r is not significant (between the critical values), you should not use the line to make predictions.

Example 12.6

Suppose you computed $r = 0.801$ using $n = 10$ data points. The degrees of freedom would be 8 ($df = n - 2 = 10 - 2 = 8$). Using **Table 12.9** with df = 8, we find that the critical value is 0.632. This means the critical values are really ±0.632. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you to see that r is not between the two critical values.

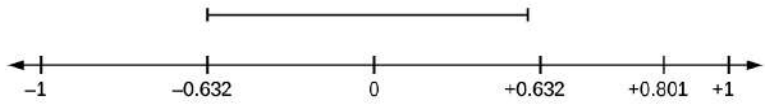


Figure 12.11 r is not between –0.632 and 0.632, so r is significant.

Try It

12.6 For a given line of best fit, you computed that $r = 0.6501$ using $n = 12$ data points, and the critical value found on the table is 0.576. Can the line be used for prediction? Why or why not?

Example 12.7

Suppose you computed $r = -0.624$ with 14 data points, where $df = 14 - 2 = 12$. The critical values are –0.532 and 0.532. Since $-0.624 < -0.532$, r is significant and the line can be used for prediction.

Figure 12.12 $r = –0.624$ and $–0.624 < –0.532$. Therefore, r is significant.

Try It

12.7 For a given line of best fit, you compute that $r = 0.5204$ using $n = 9$ data points, and the critical values are ±0.666. Can the line be used for prediction? Why or why not?

Example 12.8

Suppose you computed $r = 0.776$ and $n = 6$, with $df = 6 - 2 = 4$. The critical values are – 0.811 and 0.811. Since 0.776 is between the two critical values, r is not significant. The line should not be used for prediction.

Figure 12.13 $–0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

Try It

12.8 For a given line of best fit, you compute that $r = –0.7204$ using $n = 8$ data points, and the critical value is 0.707. Can the line be used for prediction? Why or why not?

Third Exam vs. Final Exam Example: Critical Value Method

Consider the **third exam/final exam example**. The line of best fit is: $\hat{y} = –173.51 + 4.83x$, with $r = .6631$, and there are $n = 11$ data points. Can the regression line be used for prediction? Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?

$H_0: \rho = 0$
$H_a: \rho \neq 0$
$\alpha = 0.05$

- Use the 95 Percent Critical Values table for r with $df = n – 2 = 11 – 2 = 9$.
- Using the table with $df = 9$, we find that the critical value listed is 0.602. Therefore, the critical values are ±0.602.
- Since $0.6631 > 0.602$, r is significant.
- Decision: Reject the null hypothesis.
- Conclusion: There is sufficient evidence to conclude there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.

Example 12.9

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine whether *r* is significant and whether the line of best fit associated with each correlation coefficient can be used to predict a *y* value. If it helps, draw a number line.

a. $r = –0.567$ and the sample size, n, is 19.

To solve this problem, first find the degrees of freedom. $df = n - 2 = 17$.
Then, using the table, the critical values are ±0.456.
–0.567 < –0.456, or you may say that –0.567 is not between the two critical values.
r is significant and may be used for predictions.

b. $r = 0.708$ and the sample size, n, is 9.

$df = n - 2 = 7$
The critical values are ±0.666.
0.708 > 0.666.
r is significant and may be used for predictions.

c. $r = 0.134$ and the sample size, n, is 14.

$df = 14 - 2 = 12$.
The critical values are ±0.532.
0.134 is between –0.532 and 0.532.
r is not significant and may not be used for predictions.

d. $r = 0$ and the sample size, n, is 5.

It doesn''t matter what the degrees of freedom are because $r = 0$ will always be between the two critical values, so *r* is not significant and may not be used for predictions.

Try It

12.9 For a given line of best fit, you compute that $r = 0$ using $n = 100$ data points. Can the line be used for prediction? Why or why not?

Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data be satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between *x* and *y* in the sample data provides strong enough evidence that we can conclude there is a linear relationship between *x* and *y* in the population.

The regression line equation that we calculate from the sample data gives the best-fit line for our particular sample. We want to use this best-fit line for the sample as an estimate of the best-fit line for the population. Examining the scatter plot and testing the significance of the correlation coefficient helps us determine whether it is appropriate to do this.

The assumptions underlying the test of significance are as follows:

- There is a linear relationship in the population that models the sample data. Our regression line from the sample is our best estimate of this line in the population.
- The *y* values for any particular *x* value are normally distributed about the line. This implies there are more *y* values scattered closer to the line than are scattered farther away. Assumption 1 implies that these normal distributions are centered on the line; the means of these normal distributions of *y* values lie on the line.
- Normal distributions of all the *y* values have the same shape and spread about the line.
- The residual errors are mutually independent (no pattern).

- The data are produced from a well-designed, random sample or randomized experiment.

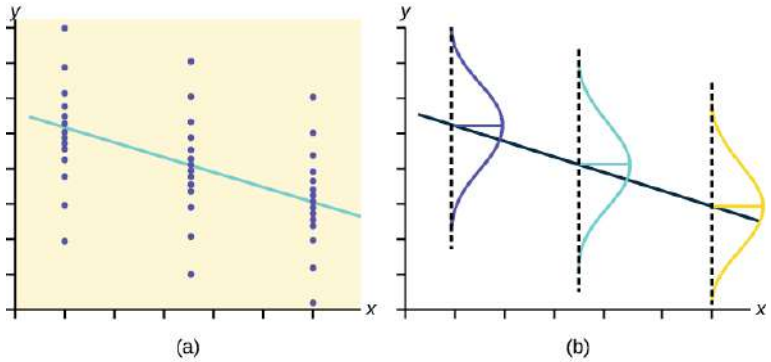


Figure 12.14 The *y* values for each *x* value are normally distributed about the line with the same standard deviation. For each *x* value, the mean of the *y* values lies on the regression line. More *y* values lie near the line than are scattered farther away from the line.

12.4 | Prediction (Optional)

Recall the **third exam/final exam example**.

We found the equation of the best-fit line for the final exam grade as a function of the grade on the third exam. We can now use the least-squares regression line for prediction.

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received a 73 on the third exam. The exam scores (*x* values) range from 65 to 75. Since 73 is between the *x* values 65 and 75, substitute $x = 73$ into the equation. Then,

$$\hat{y} = -173.51 + 4.83(73) = 179.08.$$

We predict that statistics students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example 12.10

Recall the **third exam/final exam example**.

a. What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Solution 12.10
a. 145.27

b. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Solution 12.10
b. The *x* values in the data are between 65 and 75. 90 is outside the domain of the observed *x* values in the data (independent variable), so you cannot reliably predict the final exam score for this student. Even though it is possible to enter 90 into the equation for *x* and calculate a corresponding *y* value, the *y* value that you get will not be reliable.

To understand how unreliable the prediction can be outside the *x* values observed in the data, make the substitution $x = 90$ into the equation:

$$\hat{y} = -173.51 + 4.83(90) = 261.19.$$

The final exam score is predicted to be 261.19. The most points that can be awarded for the final exam are 200.

Try It

12.10 Data are collected on the relationship between the number of hours per week practicing a musical instrument and scores on a math test. The line of best fit is as follows:

$\hat{y} = 72.5 + 2.8x$.
What would you predict the score on a math test will be for a student who practices a musical instrument for five hours a week?

12.5 | Outliers

In some data sets, there are values (observed data points) called **outliers**. Outliers are observed data points that are far from the least-squares line. They have large errors, where the error or residual is not very close to the best-fit line.

Outliers need to be examined closely. Sometimes, they should not be included in the analysis of the data, like if it is possible that an outlier is a result of incorrect data. Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to examine carefully what causes a data point to be an outlier.

Besides outliers, a sample may contain one or a few points that are called influential points. Influential points are observed data points that are far from the other observed data points in the horizontal direction. These points may have a big effect on the slope of the regression line. To begin to identify an influential point, you can remove it from the data set and determine whether the slope of the regression line is changed significantly.

You also want to examine how the correlation coefficient, r, has changed. Sometimes, it is difficult to discern a significant change in slope, so you need to look at how the strength of the linear relationship has changed. Computers and many calculators can be used to identify outliers and influential points. Regression analysis can determine if an outlier is, indeed, an influential point. The new regression will show how omitting the outlier will affect the correlation among the variables, as well as the fit of the line. A graph showing both regression lines helps determine how removing an outlier affects the fit of the model.

Identifying Outliers

We could guess at outliers by looking at a graph of the scatter plot and best-fit line. However, we would like some guideline regarding how far away a point needs to be to be considered an outlier. *As a rough rule of thumb, we can flag as an outlier any point that is located farther than two standard deviations above or below the best-fit line*. The standard deviation used is the standard deviation of the residuals or errors.

We can do this visually in the scatter plot by drawing an extra pair of lines that are two standard deviations above and below the best-fit line. Any data points outside this extra pair of lines are flagged as potential outliers. Or, we can do this numerically by calculating each residual and comparing it with twice the standard deviation. With regard to the TI-83, 83+, or 84+ calculators, the graphical approach is easier. The graphical procedure is shown first, followed by the numerical calculations. You would generally need to use only one of these methods.

Example 12.11

In the **third exam/final exam example**, you can determine whether there is an outlier. If there is an outlier, as an exercise, delete it and fit the remaining data to a new line. For this example, the new line ought to fit the remaining data better. This means the SSE (sum of the squared errors) should be smaller and the correlation coefficient ought to be closer to 1 or –1.

Solution 12.11

Graphical Identification of Outliers

With the TI-83, 83+, or 84+ graphing calculators, it is easy to identify the outliers graphically and visually. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance were equal to 2*s* or more, then we would consider the data point to be too far from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are outliers. Let's call these lines Y2 and Y3.

As we did with the equation of the regression line and the correlation coefficient, we will use technology to calculate this standard deviation for us. Using the LinRegTTest with these data, scroll down through the output screens to find $s = 16.412$.

Line Y2 = $-173.5 + 4.83x - 2(16.4)$, and line Y3 = $-173.5 + 4.83x + 2(16.4)$,
where $\hat{y} = -173.5 + 4.83x$ is the line of best fit. Y2 and Y3 have the same slope as the line of best fit.

Graph the scatter plot with the best-fit line in equation Y1, then enter the two extra lines as Y2 and Y3 in the *Y*= equation editor. Press ZOOM-9 to get a good view. You will see that the only point that is not between Y2 and Y3 is the point (65, 175). On the calculator screen, it is barely outside these lines, but it is considered an outlier because it is more than two standard deviations away from the best-fit line. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam.

Sometimes a point is so close to the lines used to flag outliers on the graph that it is difficult to tell whether the point is between or outside the lines. On a computer, enlarging the graph may help; on a small calculator screen, zooming in may make the graph clearer. Note that when the graph does not give a clear enough picture, you can use the numerical comparisons to identify outliers.

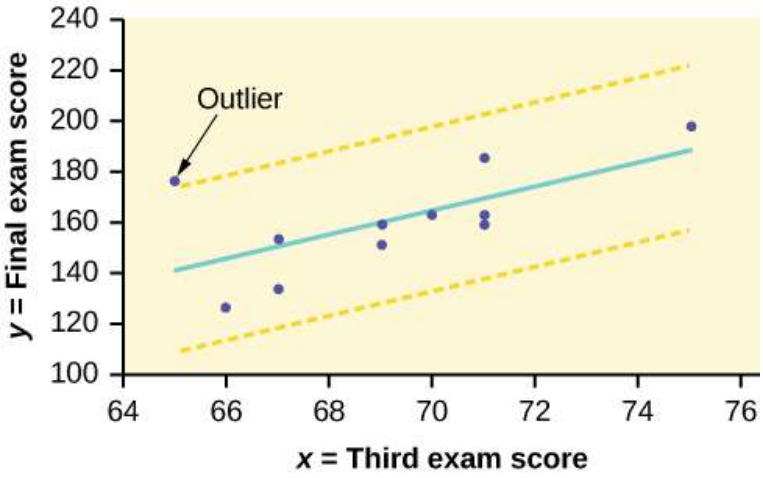


Figure 12.15

Try It

12.11 Identify the potential outlier in the scatter plot. The standard deviation of the residuals, or errors, is approximately 8.6.

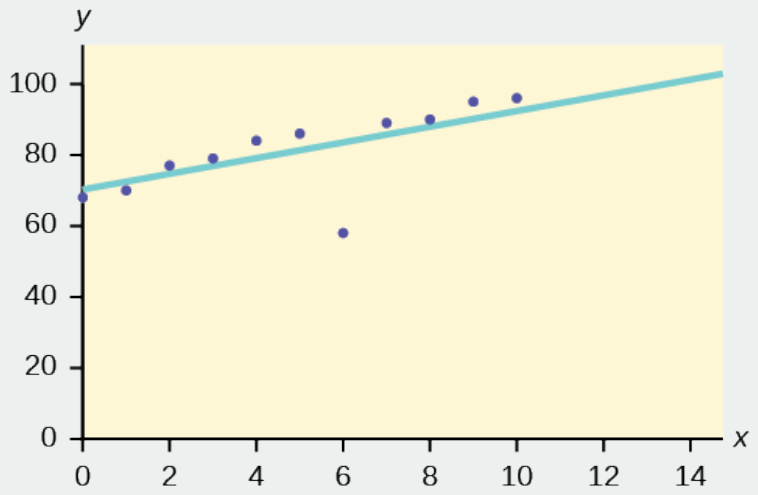


Figure 12.16

Numerical Identification of Outliers

In **Table 12.6**, the first two columns include the third exam and final exam data. The third column shows the predicted $\hat{y}$ values calculated from the line of best fit: $\hat{y} = -173.5 + 4.83x$. The residuals, or errors, that were mentioned in Section 3 of this chapter have been calculated in the fourth column of the table: Observed y value – predicted y value = $y - \hat{y}$.

s is the standard deviation of all the $y - \hat{y} = \varepsilon$ values, where n is the total number of data points. If each residual is calculated and squared, and the results are added, we get the SSE. The standard deviation of the residuals is calculated from the SSE as

$$s = \sqrt{\frac{SSE}{n-2}}.$$

NOTE

We divide by $(n - 2)$ because the regression model involves two estimates.

Rather than calculate the value of s ourselves, we can find s using a computer or calculator. For this example, the calculator function LinRegTTest found $s = 16.4$ as the standard deviation of the residuals 35; –17; 16; –6; –19; 9; 3; –1; –10; –9; –1.

x	y	$\hat{y}$	$y - \hat{y}$
65	175	140	175 – 140 = 35
67	133	150	133 – 150= –17
71	185	169	185 – 169 = 16
71	163	169	163 – 169 = –6
66	126	145	126 – 145 = –19
75	198	189	198 – 189 = 9
67	153	150	153 – 150 = 3
70	163	164	163 – 164 = –1

Table 12.6

x	y	$\hat{y}$	$y - \hat{y}$
71	159	169	159 – 169 = –10
69	151	160	151 – 160 = –9
69	159	160	159 – 160 = –1

Table 12.6

We are looking for all data points for which the residual is greater than $2s = 2(16.4) = 32.8$ or less than –32.8. Compare these values with the residuals in column four of the table. The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

How Does the Outlier Affect the Best-Fit Line?

Numerically and graphically, we have identified point (65, 175) as an outlier. Recall that recalculation of the least-squares regression line and summary statistics, following deletion of an outlier, may be used to determine whether an outlier is also an influential point. This process also allows you to compare the strength of the correlation of the variables and possible changes in the slope both before and after the omission of any outliers.

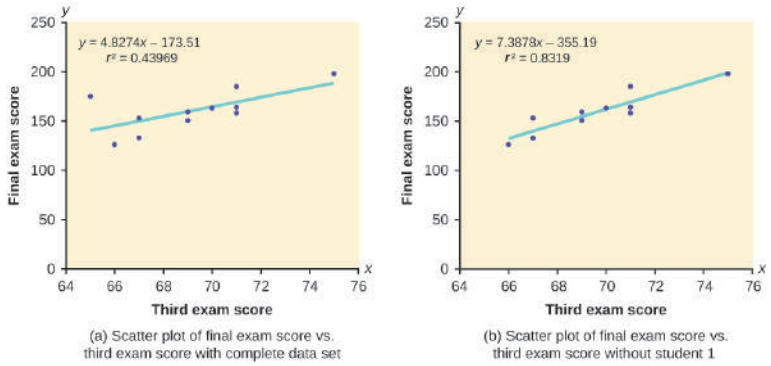
Compute a new best-fit line and correlation coefficient using the 10 remaining points.

On the TI-83, TI-83+, or TI-84+ calculators, delete the outlier from L1 and L2. Using the LinRegTTest, found under Stat and Tests, the new line of best fit and correlation coefficient are the following:

$\hat{y} = -355.19 + 7.39x$ and $r = 0.9121$.

The slope is now 7.39, compared to the previous slope of 4.83. This seems significant, but we need to look at the change in r-values as well. The new line shows $r = 0.9121$, which indicates a stronger correlation than the original line, with $r = 0.6631$, since $r = 0.9121$ is closer to 1. This means the new line is a better fit to the data values. The line can better predict the final exam score given the third exam score. It also means the outlier of (65, 175) was an influential point, since there is a sizeable difference in r-values. We must now decide whether to delete the outlier. If the outlier was recorded erroneously, it should certainly be deleted. Because it produces such a profound effect on the correlation, the new line of best fit allows for better prediction and an overall stronger model.

You may use Excel to graph the two least-squares regression lines and compare the slopes and fit of the lines to the data, as shown in **Figure 12.17**.



(a) Scatter plot of final exam score vs. third exam score with complete data set

(b) Scatter plot of final exam score vs. third exam score without student 1

Figure 12.17

You can see that the second graph shows less deviation from the line of best fit. It is clear that omission of the influential point produced a line of best fit that more closely models the data.

Numerical Identification of Outliers: Calculating *s* and Finding Outliers Manually

If you do not have the function LinRegTTest on your calculator, then you must calculate the outlier in the first example by doing the following.

First, square each $|y - \hat{y}|$.

The squares are 35^2; 17^2; 16^2; 6^2; 19^2; 9^2; 3^2; 1^2; 10^2; 9^2; 1^2.

Then, add (sum) all the $|y - \hat{y}|$ squared terms using the formula

$\sum_{i=1}^{11}(|y_i - \hat{y}_i|)^2 = \sum_{i=1}^{11}\varepsilon_i^2$ (Recall that $y_i - \hat{y}_i = \varepsilon_i$).

$= 35^2 + 17^2 + 16^2 + 6^2 + 19^2 + 9^2 + 3^2 + 1^2 + 10^2 + 9^2 + 1^2$

$= 2{,}440 = \text{SSE}$.

The result, **SSE**, is the sum of squared errors.

Next, calculate s, the standard deviation of all the $y - \hat{y} = \varepsilon$-values where n = the total number of data points.

The calculation is $s = \sqrt{\frac{\text{SSE}}{n-2}}$.

For the third exam/final exam example, $s = \sqrt{\frac{2440}{11-2}} = 16.47$.

Next, multiply s by 2:
(2)(16.47) = 32.94
32.94 is two standard deviations away from the mean of the $y - \hat{y}$ values.

If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least 2s, then we would consider the data point to be too far from the line of best fit. We call that point a **potential outlier**.

For the example, if any of the $|y - \hat{y}|$ values are *at least* 32.94, the corresponding (x, y) data point is a potential outlier.

For the third exam/final exam example, all the $|y - \hat{y}|$ values are less than 31.29 except for the first one, which is 35:

35 > 31.29. That is, $|y - \hat{y}| \geq (2)(s)$.

The point that corresponds to $|y - \hat{y}| = 35$ is (65, 175). *Therefore, the data point (65, 175) is a potential outlier.* For this example, we will delete it. (Remember, we do not always delete an outlier.)

NOTE

When outliers are deleted, the researcher should either record that data were deleted, and why, or the researcher should provide results both with and without the deleted data. If data are erroneous and the correct values are known (e.g., student 1 actually scored a 70 instead of a 65), then this correction can be made to the data.

The next step is to compute a new best-fit line using the 10 remaining points. The new line of best fit and the correlation coefficient are

$\hat{y} = -355.19 + 7.39x$ and $r = .9121$.

Example 12.12

Using this new line of best fit (based on the remaining 10 data points in the **third exam/final exam example**), what would a student who receives a 73 on the third exam expect to receive on the final exam? Is this the same as the prediction made using the original line?

Solution 12.12

Using the new line of best fit, $\hat{y} = -355.19 + 7.39(73) = 184.28$. A student who scored 73 points on the third exam would expect to earn 184 points on the final exam.

The original line predicted that $\hat{y} = -173.51 + 4.83(73) = 179.08$, so the prediction using the new line with the

outlier eliminated differs from the original prediction.

Try It

12.12 The data points for the graph from the **third exam/final exam example** are as follows: (1, 5), (2, 7), (2, 6), (3, 9), (4, 12), (4, 13), (5, 18), (6, 19), (7, 12), and (7, 21). Remove the outlier and recalculate the line of best fit. Find the value of $\hat{y}$ when $x = 10$.

Example 12.13

The consumer price index (CPI) measures the average change over time in prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price changes in the nation's economy to government, businesses, and labor forces, the CPI helps them make economic decisions. The president, U.S. Congress, and the Federal Reserve Board use CPI trends to form monetary and fiscal policies. In the following table, x is the year and y is the CPI.

x	*y*	*x*	*y*
1915	10.1	1969	36.7
1926	17.7	1975	49.3
1935	13.7	1979	72.6
1940	14.7	1980	82.4
1947	24.1	1986	109.6
1952	26.5	1991	130.7
1964	31.0	1999	166.6

Table 12.7

a. Draw a scatter plot of the data.
b. Calculate the least-squares line. Write the equation in the form $\hat{y} = a + bx$.
c. Draw the line on a scatter plot.
d. Find the correlation coefficient. Is it significant?
e. What is the average CPI for the year 1990?

Solution 12.13

a. See **Figure 12.17**.
b. Using our calculator, $\hat{y} = -3204 + 1.662x$ is the equation of the line of best fit.
c. See **Figure 12.17**.
d. $r = 0.8694$. The number of data points is $n = 14$. Use the 95 Percent Critical Values of the Sample Correlation Coefficient table at the end of Chapter 12: In this case, $df = 12$. The corresponding critical values from the table are ±0.532. Since $0.8694 > 0.532$, r is significant. We can use the predicted regression line we found above to make the prediction for $x = 1990$.
e. $\hat{y} = -3204 + 1.662(1990) = 103.4$ CPI.

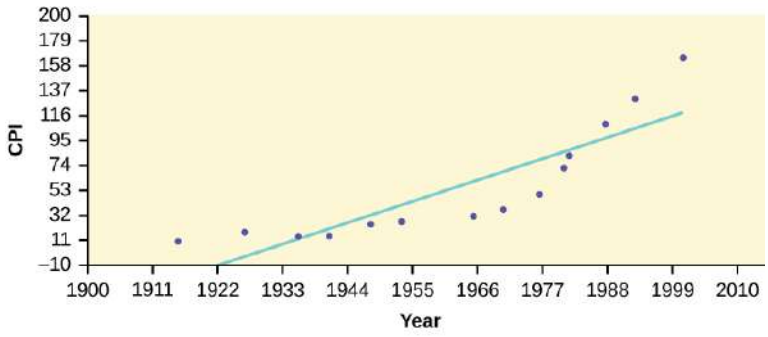


Figure 12.18

NOTE

In the example, notice the pattern of the points compared with the line. Although the correlation coefficient is significant, the pattern in the scatter plot indicates that a curve would be a more appropriate model to use than a line. In this example, a statistician would prefer to use other methods to fit a curve to these data, rather than model the data with the line we found. In addition to doing the calculations, it is always important to look at the scatter plot when deciding whether a linear model is appropriate.

If you are interested in seeing more years of data, visit the Bureau of Labor Statistics CPI website (ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt). Our data are taken from the column Annual Avg. (third column from the right). For example, you could add more current years of data. Try adding the more recent years: 2004, CPI = 188.9; 2008, CPI = 215.3; and 2011, CPI = 224.9. See how this affects the model. (Check: $\hat{y} = -4436 + 2.295x$; $r = 0.9018$. Is r significant? Is the fit better with the addition of the new points?)

Try It

12.13 The following table shows economic development measured in per capita income (PCINC).

Year	PCINC	Year	PCINC
1870	340	1920	1,050
1880	499	1930	1,170
1890	592	1940	1,364
1900	757	1950	1,836
1910	927	1960	2,132

Table 12.8

a. What are the independent and dependent variables?
b. Draw a scatter plot.
c. Use regression to find the line of best fit and the correlation coefficient.
d. Interpret the significance of the correlation coefficient.
e. Is there a linear relationship between the variables?

f. Find the coefficient of determination and interpret it.
g. What is the slope of the regression equation? What does it mean?
h. Use the line of best fit to estimate PCINC for 1900 and for 2000.
i. Determine whether there are any outliers.

95 Percent Critical Values of the Sample Correlation Coefficient Table

Degrees of Freedom: $n - 2$	Critical Values: + and –
1	0.997
2	0.950
3	0.878
4	0.811
5	0.754
6	0.707
7	0.666
8	0.632
9	0.602
10	0.576
11	0.555
12	0.532
13	0.514
14	0.497
15	0.482
16	0.468
17	0.456
18	0.444
19	0.433
20	0.423
21	0.413
22	0.404
23	0.396
24	0.388
25	0.381
26	0.374
27	0.367
28	0.361
29	0.355

Table 12.9

Degrees of Freedom: $n - 2$	Critical Values: + and –
30	0.349
40	0.304
50	0.273
60	0.250
70	0.232
80	0.217
90	0.205
100	0.195

Table 12.9

12.6 | Regression (Distance from School) (Optional)

Stats Lab

12.1 Regression (Distance From School)

Student Learning Outcomes

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine whether that relationship is significant.

Collect the Data

Use eight members of your class for the sample. Collect bivariate data (distance an individual lives from school, the cost of supplies for the current term).

1. Complete the table.

Distance from School	Cost of Supplies This Term

Table 12.10

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph *distance* vs. *cost*. Plot the points on the graph. Label both axes with words. Scale both axes.

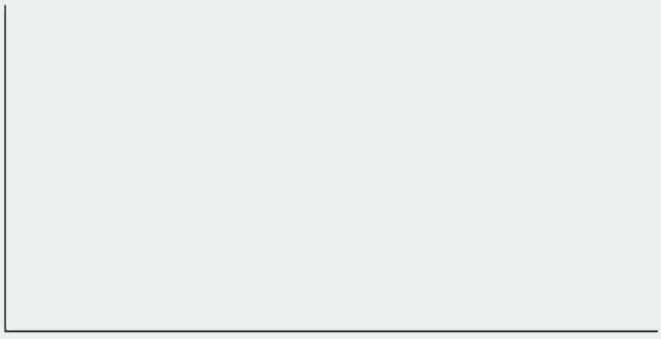


Figure 12.19

Analyze the Data

Enter your data into a calculator or computer. Write the linear equation, rounding to four decimal places.

1. Calculate the following:
   a. a = _____
   b. b = _____
   c. correlation = _____
   d. n = _____
   e. equation: $\hat{y}$ = _____
   f. Is the correlation significant? Why or why not? (Answer in one to three complete sentences.)
2. Supply an answer for the following scenarios:
   a. For a person who lives eight miles from campus, predict the total cost of supplies this term.
   b. For a person who lives 80 miles from campus, predict the total cost of supplies this term.
3. Obtain the graph on a calculator or computer. Sketch the regression line.

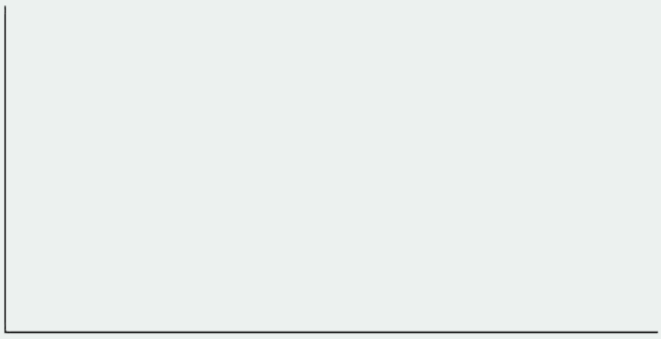


Figure 12.20

Discussion Questions

1. Answer each question in complete sentences.
   a. Does the line seem to fit the data? Why?
   b. What does the correlation imply about the relationship between distance and cost?
2. Are there any outliers? If so, which point is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

12.7 | Regression (Textbook Cost) (Optional)

Stats Lab

12.2 Regression (Textbook Cost)

Student Learning Outcomes

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine whether that relationship is significant.

Collect the Data

Survey 10 textbooks. Collect bivariate data (number of pages in a textbook, the cost of the textbook).

1. Complete the table.

Number of Pages	Cost of Textbook

Table 12.11

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph *pages vs. cost*. Plot the points on the graph in **Analyze the Data**. Label both axes with words. Scale both axes.

Analyze the Data

Enter your data into a calculator or computer. Write the linear equation, rounding to four decimal places.

1. Calculate the following:
   a. a = _____
   b. b = _____
   c. correlation = _____
   d. n = _____
   e. equation: y = _____
   f. Is the correlation significant? Why or why not? (Answer in complete sentences.)
2. Supply an answer for the following scenarios:
   a. For a textbook with 400 pages, predict the cost.
   b. For a textbook with 600 pages, predict the cost.
3. Obtain the graph on a calculator or computer. Sketch the regression line.

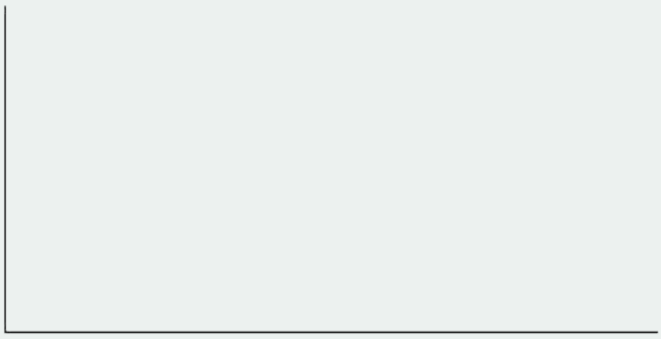


Figure 12.21

Discussion Questions

1. Answer each question in complete sentences.
    a. Does the line seem to fit the data? Why?
    b. What does the correlation imply about the relationship between the number of pages and the cost?
2. Are there any outliers? If so, which point is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

12.8 | Regression (Fuel Efficiency) (Optional)

Stats Lab

12.3 Regression (Fuel Efficiency)

Student Learning Outcomes

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine whether that relationship is significant.

Collect the Data

Find a reputable source that provides information on total fuel efficiency (in miles per gallon) and weight (in pounds) of new cars with an automatic transmission. You will use these data to determine the relationship, if any, between the fuel efficiency of a car and its weight.

1. Using your random-number generator, select 20 cars randomly from the list and record their weight and fuel efficiency into **Table 12.12**.

Weight	Fuel Efficiency

Table 12.12

2. Which variable is the dependent variable and which is the independent variable? Why?
3. By hand, draw a scatter plot of *weight vs. fuel efficiency*. Plot the points on graph paper. Label both axes with words. Scale both axes accurately.

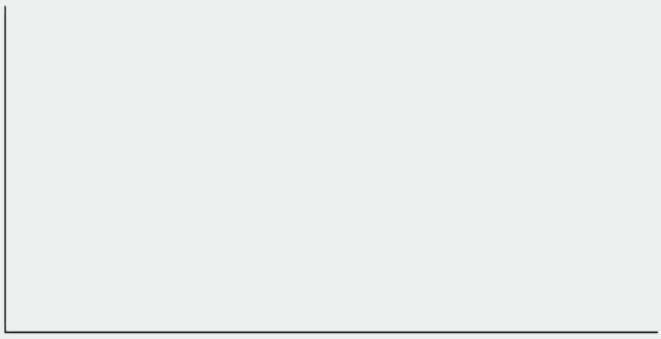


Figure 12.22

Analyze the Data

Enter your data into a calculator or computer. Write the linear equation, rounding to four decimal places.

1. Calculate the following:
    a. a = _____
    b. b = _____
    c. correlation = _____
    d. n = _____
    e. equation: $\hat{y}$ = _____
2. Obtain a graph of the regression line on a calculator. Sketch the regression line on the same axes as your scatter plot.

Discussion Questions

1. Is the correlation significant? Explain how you determined this in complete sentences.
2. Is the relationship a positive one or a negative one? Explain how you can tell and what this means in terms of weight and fuel efficiency.
3. In one or two complete sentences, what is the practical interpretation of the slope of the least-squares line in terms of fuel efficiency and weight?
4. For a car that weighs 4,000 pounds, predict its fuel efficiency. Include units.
5. Can we predict the fuel efficiency of a car that weighs 10,000 pounds using the least-squares line? Explain why or why not.
6. Answer each question in complete sentences.
    a. Does the line seem to fit the data? Why or why not?
    b. What does the correlation imply about the relationship between fuel efficiency and weight of a car? Is this what you expected?
7. Are there any outliers? If so, which point is an outlier?

KEY TERMS

coefficient of correlation a measure developed by Karl Pearson during the early 1900s that gives the strength of association between the independent variable and the dependent variable;

$$r = \frac{n\sum xy - [\sum x][\sum y]}{\sqrt{(n\sum x^2 - [\sum x]^2)(n\sum y^2 - [\sum y]^2)}}$$

where n is the number of data points
The coefficient cannot be more than 1 and less than –1. The closer the coefficient is to ±1, the stronger the evidence of a significant linear relationship between x and y.

outlier an observation that does not fit the rest of the data

CHAPTER REVIEW

12.1 Linear Equations

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used (numerically with actual or predicted data values) or graphically from a plotted curve. Lines are classified as straight curves. Algebraically, a linear equation typically takes the form $y = mx + b$, where m and b are constants, x is the independent variable, and y is the dependent variable. In a statistical context, a linear equation is written in the form $y = a + bx$, where a and b are the constants. This form is used to help you distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the x variable (b is called a *coefficient*) is called the *slope*. The slope describes the rate of change between the independent and dependent variables; in other words, the rate of change describes the change that occurs in the dependent variable as the independent variable is changed. In the equation $y = a + bx$, the constant a is called the y-intercept. Graphically, the y-intercept is the y-coordinate of the point where the graph of the line crosses the y-axis. At this point, $x = 0$.

The slope of a line is a value that describes the rate of change between the independent and dependent variables. The slope tells us how the dependent variable (y) changes for every one-unit increase in the independent (x) variable, on average. The y-intercept is used to describe the dependent variable when the independent variable equals zero. Graphically, the slope is represented by three line types in elementary statistics.

12.2 The Regression Equation

A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the x and y variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called *errors*, measure the distance from the actual value of y and the estimated value of y. The sum of squared errors, or SSE, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data but should not be used to make predictions for values outside the set of data.

The correlation coefficient, r, measures the strength of the linear association between x and y. The variable r has to be between –1 and +1. When r is positive, x and y tend to increase and decrease together. When r is negative, x increases and y decreases, or the opposite occurs: x decreases and y increases. The coefficient of determination, r^2, is equal to the square of the correlation coefficient. When expressed as a percentage, r^2 represents the percentage of variation in the dependent variable, y, that can be explained by variation in the independent variable, x, using the regression line.

12.3 Testing the Significance of the Correlation Coefficient (Optional)

Linear regression is a procedure for fitting a straight line of the form $\hat{y} = a + bx$ to data. The conditions for regression are as follows:

- Linear: In the population, there is a linear relationship that models the average value of y for different values of x.
- Independent: The residuals are assumed to be independent.
- Normal: The y values are distributed normally for any value of x.
- Equal variance: The standard deviation of the y values is equal for each x value.

- Random: The data are produced from a well-designed random sample or a randomized experiment.

The slope b and intercept a of the least-squares line estimate the slope β and intercept α of the population (true) regression line. To estimate the population standard deviation of y (σ) use the standard deviation of the residuals: $s = \sqrt{\frac{SSE}{n-2}}$. The variable ρ (rho) is the population correlation coefficient. To test the null hypothesis, H_0: ρ = hypothesized value, use a linear regression t-test. The most common null hypothesis is H_0: $\rho = 0$, which indicates there is no linear relationship between x and y in the population. The TI-83, 83+, 84, 84+ calculator function LinRegTTest can perform this test (STATS, TESTS, LinRegTTest).

12.4 Prediction (Optional)

After determining the presence of a strong correlation coefficient and calculating the line of best fit, you can use the least-squares regression line to make predictions about your data.

12.5 Outliers

To determine whether a point is an outlier, do one of the following:

1. Input the following equations into the TI 83, 83+, 84, or 84+ calculator:

   $y_1 = a + bx$
   $y_2 = a + bx + 2s$
   $y_3 = a + bx - 2s$

   where s is the standard deviation of the residuals.

   If any point is above y_2 or below y_3, then the point is considered to be an outlier.
2. Use the residuals and compare their absolute values to $2s$, where s is the standard deviation of the residuals. If the absolute value of any residual is greater than or equal to $2s$, then the corresponding point is an outlier.
3. Note: The calculator function LinRegTTest (STATS, TESTS, LinRegTTest) calculates s.

FORMULA REVIEW

12.1 Linear Equations

$y = a + bx$, where a is the y-intercept and b is the slope. The variable x is the independent variable and y is the dependent variable.

12.3 Testing the Significance of the Correlation Coefficient (Optional)

Least-Squares Line or Line of Best Fit:

$\hat{y} = a + bx$,

where a is the y-intercept and b is the slope.

Standard Deviation of the Residuals:

$s = \sqrt{\frac{SSE}{n-2}}$,

where SSE = sum of squared errors, and

n = the number of data points.

PRACTICE

12.1 Linear Equations

Use the following information to answer the next three exercises. A vacation resort rents scuba equipment to certified divers. The resort charges an up-front fee of $25 and another fee of $12.50 an hour.

1. What are the dependent and independent variables?

2. Find the equation that expresses the total fee in terms of the number of hours the equipment is rented.

3. Graph the equation from **Exercise 12.2**.

Use the following information to answer the next two exercises. A credit card company charges $10 when a payment is late and $5 a day each day the payment remains unpaid.

4. Find the equation that expresses the total fee in terms of the number of days the payment is late.

5. Graph the equation from **Exercise 12.4**.

6. Is the equation $y = 10 + 5x - 3x^2$ linear? Why or why not?

7. Which of the following equations are linear?

a. $y = 6x + 8$

b. $y + 7 = 3x$

c. $y - x = 8x^2$

d. $4y = 8$

8. Does the graph in **Figure 12.23** show a linear equation? Why or why not?

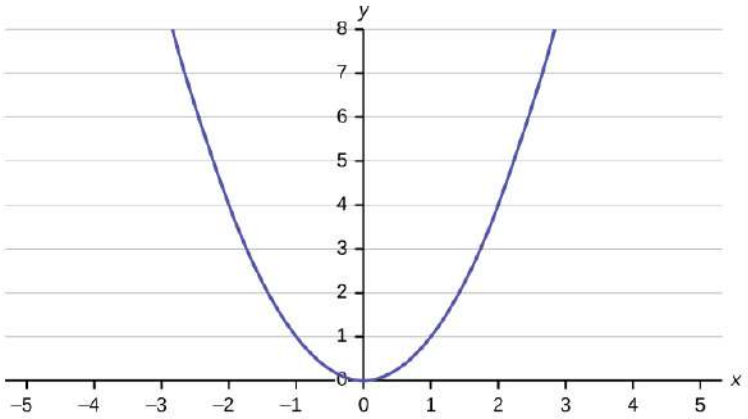


Figure 12.23

Use the following information to answer the next exercise. **Table 12.13** contains real data for the first two decades of flu reporting.

Year	Number of Flu Cases Diagnosed	Number of Flu Deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591

Table 12.13

1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Table 12.13

9. Use the columns *Year* and *Number of Flu Cases Diagnosed*. Why is year the independent variable and number of flu cases diagnosed the dependent variable (instead of the reverse)?

Use the following information to answer the next two exercises. A specialty cleaning company charges an equipment fee and an hourly labor fee. A linear equation that expresses the total amount of the fee the company charges for each session is $y = 50 + 100x$.

10. What are the independent and dependent variables?

11. What is the y-intercept, and what is the slope? Interpret them using complete sentences.

Use the following information to answer the next three questions. As a result of erosion, a river shoreline is losing several thousand pounds of soil each year. A linear equation that expresses the total amount of soil lost per year is $y = 12{,}000x$.

12. What are the independent and dependent variables?

13. How many pounds of soil does the shoreline lose in a year?

14. What is the y-intercept? Interpret its meaning.

Use the following information to answer the next two exercises. The price of a single issue of stock can fluctuate throughout the day. A linear equation that represents the price of stock for Shipment Express is $y = 15 – 1.5x$, where x is the number of hours passed in an eight-hour day of trading.

15. What are the slope and y-intercept? Interpret their meaning.

16. If you owned this stock, would you want a positive or negative slope? Why?

12.2 The Regression Equation

17. Table 12.16 below represents the relationship between the number of hours spent studying and final exam grades.

x (number of hours spent studying)	*y* (final exam grades)
3	50
5	72
1	45
2	51
6	80
8	96
4	65
7	90

Table 12.14

Fill in the following chart as a first step in finding the line of best fit, using the median–median approach.

Group	*x* (no. of hours spent studying)	*y* (final exam grades)	Median *x* Value	Median *y* Value
1				
2				
3				

Table 12.15

Use the following information to answer the next five exercises. A random sample of 10 professional athletes produced the following data, where *x* is the number of endorsements the player has and *y* is the amount of money made, in millions of dollars.

x	*y*	*x*	*y*
0	2	5	12
3	8	4	9
2	7	3	9
1	3	0	3
5	13	4	10

Table 12.16

18. Draw a scatter plot of the data.

19. Use regression to find the equation for the line of best fit.

20. Draw the line of best fit on the scatter plot.

21. What is the slope of the line of best fit? What does it represent?

22. What is the *y*-intercept of the line of best fit? What does it represent?

23. What does an *r* value of zero mean?

24. When $n = 2$ and $r = 1$, are the data significant? Explain.

25. When $n = 100$ and $r = –0.89$, is there a significant correlation? Explain.

12.3 Testing the Significance of the Correlation Coefficient (Optional)

26. When testing the significance of the correlation coefficient, what is the null hypothesis?

27. When testing the significance of the correlation coefficient, what is the alternative hypothesis?

28. If the level of significance is 0.05 and the *p*-value is 0.04, what conclusion can you draw?

12.4 Prediction (Optional)

Use the following information to answer the next two exercises. An electronics retailer used regression to find a simple model to predict sales growth in the first quarter of the new year (January through March). The model is good for 90 days, where x is the day. The model can be written as $\hat{y} = 101.32 + 2.48x$, where $\hat{y}$ is in thousands of dollars.

29. What would you predict the sales to be on day 60?

30. What would you predict the sales to be on day 90?

Use the following information to answer the next three exercises. A landscaping company is hired to mow the grass for several large properties. The total area of the properties is 1,345 acres. The rate at which one person can mow is $\hat{y} = 1350 – 1.2x$, where x is the number of hours and $\hat{y}$ represents the number of acres left to mow.

31. How many acres are left to mow after 20 hours of work?

32. How many acres are left to mow after 100 hours of work?

33. How many hours does it take to mow all the lawns, or when is $\hat{y} = 0$?

Use the following information to answer the next 14 exercises. **Table 12.17** contains real data for the first two decades of flu reporting.

Year	Number of Flu Cases Diagnosed	Number of Flu Deaths
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510

Table 12.17 Adults and Adolescents Only, United States

1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Table 12.17 Adults and Adolescents Only, United States

34. Graph year versus number of flu cases diagnosed (plot the scatter plot). Do not include pre-1981 data.

35. Perform a linear regression. What is the linear equation? Round to the nearest whole number. Find the following:

Write the equations:

- *Linear equation:* _____
- a = _____
- b = _____
- r = _____
- n = _____

36. Solve.

a. When $x = 1985$, $\hat{y}$ = _____.
b. When $x = 1990$, $\hat{y}$ = _____.
c. When $x = 1970$, $\hat{y}$ = _____. Why doesn't this answer make sense?

37. Does the line seem to fit the data? Why or why not?

38. What does the correlation imply about the relationship between time (years) and the number of diagnosed flu cases reported in the United States?

39. Plot the two points on the graph. Then, connect the two points to form the regression line.

40. Write the equation: $\hat{y}$ = _____.

41. Hand-draw a smooth curve on the graph that shows the flow of the data.

42. Does the line seem to fit the data? Why or why not?

43. Do you think a linear fit is best? Why or why not?

44. What does the correlation imply about the relationship between time (years) and the number of diagnosed flu cases reported in the United States?

45. Graph *year* vs. *number flu cases diagnosed*. Do not include pre-1981. Label both axes with words. Scale both axes.

46. Enter your data into your calculator or computer. The pre-1981 data should not be included. Why is that so?

Write the linear equation, rounding to four decimal places.

47. *Calculate the following:*

- a = _____
- b = _____
- correlation = _____
- n = _____

12.5 Outliers

48. Marcus states that all outliers are influential points. Is he correct? Explain.

Use the following information to answer the next four exercises. The scatter plot shows the relationship between hours spent studying and exam scores. The line shown is the calculated line of best fit. The correlation coefficient is 0.69.

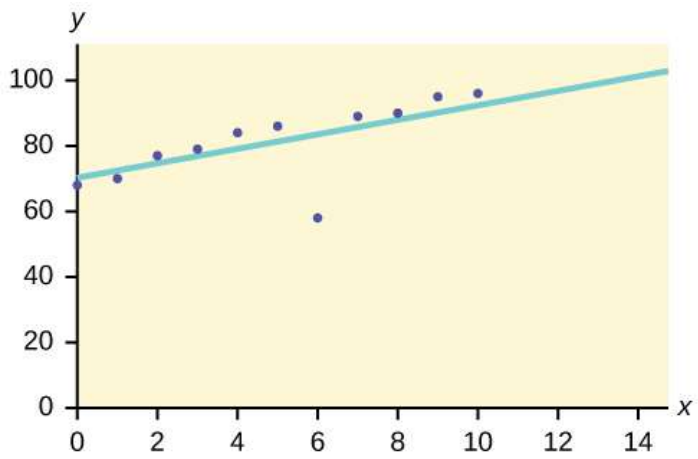


Figure 12.24

49. Do there appear to be any outliers?

50. A point is removed and the line of best fit is recalculated. The new correlation coefficient is 0.98. Does the point appear to have been an outlier? Why?

51. What effect did the potential outlier have on the line of best fit?

52. Are you more or less confident in the predictive ability of the new line of best fit?

53. The sum of squared errors (SSE) for a data set of 18 numbers is 49. What is the standard deviation?

54. The standard deviation for the SSE for a data set is 9.8. What is the cutoff for the vertical distance that a point can be from the line of best fit to be considered an outlier?

HOMEWORK

12.1 Linear Equations

55. For each of the following situations, state the independent variable and the dependent variable.

a. A study is done to determine whether elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared with the age of drivers.
b. A study is done to determine whether the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine whether a higher education reduces the crime rate in a population.

56. Piece-rate systems are widely debated incentive payment plans. In a recent study of loan officer effectiveness, the following piece-rate system was examined:

% of goal reached	< 80	80	100	120
Incentive	n/a	$4,000, with an additional $125 added per percentage point from 81% to 99%	$6,500, with an additional $125 added per percentage point from 101% to 119%	$9,500, with an additional $125 added per percentage point starting at 121%

Table 12.18

If a loan officer makes 95 percent of his or her goal, write the linear function that applies based on the incentive plan table. In context, explain the *y*-intercept and slope.

12.2 The Regression Equation

57. What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?

58. Explain what it means when a correlation has an r^2 value of .72.

59. Can a coefficient of determination be negative? Why or why not?

60. The table below represents the relationship between SAT scores on the math portion of the test and high school grade point averages (GPAs).

Use the median–-median line approach to find the equation for the line of best fit.

***x* (SAT math scores)**	***y* (GPAs)**
624	90
544	86
363	70
373	71
350	65
741	98
262	60
587	87
327	62
364	67
261	50

Table 12.19

12.4 Prediction (Optional)

61. Recently, the annual numbers of driver deaths per 100,000 people for the selected age groups are as follows:

Age (years)	Number of Driver Deaths (per 100,000 people)
16–19	38
20–24	36
25–34	24
35–54	20
55–74	18
75+	28

Table 12.20

a. For each age group, pick the midpoint of the interval for the x value. For the 75+ group, use 80.
b. Using *age* as the independent variable and *number of driver deaths per 100,000 people* as the dependent variable, make a scatter plot of the data.
c. Calculate the least-squares (best–fit) line. Put the equation in the form $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. Predict the number of deaths for ages 40 years and 60 years.
f. Based on the given data, is there a linear relationship between age of a driver and driver fatality rate?
g. What is the slope of the least-squares (best-fit) line? Interpret the slope.

62. **Table 12.21** shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy in years
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

Table 12.21

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
f. Why aren't the answers to Part E the same as the values in **Table 12.21** that correspond to those years?
g. Use the two points in Part E to plot the least-squares line on your graph from Part B.
h. Based on the data, is there a linear relationship between the year of birth and life expectancy?
i. Are there any outliers in the data?
j. Using the least-squares line, find the estimated life expectancy for an individual born in 1850. Does the least-squares line give an accurate estimate for that year? Explain why or why not.
k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

63. The maximum discount value of the Entertainment® card for the *Fine Dining* section, 10th edition, for various pages is given in **Table 12.22**.

Page Number	Maximum Value ($)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

Table 12.22

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. Find the estimated maximum values for the restaurants on page 10 and on page 70.
f. Does it appear that the restaurants giving the maximum value are placed in the beginning of the *Fine Dining* section? How did you arrive at your answer?
g. Suppose there are 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
h. Is the least-squares line valid for page 200? Why or why not?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

64. **Table 12.23** gives the gold medal times for every other Summer Olympics for the women's 100-meter freestyle in swimming.

Year	Time in seconds
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64
2000	53.8
2008	53.1

Table 12.23

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is the decrease in times significant?
f. Find the estimated gold medal time for 1932. Find the estimated time for 1984.
g. Why are the answers from Part F different from the chart values?
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Use the least-squares line to estimate the gold medal time for the next Summer Olympics. Do you think your answer is reasonable? Why or why not?

65.

State	No. of Letters in Name	Year Entered the Union	Rank for Entering the Union	Area in square miles
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

Table 12.24

We are interested in whether the number of letters in a state name depends on the year the state entered the Union.

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated number of letters (to the nearest integer) a state name would have if it entered the Union in 1900. Find the estimated number of letters a state name would have if it entered the Union in 1940.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Use the least-squares line to estimate the number of letters for a new state that enters the Union this year. Can the least-squares line be used to predict it? Why or why not?

12.5 Outliers

66. Given the information in Table 12.30, which represents the relationship between final exam math grades and final exam history grades, decide whether point (56, 95) is an influential point. Explain how you arrived at your decision. Show all work.

x (final exam math grades)	*y* (final exam history grades)
54	60
56	68
77	82
74	78
63	69
51	55
88	97
72	77
69	78
56	95

Table 12.25

67. In Table 12.31, the height (sidewalk to roof) of notable tall buildings in America is compared with the number of stories of the building (beginning at street level).

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

Table 12.26

a. Using *stories* as the independent variable and *height* as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables?
c. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. Find the estimated heights for a building that has 32 stories and for a building that has 94 stories.
f. Based on the data in **Table 12.26**, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
g. Are there any outliers in the data? If so, which point(s)?
h. What is the estimated height of a building with six stories? Does the least-squares line give an accurate estimate of height? Explain why or why not.
i. Based on the least-squares line, adding an extra story is predicted to add about how many feet to a building?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

68. Ornithologists (scientists who study birds) tag sparrow hawks in 13 different colonies to study their population. They gather data for the percentage of new sparrow hawks in each colony and the percentage of those that have returned from migration.

Percent return: 74, 66, 81, 52, 73, 62, 52, 45, 62, 46, 60, 46, 38
Percent new: 5, 6, 8, 11, 12, 15, 16, 17, 18, 18, 19, 20, 20

a. Enter the data into a calculator and make a scatter plot.
b. Use the calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from Part A.
c. Explain what the slope and *y*-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. An ecologist wants to predict how many birds will join another colony of sparrow hawks to which 70 percent of the adults from the previous year have returned. What is the prediction?

69. The following table shows data on average per capita coffee consumption and death rate from heart disease in a random sample of 10 countries.

Yearly Coffee Consumption (liters)	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
No. of Deaths from Heart Disease	221	167	131	191	220	297	71	172	211	300

Table 12.27

a. Enter the data into a calculator and make a scatter plot.
b. Use the calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from Part A.
c. Explain what the slope and *y*-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. Do the data provide convincing evidence that there is a linear relationship between the amount of coffee consumed and the heart disease death rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

70. The following table consists of one student athlete's time (in minutes) to swim 2,000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days.

Swim Time	**Heart Rate**
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

Table 12.28

a. Enter the data into a calculator and make a scatter plot.
b. Use the calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from Part A.
c. Explain what the slope and *y*-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

71. A researcher is investigating whether population impacts homicide rate. He uses demographic data from Detroit, Michigan, to compare homicide rates and the population.

Population Size	Homicide Rate per 100,000 People
558,724	8.6
538,584	8.9
519,171	8.52
500,457	8.89
482,418	13.07
465,029	14.57
448,267	21.36
432,109	28.03
416,533	31.49
401,518	37.39
387,046	46.26
373,095	47.24
359,647	52.33

Table 12.29

a. Use a calculator to construct a scatter plot of the data. What is the independent variable? Why?
b. Use the calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot.
c. Discuss what the following mean in context:
    i. The slope of the regression equation
    ii. The y-intercept of the regression equation
    iii. The correlation coefficient, r
    iv. The coefficient of determination, r^2
d. Do the data provide convincing evidence that there is a linear relationship between population size and homicide rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

72.

School	Mid-Career Salary (in thousands of U.S. dollars)	Yearly Tuition (in U.S. dollars)
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
U.S. Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Table 12.30

Use the data in the Table 12.35 to determine the linear regression line equation with the outliers removed. Is there a linear correlation for the data set with outliers removed? Justify your answer.

REFERENCES

12.1 Linear Equations

Centers for Disease Control and Prevention. (n.d.). Retrieved from https://www.cdc.gov/

National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. (n.d.). Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/nchhstp/default.htm

12.4 Prediction (Optional)

Centers for Disease Control and Prevention. (n.d.). Retrieved from https://www.cdc.gov/

National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention. (n.d.). Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/nchhstp/default.htm

National Center for Health Statistics. (n.d.). Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/nchs/index.htm

U.S. Census Bureau. (n.d.). Retrieved from http://www.census.gov/compendia/statab/cats/transportation/motor_vehicle_accidents_and_fatalities.html

12.5 Outliers

Committee on Ways and Means, U.S. House of Representatives. (n.d.). Washington, DC: U.S. Department of Health and Human Services.

Microsoft Bookshelf. (n.d.).

Physician's Desk Reference Staff. (1990). *Physician's desk reference*. Ohio: Medical Economics Company.

U.S. Bureau of Labor Statistics. (n.d.). Retrieved from https://www.bls.gov/

BRINGING IT TOGETHER: HOMEWORK

73. The average number of people in a family who attended college for various years is given in **Table 12.31**.

Year	No. of Family Members Attending College
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

Table 12.31

a. Using *year* as the independent variable and *number of family members attending college* as the dependent variable, draw a scatter plot of the data.
b. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
c. Does the *y*-intercept, *a*, have any meaning here?
d. Find the correlation coefficient. Is it significant?
e. Pick two years between 1969 and 1991 and find the estimated number of family members attending college.
f. Based on the data in **Table 12.31**, is there a linear relationship between the year and the average number of family members attending college?
g. Using the least-squares line, estimate the number of family members attending college for 1960 and 1995. Does the least-squares line give an accurate estimate for those years? Explain why or why not.
h. Are there any outliers in the data?
i. What is the estimated average number of family members attending college for 1986? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

74. The percent of female wage and salary workers who are paid hourly rates is given in **Table 12.32** for the years 1979 to 1992.

Year	Percent of Workers Paid Hourly Rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

Table 12.32

a. Using *year* as the independent variable and *percent of workers paid hourly rates* as the dependent variable, draw a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
c. Does the *y*-intercept, *a*, have any meaning here?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated percentages for 1991 and 1988.
g. Based on the data, is there a linear relationship between the year and the percentage of female wage and salary earners who are paid hourly rates?
h. Are there any outliers in the data?
i. What is the estimated percentage for the year 2050? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Use the following information to answer the next two exercises. The cost of a leading liquid laundry detergent in different sizes is given in **Table 12.33**.

Size (ounces)	Cost ($)	Cost per Ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

Table 12.33

75.

a. Using *size* as the independent variable and *cost* as the dependent variable, draw a scatter plot.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
c. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
d. Find the correlation coefficient. Is it significant?
e. If the laundry detergent were sold in a 40 oz. size, what is the estimated cost?
f. If the laundry detergent were sold in a 90 oz. size, what is the estimated cost?
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the given data?
i. Is the least-squares line valid for predicting what a 300 oz. size of the laundry detergent would cost? Why or why not?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

76.

a. Complete **Table 12.33** for the cost per ounce of the different sizes of laundry detergent.
b. Using *size* as the independent variable and *cost per ounce* as the dependent variable, draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. If the laundry detergent were sold in a 40 oz. size, what is the estimated cost per ounce?
g. If the laundry detergent were sold in a 90 oz. size, what is the estimated cost per ounce?
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the the data?
j. Is the least-squares line valid for predicting what a 300 oz. size of the laundry detergent would cost per ounce? Why or why not?
k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

77. According to a flyer published by Prudential Insurance Company, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate ($)	Approximate Probate Fees and Taxes ($)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

Table 12.34

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated total cost for a net taxable estate of $1,000,000. Find the cost for $2,500,000.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. Based on these results, what would be the probate fees and taxes for an estate that does not have any assets?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

78. The following are advertised sale prices of color televisions at Anderson's:

Size (inches)	Sale Price ($)
9	147
20	197
27	297
31	447
35	1,177
40	2,177
60	2,497

Table 12.35

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated sale price for a 32-inch television. Find the cost for a 50-inch television.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

79. **Table 12.36** shows the average heights for American boys in 1990.

Age (years)	Height (centimeters)
Birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

Table 12.36

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is it significant?
f. Find the estimated average height for a 1-year-old. Find the estimated average height for an 11-year-old.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Are there any outliers in the data?
i. Use the least-squares line to estimate the average height for a 62-year-old man. Do you think that your answer is reasonable? Why or why not?
j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

80.

State	No. of Letters in Name	Year Entered the Union	Rank for Entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

Table 12.37

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

a. What are the independent and dependent variables?
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
g. Use the two points in Part F to plot the least-squares line on your graph from Part B.
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers?
j. Use the least-squares line to estimate the area of a new state that enters the Union. Can the least-squares line be used to predict it? Why or why not?
k. Delete *Hawaii* and substitute *Alaska* for it. Alaska is a state with an area of 656,424 square miles.
l. Calculate the new least-squares line.
m. Find the estimated area for Alabama. Is it closer to the actual area with this new least-squares line or with the previous one that included Hawaii? Why do you think that's the case?
n. Do you think that, in general, newer states are larger than the original states?

SOLUTIONS

1 dependent variable: fee amount independent variable: time

3

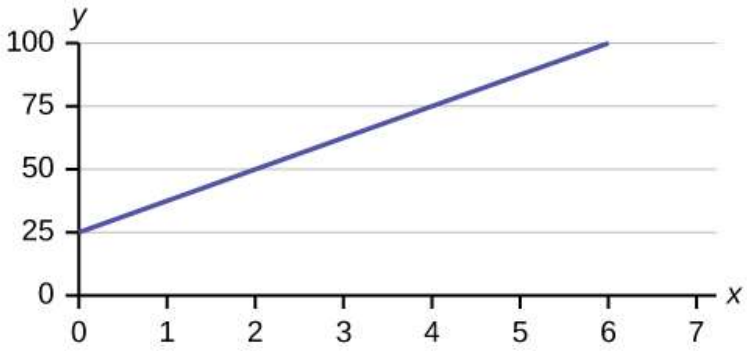


Figure 12.25

5

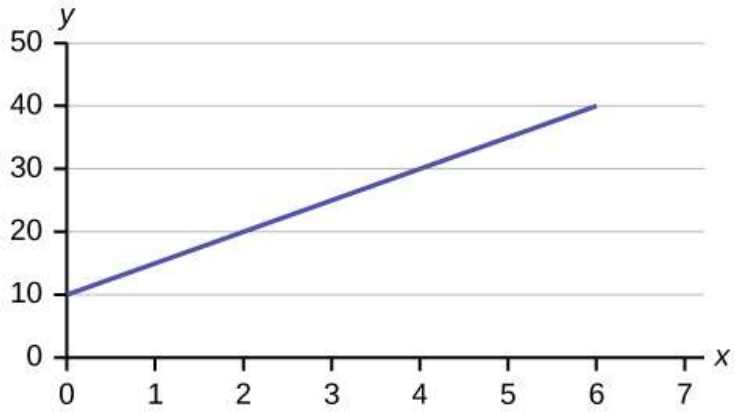


Figure 12.26

7 $y = 6x + 8$, $4y = 8$, and $y + 7 = 3x$ are all linear equations.

9 The number of flu cases depends on the year. Therefore, year becomes the independent variable and the number of flu cases is the dependent variable.

11 The y-intercept is 50 ($a = 50$). At the start of the cleaning, the company charges a one-time fee of \$50 (this is when $x = 0$). The slope is 100 ($b = 100$). For each session, the company charges \$100 for each hour they clean.

13 12,000 lb of soil

15 The slope is –1.5 ($b = –1.5$). This means the stock is losing value at a rate of \$1.50 per hour. The y-intercept is \$15 ($a = 15$). This means the price of stock before the trading day was \$15.

17

Group	*x* (no. of hours spent studying)	*y* (final exam grades)	Median *x* value	Median *y* value
1	1 2 3	45 50 51	2	50
2	4 5	65 72	4.5	68.5
3	6 7 8	80 90 96	7	90

Table 12.38

19 $\hat{y} = 2.23 + 1.99x$

21 The slope is 1.99 ($b = 1.99$). It means that for every endorsement deal a professional player gets, he gets an average of another $1.99 million in pay each year.

23 It means that there is no correlation between the data sets.

25 Yes. There are enough data points and the value of r is strong enough to show there is a strong negative correlation between the data sets.

27 H_a: $\rho \neq 0$

29 $250,120

31 1326 acres

33 1125 hours, or when $x = 1125$

35 Check student solution.

36

a. When $x = 1985$, $\hat{y} = 25{,}52$.

b. When $x = 1990$, $\hat{y} = 34{,}275$.

c. When $x = 1970$, $\hat{y} = –725$. Why doesn't this answer make sense? The range of x values was 1981 to 2002; the year 1970 is not in this range. The regression equation does not apply, because predicting for the year 1970 is extrapolation, which requires a different process. Also, a negative number does not make sense in this context, when we are predicting flu cases diagnosed.

38 Also, the correlation $r = 0.4526$. If r is compared with the value in the 95 Percent Critical Values of the Sample Correlation Coefficient Table, because $r > 0.423$, r is significant, and you would think that the line could be used for prediction. But, the scatter plot indicates otherwise.

39 Check student' solution.

40 $\hat{y} = 3{,}448{,}225 + 1750x$

42 There was an increase in flu cases diagnosed until 1993. From 1993 through 2002, the number of flu cases diagnosed declined each year. It is not appropriate to use a linear regression line to fit to the data.

44 Because there is no linear association between year and number of flu cases diagnosed, it is not appropriate to calculate a linear correlation coefficient. When there is a linear association and it is appropriate to calculate a correlation, we cannot say that one variable *causes* the other variable.

46 We don't know if the pre-1981 data were collected from a single year. So, we don't have an accurate x value for this figure. Regression equation: $\hat{y}$ (number of flu cases) = –3,448,225 + 1749.777 (year).

	Coefficients
Intercept	–3,448,225
x Variable 1	1,749.777

Table 12.39

47

- $a = -3{,}488{,}225$
- $b = 1{,}750$
- correlation = 0.4526
- $n = 22$

48 No, he is not correct. An outlier is only an influential point if it significantly impacts the slope of the least-squares regression line and the correlation coefficient, *r*. If omission of this data point from the calculation of the regression line does not show much impact on the slope or *r*-value, then the outlier is not considered an influential point. For different reasons, it still may be determined that the data point must be omitted from the data set.

49 Yes. There appears to be an outlier at (6, 58).

51 The potential outlier flattened the slope of the line of best fit because it was below the data set. It made the line of best fit less accurate as a predictor for the data.

53 $s = 1.75$

55

a. independent variable: age; dependent variable: fatalities

b. independent variable: number of family members; dependent variable: grocery bill

c. independent variable: age of applicant; dependent variable: insurance premium

d. independent variable: power consumption; dependent variable: utility

e. independent variable: higher education (years); dependent variable: crime rates

58 It means that 72 percent of the variation in the dependent variable (y) can be explained by the variation in the independent variable (x).

60

x (SAT math scores)	*y* (GPAs)
261	50
262	60
327	62
350	65
363	70
364	67
373	71
544	86
587	87
624	90
741	98

Table 12.40

We must remember to check the order of the *y* values within each group as well. We notice that the *y* values in the second group are not in order from the least value to the greatest value; these values thus must be reordered, meaning the median *y* value for that group is 70.

Group	*x* (SAT math scores)	*y* (GPAs)	Median *x* value	Median *y* value
1	261 262 327 350	50 60 62 65	294.5	61
2	363 364 373	67 70 71	364	70
3	544 587 624 741	86 87 90 98	605.5	88.5

Table 12.41

The ordered pairs are (294.5, 61), (364, 70), and (605.5, 88.5). The slope can be calculated using the formula $m = \frac{y_3 - y_1}{x_3 - x_1}$. Substituting the median x and y values, from the first and third groups gives $m = \frac{88.5 - 61}{605.5 - 294.5}$, which simplifies to $m \approx 0.09$. The y-intercept may be found using the formula $b = \frac{\sum y - m\sum x}{3}$. The sum of the median x values is 1264, and the sum of the median y values is 219.5. Substituting these sums and the slope into the formula gives $b = \frac{219.5 - 0.09(1264)}{3}$, which simplifies to $b \approx 35.25$. The line of best fit is represented as $y = mx + b$. Thus, the equation can be written as $y = 0.09x + 35.25$.

61 b. Check student solution.

c. $\hat{y} = 35.5818045 – 0.19182491x$

d. $r = –0.57874$
For four degrees of freedom and alpha = 0.05, the LinRegTTest gives a p value of 0.2288, so we do not reject the null hypothesis; there is not a significant linear relationship between deaths and age.
Using the table of critical values for the correlation coefficient, with four degrees of freedom, the critical value is 0.811. The correlation coefficient $r = –0.57874$ is not less than –0.811, so we do not reject the null hypothesis.

f. There is not a linear relationship between the two variables, as evidenced by a p value greater than 0.05.

63 a. We wonder if the better discounts appear earlier in the book, so we select page as x and discount as y.

b. Check student solution.

c. $\hat{y} = 17.21757 – 0.01412x$

d. $r = – 0.2752$
For seven degrees of freedom and alpha = 0.05, LinRegTTest gives a p value = 0.4736, so we do not reject; there is a not a significant linear relationship between page and discount.
Using the table of critical values for the correlation coefficient, with seven gives degrees of freedom, the critical value is 0.666. The correlation coefficient $xi = –0.2752$ is not less than 0.666, so we do not reject.

f. There is not a significant linear correlation so it appears there is no relationship between the page and the amount of the discount. As the page number increases by one page, the discount decreases by $0.01412.

65 a. Year is the independent or x variable; the number of letters is the dependent or y variable.

b. Check student's solution.

c. No.

d. $\hat{y} = 47.03 – 0.0216x$

e. –0.4280. The r value indicates that there is not a significant correlation between the year the state entered the Union and the number of letters in the name.

g. No. The relationship does not appear to be linear; the correlation is not significant.

66 Using LinRegTTest, the output for the original least-squares regression line is $y = 26.14 + 0.7539x$ and $r = 0.6657$. The output for the new least-squares regression line, after omitting the outlier of (56, 95), is $\hat{y} = 6.36 + 1.0045x$ and $r = 0.9757$. The slope of the new line is quite a bit different from the slope of the original least-squares regression line, but the larger change is shown in the r-values, such that the new line has an r-value that has increased to a value that is almost equal to one. Thus, it may be stated that the outlier (56, 95) is also an influential point.

68 a. and b. Check student solution. c. The slope of the regression line is –0.3031 with a y-intercept of 31.93. In context, the y-intercept indicates that when there are no returning sparrow hawks, there will be almost 32 percent new sparrow hawks, which doesn't make sense, because if there are no returning birds, then the new percentage would have to be 100% (this is an example of why we do not extrapolate). The slope tells us that for each percentage increase in returning birds, the percentage of new birds in the colony decreases by 30.3 percent. d. If we examine r_2, we see that only 57.52 percent of the variation in the percentage of new birds is explained by the model and the correlation coefficient, $r = –.7584$ only indicates a somewhat strong correlation between returning and new percentages. e. The ordered pair (66, 6) generates the largest residual of 6.0. This means that when the observed return percentage is 66 percent, our observed new percentage, 6 percent, is almost 6 percent less than the predicted new value of 11.98 percent. If we remove this data pair, we see only an adjusted slope of –0.2789 and an adjusted intercept of 30.9816. In other words, although these data generate the largest residual, it is not an outlier, nor is the data pair an influential point. f. If there are 70 percent returning birds, we would expect to see $y =– 0.2789(70) + 30.9816 = 0.114$ or 11.4 percent new birds in the colony.

70

a. Check student solution.

b. Check student solution.

c. We have a slope of –1.4946 with a *y*-intercept of 193.88. The slope, in context, indicates that for each additional minute added to the swim time, the heart rate decreases by 1.5 beats per minute. If the student is not swimming at all, the *y*-intercept indicates that his heart rate will be 193.88 beats per minute. Although the slope has meaning (the longer it takes to swim 2000 m, the less effort the heart puts out), the *y*-intercept does not make sense. If the athlete is not swimming (resting), then his heart rate should be very low.

d. Because only 1.5 percent of the heart rate variation is explained by this regression equation, we must conclude that this association is not explained with a linear relationship.

e. Point (34.72, 124) generates the largest residual: –11.82. This means that our observed heart rate is almost 12 beats less than our predicted rate of 136 beats per minute. When this point is removed, the slope becomes –2.953, with the *y*-intercept changing to 247.1616. Although the linear association is still very weak, we see that the removed data pair can be considered an influential point in the sense that the *y*-intercept becomes more meaningful.

72 If we remove the two service academies (the tuition is $0.00), we construct a new regression equation of $y = –0.0009x + 160$, with a correlation coefficient of 0.71397 and a coefficient of determination of 0.50976. This allows us to say there is a fairly strong linear association between tuition costs and salaries if the service academies are removed from the data set.

73 c. No. The *y*-intercept would occur at year 0, which doesn't exist.

74

a. Check student's solution.

b. Yes.

c. No, the *y*-intercept would occur at year 0, which doesn't exist.

d. $\hat{y} = -266.8863 + 0.1656x$.

e. 0.9448, yes.

f. 62.8233, 62.3265.

g. Yes.

h. No, (1987, 62.7).

i. 72.5937, no.

j. Slope = 0.1656. As the year increases by one, the percent of workers paid hourly rates tends to increase by 0.1656.

76

a.

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	24.94
32	4.99	15.59
64	5.99	9.36
200	10.99	5.50

Table 12.42

b. Check student solution.

c. There is a linear relationship for the sizes 16 through 64, but that linear trend does not continue to the 200-oz size.

d. $\hat{y} = 20.2368 – 0.0819x$

e. $r = –.8086$

f. 40-oz: 16.96 cents/oz

g. 90-oz: 12.87 cents/oz

h. The relationship is not linear; the least-squares line is not appropriate.

i. There are no outliers.

j. No. You would be extrapolating. The 300-oz size is outside the range of x.

k. X = –0.08194. For each additional ounce in size, the cost per ounce decreases by 0.082 cents.

78

a. Size is x, the independent variable, and price is y, the dependent variable.

b. Check student solution.

c. The relationship does not appear to be linear.

d. $\hat{y} = –745.252 + 54.75569x$.

e. $r = .8944$ and yes, it is significant.

f. 32-inch: $1006.93, 50-inch: $1992.53.

g. No, the relationship does not appear to be linear. However, r is significant.

h. No, the 60-inch TV.

i. For each additional inch, the price increases by $54.76.

80

a. Rank is the independent variable and area is the dependent variable.

b. Check student solution.

c. There appears to be a linear relationship, with one outlier.

d. $\hat{y}$ (area) = 24177.06 + 1010.478x

e. r = .50047. r is not significant, so there is no relationship between the variables.

f. Alabama: 46,407.576 square miles, Colorado: 62,575.224 square miles.

g. The Alabama estimate is closer than the Colorado estimate.

h. If the outlier is removed, there is a linear relationship.

i. There is one outlier (Hawaii).

j. rank 51: 75,711.4 square miles, no.

k.

Alabama	**7**	**1819**	**22**	**52,423**
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

Table 12.43

l. $\hat{y}$ = –87065.3 + 7828.532x.

m. Alabama: 85,162.404; the prior estimate was closer. Alaska is an outlier.

n. Yes, with the exception of Hawaii.

73 c. No. The y-intercept would occur at year 0, which doesn't exist.

74

a. Check student's solution.

b. Yes.

c. No, the y-intercept would occur at year 0, which doesn't exist.

d. $\hat{y}$ = −266.8863 + 0.1656x.

e. 0.9448, yes.

f. 62.8233, 62.3265.

g. Yes.

h. No, (1987, 62.7).

i. 72.5937, no.

j. Slope = 0.1656. As the year increases by one, the percent of workers paid hourly rates tends to increase by 0.1656.

76

a.

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	24.94
32	4.99	15.59
64	5.99	9.36
200	10.99	5.50

Table 12.44

b. Check student solution.

c. There is a linear relationship for the sizes 16 through 64, but that linear trend does not continue to the 200-oz size.

d. $\hat{y} = 20.2368 – 0.0819x$

e. $r = –.8086$

f. 40-oz: 16.96 cents/oz

g. 90-oz: 12.87 cents/oz

h. The relationship is not linear; the least-squares line is not appropriate.

i. There are no outliers.

j. No. You would be extrapolating. The 300-oz size is outside the range of x.

k. X = –0.08194. For each additional ounce in size, the cost per ounce decreases by 0.082 cents.

78

a. Size is x, the independent variable, and price is y, the dependent variable.

b. Check student solution.

c. The relationship does not appear to be linear.

d. $\hat{y} = –745.252 + 54.75569x$.

e. $r = .8944$ and yes, it is significant.

f. 32-inch: $1006.93, 50-inch: $1992.53.

g. No, the relationship does not appear to be linear. However, r is significant.

h. No, the 60-inch TV.

i. For each additional inch, the price increases by $54.76.

80

a. Rank is the independent variable and area is the dependent variable.

b. Check student solution.

c. There appears to be a linear relationship, with one outlier.

d. $\hat{y}$ (area) $= 24177.06 + 1010.478x$

e. $r = .50047$. r is not significant, so there is no relationship between the variables.

f. Alabama: 46,407.576 square miles, Colorado: 62,575.224 square miles.

g. The Alabama estimate is closer than the Colorado estimate.

h. If the outlier is removed, there is a linear relationship.

i. There is one outlier (Hawaii).

j. rank 51: 75,711.4 square miles, no.

k.

Alabama	**7**	**1819**	**22**	**52,423**
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

Table 12.45

l. $\hat{y} = -87065.3 + 7828.532x$.

m. Alabama: 85,162.404; the prior estimate was closer. Alaska is an outlier.

n. Yes, with the exception of Hawaii.

13 | F DISTRIBUTION AND ONE-WAY ANOVA

Figure 13.1 One-way ANOVA is used to measure information from several groups.

Introduction

Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Interpret the *F* probability distribution as the number of groups and the sample size change
- Discuss two uses for the *F* distribution: one-way ANOVA and the test of two variances
- Conduct and interpret one-way ANOVA
- Conduct and interpret hypothesis tests of two variances

Many statistical applications in psychology, social science, business administration, and the natural sciences involve several groups. For example, an environmentalist is interested in knowing if the average amount of pollution varies among several bodies of water. A sociologist is interested in knowing if the amount of income a person earns varies according to his or her

upbringing. A consumer looking for a new car might compare the average gas mileage of several models.

For hypothesis tests comparing averages across more than two groups, statisticians have developed a method called *analysis of variance* (abbreviated ANOVA). In this chapter, you will study the simplest form of ANOVA called single factor or one-way ANOVA. You will also study the *F* distribution, used for one-way ANOVA, and the test of two variances. This is a very brief overview of one-way ANOVA. You will study this topic in much greater detail in future statistics courses. One-way ANOVA, as it is presented here, relies heavily on a calculator or computer.

13.1 | One-Way ANOVA

The purpose of a one-way ANOVA test is to determine the existence of a statistically significant difference among several group means. The test uses **variances** to help determine if the means are equal or not. To perform a one-way ANOVA test, there are five basic assumptions to be fulfilled:

- Each population from which a sample is taken is assumed to be normal.
- All samples are randomly selected and independent.
- The populations are assumed to have equal standard deviations (or variances).
- The factor is a categorical variable.
- The response is a numerical variable.

The Null and Alternative Hypotheses

The null hypothesis is that all the group population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are k groups

H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

H_a: At least two of the group means $\mu_1, \mu_2, \mu_3, ..., \mu_k$ are not equal. That is, $\mu_i \neq \mu_j$ for some $i \neq j$.

The graphs, a set of box plots representing the distribution of values with the group means indicated by a horizontal line through the box, help in the understanding of the hypothesis test. In the first graph (red box plots), H_0: $\mu_1 = \mu_2 = \mu_3$ and the three populations have the same distribution if the null hypothesis is true. The variance of the combined data is approximately the same as the variance of each of the populations.

If the null hypothesis is false, then the variance of the combined data is larger, which is caused by the different means as shown in the second graph (green box plots).

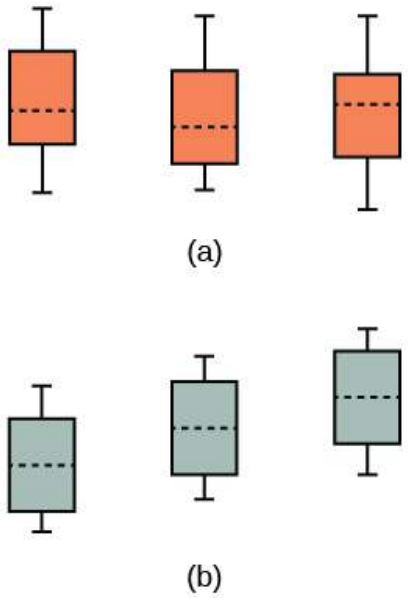


Figure 13.2 (a) We fail to reject H_0 as it may be true. All the means are about the same; the differences may be due to random variation. (b) We reject H_0 as all the means are not the same; the differences are too large to be due to random variation.

13.2 | The F Distribution and the F Ratio

The distribution used for the hypothesis test is a new one. It is called the ***F* distribution**, named after Sir Ronald Fisher, an English statistician. The *F* statistic is a ratio (a fraction). There are two sets of degrees of freedom: one for the numerator and one for the denominator.

For example, if *F* follows an *F* distribution and the number of degrees of freedom for the numerator is 4, and the number of degrees of freedom for the denominator is 10, then $F \sim F_{4,10}$.

NOTE

The *F* distribution is derived from the Student's *t*-distribution. The values of the *F* distribution are squares of the corresponding values of the *t*-distribution. One-way ANOVA expands the *t*-test for comparing more than two groups. The scope of that derivation is beyond the level of this course. It is preferable to use ANOVA when there are more than two groups instead of performing pairwise *t*-tests because performing multiple tests introduces the likelihood of making a Type 1 error.

To calculate the ***F* ratio**, two estimates of the variance are made.

1. Variance between samples: an estimate of σ^2 that is the variance of the sample means multiplied by n, when the sample sizes are the same. If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes. The variance is also called *variation due to treatment* or *explained variation.*
2. Variance within samples: an estimate of σ^2 that is the average of the sample variances, also known as a *pooled variance*. When the sample sizes are different, the variance within samples is weighted. The variance is also called the *variation due to error* or *unexplained variation.*

- $SS_{between}$ = the sum of squares that represents the variation among the different samples
- SS_{within} = the sum of squares that represents the variation within samples that is due to chance

To find a *sum of squares* mean, add together squared quantities which, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation in **Descriptive Statistics**.

MS means *mean square*. $MS_{between}$ is the variance between groups, and MS_{within} is the variance within groups.

Calculation of Sum of Squares and Mean Square

- k = the number of different groups
- n_j = the size of the j^{th} group
- s_j = the sum of the values in the j^{th} group
- n = total number of all the values combined (total sample size: $\sum n_j$)
- x = one value: $\sum x = \sum s_j$
- Sum of squares of all values from every group combined: $\sum x^2$
- Between group variability: $SS_{\text{total}} = \sum x^2 - \frac{\left(\sum x^2\right)}{n}$
- Total sum of squares: $\sum x^2 - \frac{(\sum x)^2}{n}$
- Explained variation: sum of squares representing variation among the different samples $SS_{(\text{between})} = \sum \left[\frac{(s_j)^2}{n_j}\right] - \frac{\left(\sum s_j\right)^2}{n}$
- Unexplained variation: sum of squares representing variation within samples due to chance $SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}}$
- dfs for different groups (dfs for the numerator): $df = k - 1$
- Equation for errors within samples (dfs for the denominator): $df_{\text{within}} = n - k$
- Mean square (variance estimate) explained by the different groups: $MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$
- Mean square (variance estimate) that is due to chance (unexplained): $MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$

MS_{between} and MS_{within} can be written as follows:

- $MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{SS_{\text{between}}}{k-1}$
- $MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{SS_{within}}{n-k}$

The one-way ANOVA test depends on the fact that MS_{between} can be influenced by population differences among means of the several groups. Since MS_{within} compares values of each group to its own group mean, the fact that group means might be different does not affect MS_{within}.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least two of the sample groups come from populations with different normal distributions. If the null hypothesis is true, MS_{between} and MS_{within} should both estimate the same value.

NOTE

The null hypothesis says that all the group population means are equal. The hypothesis of equal means implies that the populations have the same normal distribution because it is assumed that the populations are normal and that they have equal variances.

F Ratio or *F* Statistic

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

If MS_{between} and MS_{within} estimate the same value, following the belief that H_0 is true, then the F ratio should be approximately equal to 1. Mostly, just sampling errors would contribute to variations away from 1. As it turns out, MS_{between}

consists of the population variance plus a variance produced from the differences between the samples. MS_{within} is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, $MS_{between}$ will generally be larger than MS_{within}. Then the F ratio will be larger than 1. However, if the population effect is small, it is not unlikely that MS_{within} will be larger in a given sample.

The previous calculations were done with groups of different sizes. If the groups are the same size, the calculations simplify somewhat and the F ratio can be written as follows:

F Ratio formula when the groups are the same size

$$F = \frac{n \cdot s_{\bar{x}}^2}{s^2_{\text{pooled}}}$$

where

- n = the sample size
- $df_{\text{numerator}} = k - 1$
- $df_{\text{denominator}} = n - k$
- s^2 pooled = the mean of the sample variances (pooled variance)
- $s_{\bar{x}}^2$ = the variance of the sample means

Data is typically put into a table for easy viewing. One-way ANOVA results are often displayed in this manner by computer software.

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (Between)	*SS*(Factor)	$k - 1$	*MS*(Factor) = *SS*(Factor)/($k - 1$)	*F* = *MS*(Factor)/*MS*(Error)
Error (Within)	*SS*(Error)	$n - k$	*MS*(Error) = *SS*(Error)/($n - k$)	
Total	*SS*(Total)	$n - 1$		

Table 13.1

Example 13.1

Three different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The one-way ANOVA results are shown in **Table 13.2**.

Plan 1: $n_1 = 4$	Plan 2: $n_2 = 3$	Plan 3: $n_3 = 3$
5	3.5	8
4.5	7	4
4		3.5
3	4.5	

Table 13.2

$s_1 = 16.5$, $s_2 = 15$, $s_3 = 15.5$

Following are the calculations needed to fill in the one-way ANOVA table. The table is used to conduct a hypothesis test.

$$SS(between) = \sum\left[\frac{(s_j)^2}{n_j}\right] - \frac{\left(\sum s_j\right)^2}{n}$$

$$= \frac{s_1^2}{4} + \frac{s_2^2}{3} + \frac{s_3^2}{3} - \frac{(s_1 + s_2 + s_3)^2}{10}$$

where $n_1 = 4$, $n_2 = 3$, $n_3 = 3$, and $n = n_1 + n_2 + n_3 = 10$

$$= \frac{(16.5)^2}{4} + \frac{(15)^2}{3} + \frac{(15.5)^2}{3} - \frac{(16.5 + 15 + 15.5)^2}{10}$$

$$SS(between) = 2.2458$$

$$S(total) = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$= \left(5^2 + 4.5^2 + 4^2 + 3^2 + 3.5^2 + 7^2 + 4.5^2 + 8^2 + 4^2 + 3.5^2\right)$$

$$- \frac{(5 + 4.5 + 4 + 3 + 3.5 + 7 + 4.5 + 8 + 4 + 3.5)^2}{10}$$

$$= 244 - \frac{47^2}{10} = 244 - 220.9$$

$$SS(total) = 23.1$$

$$SS(within) = SS(total) - SS(between)$$

$$= 23.1 - 2.2458$$

$$SS(within) = 20.8542$$

Using the TI-83, 83+, 84, 84+ Calculator

One-way ANOVA Table: The formulas for *SS*(Total), *SS*(Factor) = *SS*(Between), and *SS*(Error) = *SS*(Within) as shown previously. The same information is provided by the TI calculator hypothesis test function `ANOVA` in `STAT TESTS` (syntax is `ANOVA(L1, L2, L3)` where L1, L2, L3 have the data from Plan 1, Plan 2, Plan 3, respectively).

Source of Variation	**Sum of Squares (*SS*)**	**Degrees of Freedom (*df*)**	**Mean Square (*MS*)**	***F***
Factor (Between)	*SS*(Factor) = *SS*(Between) = 2.2458	$k - 1$ = 3 groups – 1 = 2	*MS*(Factor) = *SS*(Factor)/(k – 1) = 2.2458/2 = 1.1229	F = *MS*(Factor)/*MS*(Error) = 1.1229/2.9792 = 0.3769
Error (Within)	*SS*(Error) = *SS*(Within) = 20.8542	$n - k$ = 10 total data – 3 groups = 7	*MS*(Error) = *SS*(Error)/(n – k) = 20.8542/7 = 2.9792	
Total	*SS*(Total) = 2.2458 + 20.8542 = 23.1	$n - 1$ = 10 total data – 1 = 9		

Table 13.3

Try It Σ

13.1 As part of an experiment to see how different types of soil cover would affect slicing tomato production, Marist College students grew tomato plants under different soil cover conditions. Groups of three plants each had one of the following treatments:

- Bare soil
- A commercial ground cover
- Black plastic
- Straw
- Compost

All plants grew under the same conditions and were the same variety. Students recorded the weight in grams of tomatoes produced by each of the $n = 15$ plants, as seen in **Table 13.4**.

Bare: $n_1 = 3$	Ground Cover: $n_2 = 3$	Plastic: $n_3 = 3$	Straw: $n_4 = 3$	Compost: $n_5 = 3$
2,625	5,348	6,583	7,285	6,277
2,997	5,682	8,560	6,897	7,818
4,915	5,482	3,830	9,230	8,677

Table 13.4

Create the one-way ANOVA table.

The one-way ANOVA hypothesis test is always right-tailed because larger F values are way out in the right tail of the F distribution curve and tend to make us reject H_0.

Notation

The notation for the F distribution is $F \sim F_{df(num),df(denom)}$,

where $df(num) = df_{\text{between}}$ and $df(denom) = df_{\text{within}}$.

The mean for the F distribution is $\mu = \frac{df(\text{denom})}{df(\text{denom}) - 2}$.

13.3 | Facts About the F Distribution

The following are facts about the F distribution:

- The curve is not symmetrical but skewed to the right.
- There is a different curve for each set of dfs.
- The F statistic is greater than or equal to zero.
- As the degrees of freedom for the numerator and for the denominator get larger, the curve approximates the normal.
- Other uses for the F distribution include comparing two variances and two-way analysis of variance. Two-way analysis is beyond the scope of this chapter.

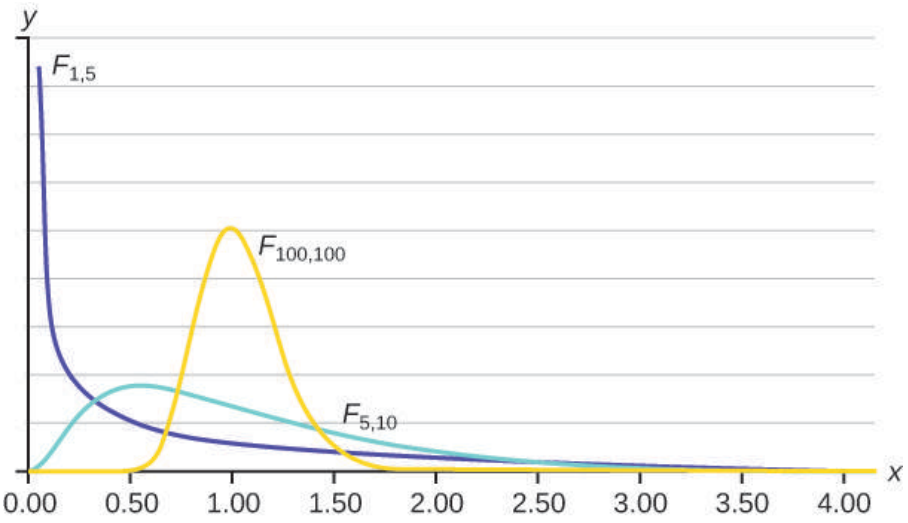


Figure 13.3

Example 13.2

Let's return to the slicing tomato exercise in **Try It**. The means of the tomato yields under the five mulching conditions are represented by μ_1, μ_2, μ_3, μ_4, μ_5. We will conduct a hypothesis test to determine if all means are the same or at least one is different. Using a significance level of 5 percent, test the null hypothesis that there is no difference in mean yields among the five groups against the alternative hypothesis that at least one mean is different from the rest.

Solution 13.2

The null and alternative hypotheses are as follows:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_a: $\mu_i \neq \mu_j$ for some $i \neq j$

The one-way ANOVA results are shown in **Table 13.4**

Source of Variation	**Sum of Squares (*SS*)**	**Degrees of Freedom (*df*)**	**Mean Square (*MS*)**	***F***
Factor (Between)	36,648,561	5 – 1 = 4	$\frac{36,648,561}{4} = 9,162,140$	$\frac{9,162,140}{2,044,672.6} = 4.4810$
Error (Within)	20,446,726	15 – 5 = 10	$\frac{20,446,726}{10} = 2,044,672.6$	
Total	57,095,287	15 – 1 = 14		

Table 13.5

Distribution for the test: $F_{4,10}$

$df(num) = 5 – 1 = 4$

$df(denom) = 15 – 5 = 10$

Test statistic: $F = 4.4810$

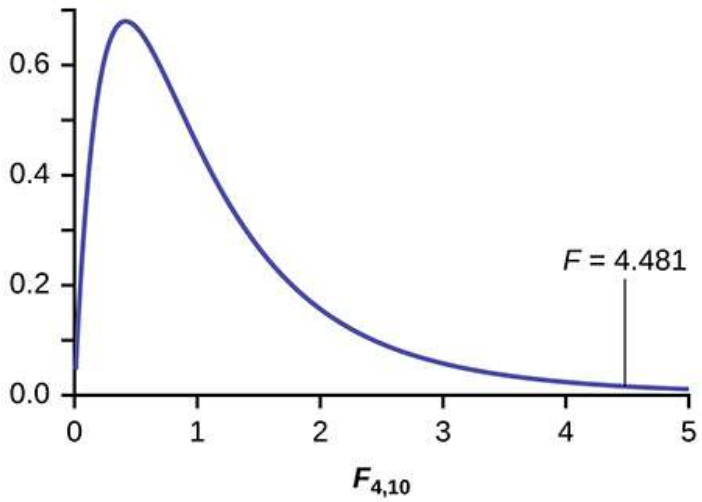


Figure 13.4

Probability statement: p-value = $P(F > 4.481) = 0.0248$

Compare α and the p-value: $\alpha = 0.05$, p-value = 0.0248

Make a decision: Since $\alpha > p$-value, we reject H_0.

Conclusion: At the 5 percent significance level, we have reasonably strong evidence that differences in mean yields for slicing tomato plants grown under different mulching conditions are unlikely to be due to chance alone. We may conclude that at least some of the mulches led to different mean yields.

Using the TI-83, 83+, 84, 84+ Calculator

To find these results on the calculator:

Press STAT. Press 1:EDIT. Put the data into the lists L1,L2,L3,L4,L5.

Press STAT, arrow over to TESTS, and arrow down to ANOVA. Press ENTER, and then enter (L1,L2,L3,L4,L5). Press ENTER. You will see that the values in the foregoing ANOVA table are easily produced by the calculator, including the test statistic and the p-value of the test.

The calculator displays:
$F = 4.4810$
$p = 0.0248$ (p-value)
Factor
$df = 4$
$SS = 36648560.9$
$MS = 9162140.23$
Error
$df = 10$
$SS = 20446726$
$MS = 2044672.6$

Try It

13.2 MRSA, or *Staphylococcus aureus*, can cause serious bacterial infections in hospital patients. **Table 13.6** shows various colony counts from different patients who may or may not have MRSA. The data from the table is plotted in **Figure 13.5**.

Conc = 0.6	Conc = 0.8	Conc = 1.0	Conc = 1.2	Conc = 1.4
9	16	22	30	27
66	93	147	199	168
98	82	120	148	132

Table 13.6

Plot of the data for the different concentrations:

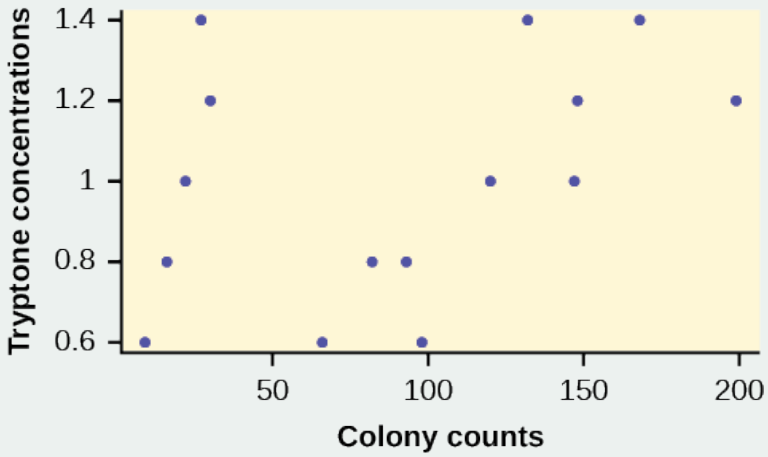


Figure 13.5

Test whether the mean numbers of colonies are the same or are different. Construct the ANOVA table by hand or by using a TI-83, 83+, or 84+ calculator, find the *p*-value, and state your conclusion. Use a 5 percent significance level.

Example 13.3

Four sororities took a random sample of sisters regarding their grade means for the past term. The results are shown in **Table 13.7**.

Sorority 1	Sorority 2	Sorority 3	Sorority 4
2.17	2.63	2.63	3.79
1.85	1.77	3.78	3.45
2.83	3.25	4.00	3.08
1.69	1.86	2.55	2.26
3.33	2.21	2.45	3.18

Table 13.7 Mean Grades for Four Sororities

Using a significance level of 1 percent, is there a difference in mean grades among the sororities?

Solution 13.3

Let μ_1, μ_2, μ_3, μ_4 be the population means of the sororities. Remember that the null hypothesis claims that the sorority groups are from the same normal distribution. The alternate hypothesis says that at least two of the sorority groups come from populations with different normal distributions. Notice that the four sample sizes are each five.

NOTE

This is an example of a *balanced design*, because each factor (i.e., sorority) has the same number of observations.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: Not all of the means μ_1, μ_2, μ_3, μ_4 are equal.

Distribution for the test: $F_{3,16}$

where $k = 4$ groups and $n = 20$ samples in total.

$df(num) = k - 1 = 4 - 1 = 3$

$df(denom) = n - k = 20 - 4 = 16$

Calculate the test statistic: $F = 2.23$

Graph

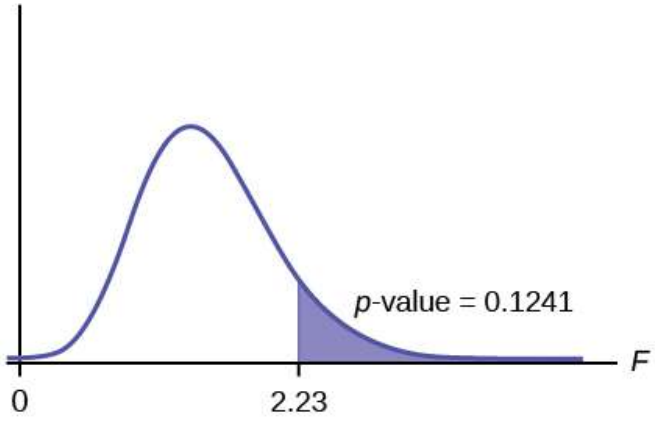


Figure 13.6

Probability statement: p-value = $P(F > 2.23) = 0.1241$

Compare α and the p-value: $\alpha = 0.01$
p-value = 0.1241
$\alpha < p$-value

Make a decision: Since $\alpha < p$-value, we cannot reject H_0.

Conclusion: There is not sufficient evidence to conclude that there is a difference among the mean grades for the sororities.

Using the TI-83, 83+, 84, 84+ Calculator

Put the data into lists L_1, L_2, L_3, and L_4. Press STAT and arrow over to TESTS. Arrow down to F:ANOVA. Press ENTER and enter (L1,L2,L3,L4).

The calculator displays the F statistic, the p-value, and the values for the one-way ANOVA table:
$F = 2.2303$
$p = 0.1241$ (p-value)
Factor
$df = 3$
$SS = 2.88732$
$MS = 0.96244$
Error
$df = 16$
$SS = 6.9044$
$MS = 0.431525$

Try It

13.3 Four sports teams took a random sample of players regarding their GPAs for the last year. The results are shown in **Table 13.8**.

Basketball	Baseball	Hockey	Lacrosse
3.6	2.1	4.0	2.0
2.9	2.6	2.0	3.6
2.5	3.9	2.6	3.9
3.3	3.1	3.2	2.7
3.8	3.4	3.2	2.5

Table 13.8 GPAs for four sports teams

Use a significance level of 5 percent and determine if there is a difference in GPA among the teams.

Example 13.4

A fourth-grade class is studying the environment. One of the assignments is to grow bean plants in different soils. Tommy chose to grow his bean plants in soil found outside his classroom mixed with dryer lint. Tara chose to grow her bean plants in potting soil bought at the local nursery. Nick chose to grow his bean plants in soil from his mother's garden. No chemicals were used on the plants, only water. They were grown inside the classroom next to a large window. Each child grew five plants. At the end of the growing period, each plant was measured, producing the data in inches in **Table 13.9**.

Tommy's Plants	Tara's Plants	Nick's Plants
24	25	23
21	31	27
23	23	22
30	20	30
23	28	20

Table 13.9

Does it appear that the three soils in which the bean plants were grown produce the same mean height? Test at a 3 percent level of significance.

Solution 13.4

This time, we will perform the calculations that lead to the F' statistic. Notice that each group has the same number of plants, so we will use the formula $F' = \frac{n \cdot s_{\bar{x}}^2}{s^2_{\text{pooled}}}$.

First, calculate the sample mean and sample variance of each group.

	Tommy's Plants	Tara's Plants	Nick's Plants
Sample Mean	24.2	25.4	24.4
Sample Variance	11.7	18.3	16.3

Table 13.10

Next, calculate the variance of the three group means by calculating the variance of 24.2, 25.4, and 24.4. Variance of the group means = 0.413 = $s_{\bar{x}}^2$,

then $MS_{between} = ns_{\bar{x}}^2 = (5)(0.413)$ where $n = 5$ is the sample size (number of plants each child grew).

Calculate the mean of the three sample variances (11.7, 18.3, and 16.3). Mean of the sample variances = 15.433 = s^2_{pooled},

then $MS_{within} = s^2_{pooled} = 15.433$.

The F statistic (or F ratio) is $F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{ns_{\bar{x}}^2}{s^2_{pooled}} = \frac{(5)(0.413)}{15.433} = 0.134$.

The dfs for the numerator = the number of groups – 1 = 3 – 1 = 2.

The dfs for the denominator = the total number of samples – the number of groups = 15 – 3 = 12.

The distribution for the test is $F_{2,12}$ and the F statistic is $F = 0.134$.

The p-value is $P(F > 0.134) = 0.8759$.

Decision: Since $\alpha = 0.03$ and the p-value = 0.8759, we do not reject H_0. Why?

Conclusion: With a 3 percent level of significance from the sample data, the evidence is not sufficient to conclude that the mean heights of the bean plants are different.

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the p-value:

•Press `2nd DISTR`,

•Arrow down to `Fcdf` and press `ENTER`,

•Enter `0.134`, `E99`, `2`, `12`, and

•Press `ENTER`.

The p-value is 0.8759.

Try It

13.4 Another fourth grader also grew bean plants, but in a jelly-like mass. The heights were (in inches) 24, 28, 25, 30, and 32. Do a one-way ANOVA test on the four groups. Are the heights of the bean plants different? Use the same method as shown in **Example 13.4**.

Collaborative Exercise

From the class, create four groups of the same size as follows: men under 22, men at least 22, women under 22, women at least 22. Have each member of each group record the number of states in the United States he or she has visited. Run an ANOVA test to determine if the average number of states visited in the four groups are the same. Test at a 1 percent level of significance. Use one of the solution sheets in **Appendix E**.

13.4 | Test of Two Variances

Another use of the F distribution is testing two variances. It is often desirable to compare two variances rather than two averages. For instance, college administrators would like two college professors grading exams to have the same variation in their grading. For a lid to fit a container, the variation in the lid and the container should be the same. A supermarket might be interested in the variability of check-out times for two checkers.

To perform a F test of two variances, it is important that the following are true:

- The populations from which the two samples are drawn are normally distributed.
- The two populations are independent of each other.

Unlike most other tests in this book, the F test for equality of two variances is very sensitive to deviations from normality. If the two distributions are not normal, the test can give higher p-values than it should, or lower ones, in ways that are unpredictable. Many texts suggest that students not use this test at all, but in the interest of completeness we include it here.

Suppose we sample randomly from two independent normal populations. Let σ_1^2 and σ_2^2 be the population variances and

s_1^2 and s_2^2 be the sample variances. Let the sample sizes be n_1 and n_2. Since we are interested in comparing the two sample variances, we use the F ratio

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]}.$$

F has the distribution $F \sim F(n_1 - 1, n_2 - 1)$,

where $n_1 - 1$ are the degrees of freedom for the numerator and $n_2 - 1$ are the degrees of freedom for the denominator.

If the null hypothesis is $\sigma_1^2 = \sigma_2^2$, then the F ratio becomes $F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2}$.

NOTE

The F ratio could also be $\frac{(s_2)^2}{(s_1)^2}$. It depends on H_a and on which sample variance is larger.

If the two populations have equal variances, then s_1^2 and s_2^2 are close in value and $F = \frac{(s_1)^2}{(s_2)^2}$ is close to 1. But if the two population variances are very different, s_1^2 and s_2^2 tend to be very different, too. Choosing s_1^2 as the larger sample variance causes the ratio $\frac{(s_1)^2}{(s_2)^2}$ to be greater than 1. If s_1^2 and s_2^2 are far apart, then $F = \frac{(s_1)^2}{(s_2)^2}$ is a large number.

Therefore, if F is close to 1, the evidence favors the null hypothesis (the two population variances are equal). But if F is much larger than 1, then the evidence is against the null hypothesis. A test of two variances may be left-tailed, right-tailed, or two-tailed.

Example 13.5

Two college instructors are interested in whethe there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a variance of 89.9. Test the claim that the first instructor's variance is smaller. In most colleges, it is desirable for the variances of exam grades to be nearly the same among instructors. The level of significance is 10 percent.

Solution 13.5

Let 1 and 2 be the subscripts that indicate the first and second instructor, respectively.

$n_1 = n_2 = 30$.

H_0: $\sigma_1^2 = \sigma_2^2$ and H_a: $\sigma_1^2 < \sigma_2^2$.

Calculate the test statistic: By the null hypothesis $(\sigma_1^2 = \sigma_2^2)$, the F statistic is

$$F = \frac{\left[\frac{(s_1)^2}{(\sigma_1)^2}\right]}{\left[\frac{(s_2)^2}{(\sigma_2)^2}\right]} = \frac{(s_1)^2}{(s_2)^2} = \frac{52.3}{89.9} = 0.5818.$$

Distribution for the test: $F_{29,29}$ where $n_1 - 1 = 29$ and $n_2 - 1 = 29$.

Graph: This test is left-tailed.

Draw the graph, labeling and shading appropriately.

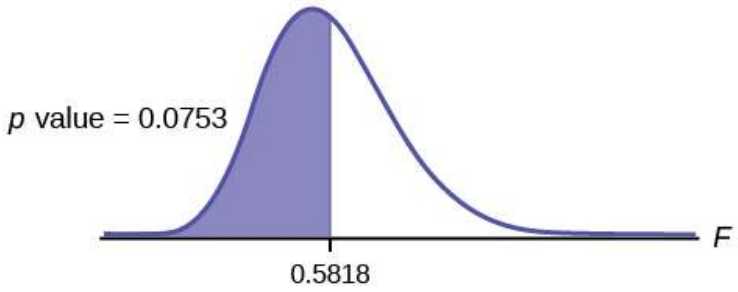


Figure 13.7

Probability statement: p-value = $P(F < 0.5818) = 0.0753$.

Compare α and the p-value: $\alpha = 0.10$; $\alpha > p$-value.

Make a decision: Since $\alpha > p$-value, reject H_0.

Conclusion: With a 10 percent level of significance from the data, there is sufficient evidence to conclude that the variance in grades for the first instructor is smaller.

Using the TI-83, 83+, 84, 84+ Calculator

Press STAT and arrow over to TESTS. Arrow down to D:2-SampFTest. Press ENTER. Arrow to Stats and press ENTER. For Sx1, n1, Sx2, and n2, enter $\sqrt{(52.3)}$, 30, $\sqrt{(89.9)}$, and 30. Press ENTER after each. Arrow to σ1: and < σ2. Press ENTER. Arrow down to Calculate and press ENTER. $F = 0.5818$ and p-value = 0.0753. Do the procedure again and try Draw instead of Calculate.

Try It Σ

13.5 The New York Choral Society divides male singers into four categories from highest voices to lowest: Tenor1, Tenor2, Bass1, and Bass2. In the table are heights of the men in the Tenor1 and Bass2 groups. One suspects that taller men will have lower voices, and that the variance of height may go up with the lower voices as well. Do we have good evidence that the variance of the heights of singers in each of these two groups (Tenor1 and Bass2) are different?

Tenor1	Bass2	Tenor1	Bass2	Tenor1	Bass2
69	72	67	72	68	67
72	75	70	74	67	70
71	67	65	70	64	70
66	75	72	66		69
76	74	70	68		72
74	72	68	75		71
71	72	64	68		74
66	74	73	70		75
68	72	66	72		

Table 13.11

13.5 | Lab: One-Way ANOVA

Stats Lab

13.1 One-Way ANOVA

Student Learning Outcome

- The student will conduct a simple one-way ANOVA test involving three variables.

Collect the Data

1. Record the price per pound of eight fruits, eight vegetables, and eight breads in your local supermarket.

Fruits	Vegetables	Breads

Table 13.12

2. Explain how you could try to collect the data randomly.

Analyze the Data and Conduct a Hypothesis Test

1. State the null hypothesis and the alternative hypothesis.
2. Compute the following:
   a. Fruit
      i. $\bar{x}$ = _____
      ii. s_x = _____
      iii. n = _____
   b. Vegetables
      i. $\bar{x}$ = _____
      ii. s_x = _____
      iii. n = _____
   c. Bread
      i. $\bar{x}$ = _____
      ii. s_x = _____
      iii. n = _____
3. Find the following:
   a. $df(num)$ = _____

b. $df(denom)$ = _____

4. State the approximate distribution for the test.
5. Test statistic: F = _____
6. Sketch a graph of this situation. Clearly label and scale the horizontal axis and shade the region(s) corresponding to the p-value.
7. p-value = _____
8. Test at $\alpha = 0.05$. State your decision and conclusion.
9. a. Decision: why did you make this decision?
   b. Conclusion (write a complete sentence):
   c. Based on the results of your study, is there a need to investigate any of the food groups's prices? Why or why not?

KEY TERMS

analysis of variance also referred to as ANOVA; a method of testing whether the means of three or more populations are equal
The method is applicable if

- all populations of interest are normally distributed,
- the populations have equal standard deviations, and
- samples (not necessarily of the same size) are randomly and independently selected from each population.

The test statistic for analysis of variance is the *F* ratio.

one-way ANOVA a method of testing whether the means of three or more populations are equal; the method is applicable if

- all populations of interest are normally distributed,
- the populations have equal standard deviations,
- samples (not necessarily of the same size) are randomly and independently selected from each population, and
- there is one independent variable and one dependent variable.

The test statistic for analysis of variance is the *F* ratio

variance mean of the squared deviations from the mean; the square of the standard deviation
For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean.
The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1.

CHAPTER REVIEW

13.1 One-Way ANOVA

Analysis of variance extends the comparison of two groups to several, each a level of a categorical variable (factor). Samples from each group are independent and must be randomly selected from normal populations with equal variances. We test the null hypothesis of equal means of the response in every group versus the alternative hypothesis of one or more group means being different from the others. A one-way ANOVA hypothesis test determines if several population means are equal. The distribution for the test is the *F* distribution with two different degrees of freedom.

Assumptions:

- Each population from which a sample is taken is assumed to be normal.
- All samples are randomly selected and independent.
- The populations are assumed to have equal standard deviations (or variances).

13.2 The F Distribution and the F Ratio

Analysis of variance compares the means of a response variable for several groups. ANOVA compares the variation within each group to the variation of the mean of each group. The ratio of these two is the *F* statistic from an *F* distribution with (number of groups – 1) as the numerator degrees of freedom and (number of observations – number of groups) as the denominator degrees of freedom. These statistics are summarized in the ANOVA table.

13.3 Facts About the F Distribution

The graph of the *F* distribution is always positive and skewed right, though the shape can be mounded or exponential depending on the combination of numerator and denominator degrees of freedom. The *F* statistic is the ratio of a measure of the variation in the group means to a similar measure of the variation within the groups. If the null hypothesis is correct, then the numerator should be small compared to the denominator. A small *F* statistic will result, and the area under the *F* curve to the right will be large, representing a large *p*-value. When the null hypothesis of equal group means is incorrect, then the numerator should be large compared to the denominator, giving a large *F* statistic and a small area (small *p*-value)

to the right of the statistic under the F curve.

When the data have unequal group sizes (unbalanced data), then techniques from **The F Distribution and the F Ratio** need to be used for hand calculations. In the case of balanced data, where the groups are the same size, simplified calculations based on group means and variances may be used. In practice, software is usually employed in the analysis. As in any analysis, graphs of various sorts should be used in conjunction with numerical techniques. Always look at your data!

13.4 Test of Two Variances

The F test for the equality of two variances rests heavily on the assumption of normal distributions. The test is unreliable if this assumption is not met. If both distributions are normal, then the ratio of the two sample variances is distributed as an F statistic, with numerator and denominator degrees of freedom that are one less than the samples sizes of the corresponding two groups. A *test of two variances* hypothesis test determines if two variances are the same. The distribution for the hypothesis test is the F distribution with two different degrees of freedom.

Assumptions:

- The populations from which the two samples are drawn are normally distributed.
- The two populations are independent of each other.

FORMULA REVIEW

13.2 The F Distribution and the F Ratio

$$SS_{\text{between}} = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{\left(\sum s_j\right)^2}{n}$$

$$SS_{\text{total}} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}}$$

$df_{\text{between}} = df(num) = k - 1$

$df_{\text{within}} = df(denom) = n - k$

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$$

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

F ratio when the groups are the same size: $F = \frac{n s_{\bar{x}}^2}{s^2_{pooled}}$

Mean of the F distribution: $\mu = \frac{df(num)}{df(denom) - 1}$

where

- k = the number of groups
- n_j = the size of the j^{th} group
- s_j = the sum of the values in the j^{th} group
- n = the total number of all values (observations) combined
- x = one value (one observation) from the data
- $s_{\bar{x}}^2$ = the variance of the sample means
- s^2_{pooled} = the mean of the sample variances (pooled variance)

13.4 Test of Two Variances

F has the distribution $F \sim F(n_1 - 1, n_2 - 1)$

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$$

If $\sigma_1 = \sigma_2$, then $F = \frac{s_1^2}{s_2^2}$

PRACTICE

13.1 One-Way ANOVA

Use the following information to answer the next five exercises. There are five basic assumptions that must be fulfilled to perform a one-way ANOVA test. What are they?

1. Write one assumption.

2. Write another assumption.

3. Write a third assumption.

4. Write a fourth assumption.

5. Write the final assumption.

6. State the null hypothesis for a one-way ANOVA test if there are four groups.

7. State the alternative hypothesis for a one-way ANOVA test if there are three groups.

8. When do you use an ANOVA test?

13.2 The F Distribution and the F Ratio

Use the following information to answer the next seven exercises. Groups of men from three different areas of the country are to be tested for mean weight. The entries in **Table 13.13** are the weights for the different groups.

Group 1	Group 2	Group 3
216	202	170
198	213	165
240	284	182
187	228	197
176	210	201

Table 13.13

9. What is the sum of squares factor?

10. What is the sum of squares error?

11. What is the *df* for the numerator?

12. What is the *df* for the denominator?

13. What is the mean square factor?

14. What is the mean square error?

15. What is the *F* statistic?

Use the following information to answer the next eight exercises. Girls from four different soccer teams are to be tested for mean goals scored per game. The entries in **Table 13.14** are the goals per game for the different teams.

Team 1	Team 2	Team 3	Team 4
1	2	0	3
2	3	1	4
0	2	1	4

Table 13.14

Team 1	Team 2	Team 3	Team 4
3	4	0	3
2	4	0	2

Table 13.14

16. What is $SS_{between}$?

17. What is the *df* for the numerator?

18. What is $MS_{between}$?

19. What is SS_{within}?

20. What is the *df* for the denominator?

21. What is MS_{within}?

22. What is the *F* statistic?

23. Judging by the *F* statistic, do you think it is likely or unlikely that you will reject the null hypothesis?

13.3 Facts About the F Distribution

24. An *F* statistic can have what values?

25. What happens to the curves as the degrees of freedom for the numerator and the denominator get larger?

Use the following information to answer the next seven exercises. Four basketball teams took a random sample of players regarding how high each player can jump (in inches). The results are shown in **Table 13.15**.

Team 1	Team 2	Team 3	Team 4	Team 5
36	32	48	38	41
42	35	50	44	39
51	38	39	46	40

Table 13.15

26. What is the *df(num)*?

27. What is the *df(denom)*?

28. What are the sum of squares and mean squares factors?

29. What are the sum of squares and mean squares errors?

30. What is the *F* statistic?

31. What is the *p*-value?

32. At the 5 percent significance level, is there a difference in the mean jump heights among the teams?

Use the following information to answer the next seven exercises. A video game developer is testing a new game on three different groups. Each group represents a different target market for the game. The developer collects scores from a random sample from each group. The results are shown in **Table 13.16**.

Group A	Group B	Group C
101	151	101
108	149	109

Table 13.16

Group A	Group B	Group C
98	160	198
107	112	186
111	126	160

Table 13.16

33. What is the *df(num)*?

34. What is the *df(denom)*?

35. What are the $SS_{between}$ and $MS_{between}$?

36. What are the SS_{within} and MS_{within}?

37. What is the *F* Statistic?

38. What is the *p*-value?

39. At the 10 percent significance level, are the scores among the different groups different?

Use the following information to answer the next three exercises. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\bar{x} =$	_____	_____	_____	_____	_____
$s^2 =$	_____	_____	_____	_____	_____

Table 13.17

Enter the data into your calculator or computer.

40. *p*-value = _____

State the decisions and conclusions (in complete sentences) for the following preconceived levels of α.

41. $\alpha = 0.05$

a. Decision: _____

b. Conclusion: _____

42. $\alpha = 0.01$

a. Decision: _____

b. Conclusion: _____

13.4 Test of Two Variances

Use the following information to answer the next two exercises. There are two assumptions that must be true to perform an *F* test of two variances.

43. Name one assumption that must be true.

44. What is the other assumption that must be true?

Use the following information to answer the next seven exercises. Two coworkers commute from the same building. They are interested in whether there is any variation in the time it takes them to drive to work. They each record their times for 20 commutes. The first worker's times have a variance of 12.1. The second worker's times have a variance of 16.9. The first worker thinks that he is more consistent with his commute times. Test the claim at the 10 percent level. Assume that commute times are normally distributed.

45. State the null and alternative hypotheses.

46. What is s_1 in this problem?

47. What is s_2 in this problem?

48. What is n?

49. What is the F statistic?

50. What is the p-value?

51. Is the claim accurate?

Use the following information to answer the next four exercises. Two students are interested in whether there is variation in their test scores for math class. There are 15 total math tests they have taken so far. The first student's grades have a standard deviation of 38.1. The second student's grades have a standard deviation of 22.5. The second student thinks his scores are more consistent.

52. State the null and alternative hypotheses.

53. What is the F statistic?

54. What is the p-value?

55. At the 5 percent significance level, do we reject the null hypothesis?

Use the following information to answer the next three exercises. Two cyclists are comparing the variances of their overall paces going uphill. Each cyclist records his or her speeds going up 35 hills. The first cyclist has a variance of 23.8, and the second cyclist has a variance of 32.1. The cyclists want to see if their variances are the same or different. Assume that speeds are normally distributed.

56. State the null and alternative hypotheses.

57. What is the F statistic?

58. At the 5 percent significance level, what can we say about the cyclists' variances?

HOMEWORK

13.1 One-Way ANOVA

59. Three different traffic routes are tested for mean driving time. The entries in the **Table 13.18** are the driving times in minutes on the three different routes.

Route 1	Route 2	Route 3
30	27	16
32	29	41
27	28	22
35	36	31

Table 13.18

State $SS_{between}$, SS_{within}, and the F statistic.

60. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8
$\bar{x} =$	_____	_____	_____	_____	_____
$s^2 =$	_____	_____	_____	_____	_____

Table 13.19

State the hypotheses.

H_0: _____

H_a: _____

13.2 The F Distribution and the F Ratio

Use the following information to answer the next three exercises. Suppose a group is interested in determining whether teenagers obtain their drivers licenses at approximately the same average age across the country. Suppose that the following data are randomly collected from five teenagers in each region of the country. The numbers represent the age at which teenagers obtained their drivers licenses.

	Northeast	South	West	Central	East
	16.3	16.9	16.4	16.2	17.1
	16.1	16.5	16.5	16.6	17.2
	16.4	16.4	16.6	16.5	16.6
	16.5	16.2	16.1	16.4	16.8

Table 13.20

	Northeast	South	West	Central	East
$\bar{x} =$	_____	_____	_____	_____	_____
$s^2 =$	_____	_____	_____	_____	_____

Table 13.20

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

$H\alpha$: At least any two of the group means $\mu_1, \mu_2, \ldots, \mu_5$ are not equal.

61. degrees of freedom – numerator: $df(num)$ = _____

62. degrees of freedom – denominator: $df(denom)$ = _____

63. F statistic = _____

13.3 Facts About the F Distribution

DIRECTIONS

Use a solution sheet to conduct the following hypothesis tests. The solution sheet can be found in **Appendix E**.

64. Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again, and the net gain in grams is recorded. Using a significance level of 10 percent, test the hypothesis that the three formulas produce the same mean weight gain.

Linda's Rats (g)	Tuan's Rats (g)	Javier's Rats (g)
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

Table 13.21

65. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are in **Table 13.22**. Using a 5 percent significance level, test the hypothesis that the three mean commuting mileages are the same.

Working-Class	Professional (middle incomes)	Professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

Table 13.22

Use the following information to answer the next two exercises. **Table 13.23** lists the number of pages in four different types of magazines.

Home Decorating	News	Health	Computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207
197	101	96	146

Table 13.23

66. Using a significance level of 5 percent, test the hypothesis that the four magazine types have the same mean length.

67. Eliminate one magazine type that you now feel has a mean length different from the others. Redo the hypothesis test, testing that the remaining three means are statistically the same. Use a new solution sheet. Based on this test, are the mean lengths for the remaining three magazines statistically the same?

68. A researcher wants to know if the mean times (in minutes) that people watch their favorite news station are the same. Suppose that **Table 13.24** shows the results of a study.

CNN	FOX	Local
45	15	72
12	43	37
18	68	56
38	50	60
23	31	51
35	22	

Table 13.24

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

69. Are the means for the final exams the same for all statistics class delivery types? **Table 13.25** shows the scores on final exams from several randomly selected classes that used the different delivery types.

Online	Hybrid	Face-to-Face
72	83	80
84	73	78
77	84	84
80	81	81
81		86
		79
		82

Table 13.25

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

70. Are the mean number of times a month a person eats out the same for whites, blacks, Hispanics, and Asians? Suppose that **Table 13.26** shows the results of a study.

White	Black	Hispanic	Asian
6	4	7	8
8	1	3	3
2	5	5	5
4	2	4	1
6		6	7

Table 13.26

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

71. Are the mean numbers of daily visitors to a ski resort the same for the three types of snow conditions? Suppose that **Table 13.27** shows the results of a study.

Powder	Machine Made	Hard Packed
1,210	2,107	2,846
1,080	1,149	1,638
1,537	862	2,019
941	1,870	1,178
	1,528	2,233
	1,382	

Table 13.27

Assume that all distributions are normal, the four population standard deviations are approximately the same, and the data were collected independently and randomly. Use a level of significance of 0.05.

72. Sanjay made identical paper airplanes out of three different weights of paper: light, medium, and heavy. He made four airplanes from each of the weights and launched them himself across the room. Here are the distances (in meters) that his planes flew.

Paper Type/Trial	Trial 1	Trial 2	Trial 3	Trial 4
Heavy	5.1 meters	3.1 meters	4.7 meters	5.3 meters
Medium	4 meters	3.5 meters	4.5 meters	6.1 meters
Light	3.1 meters	3.3 meters	2.1 meters	1.9 meters

Table 13.28

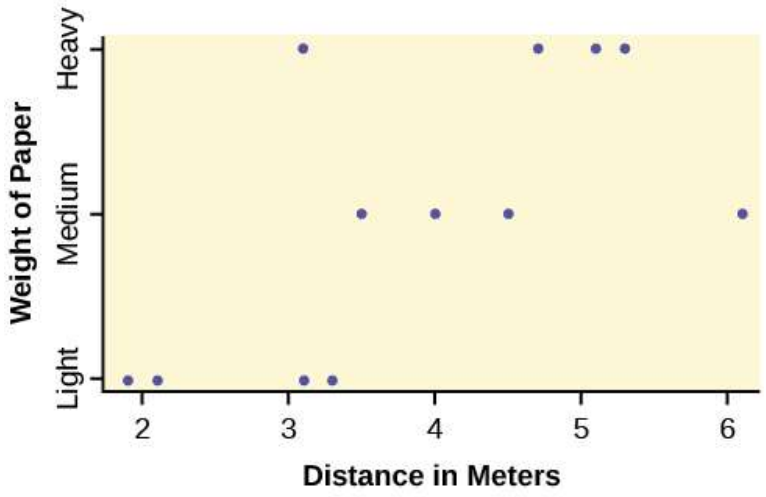


Figure 13.8

a. Take a look at the data in the graph. Look at the spread of data for each group (light, medium, heavy). Does it seem reasonable to assume a normal distribution with the same variance for each group?
b. Why is this a balanced design?
c. Calculate the sample mean and sample standard deviation for each group.
d. Does the weight of the paper have an effect on how far the plane will travel? Use a 1 percent level of significance. Complete the test using the method shown in the bean plant example in **Example 13.4**.
   - Variance of the group means _____
   - $MS_{between}$= _____
   - Mean of the three sample variances _____
   - MS_{within} = _____
   - F statistic = _____
   - $df(num)$ = _____, $df(denom)$ = _____
   - Number of groups _____
   - Number of observations _____
   - p-value = _____ ($P(F >$ _____) = _____)
   - Graph the p-value.
   - Decision: _____
   - Conclusion: _____

73. DDT is a pesticide that has been banned from use in the United States and most other areas of the world. It is quite effective but persisted in the environment and over time proved to be harmful to higher-level organisms. Famously, egg shells of eagles and other raptors were believed to be thinner and prone to breakage in the nest because of ingestion of DDT in the food chain of the birds.

An experiment was conducted on the number of eggs (fecundity) laid by female fruit flies. There are three groups of flies. One group was bred to be resistant to DDT (the RS group). Another was bred to be especially susceptible to DDT (SS). The third group was a control line of nonselected or typical fruit flies (NS). Here are the data:

RS	SS	NS	RS	SS	NS
12.8	38.4	35.4	22.4	23.1	22.6
21.6	32.9	27.4	27.5	29.4	40.4
14.8	48.5	19.3	20.3	16	34.4
23.1	20.9	41.8	38.7	20.1	30.4
34.6	11.6	20.3	26.4	23.3	14.9
19.7	22.3	37.6	23.7	22.9	51.8
22.6	30.2	36.9	26.1	22.5	33.8
29.6	33.4	37.3	29.5	15.1	37.9
416.4	26.7	228.2	38.6	31	29.5
20.3	39	23.4	44.4	16.9	42.4
29.3	12.8	33.7	23.2	16.1	36.6
914.9	14.6	29.2	23.6	10.8	47.4
27.3	12.2	41.7			

Table 13.29

The values are the average number of eggs laid daily for each of 75 flies (25 in each group) over the first 14 days of their lives. Using a 1 percent level of significance, are the mean rates of egg selection for the three strains of fruit fly different? If so, in what way? Specifically, the researchers were interested in whether the selectively bred strains were different from the nonselected line, and whether the two selected lines were different from each other.

Here is a chart of the three groups:

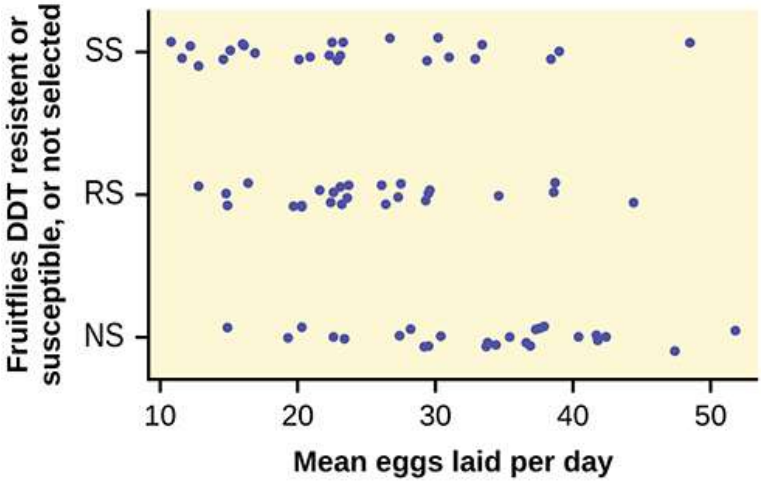


Figure 13.9

74. The data shown is the recorded body temperatures of 130 subjects as estimated from available histograms.

Traditionally, we are taught that the normal human body temperature is 98.6 °F. This is not quite correct for everyone. Are the mean temperatures among the four groups different?

Calculate 95 percent confidence intervals for the mean body temperature in each group and comment about the confidence intervals.

FL	FH	ML	MH	FL	FH	ML	MH
96.4	96.8	96.3	96.9	98.4	98.6	98.1	98.6
96.7	97.7	96.7	97	98.7	98.6	98.1	98.6
97.2	97.8	97.1	97.1	98.7	98.6	98.2	98.7
97.2	97.9	97.2	97.1	98.7	98.7	98.2	98.8
97.4	98	97.3	97.4	98.7	98.7	98.2	98.8
97.6	98	97.4	97.5	98.8	98.8	98.2	98.8
97.7	98	97.4	97.6	98.8	98.8	98.3	98.9
97.8	98	97.4	97.7	98.8	98.8	98.4	99
97.8	98.1	97.5	97.8	98.8	98.9	98.4	99
97.9	98.3	97.6	97.9	99.2	99	98.5	99
97.9	98.3	97.6	98	99.3	99	98.5	99.2
98	98.3	97.8	98		99.1	98.6	99.5
98.2	98.4	97.8	98		99.1	98.6	
98.2	98.4	97.8	98.3		99.2	98.7	
98.2	98.4	97.9	98.4		99.4	99.1	
98.2	98.4	98	98.4		99.9	99.3	
98.2	98.5	98	98.6		100	99.4	
98.2	98.6	98	98.6		100.8		

Table 13.30

13.4 Test of Two Variances

75. Three students, Linda, Tuan, and Javier, are given five laboratory rats each for a nutritional experiment. Each rat's weight is recorded in grams. Linda feeds her rats Formula A, Tuan feeds his rats Formula B, and Javier feeds his rats Formula C. At the end of a specified time period, each rat is weighed again and the net gain in grams is recorded.

Linda's Rats	Tuan's Rats	Javier's Rats
43.5	47.0	51.2
39.4	40.5	40.9
41.3	38.9	37.9
46.0	46.3	45.0
38.2	44.2	48.6

Table 13.31

Determine whether the variance in weight gain is statistically the same between Javier's and Linda's rats. Test at a significance level of 10 percent.

76. A grassroots group opposed to a proposed increase in the gas tax claimed that the increase would hurt working-class people the most since they commute the farthest to work. Suppose that the group randomly surveyed 24 individuals and asked them their daily one-way commuting mileage. The results are as follows.

Working-Class	Professional (middle incomes)	Professional (wealthy)
17.8	16.5	8.5
26.7	17.4	6.3
49.4	22.0	4.6
9.4	7.4	12.6
65.4	9.4	11.0
47.1	2.1	28.6
19.5	6.4	15.4
51.2	13.9	9.3

Table 13.32

Determine whether the variance in mileage driven is statistically the same between the working class and professional (middle income) groups. Use a 5 percent significance level.

Use the following information to answer the next two exercises. The following table lists the number of pages in four different types of magazines.

Home Decorating	News	Health	Computer
172	87	82	104
286	94	153	136
163	123	87	98
205	106	103	207

Table 13.33

Home Decorating	News	Health	Computer
197	101	96	146

Table 13.33

77. Which two magazine types do you think have the same variance in length?

78. Which two magazine types do you think have different variances in length?

79. Is the variance for the amount of money, in dollars, that shoppers spend on Saturdays at the mall the same as the variance for the amount of money that shoppers spend on Sundays at the mall? Suppose that **Table 13.34** shows the results of a study.

Saturday	Sunday	Saturday	Sunday
75	44	62	137
18	58	0	82
150	61	124	39
94	19	50	127
62	99	31	141
73	60	118	73
	89		

Table 13.34

80. Are the variances for incomes on the East Coast and the West Coast the same? Suppose that **Table 13.35** shows the results of a study. Income is shown in thousands of dollars. Assume that both distributions are normal. Use a level of significance of 0.05.

East	West
38	71
47	126
30	42
82	51
75	44
52	90
115	88
67	

Table 13.35

81. Thirty men in college were taught a method of finger tapping. They were randomly assigned to three groups of 10, with each receiving one of three doses of caffeine: 0 mg, 100 mg, or 200 mg. This is approximately the amount in zero, one, or two cups of coffee. Two hours after ingesting the caffeine, the men had the rate of finger tapping per minute recorded. The experiment was double blind, so neither the recorders nor the students knew which group they were in. Does caffeine affect the rate of tapping, and if so how?

Here are the data:

0 mg	100 mg	200 mg	0 mg	100 mg	200 mg
242	248	246	245	246	248
244	245	250	248	247	252
247	248	248	248	250	250
242	247	246	244	246	248
246	243	245	242	244	250

Table 13.36

82. King Manuel I Komnenos ruled the Byzantine Empire from Constantinople (Istanbul) during the years A.D. 1145–1170. The empire was very powerful during his reign but declined significantly afterward. Coins minted during his era were found in Cyprus, an island in the eastern Mediterranean Sea. Nine coins were from his first coinage, seven from the second, four from the third, and seven from the fourth. These spanned most of his reign. We have data on the silver content of the coins:

First Coinage	Second Coinage	Third Coinage	Fourth Coinage
5.9	6.9	4.9	5.3
6.8	9.0	5.5	5.6
6.4	6.6	4.6	5.5
7.0	8.1	4.5	5.1
6.6	9.3		6.2
7.7	9.2		5.8
7.2	8.6		5.8
6.9			
6.2			

Table 13.37

Did the silver content of the coins change over the course of Manuel’s reign?

Here are the means and variances of each coinage. The data are unbalanced.

	First	Second	Third	Fourth
Mean	6.7444	8.2429	4.875	5.6143
Variance	0.2953	1.2095	0.2025	0.1314

Table 13.38

83. The American League and the National League of Major League Baseball are each divided into three divisions: East, Central, and West. Many years, fans talk about some divisions being stronger (having better teams) than other divisions. This may have consequences for the postseason. For instance, in 2012 Tampa Bay won 90 games and did not play in the postseason, while Detroit won only 88 and did play in the postseason. This may have been an oddity, but is there good evidence that in the 2012 season, the American League divisions were significantly different in overall records? Use the following data to test whether the mean number of wins per team in the three American League divisions were the same. Note that the data are not balanced, as two divisions had five teams, while one had only four.

Division	Team	Wins
East	NY Yankees	95
East	Baltimore	93
East	Tampa Bay	90
East	Toronto	73
East	Boston	69

Table 13.39

Division	Team	Wins
Central	Detroit	88
Central	Chicago Sox	85
Central	Kansas City	72
Central	Cleveland	68
Central	Minnesota	66

Table 13.40

Division	Team	Wins
West	Oakland	94
West	Texas	93
West	LA Angels	89
West	Seattle	75

Table 13.41

REFERENCES

13.2 The F Distribution and the F Ratio

Marist College School of Science. (n.d.). *Tomato data* (Unpublished student research). Marist College School of Science, Poughkeepsie, NY.

13.3 Facts About the F Distribution

ESPN. (2012). *MLB standings – 2012.* Retrieved from http://espn.go.com/mlb/standings/_/year/2012.

Hand, D. J. et al. (1994). *A Handbook of Small Datasets: Data for Fruitfly Fecundity.* London: Chapman & Hall.

Hand, D. J. et al. (1994). *A Handbook of Small Datasets.* London: Chapman & Hall, p. 50.

Hand. *A Handbook of Small Datasets.* p. 118.

Mackowiak, P. A.,Wasserman, S. S., & Levine, M. M. (1992). A critical appraisal of 98.6 degrees F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. *Journal of the American Medical Association, 268,* 1578–1580.

Private K–12 school in San Jose, CA. (1994). *Data from a fourth grade classroom.*

13.4 Test of Two Variances

ESPN. (2012). *MLB standings – 2012.* Retrieved from http://espn.go.com/mlb/standings/_/year/2012/type/vs-division/order/true.

SOLUTIONS

1 Each population from which a sample is taken is assumed to be normal.

3 The populations are assumed to have equal standard deviations (or variances).

5 The response is a numerical value.

7 H_a: At least two of the group means μ_1, μ_2, μ_3 are not equal.

9 4,939.2

11 2

13 2,469.6

15 3.7416

17 3

19 13.2

21 0.825

23 Because a one-way ANOVA test is always right-tailed, a high F statistic corresponds to a low p value, so it is likely that we will reject the null hypothesis.

25 The curves approximate the normal distribution.

27 10

29 SS = 237.33; MS = 23.73

31 0.1614

33 two

35 SS = 5,700.4; MS = 2,850.2

37 3.6101

39 Yes, there is enough evidence to show that the scores among the groups are statistically significant at the 10 percent level.

43 The populations from which the two samples are drawn are normally distributed.

45 H_0: $\sigma_1 = \sigma_2$ H_a: $\sigma_1 < \sigma_2$ or H_0: $\sigma_1^2 = \sigma_2^2$ H_a: $\sigma_1^2 < \sigma_2^2$

47 4.11

49 0.7159

51 No, at the 10 percent level of significance, we do not reject the null hypothesis and state that the data do not show that the variation in drive times for the first worker is less than the variation in drive times for the second worker.

53 2.8674

55 Reject the null hypothesis. There is enough evidence to say that the variance of the grades for the first student is higher than the variance in the grades for the second student.

57 0.7414

59 $SS_{between} = 26$
$SS_{within} = 441$
$F = 0.2653$

62 $df(denom) = 15$

64

a. H_0: $\mu_L = \mu_T = \mu_J$
b. H_a: at least any two of the means are different
c. $df(num) = 2$; $df(denom) = 12$
d. F distribution
e. 0.67
f. 0.5305
g. Check student's solution.
h. Decision: Do not reject null hypothesis.
i. Conclusion: There is insufficient evidence to conclude that the means are different.

67

a. H_a: $\mu_c = \mu_n = \mu_h$
b. At least any two of the magazines have different mean lengths.
c. $df(num) = 2$, $df(denom) = 12$
d. F distribtuion
e. $F = 15.28$
f. p-value = 0.0005
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Reject the null hypothesis.
   iii. Reason for decision: p-value < alpha
   iv. Conclusion: There is sufficient evidence to conclude that the mean lengths of the magazines are different.

69

a. H_0: $\mu_o = \mu_h = \mu_f$
b. At least two of the means are different.
c. $df(n) = 2$, $df(d) = 13$
d. $F_{2,13}$
e. 0.64
f. 0.5437
g. Check student's solution.
h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: p-value > alpha
   iv. Conclusion: The mean scores of different class delivery are not different.

71

a. H_0: $\mu_p = \mu_m = \mu_h$

b. At least any two of the means are different.

c. $df(n) = 2$, $df(d) = 12$

d. $F_{2,12}$

e. 3.13

f. 0.0807

g. Check student's solution.

h. i. Alpha: 0.05

ii. Decision: Do not reject the null hypothesis.

iii. Reason for decision: p-value > alpha

iv. Conclusion: There is not sufficient evidence to conclude that the mean numbers of daily visitors are different.

73 The data appear normally distributed from the chart and of similar spread. There do not appear to be any serious outliers, so we may proceed with our ANOVA calculations, to see if we have good evidence of a difference between the three groups. H_0: $\mu_1 = \mu_2 = \mu_3$ H_a: $\mu_i \neq \mu_j$ some $i \neq j$ Define μ_1, μ_2, μ_3, as the population mean number of eggs laid by the three groups of fruitflies. F statistic = 8.6657 p-value = 0.0004

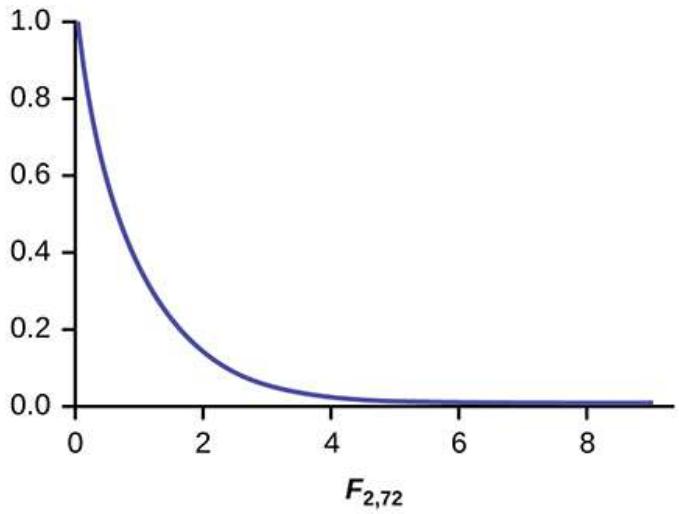


Figure 13.10

Decision: Since the p-value is less than the level of significance of 0.01, we reject the null hypothesis. **Conclusion:** We have good evidence that the average number of eggs laid during the first 14 days of life for these three strains of fruitflies are different. Interestingly, if you perform a two sample t test to compare the RS and NS groups they are significantly different (p = 0.0013). Similarly, SS and NS are significantly different (p = 0.0006). However, the two selected groups, RS and SS are *not* significantly different (p = 0.5176). Thus we appear to have good evidence that selection either for resistance or for susceptibility involves a reduced rate of egg production (for these specific strains) as compared to flies that were not selected for resistance or susceptibility to DDT. Here, genetic selection has apparently involved a loss of fecundity.

75

a. $H_0 : \sigma_1^2 = \sigma_2^2$

b. $H_a : \sigma_1^2 \neq \sigma_1^2$

c. $df(num) = 4$; $df(denom) = 4$

d. $F_{4,4}$

e. 3.00

f. 2(0.1563) = 0.3126. Using the TI-83+/84+ function 2-SampFtest, you get the test statistic as 2.9986 and *p*-value directly as 0.3127. If you input the lists in a different order, you get a test statistic of 0.3335 but the *p*-value is the same because this is a two-tailed test.

g. Check student's solution.

h. Decision: Do not reject the null hypothesis.

i. Conclusion: There is insufficient evidence to conclude that the variances are different.

78 The answers may vary. Sample answer: Home decorating magazines and news magazines have different variances.

80

a. H_0: = $\sigma_1^2 = \sigma_2^2$

b. H_a: $\sigma_1^2 \neq \sigma_1^2$

c. $df(n) = 7$, $df(d) = 6$

d. $F_{7,6}$

e. 0.8117

f. 0.7825

g. Check student's solution.

h.
   i. Alpha: 0.05
   ii. Decision: Do not reject the null hypothesis.
   iii. Reason for decision: *p*-value > alpha
   iv. Conclusion: There is not sufficient evidence to conclude that the variances are different.

82 Here is a strip chart of the silver content of the coins:

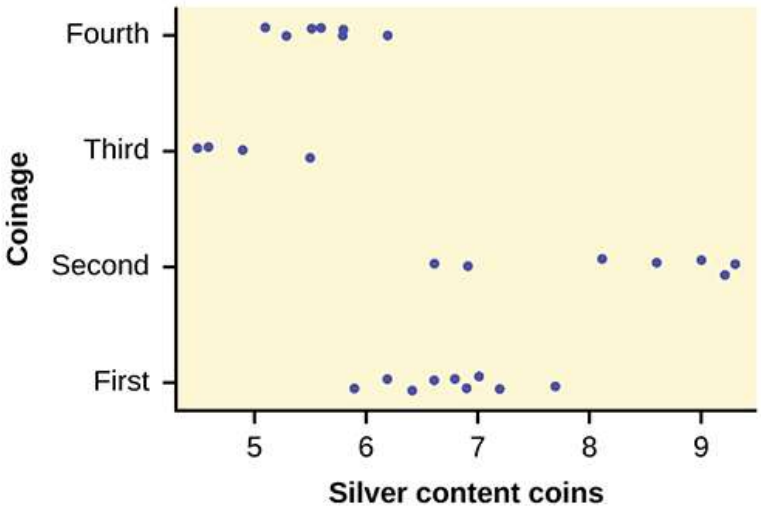


Figure 13.11

While there are differences in spread, it is not unreasonable to use ANOVA techniques. Here is the completed ANOVA table:

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (between)	37.748	4 – 1 = 3	12.5825	26.272
Error (within)	11.015	27 – 4 = 23	0.4789	
Total	48.763	27 – 1 = 26		

Table 13.42

$P(F > 26.272) = 0$. Reject the null hypothesis for any alpha. There is sufficient evidence to conclude that the mean silver content among the four coinages are different. From the strip chart, it appears that the first and second coinages had higher silver contents than the third and fourth.

83 Here is a stripchart of the number of wins for the 14 teams in the AL for the 2012 season.

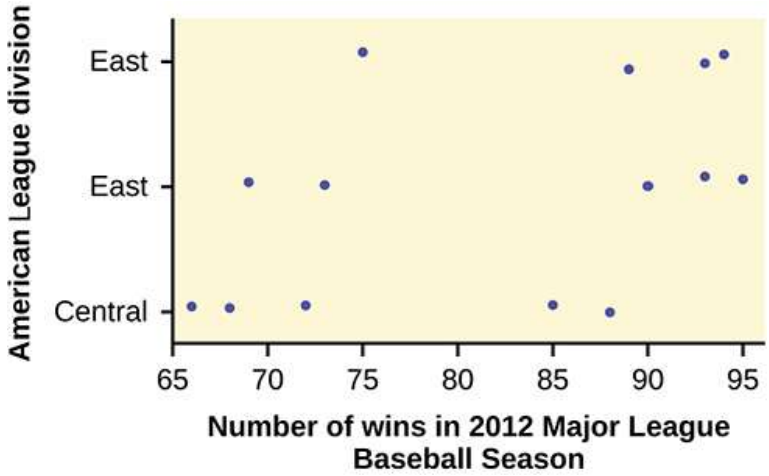


Figure 13.12

While the spread seems similar, there may be some question about the normality of the data, given the wide gaps in the middle near the 0.500 mark of 82 games (teams play 162 games each season in MLB). However, one-way ANOVA is robust. Here is the ANOVA table for the data:

Source of Variation	Sum of Squares (*SS*)	Degrees of Freedom (*df*)	Mean Square (*MS*)	*F*
Factor (between)	344.16	3 – 1 = 2	172.08	
Error (within)	1,219.55	14 – 3 = 11	110.87	1.5521
Total	1,563.71	14 – 1 = 13		

Table 13.43

$P(F > 1.5521) = 0.2548$

Since the *p*-value is so large, there is not good evidence against the null hypothesis of equal means. We decline to reject the null hypothesis. Thus, for 2012, there is not any good evidence of a significant difference in mean number of wins between the divisions of the American League.

APPENDIX A: APPENDIX A REVIEW EXERCISES (CH 3–13)

These review exercises are designed to provide extra practice on concepts learned before a particular chapter. For example, the review exercises for Chapter 3 cover material learned in Chapters 1 and 2.

Chapter 3

Use the following information to answer the next six exercises. In a survey of 100 stocks on NASDAQ, the average percent increase for the past year was 9 percent for NASDAQ stocks.

1. The *average increase* for all NASDAQ stocks is the —

A. population
B. statistic
C. parameter
D. sample
E. variable

2. All of the NASDAQ stocks are —

A. population
B. statistics
C. parameter
D. sample
E. variable

3. Nine percent is —

A. population
B. statistics
C. parameter
D. sample
E. variable

4. The 100 NASDAQ stocks in the survey are —

A. population
B. statistic
C. parameter
D. sample
E. variable

5. The percent increase for one stock in the survey is —

A. population

B. statistic

C. parameter

D. sample

E. variable

6. Would the data collected by qualitative, quantitative discrete, or quantitative continuous?

Use the following information to answer the next two exercises. Thirty people spent two weeks around Mardi Gras in New Orleans. Their two-week weight gain is below. Note—a loss is shown by a negative weight gain.

Weight Gain	Frequency
–2	3
–1	5
0	2
1	4
4	13
6	2
11	1

Table A1

7. Calculate the following values:

A. The average weight gain for the two weeks

B. The standard deviation

C. The first, second, and third quartiles

8. Construct a histogram and box plot of the data.

Chapter 4

Use the following information to answer the next two exercises. A recent poll concerning credit cards found that 35 percent of respondents use a credit card that gives them a mile of air travel for every dollar they charge. Thirty percent of the respondents charge more than $2,000 per month. Of those respondents who charge more than $2,000, 80 percent use a credit card that gives them a mile of air travel for every dollar they charge.

9. What is the probability that a randomly selected respondent will spend more than $2,000 *and* use a credit card that gives them a mile of air travel for every dollar they charge?

A. (.30)(.35)

B. (.80)(.35)

C. (.80)(.30)

D. (.80)

10. Are using a credit card that gives a mile of air travel for each dollar spent *and* charging more than $2,000 per month independent events?

A. Yes

B. No, and they are not mutually exclusive either

C. No, but they are mutually exclusive

D. Not enough information given to determine the answer

11. A sociologist wants to know the opinions of employed adult women about government funding for day care. She obtains a list of 520 members of a local business and professional women's club and mails a questionnaire to 100 of these women selected at random. Sixty-eight questionnaires are returned. What is the population in this study?

A. All employed adult women

B. All the members of a local business and professional women's club

C. The 100 women who received the questionnaire

D. All employed women with children

Use the following information to answer the next two exercises. An article from the *San Jose Mercury News* was concerned with the racial mix of the 1,500 students at Prospect High School in Saratoga, CA. The table summarizes the results. Male and female values are approximate. Suppose one Prospect High School student is randomly selected.

Gender/Ethnic Group	White	Asian	Hispanic	Black	American Indian
Male	400	468	115	35	16
Female	440	132	140	40	14

Table A2

12. Find the probability that a student is Asian or male.

13. Find the probability that a student is black given that the student is female.

14. A sample of pounds lost, in a certain month, by individual members of a weight reducing clinic produced the following statistics:

- Mean = 5 lbs
- Median = 4.5 lbs
- Mode = 4 lbs
- Standard deviation = 3.8 lbs
- First quartile = 2 lbs
- Third quartile = 8.5 lbs

What is the correct statement?

A. One fourth of the members lost exactly two pounds.

B. The middle 50 percent of the members lost from two to 8.5 lbs.

C. Most people lost 3.5 to 4.5 lbs.

D. All of the choices above are correct.

15. What does it mean when a data set has a standard deviation equal to zero?

A. All values of the data appear with the same frequency.

B. The mean of the data is also zero.

C. All of the data have the same value.

D. There are no data to begin with.

16. Which statement describes the illustration?

Figure A1

A. The mean is equal to the median.

B. There is no first quartile.

C. The lowest data value is the median.

D. The median equals $\frac{Q_1 + Q_3}{2}$.

17. According to a recent article in the *San Jose Mercury News* the average number of babies born with significant hearing loss—deafness—is approximately 2 per 1,000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1,000 babies in an intensive care nursery. Suppose that 1,000 babies from healthy baby nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

18. A *friend* offers you the following *deal*: For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Based upon the financial gain or loss over the long run, should you play the game?

A. Yes, I expect to come out ahead in money.

B. No, I expect to come out behind in money.

C. It doesn't matter. I expect to break even.

Use the following information to answer the next four exercises. Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he/she truly has the flu—and not just a nasty cold—is only about 4 percent. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

19. Define the random variable and list its possible values.

20. State the distribution of X.

21. Find the probability that at least four of the 25 patients actually have the flu.

22. On average, for every 25 patients calling in, how many do you expect to have the flu?

Use the following information to answer the next two exercises. Different types of writing can sometimes be distinguished by the number of letters in the words used. A student interested in this fact wants to study the number of letters of words used by Tom Clancy in his novels. She opens a Clancy novel at random and records the number of letters of the first 250 words on the page.

23. What kind of data was collected?

A. Qualitative

B. Quantitative continuous

C. Quantitative discrete

24. What is the population under study?

Chapter 5

Use the following information to answer the next five exercises. A recent study of mothers of junior high school children in Santa Clara County reported that 76 percent of the mothers are employed in paid positions. Of those mothers who are employed, 64 percent work full-time—more than 35 hours per week—and 36 percent work part-time. However, out of all of the mothers in the population, 49 percent work full-time. The population under study is made up of mothers of junior high school children in Santa Clara County. Let E = employed and F = full-time employment.

25.

A. Find the percent of all mothers in the population that are *not* employed.

B. Find the percent of mothers in the population that are employed part-time.

26. The *type of employment* is considered to be what type of data?

27. Find the probability that a randomly selected mother works part-time given that she is employed.

28. Find the probability that a randomly selected person from the population will be employed or work full-time.

29. Being employed and working part-time—

A. mutually exclusive events? Why or why not?

B. independent events? Why or why not?

Use the following additional information to answer the next two exercises. We randomly pick 10 mothers from the above population. We are interested in the number of the mothers that are employed. Let X = number of mothers that are employed.

30. State the distribution for X.

31. Find the probability that at least six are employed.

32. We expect the statistics discussion board to have, on average, 14 questions posted to it per week. We are interested in the number of questions posted to it per day.

A. Define X.

B. What are the values that the random variable may take on?

C. State the distribution for X.

D. Find the probability that from 10 to 14—inclusive—questions are posted to the listserv on a randomly picked day.

33. A person invests $1,000 into stock of a company that hopes to go public in one year. The probability that the person will lose all his money after one year, that is, his stock will be worthless, is 35 percent. The probability that the person's stock will still have a value of $1,000 after one year, that is, no profit and no loss, is 60 percent. The probability that the person's stock will increase in value by $10,000 after one year, that is, will be worth $11,000, is 5 percent. Find the expected profit after one year.

34. Rachel's piano cost $3,000. The average cost for a piano is $4,000 with a standard deviation of $2,500. Becca's guitar cost $550. The average cost for a guitar is $500 with a standard deviation of $200. Matt's drums cost $600. The average cost for drums is $700 with a standard deviation of $100. Whose cost was lowest when compared to his or her own instrument?

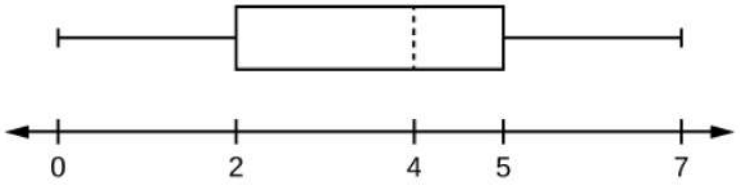


Figure A2

35. Explain why each statement is either true or false given the box plot in **Figure A2**.

A. Twenty-five percent of the data are at most five.

B. There is the same amount of data from 4–5 as there is from 5–7.

C. There are no data values of three.

D. Fifty percent of the data are four.

Using the following information to answer the next two exercises. 64 faculty members were asked the number of cars they owned—including spouse and children's cars. The results are given in the following graph.

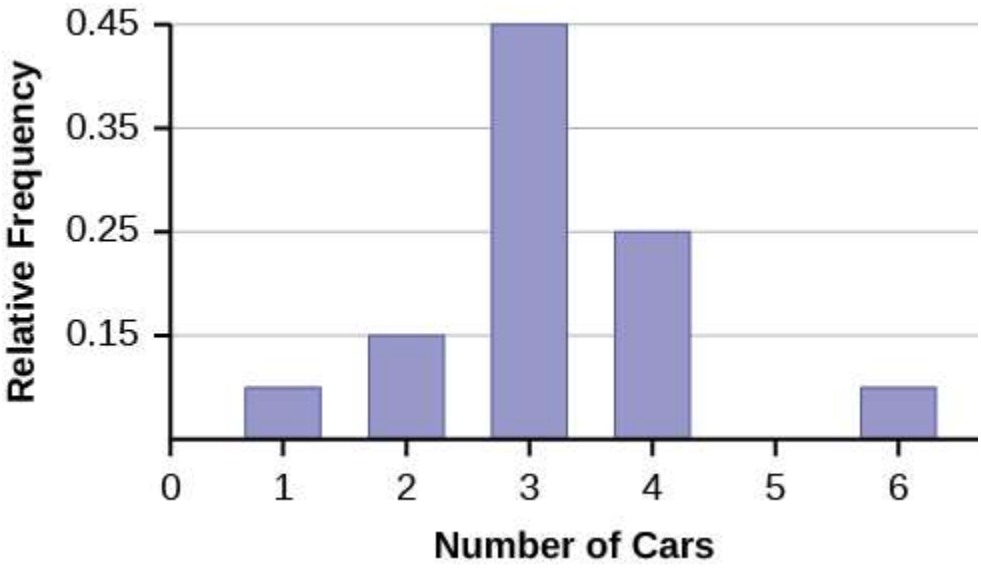


Figure A3

36. Find the approximate number of responses that were three.

37. Find the first, second, and third quartiles. Use them to construct a box plot of the data.

Use the following information to answer the next three exercises. **Table A3** shows data gathered from 15 girls on the Snow Leopard soccer team when they were asked how they liked to wear their hair. Supposed one girl from the team is randomly selected.

Hair Style/Hair Color	Blond	Brown	Black
Ponytail	3	2	5
Plain	2	2	1

Table A3

38. Find the probability that the girl has black hair GIVEN that she wears a ponytail.

39. Find the probability that the girl wears her hair plain OR has brown hair.

40. Find the probability that the girl has blond hair AND that she wears her hair plain.

Chapter 6

Use the following information to answer the next two exercises. $X \sim U(3, 13)$

41. Explain which of the following are false and which are true.

A. $f(x) = \frac{1}{10}, 3 \leq x \leq 13$

B. There is no mode.

C. The median is less than the mean.

D. $P(x > 10) = P(x \leq 6)$

42. Calculate

A. the mean,

B. the median, and

C. the 65^{th} percentile.

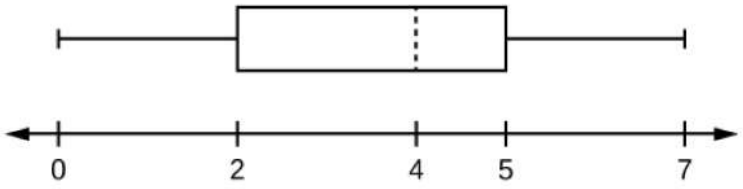


Figure A4

43. Which of the following is true for the box plot in **Figure A4**?

A. Twenty-five percent of the data are at most five.

B. There is about the same amount of data from 4–5 as there is from 5–7.

C. There are no data values of three.

D. Fifty percent of the data are four.

44. If $P(G|H) = P(G)$, then which of the following is correct?

A. G and H are mutually exclusive events.

B. $P(G) = P(H)$

C. Knowing that H has occurred will affect the chance that G will happen.

D. G and H are independent events.

45. If $P(J) = .3$, $P(K) = .63$, and J and K are independent events, then explain which are correct and which are incorrect.

A. $P(J \text{ AND } K) = 0$

B. $P(J \text{ OR } K) = .9$

C. $P(J \text{ OR } K) = .72$

D. $P(J) \neq P(J|K)$

46. On average, five students from each high school class get full scholarships to four-year colleges. Assume that most high school classes have about 500 students. X = the number of students from a high school class that get full scholarships to four-year schools. Which of the following is the distribution of X?

A. $P(5)$

B. $B(500, 5)$

C. $Exp\left(\frac{1}{5}\right)$

D. $N\left(5, \frac{(.01)(.99)}{500}\right)$

Chapter 7

Use the following information to answer the next three exercises. Richard's Furniture Company delivers furniture from 10 a.m. to 2 p.m. continuously and uniformly. We are interested in how long—in hours—past the 10 a.m. start time that individuals wait for their delivery.

47. $X \sim$ _____

A. $U(0, 4)$

B. $U(10, 20)$

C. $Exp(2)$

D. $N(2, 1)$

48. The average wait time is —

A. one hour

B. two hours

C. two and a half hours

D. four hours

49. Suppose that it is now past noon on a delivery day. The probability that a person must wait at least 1.5 more hours is —

A. $\frac{1}{4}$

B. $\frac{1}{2}$

C. $\frac{3}{4}$

D. $\frac{3}{8}$

50. Given $X \sim Exp\left(\frac{1}{3}\right)$

A. Find $P(x > 1)$.

B. Calculate the minimum value for the upper quartile.

C. Find $P\left(x = \frac{1}{3}\right)$

51.

- Forty percent of full-time students took four years to graduate.
- Thirty percent of full-time students took five years to graduate.
- Twenty percent of full-time students took six years to graduate.
- Ten percent of full-time students took seven years to graduate.

The expected time for full-time students to graduate is —

A. four years
B. four and a half years
C. five years
D. five and a half years

52. Which of the following distributions is described by the following example?
Many people can run a short distance of under two miles, but as the distance increases, fewer people can run that far.

A. binomial
B. uniform
C. exponential
D. normal

53. The length of time to brush one's teeth is generally thought to be exponentially distributed with a mean of $\frac{3}{4}$ minutes. Find the probability that a randomly selected person brushes his or her teeth less than $\frac{3}{4}$ minutes.

A. .5
B. $\frac{3}{4}$
C. .43
D. .63

54. Which distribution accurately describes the following situation?
The chance that a teenage boy regularly gives his mother a kiss goodnight is about 20 percent. Fourteen teenage boys are randomly surveyed. Let X = the number of teenage boys that regularly give their mother a kiss goodnight.

A. $B(14,.20)$
B. $P(2.8)$
C. $N(2.8,2.24)$
D. $Exp\left(\frac{1}{.20}\right)$

55. A 2008 report on technology use states that approximately 20 percent of U.S. households have never sent an email. Suppose that we select a random sample of fourteen U.S. households. Let X = the number of households in a 2008 sample of 14 households that have never sent an email.

A. $B(14,.20)$
B. $P(2.8)$
C. $N(2.8,2.24)$

D. $Exp\left(\frac{1}{.20}\right)$

Chapter 8

Use the following information to answer the next three exercises. Suppose that a sample of 15 randomly chosen people were put on a special weight-loss diet. The amount of weight lost, in pounds, follows an unknown distribution with mean equal to 12 pounds and standard deviation equal to three pounds. Assume that the distribution for the weight loss is normal.

56. To find the probability that the mean amount of weight lost by 15 people is no more than 14 pounds, the random variable should be _____.

A. number of people who lost weight on the special weight-loss diet

B. the number of people who were on the diet

C. the mean amount of weight lost by 15 people on the special weight-loss diet

D. the total amount of weight lost by 15 people on the special weight-loss diet

57. Find the probability asked for in **Question 56**.

58. Find the 90^{th} percentile for the mean amount of weight lost by 15 people.

Using the following information to answer the next three exercises. The time of occurrence of the first accident during rush-hour traffic at a major intersection is uniformly distributed between the three hour interval 4 p.m. to 7 p.m. Let X = the amount of time—hours—it takes for the first accident to occur.

59. What is the probability that the time of occurrence is within the first half-hour or the last hour of the period from 4 to 7 p.m.?

A. It cannot be determined from the information given.

B. $\frac{1}{6}$

C. $\frac{1}{2}$

D. $\frac{1}{3}$

60. The 20^{th} percentile occurs after how many hours?

A. .20

B. .60

C. .50

D. 1

61. Assume Ramon has kept track of the times for the first accidents to occur for 40 different days. Let C = the total cumulative time. Then C follows which distribution?

A. $U(0,3)$

B. $Exp(13)$

C. $N(60, 5.477)$

D. $N(1.5, .01875)$

62. Using the information in **Question 61**, find the probability that the total time for all first accidents to occur is more than 43 hours.

Use the following information to answer the next two exercises. The length of time a parent must wait for his children to

clean their rooms is uniformly distributed in the time interval from one to 15 days.

63. How long must a parent expect to wait for his children to clean their rooms?

A. 8 days

B. 3 days

C. 14 days

D. 6 days

64. What is the probability that a parent will wait more than six days given that the parent has already waited more than three days?

A. .5174

B. .0174

C. .7500

D. .2143

Use the following information to answer the next five exercises. Twenty percent of the students at a local community college live in within five miles of the campus. Thirty percent of the students at the same community college receive some kind of financial aid. Of those who live within five miles of the campus, 75 percent receive some kind of financial aid.

65. Find the probability that a randomly chosen student at the local community college does not live within five miles of the campus.

A. 80 percent

B. 20 percent

C. 30 percent

D. Cannot be determined

66. Find the probability that a randomly chosen student at the local community college lives within five miles of the campus or receives some kind of financial aid.

A. 50 percent

B. 35 percent

C. 27.5 percent

D. 75 percent

67. Are living in student housing within five miles of the campus and receiving some kind of financial aid mutually exclusive?

A. Yes

B. No

C. Cannot be determined

68. The interest rate charged on the financial aid is _____ data.

A. Quantitative discrete

B. Quantitative continuous

C. Qualitative discrete

D. Qualitative

69. The following information is about the students who receive financial aid at the local community college.

- 1st quartile = $250
- 2nd quartile = $700
- 3rd quartile = $1,200

These amounts are for the school year. If a sample of 200 students is taken, how many are expected to receive $250 or more?

A. 50
B. 250
C. 150
D. Cannot be determined

Use the following information to answer the next two exercises. $P(A) = .2$, $P(B) = .3$; A and B are independent events.

70. $P(A \text{ AND } B) =$ —

A. .5
B. .6
C. 0
D. .06

71. $P(A \text{ OR } B) =$ —

A. .56
B. .5
C. .44
D. 1

72. If H and D are mutually exclusive events, $P(H) = .25$, $P(D) = .15$, then $P(H|D)$.

A. 1
B. 0
C. .40
D. .0375

Chapter 9

73. Rebecca and Matt are 14 year old twins. Matt's height is two standard deviations below the mean for 14 year old boys' height. Rebecca's height is .10 standard deviations above the mean for 14 year old girls' height. Interpret this.

A. Matt is 2.1 inches shorter than Rebecca.
B. Rebecca is very tall compared to other 14 year old girls.
C. Rebecca is taller than Matt.
D. Matt is shorter than the average 14 year old boy.

74. Construct a histogram of the IPO data (see **Appendix C**).

Use the following information to answer the next three exercises. Ninety homeowners were asked the number of estimates they obtained before having their homes fumigated. Let X = the number of estimates.

x	Relative Frequency	Cumulative Relative Frequency
1	.3	
2	.2	
4	.4	
5	.1	

Table A4

75. Complete the cumulative frequency column.

76. Calculate the sample mean (a), the sample standard deviation (b), and the percent of the estimates that fall at or below four (c).

77. Calculate the median, M, the first quartile, Q_1, and the third quartile Q_3. Then construct a box plot of the data.

78. The middle 50 percent of the data are between _____ and _____.

Use the following information to answer the next three exercises. Seventy fifth and sixth graders were asked their favorite dinner.

	Pizza	Hamburgers	Spaghetti	Fried Shrimp
5th Grader	15	6	9	0
6th Grader	15	7	10	8

Table A5

79. Find the probability that one randomly chosen child is in the 6th grade and prefers fried shrimp.

A. $\frac{32}{70}$

B. $\frac{8}{32}$

C. $\frac{8}{8}$

D. $\frac{8}{70}$

80. Find the probability that a child does not prefer pizza.

A. $\frac{30}{70}$

B. $\frac{30}{40}$

C. $\frac{40}{70}$

D. 1

81. Find the probability a child is in the fifth grade given that the child prefers spaghetti.

A. $\frac{9}{19}$

B. $\frac{9}{70}$

C. $\frac{9}{30}$

D. $\frac{19}{70}$

82. A sample of convenience is a random sample.

A. True

B. False

83. A statistic is a number that is a property of the population.

A. True

B. False

84. You should always throw out any data that are outliers.

A. True

B. False

85. Lee bakes pies for a small restaurant in Felton, CA. She generally bakes 20 pies in a day, on average. Of interest is the number of pies she bakes each day.

A. Define the random variable X.

B. State the distribution for X.

C. Find the probability that Lee bakes more than 25 pies in any given day.

86. Six different brands of Italian salad dressing were randomly selected at a supermarket. The grams of fat per serving are 7, 7, 9, 6, 8, and 5. Assume that the underlying distribution is normal. Calculate a 95 percent confidence interval for the population mean grams of fat per serving of Italian salad dressing sold in supermarkets.

87. Given: uniform, exponential, normal distributions. Match each to a statement below.

A. mean = median ≠ mode

B. mean > median > mode

C. mean = median = mode

Chapter 10

Use the following information to answer the next three exercises. In a survey at Kirkwood Ski Resort the following information was recorded.

	0–10	**11–20**	**21–40**	**40+**
Ski	10	12	30	8
Snowboard	6	17	12	5

Table A6

Suppose that one person from **Table A6** was randomly selected.

88. Find the probability that the person was a skier or was age 11–20.

89. Find the probability that the person was a snowboarder given he or she was age 21–40.

90. Explain which of the following are true and which are false.

A. Sport and age are independent events.

B. Ski and age 11–20 are mutually exclusive events.

C. $P(\text{Ski AND age 21–40}) < P(\text{Ski}|\text{age 21–40})$

D. $P(\text{Snowboard OR age 0–10}) < P(\text{Snowboard}|\text{age 0–10})$

91. The average length of time a person with a broken leg wears a cast is approximately six weeks. The standard deviation is about three weeks. Thirty people who had recently healed from broken legs were interviewed. State the distribution that most accurately reflects total time to heal for the 30 people.

92. The distribution for X is uniform. What can we say for certain about the distribution for $\bar{X}$ when $n = 1$?

A. The distribution for $\bar{X}$ is still uniform with the same mean and standard deviation as the distribution for X.

B. The distribution for $\bar{X}$ is normal with the different mean and a different standard deviation as the distribution for X.

C. The distribution for $\bar{X}$ is normal with the same mean but a larger standard deviation than the distribution for X.

D. The distribution for $\bar{X}$ is normal with the same mean but a smaller standard deviation than the distribution for X.

93. The distribution for X is uniform. What can we say for certain about the distribution for $\sum X$ when $n = 50$?

A. The distribution for $\sum X$ is still uniform with the same mean and standard deviation as the distribution for X.

B. The distribution for $\sum X$ is normal with the same mean but a larger standard deviation as the distribution for X.

C. The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for X.

D. The distribution for $\sum X$ is normal with the same mean but a smaller standard deviation than the distribution for X.

Use the following information to answer the next three exercises. A group of students measured the lengths of all the carrots in a five-pound bag of baby carrots. They calculated the average length of baby carrots to be 2.0 inches with a standard deviation of 0.25 inches. Suppose we randomly survey 16 five-pound bags of baby carrots.

94. State the approximate distribution for $\bar{X}$, the distribution for the average lengths of baby carrots in 16 five-pound bags.

$\bar{X} \sim$ _____.

95. Explain why we cannot find the probability that one individual randomly chosen carrot is greater than 2.25 inches.

96. Find the probability that $\bar{x}$ is between 2.0 and 2.25 inches.

Use the following information to answer the next three exercises. At the beginning of the term, the amount of time a student waits in line at the campus store is normally distributed with a mean of five minutes and a standard deviation of two minutes.

97. Find the 90^{th} percentile of waiting time in minutes.

98. Find the median waiting time for one student.

99. Find the probability that the average waiting time for 40 students is at least 4.5 minutes.

Chapter 11

Use the following information to answer the next four exercises. Suppose that the time that owners keep their cars—purchased new—is normally distributed with a mean of seven years and a standard deviation of two years. We are interested in how long an individual keeps his car—purchased new. Our population is people who buy their cars new.

100. Sixty percent of individuals keep their cars *at most* how many years?

101. Suppose that we randomly survey one person. Find the probability that person keeps his or her car *less than* 2.5 years.

102. If we are to pick individuals 10 at a time, find the distribution for the *mean* car length ownership.

103. If we are to pick 10 individuals, find the probability that the *sum* of their ownership time is more than 55 years.

104. For which distribution is the median not equal to the mean?

A. Uniform

B. Exponential

C. Normal

D. Student *t*

105. Compare the standard normal distribution to the Student's *t* distribution, centered at zero. Explain which of the following are true and which are false.

A. As the number surveyed increases, the area to the left of –1 for the Student's *t* distribution approaches the area for the standard normal distribution.

B. As the degrees of freedom decrease, the graph of the Student's *t* distribution looks more like the graph of the standard normal distribution.

C. If the number surveyed is 15, the normal distribution should never be used.

Use the following information to answer the next five exercises. We are interested in the checking account balance of 24-old college students. We randomly survey 16 20-year-old college students. We obtain a sample mean of $640 and a sample standard deviation of $150. Let X = checking account balance of an individual 20-year-old college student.

106. Explain why we cannot determine the distribution of X.

107. If you were to create a confidence interval or perform a hypothesis test for the population mean checking account balance of 20-year-old college students, what distribution would you use?

108. Find the 95 percent confidence interval for the true mean checking account balance of a 20-year-old college student.

109. What type of data is the balance of the checking account considered to be?

110. What type of data is the number of 20-year-olds considered to be?

111. On average, a busy emergency room gets a patient with a shotgun wound about once per week. We are interested in the number of patients with a shotgun wound the emergency room gets per 28 days.

A. Define the random variable X.

B. State the distribution for X.

C. Find the probability that the emergency room gets no patients with shotgun wounds in the next 28 days.

Use the following information to answer the next two exercises. The probability that a certain slot machine will pay back money when a quarter is inserted is .30. Assume that each play of the slot machine is independent from each other. A person puts in 15 quarters for 15 plays.

112. Is the expected number of plays of the slot machine that will pay back money greater than, less than, or the same as the median? Explain your answer.

113. Is it likely that exactly eight of the 15 plays would pay back money? Justify your answer numerically.

114. A game is played with the following rules:

- It costs $10 to enter.
- A fair coin is tossed four times.
- If you do not get four heads or four tails, you lose your $10.
- If you get four heads or four tails, you get back your $10, plus $30 more.

Over the long run of playing this game, what are your expected earnings?

115.

- The mean grade on a math exam in Rachel's class was 74, with a standard deviation of five. Rachel earned an 80.
- The mean grade on a math exam in Becca's class was 47, with a standard deviation of two. Becca earned a 51.
- The mean grade on a math exam in Matt's class was 70, with a standard deviation of eight. Matt earned an 83.

Find whose score was the best, compared to his or her own class. Justify your answer numerically.

Use the following information to answer the next two exercises. A random sample of 70 compulsive gamblers were asked the number of days they go to casinos per week. The results are given in the following graph.

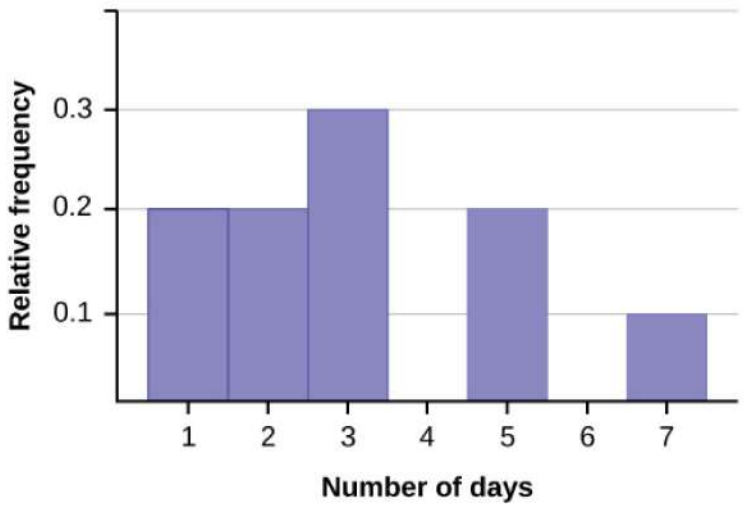


Figure A5

116. Find the number of responses that were five.

117. Find the mean, standard deviation, the median, the first quartile, the third quartile, and the *IQR*.

118. Based upon research at De Anza College, it is believed that about 19 percent of the student population speaks a language other than English at home. Suppose that a study was done this year to see if that percent has decreased. Ninety-eight students were randomly surveyed with the following results: Fourteen said that they speak a language other than English at home.

A. State an appropriate null hypothesis.
B. State an appropriate alternative hypothesis.
C. Define the random variable, P'.
D. Calculate the test statistic.
E. Calculate the *p*-value.
F. At the 5 percent level of decision, what is your decision about the null hypothesis?
G. What is the Type I error?
H. What is the Type II error?

119. Assume that you are an emergency paramedic called in to rescue victims of an accident. You need to help a patient who is bleeding profusely. The patient is also considered to be a high risk for contracting a blood-borne illness. Assume that the null hypothesis is that the patient does *not* have the a blood-borne illness. What is a Type I error?

120. It is often said that Californians are more casual than the rest of Americans. Suppose that a survey was done to see if the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals. Fifty of each was surveyed with the following results: Fifteen Californians wear jeans to work and six non-

Californians wear jeans to work.
Let C = Californian professional; NC = non-Californian professional

A. State appropriate null and alternate hypotheses.

B. Define the random variable.

C. Calculate the test statistic and p-value.

D. At the 5 percent significance level, what is your decision?

E. What is the Type I error?

F. What is the Type II error?

Use the following information to answer the next two exercises. A group of statistics students have developed a technique that they feel will lower their anxiety level on statistics exams. They measured their anxiety level at the start of the quarter and again at the end of the quarter. Recorded is the paired data in that order: (1,000, 900); (1,200, 1,050); (600, 700); (1,300, 1,100); (1,000, 900); (900, 900).

121. This is a test of (pick the best answer) —

A. large samples, and independent means

B. small samples, and independent means

C. dependent means

122. State the distribution to use for the test.

Chapter 12

Use the following information to answer the next two exercises. A recent survey of U.S. teenagers was answered by 720 teenagers, age 15–18. Six percent of teenagers surveyed said they are planning on going to college in another country. We are interested in the true proportion of U.S. teens, ages 15–18, who are planning on going to college in another country.

123. Find the 95 percent confidence interval for the true proportion of U.S. teens, ages 15–19, who are planning to go to college in another country.

124. The report also stated that the results of the survey are accurate to within ±3.7 percent at the 95 percent confidence level. Suppose that a new study is to be done. It is desired to be accurate to within 2 percent of the 95 percent confidence level. What is the minimum number that should be surveyed?

125. Given $X \sim Exp\left(\frac{1}{3}\right)$. Sketch the graph that depicts: $P(x > 1)$.

Use the following information to answer the next three exercises. The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the mean amount of money a customer spends in one trip to the supermarket is $72.

126. Find the probability that one customer spends less than $72 in one trip to the supermarket?

127. Suppose five customers pool their money. How much money altogether would you expect the five customers to spend in one trip to the supermarket in dollars?

128. State the distribution to use if you want to find the probability that the mean amount spent by five customers in one trip to the supermarket is less than $60.

Chapter 13

Use the following information to answer the next two exercises. Suppose that the probability of a drought in any independent year is 20 percent. Out of those years in which a drought occurs, the probability of water rationing is 10 percent. However, in any year, the probability of water rationing is 5 percent.

129. What is the probability of both a drought *and* water rationing occurring?

130. Out of the years with water rationing, find the probability that there is a drought.

Use the following information to answer the next three exercises.

	Apple	Pumpkin	Pecan
Female	40	10	30
Male	20	30	10

Table A7

131. Suppose that one individual is randomly chosen. Find the probability that the person's favorite pie is apple *or* the person is male.

132. Suppose that one male is randomly chosen. Find the probability his favorite pie is pecan.

133. Conduct a hypothesis test to determine if favorite pie type and gender are independent.

Use the following information to answer the next two exercises. Let's say that the probability that an adult watches the news at least once per week is .60.

134. We randomly survey 14 people. On average, how many people do we expect to watch the news at least once per week?

135. We randomly survey 14 people. Of interest is the number that watch the news at least once per week. State the distribution of X. $X \sim$ _____.

136. The following histogram is most likely to be a result of sampling from which distribution?

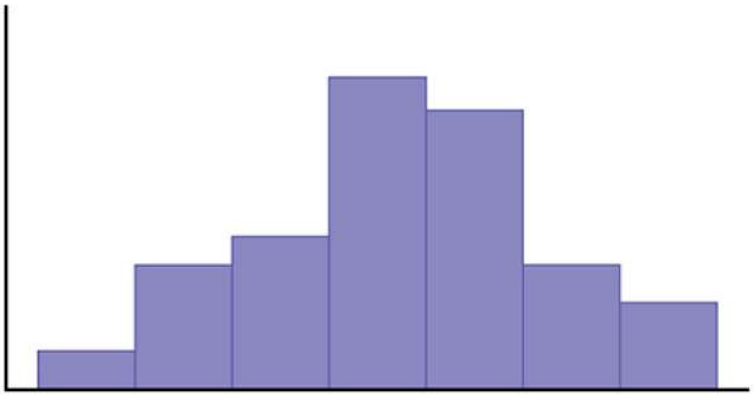


Figure A6

A. Chi-square
B. Geometric
C. Uniform
D. Binomial

137. The ages of De Anza evening students is known to be normally distributed with a population mean of 40 and a population standard deviation of six. A sample of six De Anza evening students reported their ages in years as: 28; 35; 47; 45; 30; 50. Find the probability that the mean of six ages of randomly chosen students is less than 35 years. Hint—Find the sample mean.

138. A math exam was given to all the fifth grade children attending Country School. Two random samples of scores were taken. The null hypothesis is that the mean math scores for boys and girls in fifth grade are the same. Conduct a hypothesis test.

	n	$\bar{x}$	s^2
Boys	55	82	29
Girls	60	86	46

Table A8

139. In a survey of 80 males, 55 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. Conduct a hypothesis test.

140. Which of the following is preferable when designing a hypothesis test?

A. Maximize α and minimize β

B. Minimize α and maximize β

C. Maximize α and β

D. Minimize α and β

Use the following information to answer the next three exercises. One hundred twenty people were surveyed as to their favorite beverage. The results are below.

Beverage/Age	0–9	10–19	20–29	30+	Totals
Milk	14	10	6	0	30
Soda	3	8	26	15	52
Juice	7	12	12	7	38
Totals	24	330	44	22	120

Table A9

141. Are the events of milk and 30+—

A. independent events? Justify your answer.

B. mutually exclusive events? Justify your answer.

142. Suppose that one person is randomly chosen. Find the probability that person is 10–19 given that he or she prefers juice.

143. Are Preferred Beverage and Age independent events? Conduct a hypothesis test.

144. Given the following histogram, which distribution is the data most likely to come from?

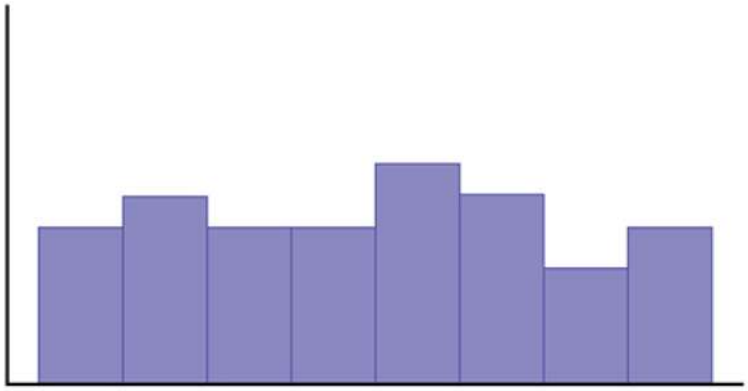


Figure A7

A. Uniform
B. Exponential
C. Normal
D. Chi-square

Solutions

Chapter 3

1. C Parameter

2. A Population

3. B Statistic

4. D Sample

5. E Variable

6. quantitative continuous

7.

A. 2.27
B. 3.04
C. –1, 4, 4

8. Answers will vary.

Chapter 4

9. C (.80)(.30)

10. B No, and they are not mutually exclusive either.

11. A All employed adult women

12. .5773

13. .0522

14. B The middle fifty percent of the members lost from 2 to 8.5 lbs.

15. C All of the data have the same value.

16. C The lowest data value is the median.

17. .279

18. B No, I expect to come out behind in money.

19. X = the number of patients calling in claiming to have the flu, who actually have the flu.
X = 0, 1, 2, …25

20. $B(25, .04)$

21. .0165

22. 1

23. C Quantitative discrete

24. all words used by Tom Clancy in his novels

Chapter 5

25.

A. 24 percent

B. 27 percent

26. qualitative

27. .36

28. .7636

29.

A. no

B. no

30. $B(10, .76)$

31. .9330

32.

A. X = the number of questions posted to the statistics listserv per day.

B. X = 0, 1, 2,…

C. $X \sim P(2)$

D. 0

33. $150

34. Matt

35.

A. False

B. True

C. False

D. False

36. 16

37. first quartile: 2
second quartile: 2
third quartile: 3

38. 0.5

39. $\frac{7}{15}$

40. $\frac{2}{15}$

Chapter 6

41.

A. True

B. True

C. False – the median and the mean are the same for this symmetric distribution.

D. True

42.

A. 8

B. 8

C. $P(x < k) = 0.65 = (k - 3)\left(\frac{1}{10}\right)$. $k = 9.5$

43.

A. False – $\frac{3}{4}$ of the data are at most five.

B. True – each quartile has 25 percent of the data.

C. False – that is unknown.

D. False – 50 percent of the data are four or less.

44. D G and H are independent events.

45.

A. False – J and K are independent so they are not mutually exclusive which would imply dependency (meaning $P(J$ AND $K)$ is not 0).

B. False – see answer c.

C. True – $P(J$ OR $K) = P(J) + P(K) - P(J$ AND $K) = P(J) + P(K) - P(J)P(K) = .3 + .6 - (.3)(.6) = .72$. Note the $P(J$ AND $K) = P(J)P(K)$ because J and K are independent.

D. False – J and K are independent so $P(J) = P(J|K)$.

46. A $P(5)$

Chapter 7

47. A $U(0, 4)$

48. B 2 hours

49. A $\frac{1}{4}$

50.

A. .7165

B. 4.16

C. 0

51. C 5 years

52. C exponential

53. .63

54. A $B(14, .20)$

55. A $B(14, .20)$

Chapter 8

56. C The mean amount of weight lost by 15 people on the special weight-loss diet.

57. .9951

58. 12.99

59. C $\frac{1}{2}$

60. B .60

61. C $N(60, 5.477)$

62. .9990

63. A eight days

64. C .7500

65. A 80 percent

66. B 35 percent

67. B no

68. B Quantitative continuous

69. C 150

70. D .06

71. C .44

72. B 0

Chapter 9

73. D Matt is shorter than the average 14 year old boy.

74. Answers will vary.

75.

x	Relative Frequency	Cumulative Relative Frequency
1	.3	.3
2	.2	.2
4	.4	.4
5	.1	.1

Table A10

76.

A. 2.8

B. 1.48

C. 90 percent

77. $M = 3$; $Q_1 = 1$; $Q_3 = 4$

78. 1 and 4

79. D $\frac{8}{70}$

80. C $\frac{40}{70}$

81. A $\frac{9}{19}$

82. B False

83. B False

84. B False

85.

A. X = the number of pies Lee bakes every day.

B. $P(20)$

C. .1122

86. CI: (5.25, 8.48)

87.

A. uniform

B. exponential

C. normal

Chapter 10

88. $\frac{77}{100}$

89. $\frac{12}{42}$

90.

A. False

B. False

C. True

D. False

91. $N(180, 16.43)$

92. A The distribution for $\bar{X}$ is still uniform with the same mean and standard deviation as the distribution for X.

93. C The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for X.

94. $N\left(2, \frac{.25}{\sqrt{16}}\right)$

95. Answers will vary.

96. .5000

97. 7.6

98. 5

99. .9431

Chapter 11

100. 7.5

101. .0122

102. $N(7, .63)$

103. .9911

104. B exponential

105.

A. True

B. False

C. False

106. Answers will vary.

107. Student's t with $df = 15$

108. (560.07, 719.93)

109. quantitative continuous data

110. quantitative discrete data

111.

A. X = the number of patients with a shotgun wound the emergency room gets per 28 days.

B. $P(4)$

C. .0183

112. greater than

113. no; $P(x = 8) = .0348$

114. You will lose $5.

115. Becca

116. 14

117. sample mean = 3.2
sample standard deviation = 1.85
median = 3
$Q_1 = 2$
$Q_3 = 5$
$IQR = 3$

118. d. $z = –1.19$
e. .1171
f. Do not reject the null hypothesis.

119. We conclude that the patient does have the illness when, in fact, the patient does not.

120. c. $z = 2.21$; $p = .0136$
d. Reject the null hypothesis.
e. We conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is not greater.
f. We cannot conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is greater.

121. C dependent means

122. t_5

Chapter 12

123. (.0424, .0770)

124. 2,401

125. Check student's solution.

126. .6321

127. $360

128. $N\left(72, \frac{72}{\sqrt{5}}\right)$

Chapter 13

129. .02

130. .40

131. $\frac{100}{140}$

132. $\frac{10}{60}$

133. *p*-value = 0; reject the null hypothesis; conclude that they are dependent events

134. 8.4

135. $B(14, .60)$

136. D Binomial

137. .3669

138. *p*-value = .0006; reject the null hypothesis; conclude that the averages are not equal

139. *p*-value = 0; reject the null hypothesis; conclude that the proportion of males is higher

140. minimize α and β

141.

A. no

B. yes, $P(M \text{ AND } 30+) = 0$

142. $\frac{12}{38}$

143. no; *p*-value = 0

144. A uniform

References

Baran, D. (2010). Twenty percent of Americans have never used email. Retrieved from http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email.

Parade Magazine. (n.d.). Retrieved from https://parade.com/.

San Jose Mercury News. (n.d.). Retrieved from http://www.mercurynews.com/.

APPENDIX B: APPENDIX B PRACTICE TESTS (1–4) AND FINAL EXAMS

Practice Test 1

1.1: Definitions of Statistics, Probability, and Key Terms

Use the following information to answer the next three exercises. A grocery store is interested in how much money, on average, their customers spend each visit in the produce department. Using their store records, they draw a sample of 1,000 visits and calculate each customer's average spending on produce.

1. Identify the population, sample, parameter, statistic, variable, and data for this example.

A. population
B. sample
C. parameter
D. statistic
E. variable
F. data

2. What kind of data is *amount of money spent on produce per visit*?

A. Qualitative
B. Quantitative-continuous
C. Quantitative-discrete

3. The study finds that the mean amount spent on produce per visit by the customers in the sample is $12.84. This is an example of a

A. Population
B. Sample
C. Parameter
D. Statistic
E. Variable

1.2: Data, Sampling, and Variation in Data and Sampling

Use the following information to answer the next two exercises. A health club is interested in knowing how many times a typical member uses the club in a week. They decide to ask every tenth customer on a specified day to complete a short survey, including information about how many times they have visited the club in the past week.

4. What kind of a sampling design is this?

A. Cluster
B. Stratified

C. Simple random

D. Systematic

5. *Number of visits per week* is what kind of data?

A. Qualitative

B. Quantitative-continuous

C. Quantitative-discrete

6. Describe a situation in which you would calculate a parameter, rather than a statistic.

7. The U.S. federal government conducts a survey of high school seniors concerning their plans for future education and employment. One question asks whether they are planning to attend a four-year college or university in the following year. Fifty percent answer yes to this question. That 50 percent is a

A. Parameter

B. Statistic

C. Variable

D. Data

8. Imagine that the U.S. federal government had the means to survey all high school seniors in the United States concerning their plans for future education and employment, and found that 50 percent were planning to attend a four-year college or university in the following year. This 50 percent is an example of a

A. Parameter

B. Dtatistic

C. Variable

D. Data

Use the following information to answer the next three exercises. A survey of a random sample of 100 nurses working at a large hospital asked how many years they had been working in the profession. Their answers are summarized in the following (incomplete) table.

9. Fill in the blanks in the table and round your answers to two decimal places for the Relative Frequency and Cumulative Relative Frequency cells.

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
< 5	25		
5–10	30		
> 10	empty		

Table B1

10. What proportion of nurses have five or more years of experience?

11. What proportion of nurses have 10 or fewer years of experience?

12. Describe how you might draw a random sample of 30 students from a lecture class of 200 students.

13. Describe how you might draw a stratified sample of students from a college, where the strata are the students' class standing (freshman, sophomore, junior, or senior).

14. A manager wants to draw a sample, without replacement, of 30 employees from a workforce of 150. Describe how the chance of being selected will change over the course of drawing the sample.

15. The manager of a department store decides to measure employee satisfaction by selecting four departments at random, and conducting interviews with all the employees in those four departments. What type of survey design is this?

A. Cluster

B. Stratified

C. Simple random

D. Systematic

16. A popular American television sports program conducts a poll of viewers to see which team they believe will win the National Football League (NFL) championship this year. Viewers vote by calling a number displayed on the television screen and telling the operator which team they think will win. Do you think that those who participate in this poll are representative of all football fans in America?

17. Two researchers studying vaccination rates independently draw samples of 50 children, aged three–18 months, from a large urban area, and determine if they are up to date on their vaccinations. One researcher finds that 84 percent of the children in her sample are up to date, and the other finds that 86 percent in his sample are up to date. Assuming both followed proper sampling procedures and did their calculations correctly, what is a likely explanation for this discrepancy?

18. A high school increased the length of the school day from 6.5 to 7.5 hours. Students who wished to attend this high school were required to sign contracts pledging to put forth their best effort on their school work and to obey the school rules; if they did not wish to do so, they could attend another high school in the district. At the end of one year, student performance on statewide tests had increased by 10 percentage points over the previous year. Does this prove that a longer school day improves student achievement?

19. You read a newspaper article reporting that eating almonds leads to increased life satisfaction. The study was conducted by the Almond Growers Association, and was based on a randomized survey asking people about their consumption of various foods, including almonds, and also about their satisfaction with different aspects of their life. Does anything about this poll lead you to question its conclusion?

20. Why is non-response a problem in surveys?

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. Compute the mean of the following numbers, and report your answer using one more decimal place than is present in the original data:
14, 5, 18, 23, 6

1.4: Experimental Design and Ethics

22. A psychologist is interested in whether the size of tableware (bowls, plates, etc.) influences how much college students eat. He randomly assigns 100 college students to one of two groups. The first is served a meal using normal-sized tableware, while the second is served the same meal but using tableware that it 20 percent smaller than normal. He records how much food is consumed by each group. Identify the following components of this study.

A. population

B. sample

C. experimental units

D. explanatory variable

E. treatment

F. response variable

23. A researcher analyzes the results of the Scholastic Aptitude Test (SAT) over a five-year period and finds that male students on average score higher on the math section, and female students on average score higher on the verbal section. She concludes that these observed differences in test performance are due to genetic factors. Explain how lurking variables could offer an alternative explanation for the observed differences in test scores.

24. Explain why it would not be possible to use random assignment to study the health effects of exercise.

25. A professor conducts a telephone survey of a city's population by drawing a sample of numbers from the phone book and having her student assistants call each of the selected numbers once to administer the survey. What are some sources of

bias with this survey?

26. A professor offers extra credit to students who take part in her research studies. What is an ethical problem with this method of recruiting subjects?

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

Use the following information to answer the next four exercises. The midterm grades on a chemistry exam, graded on a scale of 0 to 100, were
62, 64, 65, 65, 68, 70, 72, 72, 74, 75, 75, 75, 76, 78, 78, 81, 83, 83, 84, 85, 87, 88, 92, 95, 98, 98, 100, 100, 740

27. Do you see any outliers in this data? If so, how would you address the situation?

28. Construct a stem plot for this data, using only the values in the range zero–100.

29. Describe the distribution of exam scores.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. In a class of 35 students, seven students received scores in the 70–79 range. What is the relative frequency of scores in this range?

Use the following information to answer the next three exercises. You conduct a poll of 30 students to see how many classes they are taking this term. Your results are
1; 1; 1; 1
2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4; 4; 4; 4
5; 5; 5; 5

31. You decide to construct a histogram of this data. What will be the range of your first bar, and what will be the central point?

32. What will be the widths and central points of the other bars?

33. Which bar in this histogram will be the tallest, and what will be its height?

34. You get data from the U.S. Census Bureau on the median household income for your city, and decide to display it graphically. Which is the better choice for this data, a bar graph or a histogram?

35. You collect data on the color of cars driven by students in your statistics class, and want to display this information graphically. Which is the better choice for this data, a bar graph or a histogram?

2.3: Measures of the Location of the Data

36. Your daughter brings home test scores showing that she scored in the 80^{th} percentile in math and the 76^{th} percentile in reading for her grade. Interpret these scores.

37. You have to wait 90 minutes in the emergency room of a hospital before you can see a doctor. You learn that your wait time was in the 82^{nd} percentile of all wait times. Explain what this means, and whether you think it is good or bad.

2.4: Box Plots

Use the following information to answer the next three exercises. 1; 1; 2; 3; 4; 4; 5; 5; 6; 7; 7; 8; 9

38. What is the median for this data?

39. What is the first quartile for this data?

40. What is the third quartile for this data?

Use the following information to answer the next four exercises. This box plot represents scores on the final exam for a physics class.

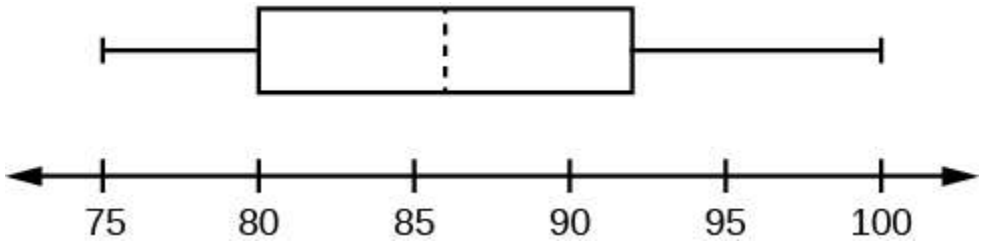


Figure B1

41. What is the median for this data, and how do you know?

42. What are the first and third quartiles for this data, and how do you know?

43. What is the interquartile range for this data?

44. What is the range for this data?

2.5: Measures of the Center of the Data

45. In a marathon, the median finishing time was 3:35:04 (three hours, 35 minutes, and four seconds). You finished in 3:34:10. Interpret the meaning of the median time, and discuss your time in relation to it.

Use the following information to answer the next three exercises. The values, in thousands of dollars, for houses on a block, are 45; 47; 47.5; 51; 53.5; 125.

46. Calculate the mean for this data.

47. Calculate the median for this data.

48. Which do you think better reflects the average value of the homes on this block?

2.6: Skewness and the Mean, Median, and Mode

49. In a left-skewed distribution, which is greater?

A. The mean

B. The media

C. The mode

50. In a right-skewed distribution, which is greater?

A. The mean

B. The median

C. The mode

51. In a symmetrical distribution, what will be the relationship among the mean, median, and mode?

2.7: Measures of the Spread of the Data

Use the following information to answer the next four exercises. 10; 11; 15; 15; 17; 22

52. Compute the mean and standard deviation for this data; use the sample formula for the standard deviation.

53. What number is two standard deviations above the mean of this data?

54. Express the number 13.7 in terms of the mean and standard deviation of this data.

55. In a biology class, the scores on the final exam were normally distributed, with a mean of 85 and a standard deviation of five. Susan got a final exam score of 95. Express her exam result as a *z* score, and interpret its meaning.

3.1: Terminology

Use the following information to answer the next two exercises. You have a jar full of marbles: 50 are red, 25 are blue, and 15 are yellow. Assume you draw one marble at random for each trial and replace it before the next trial.
Let $P(R)$ = the probability of drawing a red marble.
Let $P(B)$ = the probability of drawing a blue marble.

Let $P(Y)$ = the probability of drawing a yellow marble.

56. Find $P(B)$.

57. Which is more likely, drawing a red marble or a yellow marble? Justify your answer numerically.

Use the following information to answer the next two exercises. The following are probabilities describing a group of college students.
Let $P(M)$ = the probability that the student is male
Let $P(F)$ = the probability that the student is female
Let $P(E)$ = the probability the student is majoring in education
Let $P(S)$ = the probability the student is majoring in science

58. Write the symbols for the probability that a student, selected at random, is both female and a science major.

59. Write the symbols for the probability that the student is an education major, given that the student is male.

3.2: Independent and Mutually Exclusive Events

60. Events A and B are independent.
If $P(A) = 0.3$ and $P(B) = 0.5$, find $P(A \text{ AND } B)$.

61. C and D are mutually exclusive events.
If $P(C) = 0.18$ and $P(D) = 0.03$, find $P(C \text{ OR } D)$.

3.3: Two Basic Rules of Probability

62. In a high school graduating class of 300, 200 students are going to college, 40 are planning to work full-time, and 80 are taking a gap year. Are these events mutually exclusive?

Use the following information to answer the next two exercises. An archer hits the center of the target (the bullseye) 70 percent of the time. However, she is a streak shooter, and if she hits the center on one shot, her probability of hitting it on the shot immediately following is 0.85. Written in probability notation
$P(A) = P(B) = P(\text{hitting the center on one shot}) = 0.70$
$P(B|A) = \text{P(hitting the center on a second shot, given that she hit it on the first)} = 0.85$

63. Calculate the probability that she will hit the center of the target on two consecutive shots.

64. Are $P(A)$ and $P(B)$ independent in this example?

3.4: Contingency Tables

Use the following information to answer the next three exercises. The following contingency table displays the number of students who report studying at least 15 hours per week, and how many made the honor roll in the past semester.

	Honor Roll	No Honor Roll	Total
Study at least 15 hours/week		200	
Study less than 15 hours/week	125	193	
Total			1,000

Table B2

65. Complete the table.

66. Find P (honor roll|study at least 15 hours per week).

67. What is the probability a student studies less than 15 hours per week?

68. Are the events *study at least 15 hours per week* and *makes the honor roll* independent? Justify your answer numerically.

3.5: Tree and Venn Diagrams

69. At a high school, some students play on the tennis team and some play on the soccer team, but neither plays both tennis and soccer. Draw a Venn diagram illustrating this.

70. At a high school, some students play tennis, some play soccer, and some play both. Draw a Venn diagram illustrating this.

Practice Test 1 Solutions

1.1: Definitions of Statistics, Probability, and Key Terms

1.

A. population: all the shopping visits by all the store's customers

B. sample: the 1,000 visits drawn for the study

C. parameter: the average expenditure on produce per visit by all the store's customers

D. statistic: the average expenditure on produce per visit by the sample of 1,000

E. variable: the expenditure on produce for each visit

F. data: the dollar amounts spent on produce; for instance, $15.40, $11.53, etc.

2. C

3. D

1.2: Data, Sampling, and Variation in Data and Sampling

4. D

5. C

6. Answers will vary.
Sample Answer: Any solution in which you use data from the entire population is acceptable. For instance, a professor might calculate the average exam score for her class: Because the scores of all members of the class were used in the calculation, the average is a parameter.

7. B

8. A

9.

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
< 5	25	0.25	0.25
5–10	30	0.30	0.55
> 10	45	0.45	1

Table B3

10. 0.75

11. 0.55

12. Answers will vary.
Sample Answer: One possibility is to obtain the class roster and assign each student a number from 1 to 200. Then, use a random number generator or table of random number to generate 30 numbers between 1 and 200, and select the students matching the random numbers. It would also be acceptable to write each student's name on a card, shuffle them in a box, and draw 30 names at random.

13. One possibility would be to obtain a roster of students enrolled in the college, including the class standing for each student. Then, you would draw a proportionate random sample from within each class. For instance, if 30 percent of the students in the college are freshman, then 30 percent of your sample would be drawn from the freshman class.

14. For the first person picked, the chance of any individual being selected is one in 150. For the second person, it is one in 149, for the third it is one in 148, and so on. For the 30th person selected, the chance of selection is one in 121.

15. A

16. No. There are at least two chances for bias. First, the viewers of this particular program may not be representative of American football fans as a whole. Second, the sample will be self-selected, because people have to make a phone call in order to take part, and those people are probably not representative of the American football fan population as a whole.

17. These results (84 percent in one sample, 86 percent in the other) are probably due to sampling variability. Each researcher drew a different sample of children, and you would not expect them to get exactly the same result, although you would expect the results to be similar, as they are in this case.

18. No. The improvement could also be due to self-selection: Only motivated students were willing to sign the contract, and they would have done well even in a school with 6.5 hour days. Because both changes were implemented at the same time, it is not possible to separate out their influence.

19. At least two aspects of this poll are troublesome. The first is that it was conducted by a group who would benefit by the result—almond sales are likely to increase if people believe that eating almonds will make them happier. The second is that this poll found that almond consumption and life satisfaction are correlated, but it does not establish that eating almonds causes satisfaction. It is equally possible, for instance, that people with higher incomes are more likely to eat almonds and are also more satisfied with their lives.

20. You want the sample of people who take part in a survey to be representative of the population from which they are drawn. People who refuse to take part in a survey often have different views than those who do participate, and so even a random sample may produce biased results if a large percentage of those selected refuse to participate in a survey.

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. 13.2

1.4: Experimental Design and Ethics

22.

A. population: all college students
B. sample: the 100 college students in the study
C. experimental units: each individual college student who participated
D. explanatory variable: the size of the tableware
E. treatment: tableware that is 20 percent smaller than normal
F. response variable: the amount of food eaten

23. There are many lurking variables that could influence the observed differences in test scores. Perhaps the boys, on average, have taken more math courses than the girls, and the girls have taken more English classes than the boys. Perhaps the boys have been encouraged by their families and teachers to prepare for a career in math and science, and thus have put more effort into studying math, while the girls have been encouraged to prepare for fields like communication and psychology that are more focused on language use. A study design would have to control for these and other potential lurking variables (anything that could explain the observed difference in test scores, other than the genetic explanation) in order to draw a scientifically sound conclusion about genetic differences.

24. To use random assignment, you would have to be able to assign people to either exercise or not exercise. Because exercise has many beneficial effects, this would not be an ethical experiment. We will study people who chose to exercise and compare them to people who chose not to exercise, and try to control for the other ways those two groups may differ (lurking variables).

25. Sources of bias include the fact that not everyone has a telephone, that cell phone numbers are often not listed in published directories, and that an individual might not be at home at the time of the phone call; all these factors make it likely that the respondents to the survey will not be representative of the population as a whole.

26. Research subjects should not be coerced into participation, and offering extra credit in exchange for participation could be construed as coercion. In addition, this method will result in a volunteer sample, which cannot be assumed to be representative of the population as a whole.

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

27. The value 740 is an outlier, because the exams were graded on a scale of zero to 100, and 740 is far outside that range. It may be a data entry error, with the actual score being 74, so the professor should check that exam again to see what the actual score was.

28.

Stem	Leaf
6	2 4 5 5 8
7	0 2 2 4 5 5 5 6 8 8
8	1 3 3 4 5 7 8
9	2 5 8 8
10	0 0

Table B4

29. Most scores on this exam were in the range of 70–89, with a few scoring in the 60–69 range, and a few in the 90–100 range.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. $RF = \frac{7}{35} = 0.2$

31. The range will be 0.5–1.5, and the central point will be 1.

32. Range 1.5–2.5, central point 2; range 2.5–3.5, central point 3; range 3.5–4.5, central point 4; range 4.5–5.5, central point 5.

33. The bar from 3.5 to 4.5, with a central point of 4, will be tallest; its height will be nine, because there are nine students taking four courses.

34. The histogram is a better choice, because income is a continuous variable.

35. A bar graph is the better choice, because this data is categorical rather than continuous.

2.3: Measures of the Location of the Data

36. Your daughter scored better than 80 percent of the students in her grade on math and better than 76 percent of the students in reading. Both scores are very good, and place her in the upper quartile, but her math score is slightly better in relation to her peers than her reading score.

37. You had an unusually long wait time, which is bad: 82 percent of patients had a shorter wait time than you, and only 18 percent had a longer wait time.

2.4: Box Plots

38. 5

39. 3

40. 7

41. The median is 86, as represented by the vertical line in the box.

42. The first quartile is 80, and the third quartile is 92, as represented by the left and right boundaries of the box.

43. *IQR* = 92 – 80 = 12

44. Range = 100 – 75 = 25

2.5: Measures of the Center of the Data

45. Half the runners who finished the marathon ran a time faster than 3:35:04, and half ran a time slower than 3:35:04. Your time is faster than the median time, so you did better than more than half of the runners in this race.

46. 61.5, or $61,500

47. 49.25, or $49,250

48. The median, because the mean is distorted by the high value of one house.

2.6: Skewness and the Mean, Median, and Mode

49. C

50. A

51. They will all be fairly close to one another.

2.7: Measures of the Spread of the Data

52. Mean: 15
Standard deviation: 4.3

$$\mu = \frac{10 + 11 + 15 + 15 + 17 + 22}{6} = 15$$

$$s = \sqrt{\frac{\sum \left(x - \bar{x}\right)^2}{n - 1}} = \sqrt{\frac{94}{5}} = 4.3$$

53. 15 + (2)(4.3) = 23.6

54. 13.7 is one standard deviation below the mean of this data, because 15 – 4.3 = 10.7

55. $z = \frac{95 - 85}{5} = 2.0$

Susan's *z* score was 2.0, meaning she scored two standard deviations above the class mean for the final exam.

3.1: Terminology

56. $P(B) = \frac{25}{90} = 0.28$

57. Drawing a red marble is more likely.

$P(R) = \frac{50}{80} = 0.62$

$P(Y) = \frac{15}{80} = 0.19$

58. $P(F \text{ AND } S)$

59. $P(E|M)$

3.2: Independent and Mutually Exclusive Events

60. $P(A \text{ AND } B) = (0.3)(0.5) = 0.15$

61. $P(C \text{ OR } D) = 0.18 + 0.03 = 0.21$

3.3: Two Basic Rules of Probability

62. No, they cannot be mutually exclusive, because they add up to more than 300. Therefore, some students must fit into two or more categories (e.g., both going to college and working full time).

63. $P(A \text{ and } B) = (P(B|A))(P(A)) = (0.85)(0.70) = 0.595$

64. No. If they were independent, $P(B)$ would be the same as $P(B|A)$. We know this is not the case, because $P(B) = 0.70$ and $P(B|A) = 0.85$.

3.4: Contingency Tables

65.

	Honor roll	No honor roll	Total
Study at least 15 hours/week	482	200	682
Study less than 15 hours/week	125	193	318
Total	607	393	1,000

Table B5

66. $P(\text{honor roll}|\text{study at least 15 hours word per week}) = \frac{482}{1{,}000} = 0.482$

67. $P(\text{study less than 15 hours word per week}) = \frac{125 + 193}{1{,}000} = 0.318$

68. Let $P(S)$ = study at least 15 hours per week
Let $P(H)$ = make the honor roll
From the table, $P(S) = 0.682$, $P(H) = 0.607$, and $P(S \text{ AND } H) = 0.482$.
If $P(S)$ and $P(H)$ were independent, then $P(S \text{ AND } H)$ would equal $(P(S))(P(H))$.
However, $(P(S))(P(H)) = (0.682)(0.607) = 0.414$, while $P(S \text{ AND } H) = 0.482$.
Therefore, $P(S)$ and $P(H)$ are not independent.

3.5: Tree and Venn Diagrams

69.

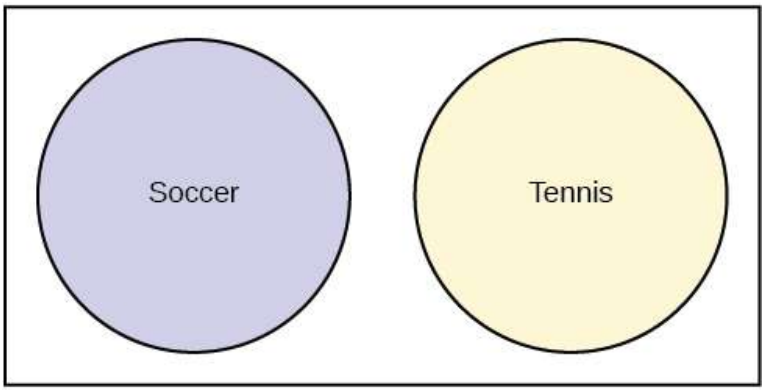


Figure B2

70.

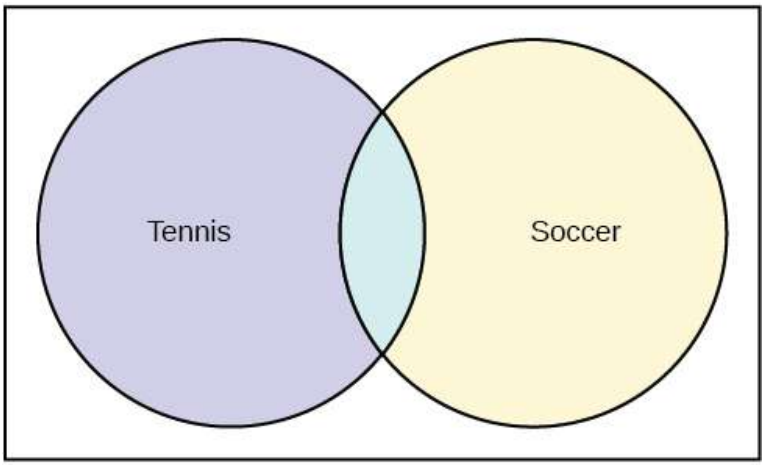


Figure B3

Practice Test 2

4.1: Probability Distribution Function (PDF) for a Discrete Random Variable

Use the following information to answer the next five exercises. You conduct a survey among a random sample of students at a particular university. The data collected includes their major, the number of classes they took the previous semester, and the amount of money they spent on books purchased for classes in the previous semester.

1. If X = student's major, then what is the domain of X?

2. If Y = the number of classes taken in the previous semester, what is the domain of Y?

3. If Z = the amount of money spent on books in the previous semester, what is the domain of Z?

4. Why are X, Y, and Z in the previous example random variables?

5. After collecting data, you find that, for one case, $z = -7$. Is this a possible value for Z?

6. What are the two essential characteristics of a discrete probability distribution?

Use this discrete probability distribution represented in this table to answer the following six questions. The university library records the number of books checked out by each patron over the course of one day, with the following result:

x	$P(x)$
0	0.20
1	0.45
2	0.20
3	0.10
4	0.05

Table B6

7. Define the random variable X for this example.

8. What is $P(x > 2)$?

9. What is the probability a patron will check out at least one book?

10. What is the probability a patron will take out no more than three books?

11. If the table listed $P(x)$ as 0.15, how would you know that there was a mistake?

12. What is the average number of books taken out by a patron?

4.2: Mean or Expected Value and Standard Deviation

Use the following information to answer the next four exercises. Three jobs are open in a company: one in the accounting department, one in the human resources department, and one in the sales department. The accounting job receives 30 applicants, and the human resources and sales department 60 applicants.

13. If X = the number of applications for a job, use this information to fill in **Table B7**.

x	$P(x)$	$xP(x)$

Table B7

14. What is the mean number of applicants?

15. What is the PDF for X?

16. Add a fourth column to the table, for $(x - \mu)^2 P(x)$.

17. What is the standard deviation of X?

4.3: Binomial Distribution

18. In a binomial experiment, if $p = 0.65$, what does q equal?

19. What are the required characteristics of a binomial experiment?

20. Joe conducts an experiment to see how many times he has to flip a coin before he gets four heads in a row. Does this qualify as a binomial experiment?

Use the following information to answer the next three exercises. In a particular community, 65 percent of households include at least one person who has graduated from college. You randomly sample 100 households in this community. Let X = the number of households including at least one college graduate.

21. Describe the probability distribution of X.

22. What is the mean of X?

23. What is the standard deviation of X?

Use the following information to answer the next four exercises. Joe is the star of his school's baseball team. His batting average is 0.400, meaning that for every 10 times he comes to bat (an at-bat), four of those times he gets a hit. You decide to track his batting performance for his next 20 at-bats.

24. Define the random variable X in this experiment.

25. Assuming Joe's probability of getting a hit is independent and identical across all 20 at-bats, describe the distribution of X.

26. Given this information, what number of hits do you predict Joe will get?

27. What is the standard deviation of X?

4.4: Geometric Distribution

28. What are the three major characteristics of a geometric experiment?

29. You decide to conduct a geometric experiment by flipping a coin until it comes up heads. This takes five trials. Represent the outcomes of this trial, using H for heads and T for tails.

30. You are conducting a geometric experiment by drawing cards from a normal 52-card pack, with replacement, until you draw the Queen of Hearts. What is the domain of X for this experiment?

31. You are conducting a geometric experiment by drawing cards from a normal 52-card deck, without replacement, until you draw a red card. What is the domain of X for this experiment?

Use the following information to answer the next three exercises. In a particular university, 27 percent of students are engineering majors. You decide to select students at random until you choose one that is an engineering major. Let X = the number of students you select until you find one that is an engineering major.

32. What is the probability distribution of X?

33. What is the mean of X?

34. What is the standard deviation of X?

4.5: Hypergeometric Distribution

35. You draw a random sample of 10 students to participate in a survey, from a group of 30, consisting of 16 boys and 14 girls. You are interested in the probability that seven of the students chosen will be boys. Does this qualify as a hypergeometric experiment? List the conditions and whether or not they are met.

36. You draw five cards, without replacement, from a normal 52-card deck of playing cards, and are interested in the probability that two of the cards are spades. What are the group of interest, size of the group of interest, and sample size for this example?

4.6: Poisson Distribution

37. What are the key characteristics of the Poisson distribution?

Use the following information to answer the next three exercises. The number of drivers to arrive at a toll booth in an hour can be modeled by the Poisson distribution.

38. If X = the number of drivers, and the average numbers of drivers per hour is four, how would you express this distribution?

39. What is the domain of X?

40. What are the mean and standard deviation of X?

5.1: Continuous Probability Functions

41. You conduct a survey of students to see how many books they purchased the previous semester, the total amount they paid for those books, the number they sold after the semester was over, and the amount of money they received for the books they sold. Which variables in this survey are discrete, and which are continuous?

42. With continuous random variables, we never calculate the probability that X has a particular value, but we always speak in terms of the probability that X has a value within a particular range. Why is this?

43. For a continuous random variable, why are $P(x < c)$ and $P(x \leq c)$ equivalent statements?

44. For a continuous probability function, $P(x < 5) = 0.35$. What is $P(x > 5)$, and how do you know?

45. Describe how you would draw the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$. What type of a distribution is this?

46. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$. what is the $P(0 < x < 4)$?

5.2: The Uniform Distribution

47. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$, what is the $P(2 < x < 5)$?

Use the following information to answer the next four exercises. The number of minutes that a patient waits at a medical clinic to see a doctor is represented by a uniform distribution between zero and 30 minutes, inclusive.

48. If X equals the number of minutes a person waits, what is the distribution of X?

49. Write the probability density function for this distribution.

50. What is the mean and standard deviation for waiting time?

51. What is the probability that a patient waits less than 10 minutes?

5.3: The Exponential Distribution

52. The distribution of the variable X, representing the average time to failure for an automobile battery, can be written as $X \sim Exp(m)$. Describe this distribution in words.

53. If the value of m for an exponential distribution is 10, what are the mean and standard deviation for the distribution?

54. Write the probability density function for a variable distributed as $X \sim Exp(0.2)$.

6.1: The Standard Normal Distribution

55. Translate this statement about the distribution of a random variable X into words: $X \sim (100, 15)$.

56. If the variable X has the standard normal distribution, express this symbolically.

Use the following information for the next six exercises. According to the World Health Organization, distribution of height in centimeters for girls aged five years and zero months has the distribution $X \sim N(109, 4.5)$.

57. What is the z score for a height of 112 inches?

58. What is the z score for a height of 100 centimeters?

59. Find the z score for a height of 105 centimeters and explain what that means in the context of the population.

60. What height corresponds to a z score of 1.5 in this population?

61. Using the empirical rule, we expect about 68 percent of the values in a normal distribution to lie within one standard deviation above or below the mean. What does this mean, in terms of a specific range of values, for this distribution?

62. Using the empirical rule, about what percentage of heights in this distribution do you expect to be between 95.5 cm and 122.5 cm?

6.2: Using the Normal Distribution

Use the following information to answer the next four exercises. The distributor of raffle tickets claims that 20 percent of the tickets are winners. You draw a sample of 500 tickets to test this proposition.

63. Can you use the normal approximation to the binomial for your calculations? Why or why not.

64. What are the expected mean and standard deviation for your sample, assuming the distributor's claim is true?

65. What is the probability that your sample will have a mean greater than 100?

66. If the *z* score for your sample result is –2, explain what this means, using the empirical rule.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. What does the central limit theorem state with regard to the distribution of sample means?

68. The distribution of results from flipping a fair coin is uniform: Heads and tails are equally likely on any flip, and over a large number of trials, you expect about the same number of heads and tails. Yet if you conduct a study by flipping 30 coins and recording the number of heads, and repeat this 100 times, the distribution of the mean number of heads will be approximately normal. How is this possible?

69. The mean of a normally-distributed population is 50, and the standard deviation is four. If you draw 100 samples of size 40 from this population, describe what you would expect to see in terms of the sampling distribution of the sample mean.

70. *X* is a random variable with a mean of 25 and a standard deviation of two. Write the distribution for the sample mean of samples of size 100 drawn from this population.

71. Your friend is doing an experiment drawing samples of size 50 from a population with a mean of 117 and a standard deviation of 16. This sample size is large enough to allow use of the central limit theorem, so he says the standard deviation of the sampling distribution of sample means will also be 16. Explain why this is wrong, and calculate the correct value.

72. You are reading a research article that refers to *the standard error of the mean*. What does this mean, and how is it calculated?

Use the following information to answer the next six exercises. You repeatedly draw samples of $n = 100$ from a population with a mean of 75 and a standard deviation of 4.5.

73. What is the expected distribution of the sample means?

74. One of your friends tries to convince you that the standard error of the mean should be 4.5. Explain what error your friend made.

75. What is the *z* score for a sample mean of 76?

76. What is the *z* score for a sample mean of 74.7?

77. What sample mean corresponds to a *z* score of 1.5?

78. If you decrease the sample size to 50, will the standard error of the mean be smaller or larger? What would be its value?

Use the following information to answer the next two questions. We use the empirical rule to analyze data for samples of size 60 drawn from a population with a mean of 70 and a standard deviation of 9.

79. What range of values would you expect to include 68 percent of the sample means?

80. If you increased the sample size to 100, what range would you expect to contain 68 percent of the sample means, applying the empirical rule?

7.2: The Central Limit Theorem for Sums

81. How does the central limit theorem apply to sums of random variables?

82. Explain how the rules applying the central limit theorem to sample means, and to sums of a random variable, are similar.

83. If you repeatedly draw samples of size 50 from a population with a mean of 80 and a standard deviation of four, and calculate the sum of each sample, what is the expected distribution of these sums?

Use the following information to answer the next four exercises. You draw one sample of size 40 from a population with a mean of 125 and a standard deviation of seven.

84. Compute the sum. What is the probability that the sum for your sample will be less than 5,000?

85. If you drew samples of this size repeatedly, computing the sum each time, what range of values would you expect to contain 95 percent of the sample sums?

86. What value is one standard deviation below the mean?

87. What value corresponds to a *z* score of 2.2?

7.3: Using the Central Limit Theorem

88. What does the law of large numbers say about the relationship between the sample mean and the population mean?

89. Applying the law of large numbers, which sample mean would you expect to be closer to the population mean: a sample of size 10 or a sample of size 100?

Use this information for the next three questions. A manufacturer makes screws with a mean diameter of 0.15 cm (centimeters) and a range of 0.10 cm to 0.20 cm; within that range, the distribution is uniform.

90. If X = the diameter of one screw, what is the distribution of X?

91. Suppose you repeatedly draw samples of size 100 and calculate their mean. Applying the central limit theorem, what is the distribution of these sample means?

92. Suppose you repeatedly draw samples of 60 and calculate their sum. Applying the central limit theorem, what is the distribution of these sample sums?

Practice Test 2 Solutions

Probability Distribution Function (PDF) for a Discrete Random Variable

1. The domain of X = {English, Mathematics, . . .}, i.e., a list of all the majors offered at the university, plus *undeclared*.

2. The domain of Y = {0, 1, 2, . . .}; i.e., the integers from zero to the upper limit of classes allowed by the university.

3. The domain of Z = any amount of money from zero upwards.

4. Because they can take any value within their domain, and their value for any particular case is not known until the survey is completed.

5. No, because the domain of Z includes only positive numbers (you cannot spend a negative amount of money). Possibly the value –7 is a data entry error, or a special code to indicate that the student did not answer the question.

6. The probabilities must sum to 1.0, and the probabilities of each event must be between 0 and 1, inclusive.

7. Let X = the number of books checked out by a patron.

8. $P(x > 2) = 0.10 + 0.05 = 0.15$

9. $P(x \geq 0) = 1 - 0.20 = 0.80$

10. $P(x \leq 3) = 1 - 0.05 = 0.95$

11. The probabilities would sum to 1.10, and the total probability in a distribution must always equal 1.0.

12. $\bar{x} = 0(0.20) + 1(0.45) + 2(0.20) + 3(0.10) + 4(0.05) = 1.35$

Mean or Expected Value and Standard Deviation

13.

x	$P(x)$	$xP(x)$
30	0.33	9.90
40	0.33	13.20
60	0.33	19.80

Table B8

14. $\bar{x} = 9.90 + 13.20 + 19.80 = 42.90$

15. $P(x = 30) = 0.33$
$P(x = 40) = 0.33$
$P(x = 60) = 0.33$

16.

x	$P(x)$	$xP(x)$	$(x-\mu)^2P(x)$
30	0.33	9.90	$(30-42.90)^2(0.33) = 54.91$
40	0.33	13.20	$(40-42.90)^2(0.33) = 2.78$
60	0.33	19.90	$(60-42.90)^2(0.33) = 96.49$

Table B9

17. $\sigma_x = \sqrt{54.91 + 2.78 + 96.49} = 12.42$

Binomial Distribution

18. $q = 1 - 0.65 = 0.35$

19.

1. There are a fixed number of trials.
2. There are only two possible outcomes, and they add up to one.
3. The trials are independent and conducted under identical conditions.

20. No, because there are not a fixed number of trials

21. $X \sim B(100, 0.65)$

22. $\mu = np = 100(0.65) = 65$

23. $\sigma_x = \sqrt{npq} = \sqrt{100(0.65)(0.35)} = 4.77$

24. X = Joe gets a hit in one at-bat (in one occasion of his coming to bat)

25. $X \sim B(20, 0.4)$

26. $\mu = np = 20(0.4) = 8$

27. $\sigma_x = \sqrt{npq} = \sqrt{20(0.40)(0.60)} = 2.19$

4.4: Geometric Distribution

28.

1. A series of Bernoulli trials are conducted until one is a success, and then the experiment stops.
2. At least one trial is conducted, but there is no upper limit to the number of trials.
3. The probability of success or failure is the same for each trial.

29. $T\ T\ T\ T\ H$

30. The domain of $X = \{1, 2, 3, 4, 5, \ldots n\}$. Because you are drawing with replacement, there is no upper bound to the number of draws that may be necessary.

31. The domain of $X = \{1, 2, 3, 4, 5, 6, 7, 8., 9, 10, 11, 12, \ldots 27\}$. Because you are drawing without replacement, and 26 of the 52 cards are red, you have to draw a red card within the first 17 draws.

32. $X \sim G(0.24)$

33. $\mu = \frac{1}{p} = \frac{1}{0.27} = 3.70$

34. $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.27}{0.27^2}} = 3.16$

4.5: Hypergeometric Distribution

35. Yes, because you are sampling from a population composed of two groups (boys and girls), have a group of interest (boys), and are sampling without replacement (hence, the probabilities change with each pick, and you are not performing Bernoulli trials).

36. The group of interest is the cards that are spades, the size of the group of interest is 13, and the sample size is five.

4.6: Poisson Distribution

37. A Poisson distribution models the number of events occurring in a fixed interval of time or space, when the events are independent and the average rate of the events is known.

38. $X \sim P(4)$

39. The domain of $X = \{0, 1, 2, 3, \ldots\}$; i.e., any integer from 0 upwards.

40. $\mu = 4$

$\sigma = \sqrt{4} = 2$

5.1: Continuous Probability Functions

41. The discrete variables are the number of books purchased, and the number of books sold after the end of the semester. The continuous variables are the amount of money spent for the books, and the amount of money received when they were sold.

42. Because for a continuous random variable, $P(x = c) = 0$, where c is any single value. Instead, we calculate $P(c < x < d)$; i.e., the probability that the value of x is between the values c and d.

43. Because $P(x = c) = 0$ for any continuous random variable.

44. $P(x > 5) = 1 - 0.35 = 0.65$, because the total probability of a continuous probability function is always 1.

45. This is a uniform probability distribution. You would draw it as a rectangle with the vertical sides at 0 and 20, and the horizontal sides at $\frac{1}{10}$ and 0.

46. $P(0 < x < 4) = (4 - 0)\left(\frac{1}{10}\right) = 0.4$

5.2: The Uniform Distribution

47. $P(2 < x < 5) = (5 - 2)\left(\frac{1}{10}\right) = 0.3$

48. $X \sim U(0, 15)$

49. $f(x) = \frac{1}{b - a}$ for $(a \leq x \leq b)$ so $f(x) = \frac{1}{30}$ for $(0 \leq x \leq 30)$

50. $\mu = \frac{a + b}{2} = \frac{0 + 30}{5} = 15.0$

$\sigma = \sqrt{\frac{(b - a)^2}{12}} = \sqrt{\frac{(30 - 0)^2}{12}} = 8.66$

51. $P(x < 10) = (10)\left(\frac{1}{30}\right) = 0.33$

5.3: The Exponential Distribution

52. X has an exponential distribution with decay parameter m and mean and standard deviation $\frac{1}{m}$. In this distribution, there will be relatively large numbers of small values, with values becoming less common as they become larger.

53. $\mu = \sigma = \frac{1}{m} = \frac{1}{10} = 0.1$

54. $f(x) = 0.2e^{-0.2x}$ where $x \geq 0$.

6.1: The Standard Normal Distribution

55. The random variable X has a normal distribution with a mean of 100 and a standard deviation of 15.

56. $X \sim N(0,1)$

57. $z = \frac{x - \mu}{\sigma}$ so $z = \frac{112 - 109}{4.5} = 0.67$

58. $z = \frac{x - \mu}{\sigma}$ so $z = \frac{100 - 109}{4.5} = -2.00$

59. $z = \frac{105 - 109}{4.5} = -0.89$

This girl is shorter than average for her age, by 0.89 standard deviations.

60. 109 + (1.5)(4.5) = 115.75 cm

61. We expect about 68 percent of the heights of girls aged five years and zero months to be between 104.5 cm and 113.5 cm.

62. We expect 99.7 percent of the heights in this distribution to be between 95.5 cm and 122.5 cm, because that range represents the values three standard deviations above and below the mean.

6.2: Using the Normal Distribution

63. Yes, because both np and nq are greater than five.
np = (500)(0.20) = 100 and nq = 500(0.80) = 400

64. $\mu = np = (500)(0.20) = 100$

$\sigma = \sqrt{npq} = \sqrt{500(0.20)(0.80)} = 8.94$

65. Fifty percent, because in a normal distribution, half the values lie above the mean.

66. The results of our sample were two standard deviations below the mean, suggesting it is unlikely that 20 percent of the raffle tickets are winners, as claimed by the distributor, and that the true percentage of winners is lower. Applying the Empirical Rule, if that claim were true, we would expect to see a result this far below the mean only about 2.5 percent of the time.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. The central limit theorem states that if samples of sufficient size are drawn from a population, the distribution of sample means will be normal, even if the distribution of the population is not normal.

68. The sample size of 30 is sufficiently large in this example to apply the central limit theorem. This theorem states that, for samples of sufficient size drawn from a population, the sampling distribution of the sample mean will approach normality, regardless of the distribution of the population from which the samples were drawn.

69. You would not expect each sample to have a mean of 50, because of sampling variability. However, you would expect the sampling distribution of the sample means to cluster around 50, with an approximately normal distribution, so that values close to 50 are more common than values further removed from 50.

70. $\bar{X} \sim N(25, 0.2)$ because $\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$

71. The standard deviation of the sampling distribution of the sample means can be calculated using the formula $\left(\frac{\sigma_x}{\sqrt{n}}\right)$, which in this case is $\left(\frac{16}{\sqrt{50}}\right)$. The correct value for the standard deviation of the sampling distribution of the sample means is therefore 2.26.

72. The standard error of the mean is another name for the standard deviation of the sampling distribution of the sample mean. Given samples of size n drawn from a population with standard deviation σ_x, the standard error of the mean is $\left(\frac{\sigma_x}{\sqrt{n}}\right)$.

73. $X \sim N(75, 0.45)$

74. Your friend forgot to divide the standard deviation by the square root of n.

75. $z = \frac{\bar{x} - \mu_x}{\sigma_x} = \frac{76 - 75}{4.5} = 2.2$

76. $z = \frac{\bar{x} - \mu_x}{\sigma_x} = \frac{74.7 - 75}{4.5} = -0.67$

77. 75 + (1.5)(0.45) = 75.675

78. The standard error of the mean will be larger, because you will be dividing by a smaller number. The standard error of the mean for samples of size $n = 50$ is
$\left(\frac{\sigma_x}{\sqrt{n}}\right) = \frac{4.5}{\sqrt{50}} = 0.64$

79. You would expect this range to include values up to one standard deviation above or below the mean of the sample means. In this case:
$70 + \frac{9}{\sqrt{60}} = 71.16$ and $70 - \frac{9}{\sqrt{60}} = 68.84$ so you would expect 68 percent of the sample means to be between 68.84 and 71.16.

80. $70 + \frac{9}{\sqrt{100}} = 70.9$ and $70 - \frac{9}{\sqrt{100}} = 69.1$ so you would expect 68 percent of the sample means to be between 69.1 and 70.9. Note that this is a narrower interval due to the increased sample size.

7.2: The Central Limit Theorem for Sums

81. For a random variable X, the random variable ΣX will tend to become normally distributed as the size n of the samples used to compute the sum increases.

82. Both rules state that the distribution of a quantity (the mean or the sum) calculated on samples drawn from a population will tend to have a normal distribution as the sample size increases, regardless of the distribution of population from which the samples are drawn.

83. $\Sigma X \sim N(n\mu_x, (\sqrt{n})(\sigma_x))$ so $\Sigma X \sim N(4{,}000, 28.3)$

84. The probability is 0.50, because 5,000 is the mean of the sampling distribution of sums of size 40 from this population. Sums of random variables computed from a sample of sufficient size are normally distributed, and in a normal distribution, half the values lie below the mean.

85. Using the empirical rule, you would expect 95 percent of the values to be within two standard deviations of the mean. Using the formula for the standard deviation is for a sample sum $(\sqrt{n})(\sigma_x) = (\sqrt{40})(7) = 44.3$, so you would expect 95 percent of the values to be between 5,000 + (2)(44.3) and 5,000 – (2)(44.3), or between 4,911.4 and 588.6.

86. $\mu - (\sqrt{n})(\sigma_x) = 5{,}000 - (\sqrt{40})(7) = 4{,}955.7$

87. $5{,}000 + (2.2)(\sqrt{40})(7) = 5097.4$

7.3: Using the Central Limit Theorem

88. The law of large numbers says that, as sample size increases, the sample mean tends to get nearer and nearer to the population mean.

89. You would expect the mean from a sample of size 100 to be nearer to the population mean, because the law of large numbers says that, as sample size increases, the sample mean tends to approach the population mean.

90. $X \sim N(0.10, 0.20)$

91. $\bar{X} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$ and the standard deviation of a uniform distribution is $\frac{b-a}{\sqrt{12}}$. In this example, the standard deviation of the distribution is $\frac{b-a}{\sqrt{12}} = \frac{0.10}{\sqrt{12}} = 0.03$

so $\bar{X} \sim N(0.15, 0.003)$

92. $\Sigma X \sim N((n)(\mu_x), (\sqrt{n})(\sigma_x))$ so $\Sigma X \sim N(9.0, 0.23)$

Practice Test 3

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

Use the following information to answer the next seven exercises. You draw a sample of size 30 from a normally distributed population with a standard deviation of four.

1. What is the standard error of the sample mean in this scenario, rounded to two decimal places?

2. What is the distribution of the sample mean?

3. If you want to construct a two-sided 95 percent confidence interval, how much probability will be in each tail of the distribution?

4. What is the appropriate z score and error bound or margin of error (*EBM*) for a 95 percent confidence interval for this data?

5. Rounding to two decimal places, what is the 95 percent confidence interval if the sample mean is 41?

6. What is the 90 percent confidence interval if the sample mean is 41? Round to two decimal places

7. Suppose the sample size in this study had been 50, rather than 30. What would the 95 percent confidence interval be if the sample mean is 41? Round your answer to two decimal places.

8. For any given data set and sampling situation, which would you expect to be wider: a 95 percent confidence interval or a 99 percent confidence interval?

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's *t*

9. Comparing graphs of the standard normal distribution (z distribution) and a t distribution with 15 degrees of freedom (*df*), how do they differ?

10. Comparing graphs of the standard normal distribution (z distribution) and a t distribution with 15 degrees of freedom (*df*), how are they similar?

Use the following information to answer the next five exercises. Body temperature is known to be distributed normally among healthy adults. Because you do not know the population standard deviation, you use the t distribution to study body temperature. You collect data from a random sample of 20 healthy adults and find that your sample temperatures have a mean of 98.4 and a sample standard deviation of 0.3 (both in degrees Fahrenheit).

11. What are the degrees of freedom (*df*) for this study?

12. For a two-tailed 95 percent confidence interval, what is the appropriate t value to use in the formula?

13. What is the 95 percent confidence interval?

14. What is the 99 percent confidence interval? Round to two decimal places.

15. Suppose your sample size had been 30 rather than 20. What would the 95 percent confidence interval be then? Round to two decimal places

8.3: Confidence Interval for a Population Proportion

Use this information to answer the next four exercises. You conduct a poll of 500 randomly selected city residents, asking them if they own an automobile. Of the respondents, 280 say they own an automobile, and 220 say they do not.

16. Find the sample proportion and sample standard deviation for this data.

17. What is the 95 percent two-sided confidence interval? Round to four decimal places.

18. Calculate the 90 percent confidence interval. Round to four decimal places.

19. Calculate the 99 percent confidence interval. Round to four decimal places.

Use the following information to answer the next three exercises. You are planning to conduct a poll of community members aged 65 and older, to determine how many own mobile phones. You want to produce an estimate whose 95 percent confidence interval will be within four percentage points (plus or minus) of the true population proportion. Use an estimated population proportion of 0.5.

20. What sample size do you need?

21. Suppose you knew from prior research that the population proportion was 0.6. What sample size would you need?

22. Suppose you wanted a 95 percent confidence interval within three percentage points of the population. Assume the population proportion is 0.5. What sample size do you need?

9.1: Null and Alternate Hypotheses

23. In your state, 58 percent of registered voters in a community are registered as republicans. You want to conduct a study to see if this also holds up in your community. State the null and alternative hypotheses to test this.

24. You believe that at least 58 percent of registered voters in a community are registered as republicans. State the null and alternative hypotheses to test this.

25. The mean household value in a city is $268,000. You believe that the mean household value in a particular neighborhood is lower than the city average. Write the null and alternative hypotheses to test this.

26. State the appropriate alternative hypothesis to this null hypothesis: H_0: $\mu = 107$

27. State the appropriate alternative hypothesis to this null hypothesis: H_0: $p < 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. If you reject H_0 when H_0 is correct, what type of error is this?

29. If you fail to reject H_0 when H_0 is false, what type of error is this?

30. What is the relationship between the Type II error and the power of a test?

31. A new blood test is being developed to screen patients for cancer. Positive results are followed up by a more accurate (and expensive) test. It is assumed that the patient does not have cancer. Describe the null hypothesis and the Type I and Type II errors for this situation, and explain which type of error is more serious.

32. Explain in words what it means that a screening test for TB has an α level of 0.10. The null hypothesis is that the patient does not have TB.

33. Explain in words what it means that a screening test for TB has a β level of 0.20. The null hypothesis is that the patient does not have TB.

34. Explain in words what it means that a screening test for TB has a power of 0.80.

9.3: Distribution Needed for Hypothesis Testing

35. If you are conducting a hypothesis test of a single population mean, and you do not know the population variance, what test will you use if the sample size is 10 and the population is normal?

36. If you are conducting a hypothesis test of a single population mean, and you know the population variance, what test will you use?

37. If you are conducting a hypothesis test of a single population proportion, with np and nq greater than or equal to five, what test will you use, and with what parameters?

38. Published information indicates that, on average, college students spend less than 20 hours studying per week. You draw a sample of 25 students from your college and find the sample mean to be 18.5 hours, with a standard deviation of 1.5 hours. What distribution will you use to test whether study habits at your college are the same as the national average, and why?

39. A published study says that 95 percent of American children are vaccinated against a disease, with a standard deviation of 1.5 percent. You draw a sample of 100 children from your community and check their vaccination records to see if the vaccination rate in your community is the same as the national average. What distribution will you use for this test, and why?

9.4: Rare Events, the Sample, Decision, and Conclusion

40. You are conducting a study with an α level of 0.05. If you get a result with a p-value of 0.07, what will be your decision?

41. You are conducting a study with $\alpha = 0.01$. If you get a result with a p-value of 0.006, what will be your decision?

Use the following information to answer the next five exercises. According to the World Health Organization, the average height of a one-year-old child is 29”. You believe children with a particular disease are smaller than average, so you draw a sample of 20 children with this disease and find a mean height of 27.5” and a sample standard deviation of 1.5”.

42. What are the null and alternative hypotheses for this study?

43. What distribution will you use to test your hypothesis, and why?

44. What is the test statistic and the *p*-value?

45. Based on your sample results, what is your decision?

46. Suppose the mean for your sample was 25. Redo the calculations and describe what your decision would be.

9.5: Additional Information and Full Hypothesis Test Examples

47. You conduct a study using $\alpha = 0.05$. What is the level of significance for this study?

48. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
H_0: $\mu = 35.5$
H_a: $\mu \neq 35.5$
Will you conduct a one-tailed or two-tailed test?

49. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
H_0: $\mu \geq 35.5$
H_a: $\mu < 35.5$
Will you conduct a one-tailed or two-tailed test?

Use the following information to answer the next three exercises. Nationally, 80 percent of adults own an automobile. You are interested in whether the same proportion in your community own cars. You draw a sample of 100 and find that 75 percent own cars.

50. What are the null and alternative hypotheses for this study?

51. What test will you use, and why?

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. You conduct a poll of political opinions, interviewing both members of 50 married couples. Are the groups in this study independent or matched?

53. You are testing a new drug to treat insomnia. You randomly assign 80 volunteer subjects to either the experimental (new drug) or control (standard treatment) conditions. Are the groups in this study independent or matched?

54. You are investigating the effectiveness of a new math textbook for high school students. You administer a pretest to a group of students at the beginning of the semester, and a posttest at the end of a year's instruction using this textbook, and compare the results. Are the groups in this study independent or matched?

Use the following information to answer the next two exercises. You are conducting a study of the difference in time at two colleges for undergraduate degree completion. At College A, students take an average of 4.8 years to complete an undergraduate degree, while at College B, they take an average of 4.2 years. The pooled standard deviation for this data is 1.6 years.

55. Calculate Cohen's *d* and interpret it.

56. Suppose the mean time to earn an undergraduate degree at College A was 5.2 years. Calculate the effect size and interpret it.

57. You conduct an independent-samples *t* test with sample size 10 in each of two groups. If you are conducting a two-tailed hypothesis test with $\alpha = 0.01$, what *p*-values will cause you to reject the null hypothesis?

58. You conduct an independent samples *t* test with sample size 15 in each group, with the following hypotheses:
H_0: $\mu \geq 110$
H_a: $\mu < 110$
If $\alpha = 0.05$, what *t* values will cause you to reject the null hypothesis?

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

Use the following information to answer the next six exercises. College students in the sciences often complain that they must spend more on textbooks each semester than students in the humanities. To test this, you draw random samples of 50 science and 50 humanities students from your college, and record how much each spent last semester on textbooks. Consider the science students to be group one, and the humanities students to be group two.

59. What is the random variable for this study?

60. What are the null and alternative hypotheses for this study?

61. If the 50 science students spent an average of $530 with a sample standard deviation of $20, and the 50 humanities students spent an average of $380 with a sample standard deviation of $15, would you not reject or reject the null hypothesis? Use an alpha level of 0.05. What is your conclusion?

62. What would be your decision, if you were using $\alpha = 0.01$?

10.3: Comparing Two Independent Population Proportions

Use the information to answer the next six exercises. You want to know if the proportion of homes with cable television service differs between Community A and Community B. To test this, you draw a random sample of 100 for each and record whether they have cable service.

63. What are the null and alternative hypotheses for this study?

64. If 65 households in Community A have cable service, and 78 households in Community B, what is the pooled proportion?

65. At $\alpha = 0.03$, will you reject the null hypothesis? What is your conclusion? Sixty-five households in Community A have cable service, and 78 households in community B. One hundred households in each community were surveyed.

66. Using an alpha value of 0.01, would you reject the null hypothesis? What is your conclusion? Sixty-five households in Community A have cable service, and 78 households in Community B. One hundred households in each community were surveyed.

10.4: Matched or Paired Samples

Use the following information to answer the next five exercises. You are interested in whether a particular exercise program helps people run a mile faster. You conduct a study in which you weigh the participants at the start of the study, and again at the conclusion, after they have participated in the exercise program for six months. You compare the results using a matched-pairs *t* test, in which the data is {time to run a mile at conclusion, time at start}. You believe that, on average, the participants will be able to run a mile faster after six months on the exercise program.

67. What are the null and alternative hypotheses for this study?

68. Calculate the test statistic, assuming that $\bar{x}_d = -5$, $s_d = 6$, and $n = 30$ (pairs).

69. What are the degrees of freedom for this statistic?

70. Using $\alpha = 0.05$, what is your decision regarding the effectiveness of this program in improving running speed? What is the conclusion?

71. What would it mean if the *t* statistic had been 4.56, and what would have been your decision in that case?

11.1: Facts About the Chi-Square Distribution

72. What is the mean and standard deviation for a chi-square distribution with 20 degrees of freedom?

11.2: Goodness-of-Fit Test

Use the following information to answer the next four exercises. Nationally, about 66 percent of high school graduates enroll in higher education. You perform a chi-square goodness of fit test to see if this same proportion applies to your high school's most recent graduating class of 200. Your null hypothesis is that the national distribution also applies to your high school.

73. What are the expected numbers of students from your high school graduating class enrolled and not enrolled in higher education?

74. Fill out the rest of this table.

	Observed (*O*)	**Expected (*E*)**	***O* – *E***	**(*O* – *E*)2**	$\frac{(O-E)^2}{z}$
Enrolled	145				

Table B10

	Observed (*O*)	Expected (*E*)	*O* – *E*	(*O* – *E*)2	$\frac{(O-E)^2}{z}$
Not enrolled	55				

Table B10

75. What are the degrees of freedom for this chi-square test?

76. What is the chi-square test statistic and the *p*-value? At the five percent significance level, what do you conclude?

77. For a chi-square distribution with 92 degrees of freedom, the curve _____.

78. For a chi-square distribution with five degrees of freedom, the curve is _____.

11.3: Test of Independence

Use the following information to answer the next four exercises. You are considering conducting a chi-square test of independence for the data in this table, which displays data about cell phone ownership for freshman and seniors at a high school. Your null hypothesis is that cell phone ownership is independent of class standing.

79. Compute the expected values for the cells.

	Cell = Yes	Cell = No
Freshman	100	150
Senior	200	50

Table B11

80. Compute $\frac{(O-E)^2}{z}$ for each cell, where O = observed and E = expected.

81. What is the chi-square statistic and degrees of freedom for this study?

82. At the $\alpha = 0.5$ significance level, what is your decision regarding the null hypothesis?

11.4: Test of Homogeneity

83. You conduct a chi-square test of homogeneity for data in a five-by-two table. What are the degrees of freedom for this test?

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. A 2013 poll in the State of California surveyed people about a tax. The results are presented in the following table, and are classified by ethnic group and response type. Are the poll responses independent of the participants' ethnic group? Conduct a hypothesis test at the five percent significance level.

Ethnic Group/Response Type	Favor	Oppose	No Opinion	Row Total
White/Non-Hispanic	234	433	43	710
Latino	147	106	19	272
African American	24	41	6	71
Asian American	54	48	16	118
Column Total	459	628	84	1171

Table B12

85. In a test of homogeneity, what must be true about the expected value of each cell?

86. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of independence?

87. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of homogeneity?

11.6: Test of a Single Variance

88. A lab test claims to have a variance of no more than five. You believe the variance is greater. What are the null and alternative hypotheses to test this?

Practice Test 3 Solutions

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

1. $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{30}} = 0.73$

2. normal

3. 0.025 or 2.5 percent; A 95 percent confidence interval contains 95 percent of the probability, and excludes 5 percent, and the 5 percent excluded is split evenly between the upper and lower tails of the distribution.

4. z score = 1.96; $EBM = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = (1.96)(0.73) = 1.4308$

5. 41 ± 1.43 = (39.57, 42.43); using the calculator function ZInterval, answer is (40.74, 41.26). Answers differ due to rounding.

6. The z-value for a 90 percent confidence interval is 1.645, so EBM = 1.645(0.73) = 1.20085.
The 90 percent confidence interval is 41 ± 1.20 = (39.80, 42.20).
The calculator function ZInterval answer is (40.78, 41.23). Answers differ due to rounding.

7. The standard error of measurement is $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{50}} = 0.57$

$EBM = z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = (1.96)(0.57) = 1.12$

The 95 percent confidence interval is 41 ± 1.12 = (39.88, 42.12).
The calculator function ZInterval answer is (40.84, 41.16). Answers differ due to rounding.

8. The 99 percent confidence interval, because it includes all but one percent of the distribution. The 95 percent confidence interval will be narrower, because it excludes five percent of the distribution.

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's *t*

9. The t distribution will have more probability in its tails (*thicker tails*) and less probability near the mean of the distribution (*shorter in the center*).

10. Both distributions are symmetrical and centered at zero.

11. $df = n - 1 = 20 - 1 = 19$

12. You can get the t value from a probability table or a calculator. In this case, for a t distribution with 19 degrees of freedom and a 95 percent two-sided confidence interval, the value is 2.093; i.e., $t_{\frac{\alpha}{2}} = 2.093$. The calculator function is invT(0.975, 19).

13. $EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = (2.093)\left(\frac{0.3}{\sqrt{20}}\right) = 0.140$

98.4 ± 0.14 = (98.26, 98.54).
The calculator function TInterval answer is (98.26, 98.54).

14. $t_{\frac{\alpha}{2}} = 2.861$. The calculator function is invT(0.995, 19).

$EBM = t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) = (2.861)\left(\frac{0.3}{\sqrt{20}}\right) = 0.192$

98.4 ± 0.19 = (98.21, 98.59). The calculator function TInterval answer is (98.21, 98.59).

15. $df = n - 1 = 30 - 1 = 29$. $t_{\frac{\alpha}{2}} = 2.045$

$$EBM = z_t\left(\frac{s}{\sqrt{n}}\right) = (2.045)\left(\frac{0.3}{\sqrt{30}}\right) = 0.112$$

98.4 ± 0.11 = (98.29, 98.51). The calculator function TInterval answer is (98.29, 98.51).

8.3: Confidence Interval for a Population Proportion

16. $p' = \frac{280}{500} = 0.56$

$q' = 1 - p' = 1 - 0.56 = 0.44$

$$s = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.56(0.44)}{500}} = 0.0222$$

17. Because you are using the normal approximation to the binomial, $z_{\frac{\alpha}{2}} = 1.96$.

Calculate the error bound for the population (*EBP*):

$$EBP = z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 1.96(0.222) = 0.0435$$

Calculate the 95 percent confidence interval:
0.56 ± 0.0435 = (0.5165, 0.6035).
The calculator function 1-PropZint answer is (0.5165, 0.6035).

18. $z_{\frac{\alpha}{2}} = 1.64$

$$EBP = z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 1.64(0.0222) = 0.0364$$

0.56 ± 0.03 = (0.5236, 0.5964). The calculator function 1-PropZint answer is (0.5235, 0.5965).

19. $z_{\frac{\alpha}{2}} = 2.58$

$$EBP = z_{\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} = 2.58(0.0222) = 0.0573$$

0.56 ± 0.05 = (0.5127, 0.6173).
The calculator function 1-PropZint answer is (0.5028, 0.6172).

20. *EBP* = 0.04 (because 4 percent = 0.04)
$z_{\frac{\alpha}{2}} = 1.96$ for a 95 percent confidence interval.

$$n = \frac{z^2 pq}{EBP^2} = \frac{1.96^2(0.5)(0.5)}{0.04^2} = \frac{0.9604}{0.0016} = 600.25$$

You need 601 subjects (rounding upward from 600.25).

21. $$n = \frac{n^2 pq}{EBP^2} = \frac{1.96^2(0.6)(0.4)}{0.04^2} = \frac{0.9220}{0.0016} = 576.24$$

You need 577 subjects (rounding upward from 576.24).

22. $$n = \frac{n^2 pq}{EBP^2} = \frac{1.96^2(0.5)(0.5)}{0.03^2} = \frac{0.9604}{0.0009} = 1067.11$$

You need 1,068 subjects (rounding upward from 1,067.11).

9.1: Null and Alternate Hypotheses

23. H_0: $p = 0.58$
H_a: $p \neq 0.58$

24. H_0: $p \geq 0.58$
H_a: $p < 0.58$

25. H_0: $\mu \geq \$268{,}000$

H_a: $\mu <$ \$268,000

26. H_a: $\mu \neq 107$

27. H_a: $p \geq 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. a Type I error

29. a Type II error

30. Power = $1 - \beta = 1 - P$(Type II error).

31. The null hypothesis is that the patient does not have cancer. A Type I error would be detecting cancer when it is not present. A Type II error would be not detecting cancer when it is present. A Type II error is more serious, because failure to detect cancer could keep a patient from receiving appropriate treatment.

32. The screening test has a 10 percent probability of a Type I error, meaning that 10 percent of the time, it will detect TB when it is not present.

33. The screening test has a 20 percent probability of a Type II error, meaning that 20 percent of the time, it will fail to detect TB when it is in fact present.

34. Eighty percent of the time, the screening test will detect TB when it is actually present.

9.3: Distribution Needed for Hypothesis Testing

35. The Student's *t* test.

36. The normal distribution or *z* test.

37. The normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$

38. t_{24}. You use the *t* distribution because you do not know the population standard deviation, and the degrees of freedom are 24 because $df = n - 1$.

39. $\bar{X} \sim N\left(0.95, \frac{0.051}{\sqrt{100}}\right)$

Because you know the population standard deviation and have a large sample, you can use the normal distribution.

9.4: Rare Events, the Sample, Decision, and Conclusion

40. Fail to reject the null hypothesis, because $\alpha \leq p$.

41. Reject the null hypothesis, because $\alpha \geq p$.

42. H_0: $\mu \geq 29.0$"
H_a: $\mu < 29.0$"

43. t_{19}. Because you do not know the population standard deviation, use the *t* distribution. The degrees of freedom are 19, because $df = n - 1$.

44. The test statistic is –4.4721 and the *p*-value is 0.00013 using the calculator function TTEST.

45. With $\alpha = 0.05$, reject the null hypothesis.

46. With $\alpha = 0.05$, the *p*-value is almost zero using the calculator function TTEST, so reject the null hypothesis.

9.5: Additional Information and Full Hypothesis Test Examples

47. The level of significance is five percent.

48. two-tailed

49. one-tailed

50. H_0: $p = 0.8$
H_a: $p \neq 0.8$

51. You will use the normal test for a single population proportion because *np* and *nq* are both greater than five.

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. They are matched (paired), because you interviewed married couples.

53. They are independent, because participants were assigned at random to the groups.

54. They are matched (paired), because you collected data twice from each individual.

55. $d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}} = \frac{4.8 - 4.2}{1.6} = 0.375$

This is a small effect size, because 0.375 falls between Cohen's small (0.2) and medium (0.5) effect sizes.

56. $d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}} = \frac{5.2 - 4.2}{1.6} = 0.625$

The effect size is 0.625. By Cohen's standard, this is a medium effect size, because it falls between the medium (0.5) and large (0.8) effect sizes.

57. p-value < 0.01.

58. You will only reject the null hypothesis if you get a value significantly below the hypothesized mean of 110.

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

59. $\bar{X}_1 - \bar{X}_2$; i.e., the mean difference in amount spent on textbooks for the two groups.

60. H_0: $\bar{X}_1 - \bar{X}_2 \leq 0$

H_a: $\bar{X}_1 - \bar{X}_2 > 0$

This could also be written as

H_0: $\bar{X}_1 \leq \bar{X}_2$

H_a: $\bar{X}_1 > \bar{X}_2$

61. Using the calculator function 2-SampTTest, reject the null hypothesis. At the five percent significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

62. Using the calculator function 2-SampTTest, reject the null hypothesis. At the one percent significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

10.3: Comparing Two Independent Population Proportions

63. H_0: $p_A = p_B$
H_a: $p_A \neq p_B$

64. $p_c = \frac{x_A + x_A}{n_A + n_A} = \frac{65 + 78}{100 + 100} = 0.715$

65. Using the calculator function 2-PropZTest, the p-value = 0.0417. Reject the null hypothesis. At the three percent significance level, here is sufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

66. Using the calculator function 2-PropZTest, the p-value = 0.0417. Do not reject the null hypothesis. At the one percent significance level, there is insufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

10.4: Matched or Paired Samples

67. H_0: $\bar{x}_d \geq 0$

H_a: $\bar{x}_d < 0$

68. t = –4.5644.

69. $df = 30 - 1 = 29$.

70. Using the calculator function TTEST, the p-value = 0.00004, so reject the null hypothesis. At the five percent level, there is sufficient evidence to conclude that the participants lost weight, on average.

71. A positive t statistic would mean that participants, on average, gained weight over the six months.

11.1: Facts About the Chi-Square Distribution

72. $\mu = df = 20$

$\sigma = \sqrt{2(df)} = \sqrt{40} = 6.32$

11.2: Goodness-of-Fit Test

73. Enrolled = 200(0.66) = 132. Not enrolled = 200(0.34) = 68.

74.

	Observed (O)	Expected (E)	O – E	(O – E)2	$\frac{(O-E)^2}{z}$
Enrolled	145	132	145 – 132 = 13	169	$\frac{169}{132} = 1.280$
Not enrolled	55	68	55 – 68 = –13	169	$\frac{169}{68} = 2.485$

Table B13

75. $df = n - 1 = 2 - 1 = 1$.

76. Using the calculator function Chi-Square GOF Test (in STAT TESTS), the test statistic is 3.7656 and the p-value is 0.0523. Do not reject the null hypothesis. At the five percent significance level, there is insufficient evidence to conclude that high school most recent graduating class distribution of enrolled and not enrolled does not fit that of the national distribution.

77. approximates the normal

78. skewed right

11.3: Test of Independence

79.

	Cell = Yes	Cell = No	Total
Freshman	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Senior	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Total	300	200	500

Table B14

80. $\frac{(100-150)^2}{150} = 16.67$

$\frac{(150-100)^2}{100} = 25$

$\frac{(200-100)^2}{150} = 16.67$

$$\frac{(50-100)^2}{100} = 25$$

81. Chi-square = 16.67 + 25 + 16.67 + 25 = 83.34.
$df = (r - 1)(c - 1) = 1$.

82. p-value = P(Chi-square, 83.34) = 0.
Reject the null hypothesis.
You could also use the calculator function STAT TESTS Chi-Square Test.

11.4: Test of Homogeneity

83. The table has five rows and two columns. $df = (r - 1)(c - 1) = (4)(1) = 4$.

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. Using the calculator function (STAT TESTS) Chi-Square Test, the p-value = 0. Reject the null hypothesis. At the five percent significance level, there is sufficient evidence to conclude that the poll responses are independent of the participants' ethnic group.

85. The expected value of each cell must be at least five.

86. H_0: The variables are independent.
H_a: The variables are not independent.

87. H_0: The populations have the same distribution.
H_a: The populations do not have the same distribution.

11.6: Test of a Single Variance

88. H_0: $\sigma^2 \leq 5$
H_a: $\sigma^2 > 5$

Practice Test 4

12.1 Linear Equations

1. Which of the following equations is/are linear?

A. $y = -3x$

B. $y = 0.2 + 0.74x$

C. $y = -9.4 - 2x$

D. A and B

E. A, B, and C

2. To complete a painting job requires four hours setup time, plus one hour per 1,000 square feet. How would you express this information in a linear equation?

3. A statistics instructor is paid a per-class fee of $2,000, plus $100 for each student in the class. How would you express this information in a linear equation?

4. A tutoring school requires students to pay a one-time enrollment fee of $500, plus tuition of $3,000 per year. Express this information in an equation.

12.2: Slope and y-intercept of a Linear Equation

Use the following information to answer the next four exercises. For the labor costs of doing repairs, an auto mechanic charges a flat fee of $75 per car, plus an hourly rate of $55.

5. What are the independent and dependent variables for this situation?

6. Write the equation and identify the slope and intercept.

7. What is the labor charge for a job that takes 3.5 hours to complete?

8. One job takes 2.4 hours to complete, while another takes 6.3 hours. What is the difference in labor costs for these two jobs?

12.3: Scatter Plots

9. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.

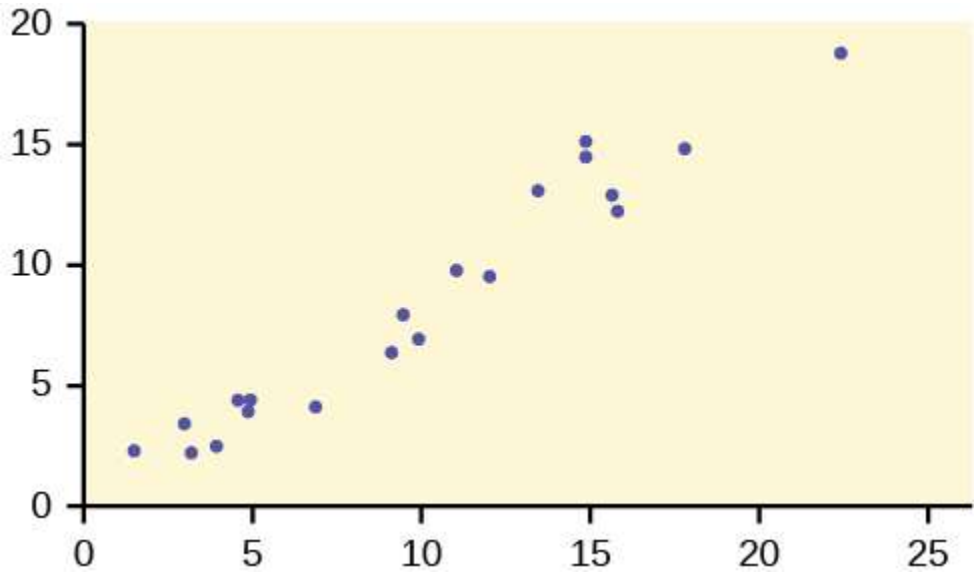


Figure B4

10. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.

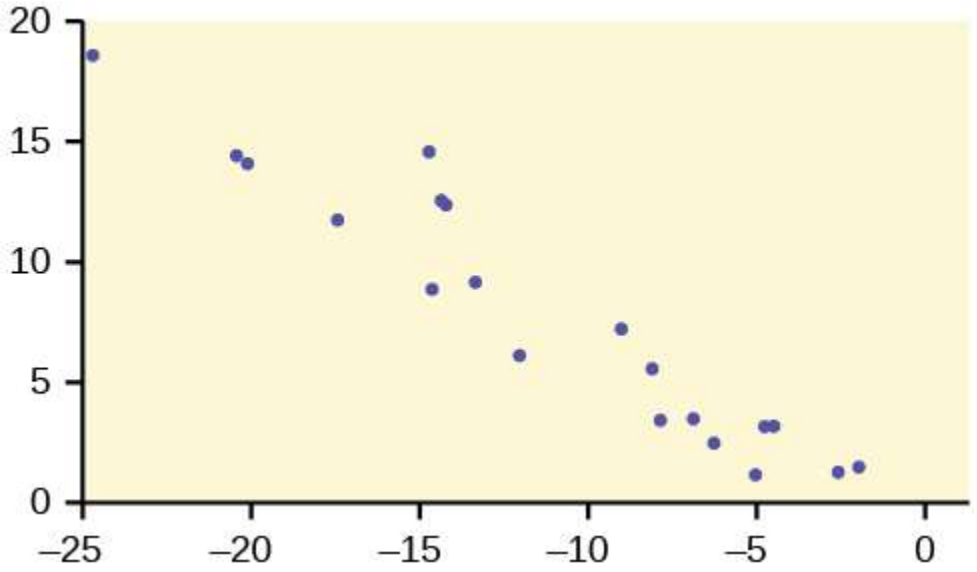


Figure B5

11. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.

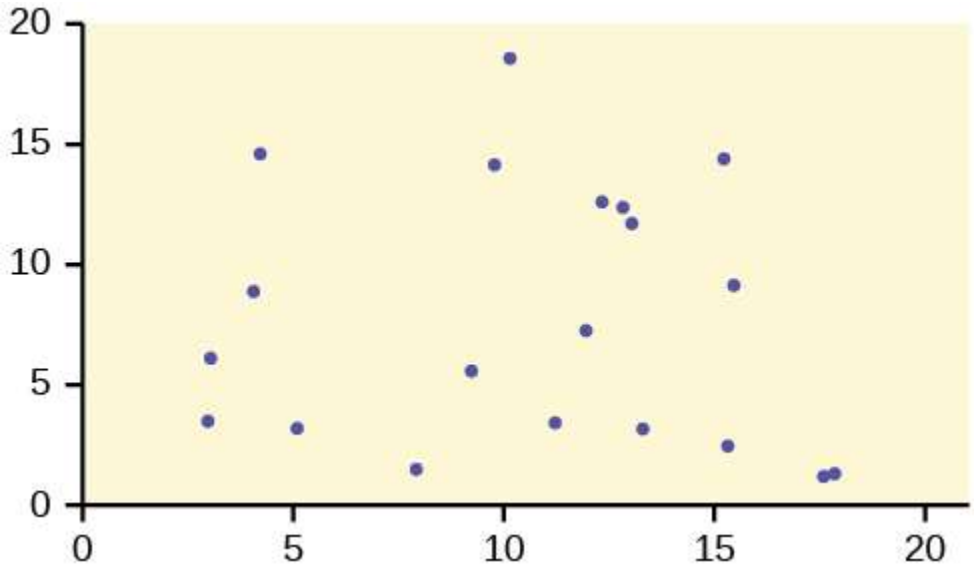


Figure B6

12. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.

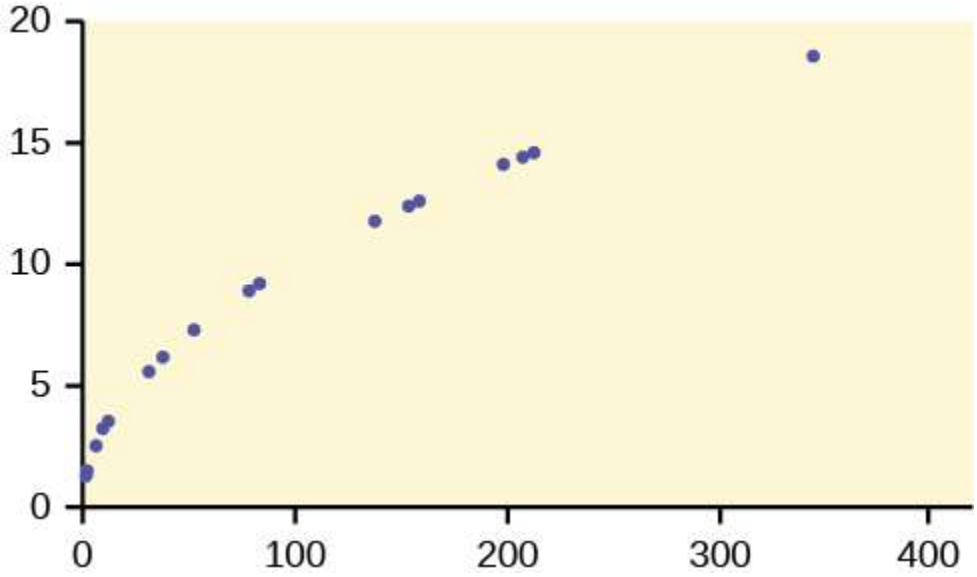


Figure B7

12.4: The Regression Equation

Use the following information to answer the next four exercises. Height (in inches) and weight (in pounds) in a sample of college freshman males have a linear relationship with the following summary statistics:

$\bar{x}$ = 68.4

$\bar{y}$ =141.6

s_x = 4.0

s_y = 9.6

r = 0.73

Let Y = weight and X = height, and write the regression equation in the form

$$\hat{y} = a + bx$$

13. What is the value of the slope?

14. What is the value of the *y*-intercept?

15. Write the regression equation predicting weight from height in this data set, and calculate the predicted weight for someone 68 inches tall.

12.5: Correlation Coefficient and Coefficient of Determination

16. The correlation between body weight and fuel efficiency (measured as miles per gallon) for a sample of 2,012 model cars is –0.56. Calculate the coefficient of determination for this data and explain what it means.

17. The correlation between high school GPA and freshman college GPA for a sample of 200 university students is 0.32. How much variation in freshman college GPA is not explained by high school GPA?

18. Rounded to two decimal places, what correlation between two variables is necessary to have a coefficient of determination of at least 0.50?

12.6: Testing the Significance of the Correlation Coefficient

19. Write the null and alternative hypotheses for a study to determine if two variables are significantly correlated.

20. In a sample of 30 cases, two variables have a correlation of 0.33. Do a *t* test to see if this result is significant at the α = 0.05 level. Use the formula

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

21. In a sample of 25 cases, two variables have a correlation of 0.45. Do a *t* test to see if this result is significant at the α = 0.05 level. Use the formula

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

12.7: Prediction

Use the following information to answer the next two exercises. A study relating the grams of potassium (*Y*) to the grams of fiber (*X*) per serving in enriched flour products (bread, rolls, etc.) produced the equation

$$\hat{y} = 25 + 16x$$

22. For a product with five grams of fiber per serving, what are the expected grams of potassium per serving?

23. Comparing two products, one with three grams of fiber per serving and one with six grams of fiber per serving, what is the expected difference in grams of potassium per serving?

12.8: Outliers

24. In the context of regression analysis, what is the definition of an outlier, and what is a rule of thumb to evaluate if a given value in a data set is an outlier?

25. In the context of regression analysis, what is the definition of an influential point, and how does an influential point differ from an outlier?

26. The least squares regression line for a data set is $\hat{y} = 5 + 0.3x$ and the standard deviation of the residuals is 0.4. Does a case with the values x = 2, y = 6.2 qualify as an outlier?

27. The least squares regression line for a data set is $\hat{y} = 2.3 - 0.1x$ and the standard deviation of the residuals is 0.13. Does a case with the values x = 4.1, y = 2.34 qualify as an outlier?

13.1: One-Way ANOVA

28. What are the five basic assumptions to be met if you want to do a one-way ANOVA?

29. You are conducting a one-way ANOVA comparing the effectiveness of four drugs in lowering blood pressure in hypertensive patients. What are the null and alternative hypotheses for this study?

30. What is the primary difference between the independent samples *t* test and one-way ANOVA?

31. You are comparing the results of three methods of teaching geometry to high school students. The final exam scores X_1, X_2, X_3, for the samples taught by the different methods have the following distributions:
$X_1 \sim N(85, 3.6)$
$X_1 \sim N(82, 4.8)$
$X_1 \sim N(79, 2.9)$
Each sample includes 100 students, and the final exam scores have a range of zero–100. Assuming the samples are independent and randomly selected, have the requirements for conducting a one-way ANOVA been met? Explain why or why not for each assumption.

32. You conduct a study comparing the effectiveness of four types of fertilizer to increase crop yield on wheat farms. When examining the sample results, you find that two of the samples have an approximately normal distribution, and two have an approximately uniform distribution. Is this a violation of the assumptions for conducting a one-way ANOVA?

13.2: The *F* Distribution

Use the following information to answer the next seven exercises. You are conducting a study of three types of feed supplements for cattle to test their effectiveness in producing weight gain among calves whose feed includes one of the supplements. You have four groups of 30 calves (one is a control group receiving the usual feed, but no supplement). You will conduct a one-way ANOVA after one year to see if there are differences in the mean weight for the four groups.

33. What is SS_{within} in this experiment, and what does it mean?

34. What is $SS_{between}$ in this experiment, and what does it mean?

35. What are k and i for this experiment?

36. If $SS_{within} = 374.5$ and $SS_{total} = 621.4$ for this data, what is $SS_{between}$?

37. What are $MS_{between}$, and MS_{within} for this experiment?

38. What is the F statistic for this data?

39. If there had been 35 calves in each group, instead of 30, with the sums of squares remaining the same, would the F statistic be larger or smaller?

13.3: Facts About the *F* Distribution

40. Which of the following numbers are possible F statistics?

A. 2.47

B. 5.95

C. –3.61

D. 7.28

E. 0.97

41. Histograms $F1$ and $F2$ below display the distribution of cases from samples from two populations, one distributed $F_{3,15}$ and one distributed $F_{5,500}$. Which sample came from which population?

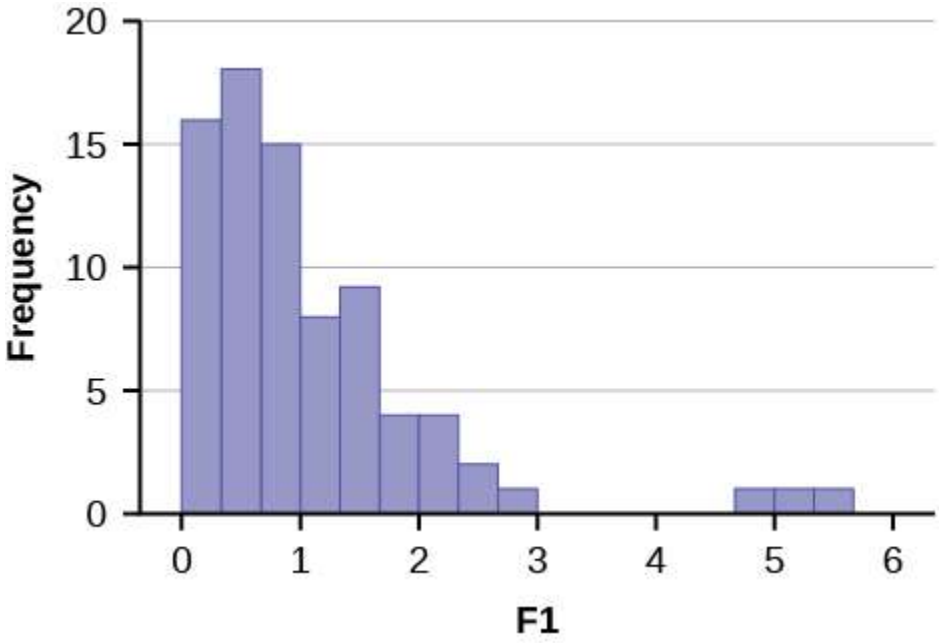


Figure B8

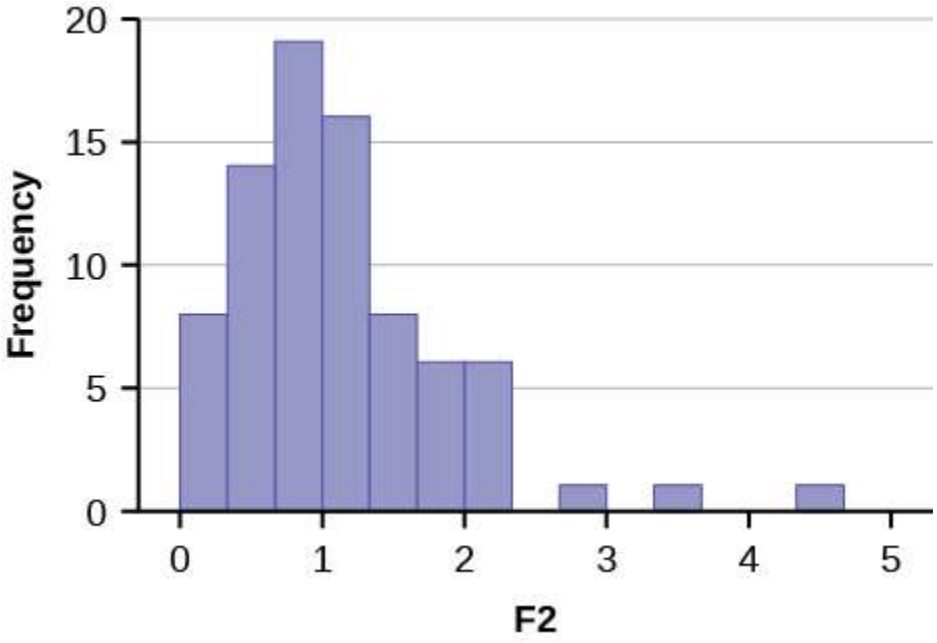


Figure B9

42. The F statistic from an experiment with $k = 3$ and $n = 50$ is 3.67. At $\alpha = 0.05$, will you reject the null hypothesis?

43. The F statistic from an experiment with $k = 4$ and $n = 100$ is 4.72. At $\alpha = 0.01$, will you reject the null hypothesis?

13.4: Test of Two Variances

44. What assumptions must be met to perform the F test of two variances?

45. You believe there is greater variance in grades given by the math department at your university than in the English department. You collect all the grades for undergraduate classes in the two departments for a semester, compute the variance of each, and conduct an F test of two variances. What are the null and alternative hypotheses for this study?

Practice Test 4 Solutions

12.1 Linear Equations

1. e. A, B, and C.
All three are linear equations of the form $y = mx + b$.

2. Let y = the total number of hours required, and x the square footage, measured in units of 1,000. The equation is $y = x + 4$

3. Let y = the total payment, and x the number of students in a class. The equation is $y = 100(x) + 2{,}000$

4. Let y = the total cost of attendance, and x the number of years enrolled. The equation is $y = 3{,}000(x) + 500$

12.2: Slope and *y*-intercept of a Linear Equation

5. The independent variable is the hours worked on a car. The dependent variable is the total labor charges to fix a car.

6. Let y = the total charge, and x the number of hours required. The equation is $y = 55x + 75$
The slope is 55 and the intercept is 75.

7. $y = 55(3.5) + 75 = 267.50$

8. Because the intercept is included in both equations, while you are only interested in the difference in costs, you do not need to include the intercept in the solution. The difference in number of hours required is 6.3 – 2.4 = 3.9.
Multiply this difference by the cost per hour: 55(3.9) = 214.5.
The difference in cost between the two jobs is $214.50.

12.3: Scatter Plots

9. The *X* and *Y* variables have a strong linear relationship. These variables would be good candidates for analysis with linear regression.

10. The *X* and *Y* variables have a strong negative linear relationship. These variables would be good candidates for analysis with linear regression.

11. There is no clear linear relationship between the *X* and *Y* variables, so they are not good candidates for linear regression.

12. The *X* and *Y* variables have a strong positive relationship, but it is curvilinear rather than linear. These variables are not good candidates for linear regression.

12.4: The Regression Equation

13. $r\left(\frac{s_y}{s_x}\right) = 0.73\left(\frac{9.6}{4.0}\right) = 1.752 \approx 1.75$

14. $a = \bar{y} - b\bar{x} = 141.6 - 1.752(68.4) = 21.7632 \approx 21.76$

15. $\hat{y} = 21.76 + 1.75(68) = 140.76$

12.5: Correlation Coefficient and Coefficient of Determination

16. The coefficient of determination is the square of the correlation, or r^2.
For this data, $r^2 = (–0.56)2 = 0.3136 \approx 0.31$ or 31 percent. This means that 31 percent of the variation in fuel efficiency can be explained by the bodyweight of the automobile.

17. The coefficient of determination = $0.32^2 = 0.1024$. This is the amount of variation in freshman college GPA that can be explained by high school GPA. The amount that cannot be explained is 1 – 0.1024 = 0.8976 ≈ 0.90. So, about 90 percent of variance in freshman college GPA in this data is not explained by high school GPA.

18. $r = \sqrt{r^2}$
$\sqrt{0.5} = 0.707106781 \approx 0.71$
You need a correlation of 0.71 or higher to have a coefficient of determination of at least 0.5.

12.6: Testing the Significance of the Correlation Coefficient

19. H_0: $\rho = 0$
H_a: $\rho \neq 0$

20. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.33\sqrt{30-2}}{\sqrt{1-0.33^2}} = 1.85$

The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{29} distribution is 2.045. Your value is less than this, so you fail to reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated.

Using the calculator function tcdf, the *p*-value is 2tcdf(1.85, 10^99, 29) = 0.0373. Do not reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated.

21. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.45\sqrt{25-2}}{\sqrt{1-0.45^2}} = 2.417$

The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{24} distribution is 2.064. Your value is greater than this, so you reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.
Using the calculator function tcdf, the *p*-value is 2tcdf(2.417, 10^99, 24) = 0.0118. Reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.

12.7: Prediction

22. $\hat{y} = 25 + 16(5) = 105$

23. Because the intercept appears in both predicted values, you can ignore it in calculating a predicted difference score. The difference in grams of fiber per serving is 6 – 3 = 3, and the predicted difference in grams of potassium per serving is (16)(3) = 48.

12.8: Outliers

24. An outlier is an observed value that is far from the least squares regression line. A rule of thumb is that a point more than two standard deviations of the residuals from its predicted value on the least squares regression line is an outlier.

25. An influential point is an observed value in a data set that is far from other points in the data set, in a horizontal direction. Unlike an outlier, an influential point is determined by its relationship with other values in the data set, not by its relationship to the regression line.

26. The predicted value for y is $\hat{y} = 5 + 0.3x = 5.6$. The value of 6.2 is less than two standard deviations from the predicted value, so it does not qualify as an outlier.
Residual for (2, 6.2): 6.2 – 5.6 = 0.6 (0.6 < 2(0.4))

27. The predicted value for y is $\hat{y}$ = 2.3 – 0.1(4.1) = 1.89. The value of 2.32 is more than two standard deviations from the predicted value, so it qualifies as an outlier.
Residual for (4.1, 2.34): 2.32 – 1.89 = 0.43 (0.43 > 2(0.13))

13.1: One-Way ANOVA

28.

1. Each sample is drawn from a normally distributed population.
2. All samples are independent and randomly selected.
3. The populations from which the samples are drawn have equal standard deviations.
4. The factor is a categorical variable.
5. The response is a numerical variable.

29. H_0: $\mu 1 = \mu 2 = \mu 3 = \mu 4$
H_a: At least two of the group means $\mu 1$, $\mu 2$, $\mu 3$, $\mu 4$ are not equal.

30. The independent samples *t* test can only compare means from two groups, while one-way ANOVA can compare means of more than two groups.

31. Each sample appears to have been drawn from normally distributed populations, the factor is a categorical variable (method), the outcome is a numerical variable (test score), and you were told the samples were independent and randomly selected, so those requirements are met. However, each sample has a different standard deviation, and this suggests that the populations from which they were drawn also have different standard deviations, which is a violation of an assumption for one-way ANOVA. Further statistical testing will be necessary to test the assumption of equal variance before proceeding with the analysis.

32. One of the assumptions for a one-way ANOVA is that the samples are drawn from normally distributed populations. Since two of your samples have an approximately uniform distribution, this casts doubt on whether this assumption has been met. Further statistical testing will be necessary to determine if you can proceed with the analysis.

13.2: The *F* Distribution

33. SS_{within} is the sum of squares within groups, representing the variation in outcome that cannot be attributed to the different feed supplements but due to individual or chance factors among the calves in each group.

34. $SS_{between}$ is the sum of squares between groups, representing the variation in outcome that can be attributed to the different feed supplements.

35. k = the number of groups = 4
n_1 = the number of cases in group 1 = 30
n = the total number of cases = 4(30) = 120

36. $SS_{total} = SS_{within} + SS_{between}$, so $SS_{between} = SS_{total} - SS_{within}$
621.4 – 374.5 = 246.9

37. The mean squares in an ANOVA are found by dividing each sum of squares by its respective degrees of freedom (df).
For SS_{total}, $df = n - 1 = 120 - 1 = 119$.
For $SS_{between}$, df = k – 1 = 4 – 1 = 3.
For SS_{within}, $df = 120 - 4 = 116$.

$MS_{between} = \frac{246.9}{3} = 82.3$

$MS_{within} = \frac{374.5}{116} = 3.23$

38. $F = \frac{MS_{between}}{MS_{within}} = \frac{82.3}{3.23} = 25.48$

39. It would be larger, because you would be dividing by a smaller number. The value of $MS_{between}$ would not change with a change of sample size, but the value of MS_{within} would be smaller, because you would be dividing by a larger number (df_{within} would be 136, not 116). Dividing a constant by a smaller number produces a larger result.

13.3: Facts About the *F* Distribution

40. All but choice c, –3.61. F Statistics are always greater than or equal to 0.

41. As the degrees of freedom increase in an F distribution, the distribution becomes more nearly normal. Histogram $F2$ is closer to a normal distribution than histogram $F1$, so the sample displayed in histogram $F1$ was drawn from the $F_{3,15}$ population, and the sample displayed in histogram $F2$ was drawn from the $F_{5,500}$ population.

42. Using the calculator function Fcdf, p-value = Fcdf(3.67, 1E, 3, 50) = 0.0182. Reject the null hypothesis.

43. Using the calculator function Fcdf, p-value = Fcdf(4.72, 1E, 4, 100) = 0.0016 Reject the null hypothesis.

13.4: Test of Two Variances

44. The samples must be drawn from populations that are normally distributed, and must be drawn from independent populations.

45. Let σ_M^2 = variance in math grades, and σ_E^2 = variance in English grades.

H_0: $\sigma_M^2 \leq \sigma_E^2$

H_a: $\sigma_M^2 > \sigma_E^2$

Practice Final Exam 1

Use the following information to answer the next two exercises. An experiment consists of tossing two, 12-sided dice (the numbers 1–12 are printed on the sides of each die).

- Let Event A = both dice show an even number.
- Let Event B = both dice show a number greater than eight

1. Events A and B are

A. Mutually exclusive

B. Independent

C. Mutually exclusive and independent

D. Neither mutually exclusive nor independent

2. Find $P(A|B)$.

A. $\frac{2}{4}$

B. $\frac{16}{144}$

C. $\frac{4}{16}$

D. $\frac{2}{144}$

3. Which of the following are TRUE when we perform a hypothesis test on matched or paired samples?

A. Sample sizes are almost never small.

B. Two measurements are drawn from the same pair of individuals or objects.

C. Two sample means are compared to each other.

D. Answer choices b and c are both true.

Use the following information to answer the next two exercises. One hundred eighteen students were asked what type of color their bedrooms were painted: light colors, dark colors, or vibrant colors. The results were tabulated according to gender.

	Light colors	Dark colors	Vibrant colors
Female	20	22	28
Male	10	30	8

Table B15

4. Find the probability that a randomly chosen student is male or has a bedroom painted with light colors.

A. $\frac{10}{118}$

B. $\frac{68}{118}$

C. $\frac{48}{118}$

D. $\frac{10}{48}$

5. Find the probability that a randomly chosen student is male given the student's bedroom is painted with dark colors.

A. $\frac{30}{118}$

B. $\frac{30}{48}$

C. $\frac{22}{118}$

D. $\frac{30}{52}$

Use the following information to answer the next two exercises. We are interested in the number of times a teenager must be reminded to do his or her chores each week. A survey of 40 mothers was conducted. **Table B16** shows the results of the survey.

x	$P(x)$
0	$\frac{2}{40}$
1	$\frac{5}{40}$
2	
3	$\frac{14}{40}$
4	$\frac{7}{40}$
5	$\frac{4}{40}$

Table B16

6. Find the probability that a teenager is reminded two times.

A. 8

B. $\frac{8}{40}$

C. $\frac{6}{40}$

D. 2

7. Find the expected number of times a teenager is reminded to do his or her chores.

A. 15

B. 2.78

C. 1.0

D. 3.13

Use the following information to answer the next two exercises. On any given day, approximately 37.5 percent of the cars parked in the De Anza parking garage are parked crookedly. We randomly survey 22 cars. We are interested in the number of cars that are parked crookedly.

8. For every 22 cars, how many would you expect to be parked crookedly, on average?

A. 8.25

B. 11

C. 18

D. 7.5

9. What is the probability that at least 10 of the 22 cars are parked crookedly?

A. 0.1263

B. 0.1607

C. 0.2870

D. 0.8393

10. Using a sample of 15 Stanford-Binet IQ scores, we wish to conduct a hypothesis test. Our claim is that the mean IQ score on the Stanford-Binet IQ test is more than 100. It is known that the standard deviation of all Stanford-Binet IQ scores is 15 points. Which of the following is the correct distribution to use for the hypothesis test?

A. Binomial

B. Student's *t*

C. Normal

D. Uniform

Use the following information to answer the next three exercises. De Anza College keeps statistics on the pass rate of students who enroll in math classes. In a sample of 1,795 students enrolled in Math 1A (1st quarter calculus), 1,428 passed the course. In a sample of 856 students enrolled in Math 1B (2nd quarter calculus), 662 passed. In general, are the pass rates of Math 1A and Math 1B statistically the same? Let A = the subscript for Math 1A and B = the subscript for Math 1B.

11. If you were to conduct an appropriate hypothesis test, the alternate hypothesis would be

A. H_a: $p_A = p_B$

B. H_a: $p_A > p_B$

C. H_o: $p_A = p_B$

D. H_a: $p_A \neq p_B$

12. The Type I error is to

A. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, the pass rates are different.

B. conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.

C. conclude that the pass rate for Math 1A is greater than the pass rate for Math 1B when, in fact, the pass rate for Math 1A is less than the pass rate for Math 1B.

D. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, they are the same.

13. The correct decision is to

A. reject H_0.

B. not reject H_0.

C. There is not enough information given to conduct the hypothesis test.

Kia, Alejandra, and Iris are runners on the track teams at three different schools. Their running times, in minutes, and the statistics for the track teams at their respective schools, for a one mile run, are given in the table below:

	Running Time	School Average Running Time	School Standard Deviation
Kia	4.9	5.2	0.15
Alejandra	4.2	4.6	0.25
Iris	4.5	4.9	0.12

Table B17

14. Which student is the BEST when compared to the other runners at her school?

A. Kia

B. Alejandra

C. Iris

D. Impossible to determine

Use the following information to answer the next two exercises. The following adult ski sweater prices are from the Gorsuch Ltd. Winter catalog: $212, $292, $278, $199, $280, $236.

Assume the underlying sweater price population is approximately normal. The null hypothesis is that the mean price of adult ski sweaters from Gorsuch Ltd. is at least $275.

15. Which of the following is the correct distribution to use for the hypothesis test?

A. Normal

B. Binomial

C. Student's *t*

D. Exponential

16. The hypothesis test

A. is two-tailed.

B. is left-tailed.

C. is right-tailed.

D. has no tails.

17. Sara, a statistics student, wanted to determine the mean number of books that college professors have in their office. She randomly selected two buildings on campus and asked each professor in the selected buildings how many books are in his or her office. Sara surveyed 25 professors. The type of sampling selected is

A. simple random sampling.

B. systematic sampling.

C. cluster sampling.

D. stratified sampling.

18. A clothing store would use which measure of the center of data when placing orders for the typical *middle* customer?

A. Mean

B. Median

C. Mode

D. IQR

19. In a hypothesis test, the *p*-value is

A. the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.

B. called the preconceived alpha.

C. compared to beta to decide whether to reject or not reject the null hypothesis.

D. Answer choices A and B are both true.

Use the following information to answer the next three exercises. A community college offers classes six days a week: Monday through Saturday. Maria conducted a study of the students in her classes to determine how many days per week the students who are in her classes come to campus for classes. In each of her five classes she randomly selected 10 students and asked them how many days they come to campus for classes. Each of her classes are the same size. The results of her survey are summarized in **Table B18**.

Number of Days on Campus	Frequency	Relative Frequency	Cumulative Relative Frequency
1	2		

Table B18

Number of Days on Campus	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	.24	
3	10	.20	
4			.98
5	0		
6	1	.02	1

Table B18

20. Combined with convenience sampling, what other sampling technique did Maria use?

A. Simple random

B. Systematic

C. Cluster

D. Stratified

21. How many students come to campus for classes four days a week?

A. 49

B. 25

C. 30

D. 13

22. What is the 60^{th} percentile for this data?

A. 2

B. 3

C. 4

D. 5

Use the following information to answer the next two exercises. The following data are the results of a random survey of 110 reservists called to active duty to increase security at California airports.

Number of Dependents	Frequency
0	11
1	27
2	33
3	20
4	19

Table B19

23. Construct a 95 percent confidence interval for the true population mean number of dependents of reservists called to active duty to increase security at California airports.

A. (1.85, 2.32)

B. (1.80, 2.36)

C. (1.97, 2.46)

D. (1.92, 2.50)

24. The 95 percent confidence interval above means:

A. Five percent of confidence intervals constructed this way will not contain the true population aveage number of dependents.

B. We are 95 percent confident the true population mean number of dependents falls in the interval.

C. Both of the above answer choices are correct.

D. None of the above.

25. $X \sim U(4, 10)$. Find the 30^{th} percentile.

A. 0.3000

B. 3

C. 5.8

D. 6.1

26. If $X \sim Exp(0.8)$, then $P(x < \mu) =$ —

A. 0.3679

B. 0.4727

C. 0.6321

D. cannot be determined

27. The lifetime of a computer circuit board is normally distributed with a mean of 2,500 hours and a standard deviation of 60 hours. What is the probability that a randomly chosen board will last at most 2,560 hours?

A. 0.8413

B. 0.1587

C. 0.3461

D. 0.6539

28. A survey of 123 reservists called to active duty as a result of the September 11, 2001, attacks was conducted to determine the proportion that were married. Eighty-six reported being married. Construct a 98 percent confidence interval for the true population proportion of reservists called to active duty that are married.

A. (0.6030, 0.7954)

B. (0.6181, 0.7802)

C. (0.5927, 0.8057)

D. (0.6312, 0.7672)

29. Winning times in 26 mile marathons run by world class runners average 145 minutes with a standard deviation of 14 minutes. A sample of the last 10 marathon winning times is collected. Let x = mean winning times for 10 marathons. The distribution for x is

A. $N\left(145, \frac{14}{\sqrt{10}}\right)$

B. $N(145,14)$

C. t_9

D. t_{10}

30. Suppose that Phi Beta Kappa honors the top 1 percent of college and university seniors. Assume that grade point means (GPA) at a certain college are normally distributed with a 2.5 mean and a standard deviation of 0.5. What would be the minimum GPA needed to become a member of Phi Beta Kappa at that college?

A. 3.99

B. 1.34

C. 3.00

D. 3.66

The number of people living on American farms has declined steadily during the 20th century. Here are data on the farm population (in millions of persons) from 1935 to 1980.

Year	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Population	32.1	30.5	24.4	23	19.1	15.6	12.4	9.7	8.9	7.2

Table B20

31. The linear regression equation is $\hat{y} = 1166.93 - 0.5868x$. What was the expected farm population in millions of persons for 1980?

A. 7.2

B. 5.1

C. 6

D. 8

32. In linear regression, which is the best possible *SSE*?

A. 13.46

B. 18.22

C. 24.05

D. 16.33

33. In regression analysis, if the correlation coefficient is close to one, what can be said about the best fit line?

A. It is a horizontal line. Therefore, we cannot use it.

B. There is a strong linear pattern. Therefore, it is most likely a good model to be used.

C. The coefficient correlation is close to the limit. Therefore, it is hard to make a decision.

D. We do not have the equation. Therefore, we cannot say anything about it.

Use the following information to answer the next three exercises. A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded.

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

Table B21 Does the data suggest that there is a relationship between the gender of students and their choice of major?

34. The distribution for the test is

A. Chi^2_8.

B. Chi^2_3.

C. t_{721}.

D. $N(0, 1)$.

35. The expected number of females who choose finance is

A. 37.

B. 61.

C. 60.

D. 70.

36. The *p*-value is 0.0127 and the level of significance is 0.05. The conclusion to the test is:

A. there is insufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.

B. there is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.

C. there is sufficient evidence to conclude that students find economics very hard.

D. there is in sufficient evidence to conclude that more females prefer administration than males.

37. An agency reported that the work force nationwide is composed of 10 percent professional, 10 percent clerical, 30 percent skilled, 15 percent service, and 35 percent semiskilled laborers. A random sample of 100 San Jose residents indicated 15 professional, 15 clerical, 40 skilled, 10 service, and 20 semiskilled laborers. At $\alpha = 0.10$, does the work force in San Jose appear to be consistent with the agency report for the nation? Which kind of test is it?

A. Chi^2 goodness of fit

B. Chi^2 test of independence

C. Independent groups proportions

D. Unable to determine

Practice Final Exam 1 Solutions

Solutions

1. B independent

2. C $\frac{4}{16}$

3. B Two measurements are drawn from the same pair of individuals or objects.

4. B $\frac{68}{118}$

5. D $\frac{30}{52}$

6. B $\frac{8}{40}$

7. B 2.78

8. A 8.25

9. C 0.2870

10. C Normal

11. D H_a: $p_A \neq p_B$

12. B conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.

13. B not reject H_0

14. C Iris

15. C Student's *t*

16. B is left-tailed.

17. C cluster sampling

18. B median

19. A the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.

20. D stratified

21. B 25

22. C 4

23. A (1.85, 2.32)

24. C Both above are correct.

25. C 5.8

26. C 0.6321

27. A 0.8413

28. A (0.6030, 0.7954)

29. A $N\left(145, \frac{14}{\sqrt{10}}\right)$

30. D 3.66

31. B 5.1

32. A 13.46

33. B There is a strong linear pattern. Therefore, it is most likely a good model to be used.

34. B Chi^2_3.

35. D 70

36. B There is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.

37. A Chi^2 goodness-of-fit

Practice Final Exam 2

1. A study was done to determine the proportion of teenagers that own a car. The population proportion of teenagers that own a car is the

A. statistic.

B. parameter.

C. population.

D. variable.

Use the following information to answer the next two exercises.

value	frequency
0	1
1	4

Table B22

value	frequency
2	7
3	9
6	4

Table B22

2. The box plot for the data is

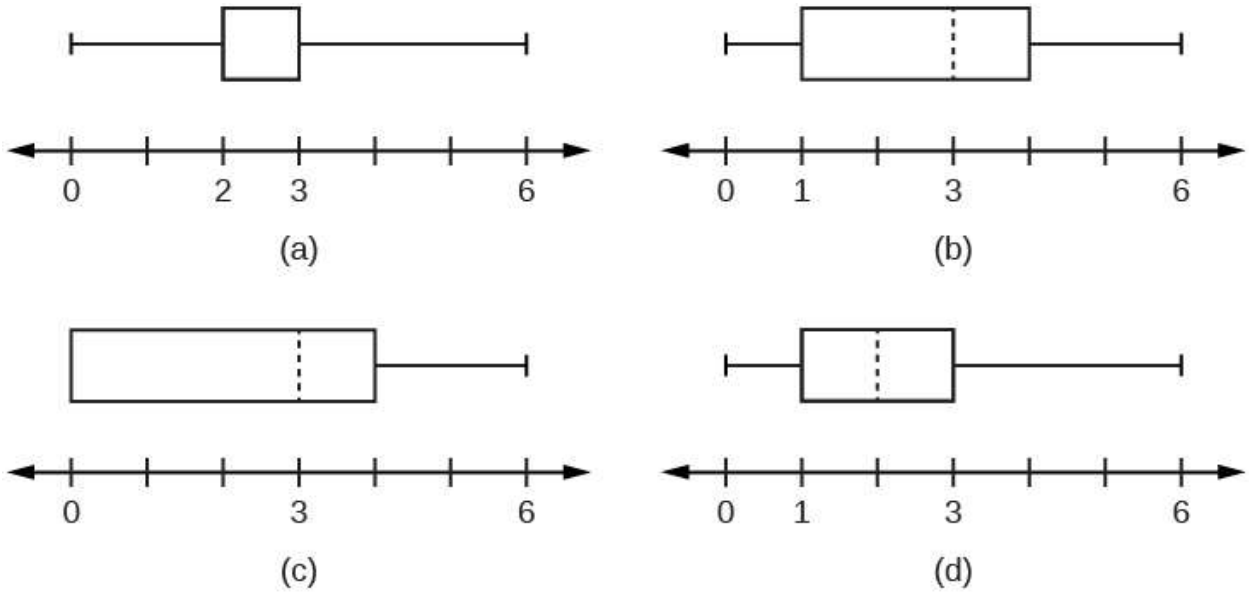


Figure B10

3. If six were added to each value of the data in the table, the 15th percentile of the new list of values is would be

A. six
B. one
C. seven
D. eight

Use the following information to answer the next two exercises. Suppose that the probability of a drought in any independent year is 20 percent. Out of those years in which a drought occurs, the probability of water rationing is 10 percent. However, in any year, the probability of water rationing is 5 percent.

4. What is the probability of both a drought and water rationing occurring?

A. 0.05
B. 0.01
C. 0.02
D. 0.30

5. Which of the following is true?

A. Drought and water rationing are independent events.
B. Drought and water rationing are mutually exclusive events.
C. None of the above.

Use the following information to answer the next two exercises. Suppose that a survey yielded the following data:

gender	apple	pumpkin	pecan
female	40	10	30
male	20	30	10

Table B23 Favorite Pie

6. Suppose that one individual is randomly chosen. The probability that the person's favorite pie is apple or the person is male is —

A. $\frac{40}{60}$

B. $\frac{60}{140}$

C. $\frac{120}{140}$

D. $\frac{100}{140}$

7. Suppose H_0 is favorite pie and gender are independent. The *p*-value is —

A. ≈ 0

B. 1

C. 0.05

D. Cannot be determined

Use the following information to answer the next two exercises. Let's say that the probability that an adult watches the news at least once per week is 0.60. We randomly survey 14 people. Of interest is the number of people who watch the news at least once per week.

8. Which of the following statements is FALSE?

A. $X \sim B(14\ 0.60)$

B. The values for x are $\{1, 2, 3, \ldots 14\}$.

C. $\mu = 8.4$

D. $P(X = 5) = 0.0408$

9. Find the probability that at least six adults watch the news at least once per week.

A. $\frac{6}{14}$

B. 0.8499

C. 0.9417

D. 0.6429

10. The following histogram is most likely to be a result of sampling from which distribution?

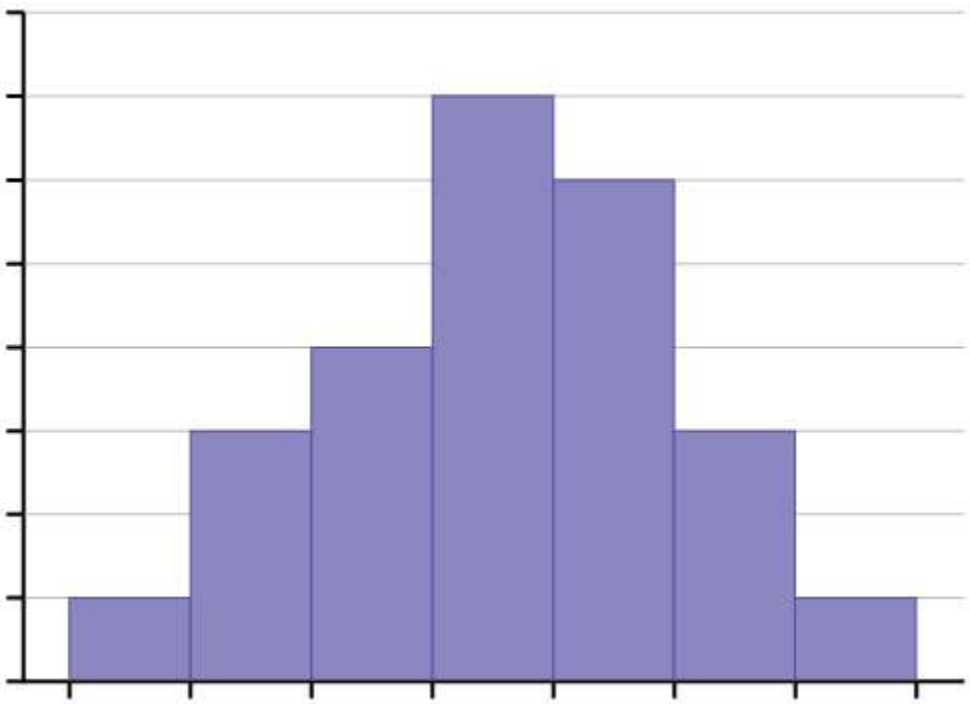


Figure B11

A. Chi-square with $df = 6$

B. Exponential

C. Uniform

D. Binomial

11. The ages of campus day and evening students is known to be normally distributed. A sample of six campus day and evening students reported their ages (in years) as {18, 35, 27, 45, 20, 20}. What is the error bound for the 90 percent confidence interval of the true average age?

A. 11.2

B. 22.3

C. 17.5

D. 8.7

12. If a normally distributed random variable has $\mu = 0$ and $\sigma = 1$, then 97.5 percent of the population values lie above

A. –1.96

B. 1.96

C. 1

D. –1

Use the following information to answer the next three exercises. The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the average amount of money a customer spends in one trip to the supermarket is $72.

13. What is the probability that one customer spends less than $72 in one trip to the supermarket?

A. 0.6321

B. 0.5000

C. 0.3714

D. 1

14. How much money altogether would you expect the next five customers to spend in one trip to the supermarket (in dollars)?

A. 72

B. $\frac{72^2}{5}$

C. 5184

D. 360

15. If you want to find the probability that the mean amount of money 50 customers spend in one trip to the supermarket is less than \$60, the distribution to use is

A. $N(72, 72)$

B. $N\left(72, \frac{72}{\sqrt{50}}\right)$

C. $Exp(72)$

D. $Exp\left(\frac{1}{72}\right)$

Use the following information to answer the next three exercises. The amount of time it takes a fourth grader to carry out the trash is uniformly distributed in the interval from one to 10 minutes.

16. What is the probability that a randomly chosen fourth grader takes more than seven minutes to take out the trash?

A. $\frac{3}{9}$

B. $\frac{7}{9}$

C. $\frac{3}{10}$

D. $\frac{7}{10}$

17. Which graph best shows the probability that a randomly chosen fourth grader takes more than six minutes to take out the trash, given that he or she has already taken more than three minutes?

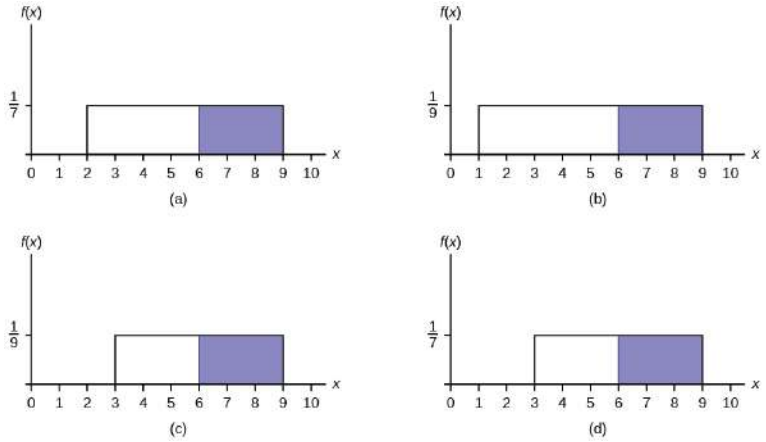


Figure B12

18. We should expect a fourth grader to take how many minutes to take out the trash?

A. 4.5

B. 5.5

C. 5

D. 10

Use the following information to answer the next three exercises. At the beginning of the quarter, the amount of time a student waits in line at the campus cafeteria is normally distributed with a mean of five minutes and a standard deviation of 1.5 minutes.

19. What is the 90th percentile of waiting times in minutes?

A. 1.28

B. 90

C. 7.47

D. 6.92

20. The median waiting time in minutes for one student is

A. 5

B. 50

C. 2.5

D. 1.5

21. Find the probability that the average wait time for ten students is at most 5.5 minutes.

A. 0.6301

B. 0.8541

C. 0.3694

D. 0.1459

22. A sample of 80 software engineers in Silicon Valley is taken, and it is found that 20 percent of them earn approximately $50,000 per year. A point estimate for the true proportion of engineers in Silicon Valley who earn $50,000 per year is

A. 16

B. 0.2

C. 1

D. 0.95

23. If $P(Z < z_\alpha) = 0.1587$ where $Z \sim N(0, 1)$, then α is equal to

A. –1

B. 0.1587

C. 0.8413

D. 1

24. A professor tested 35 students to determine their entering skills. At the end of the term, after completing the course, the same test was administered to the same 35 students to study their improvement. This would be a test of

A. independent groups

B. two proportions

C. matched pairs, dependent groups

D. exclusive groups

A math exam was given to all the third-grade children attending ABC School. Two random samples of scores were taken.

	n	$\bar{x}$	s
Boys	55	82	5
Girls	60	86	7

Table B24

25. Which of the following correctly describes the results of a hypothesis test of the claim, "There is a difference between the mean scores obtained by third-grade girls and boys at the 5 percent level of significance"?

A. Do not reject H_0. There is insufficient evidence to conclude that there is a difference in the mean scores.

B. Do not reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

C. Reject H_0. There is insufficient evidence to conclude that there is no difference in the mean scores.

D. Reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

26. In a survey of 80 males, 45 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. The correct conclusion is that

A. There is insufficient information to conclude that the proportion for males is the same as the proportion for females.

B. There is insufficient information to conclude that the proportion for males is not the same as the proportion for females.

C. There is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.

D. There is not enough information to make a conclusion.

27. From past experience, a statistics teacher has found that the average score on a midterm is 81, with a standard deviation of 5.2. This term, a class of 49 students had a standard deviation of 5 on the midterm. Do the data indicate that we should reject the teacher's claim that the standard deviation is 5.2? Use $\alpha = 0.05$.

A. Yes

B. No

C. Not enough information given to solve the problem

28. Three loading machines are being compared. Ten samples were taken for each machine. Machine I took an average of 31 minutes to load packages, with a standard deviation of two minutes. Machine II took an average of 28 minutes to load packages, with a standard deviation of 1.5 minutes. Machine III took an average of 29 minutes to load packages, with a standard deviation of one minute. Find the *p*-value when testing that the average loading times are the same.

A. *p*-value is close to zero

B. *p*-value is close to one

C. Not enough information given to solve the problem

Use the following information to answer the next three exercises. A corporation has offices in different parts of the country. It has gathered the following information concerning the number of bathrooms and the number of employees at seven sites:

Number of employees x	650	730	810	900	102	107	1150
Number of bathrooms y	40	50	54	61	82	110	121

Table B25

29. Is the correlation between the number of employees and the number of bathrooms significant?

A. Yes

B. No

C. Not enough information to answer question

30. The linear regression equation is

A. $\hat{y} = 0.0094 - 79.96x$

B. $\hat{y} = 79.96 + 0.0094x$

C. $\hat{y} = 79.96 - 0.0094x$

D. $\hat{y} = -0.0094 + 79.96x$

31. If a site has 1,150 employees, approximately how many bathrooms should it have?

A. 69

B. 91

C. 91,954

D. We should not be estimating here.

32. Suppose that a sample of size 10 was collected, with $\bar{x} = 4.4$ and $s = 1.4$. H_0: $\sigma^2 = 1.6$ vs. H_a: $\sigma^2 \neq 1.6$. Which graph best describes the results of the test?

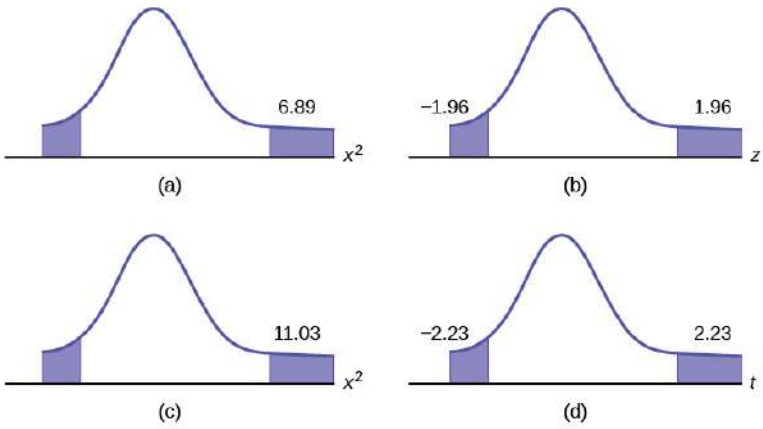


Figure B13

Sixty-four backpackers were asked the number of days since their latest backpacking trip. The number of days is given in **Table B26**.

# of days	1	2	3	4	5	6	7	8
Frequency	5	9	6	12	7	10	5	10

Table B26

33. Conduct an appropriate test to determine if the distribution is uniform.

A. The p-value is > 0.10. There is insufficient information to conclude that the distribution is not uniform.

B. The p-value is < 0.01. There is sufficient information to conclude the distribution is not uniform.

C. The p-value is between 0.01 and 0.10, but without alpha (α) there is not enough information.

D. There is no such test that can be conducted.

34. Which of the following statements is true when using one-way ANOVA?

A. The populations from which the samples are selected have different distributions.

B. The sample sizes are large.

C. The test is to determine if the different groups have the same means.

D. There is a correlation between the factors of the experiment.

Practice Final Exam 2 Solutions

Solutions

1. B parameter.

2. A

3. C seven

4. C 0.02

5. C none of the above

6. D $\frac{100}{140}$

7. A ≈ 0

8. B The values for x are: {1, 2, 3, . . . 14}

9. C 0.9417.

10. D binomial

11. D 8.7

12. A –1.96

13. A 0.6321

14. D 360

15. B $N\left(72, \frac{72}{\sqrt{50}}\right)$

16. A $\frac{3}{9}$

17. D

18. B 5.5

19. D 6.92

20. A 5

21. B 0.8541

22. B 0.2

23. A –1.

24. C matched pairs, dependent groups.

25. D Reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

26. C there is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.

27. B no

28. B p-value is close to 1.

29. B No

30. C $\hat{y} = 79.96x - 0.0094$

31. D We should not be estimating here.

32. A

33. A The p-value is > 0.10. There is insufficient information to conclude that the distribution is not uniform.

34. C The test is to determine if the different groups have the same means.

APPENDIX C: DATA SETS

Lap Times

The following tables provide lap times from Terri Vogel's log book. Times are recorded in seconds for 2.5-mile laps completed in a series of races and practice runs.

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 1	135	130	131	132	130	131	133
Race 2	134	131	131	129	128	128	129
Race 3	129	128	127	127	130	127	129
Race 4	125	125	126	125	124	125	125
Race 5	133	132	132	132	131	130	132
Race 6	130	130	130	129	129	130	129
Race 7	132	131	133	131	134	134	131
Race 8	127	128	127	130	128	126	128
Race 9	132	130	127	128	126	127	124
Race 10	135	131	131	132	130	131	130
Race 11	132	131	132	131	130	129	129
Race 12	134	130	130	130	131	130	130
Race 13	128	127	128	128	128	129	128
Race 14	132	131	131	131	132	130	130
Race 15	136	129	129	129	129	129	129
Race 16	129	129	129	128	128	129	129
Race 17	134	131	132	131	132	132	132
Race 18	129	129	130	130	133	133	127
Race 19	130	129	129	129	129	129	128
Race 20	131	128	130	128	129	130	130

Table C1 Race Lap Times (in seconds)

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 1	142	143	180	137	134	134	172
Practice 2	140	135	134	133	128	128	131
Practice 3	130	133	130	128	135	133	133

Table C2 Practice Lap Times (in seconds)

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 4	141	136	137	136	136	136	145
Practice 5	140	138	136	137	135	134	134
Practice 6	142	142	139	138	129	129	127
Practice 7	139	137	135	135	137	134	135
Practice 8	143	136	134	133	134	133	132
Practice 9	135	134	133	133	132	132	133
Practice 10	131	130	128	129	127	128	127
Practice 11	143	139	139	138	138	137	138
Practice 12	132	133	131	129	128	127	126
Practice 13	149	144	144	139	138	138	137
Practice 14	133	132	137	133	134	130	131
Practice 15	138	136	133	133	132	131	131

Table C2 Practice Lap Times (in seconds)

Stock Prices

The following table lists initial public offering (IPO) stock prices for all 1999 stocks that at least doubled in value during the first day of trading.

$17.00	$23.00	$14.00	$16.00	$12.00	$26.00
$20.00	$22.00	$14.00	$15.00	$22.00	$18.00
$18.00	$21.00	$21.00	$19.00	$15.00	$21.00
$18.00	$17.00	$15.00	$25.00	$14.00	$30.00
$16.00	$10.00	$20.00	$12.00	$16.00	$17.44
$16.00	$14.00	$15.00	$20.00	$20.00	$16.00
$17.00	$16.00	$15.00	$15.00	$19.00	$48.00
$16.00	$18.00	$9.00	$18.00	$18.00	$20.00
$8.00	$20.00	$17.00	$14.00	$11.00	$16.00
$19.00	$15.00	$21.00	$12.00	$8.00	$16.00
$13.00	$14.00	$15.00	$14.00	$13.41	$28.00
$21.00	$17.00	$28.00	$17.00	$19.00	$16.00
$17.00	$19.00	$18.00	$17.00	$15.00	
$14.00	$21.00	$12.00	$18.00	$24.00	
$15.00	$23.00	$14.00	$16.00	$12.00	
$24.00	$20.00	$14.00	$14.00	$15.00	
$14.00	$19.00	$16.00	$38.00	$20.00	

Table C3 IPO Offer Prices

$24.00	$16.00	$8.00	$18.00	$17.00	
$16.00	$15.00	$7.00	$19.00	$12.00	
$8.00	$23.00	$12.00	$18.00	$20.00	
$21.00	$34.00	$16.00	$26.00	$14.00	

Table C3 IPO Offer Prices

References

Data compiled by Jay R. Ritter of University of Florida using data from *Securities Data Co.* and *Bloomberg*.

APPENDIX D: GROUP AND PARTNER PROJECTS

Univariate Data

Student Learning Objectives

- The student will design and carry out a survey.
- The student will analyze and graphically display the results of the survey.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.
____ Decide what data you are going to study.

> Here are two examples, but you may **NOT** use them: number of M&M's per bag, number of pencils students have in their backpacks.

____ Are your data discrete or continuous? How do you know?
____ Decide how you are going to collect the data (for instance, buy 30 bags of M&M's; collect data from the World Wide Web).
____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. Which method did you use? Why did you pick that method?

____ Conduct your survey. **Your data size must be at least 30.**
____ Summarize your data in a chart with columns showing **data value, frequency, relative frequency and cumulative relative frequency.**
Answer the following (rounded to two decimal places):

a. $\hat{A}\bar{x}$ = _____
b. s = _____
c. First quartile = _____
d. Median = _____
e. 70^{th} percentile = _____

____ What value is two standard deviations above the mean?
____ What value is 1.5 standard deviations below the mean?
____ Construct a histogram displaying your data.
____ In complete sentences, describe the shape of your graph.
____ Do you notice any potential outliers? If so, what values are they? Show your work in how you used the potential outlier formula to determine whether or not the values might be outliers.
____ Construct a box plot displaying your data.
____ Does the middle 50% of the data appear to be concentrated together or spread apart? Explain how you determined this.
____ Looking at both the histogram and the box plot, discuss the distribution of your data.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

____ **Cover sheet**: name, class time, and name of your study

____ **Summary page**: This should contain paragraphs written with complete sentences. It should include answers to all the questions above. It should also include statements describing the population under study, the sample, a parameter or parameters being studied, and the statistic or statistics produced.
____ **URL** for data, if your data are from the World Wide Web
____ **Chart of data, frequency, relative frequency, and cumulative relative frequency**
____ **Page(s) of graphs:** histogram and box plot

Continuous Distributions and Central Limit Theorem

Student Learning Objectives

- The student will collect a sample of continuous data.
- The student will attempt to fit the data sample to various distribution models.
- The student will validate the central limit theorem.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

Part I: Sampling

____ Decide what **continuous** data you are going to study. (Here are two examples, but you may NOT use them: the amount of money a student spent on college supplies this term, or the length of time distance telephone call lasts.)
____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?

____ Conduct your survey. Gather **at least 150 pieces of continuous, quantitative data**.
____ Define (in words) the random variable for your data. X = _____
____ Create two lists of your data: (1) unordered data, (2) in order of smallest to largest.
____ Find the sample mean and the sample standard deviation (rounded to two decimal places).

a. $\bar{x}$ = _____
b. s = _____

____ Construct a histogram of your data containing five to ten intervals of equal width. The histogram should be a representative display of your data. Label and scale it.

Part II: Possible Distributions

____ Suppose that X followed the following theoretical distributions. Set up each distribution using the appropriate information from your data.
____ Uniform: $X \sim U$ _____ Use the lowest and highest values as a and b.
____ Normal: $X \sim N$ _____ Use $\bar{x}$ to estimate for μ and s to estimate for σ.
____ **Must** your data fit one of the above distributions? Explain why or why not.
____ **Could** the data fit two or three of the previous distributions (at the same time)? Explain.
____ Calculate the value k(an X value) that is 1.75 standard deviations above the sample mean. k = _____ (rounded to two decimal places) Note: $k = \bar{x} + (1.75)s$
____ Determine the relative frequencies (RF) rounded to four decimal places.

NOTE

$$RF = \frac{\text{frequency}}{\text{total number surveyed}}$$

a. $RF(X < k)$ = _____
b. $RF(X > k)$ = _____
c. $RF(X = k)$ = _____

NOTE

You should have one page for the uniform distribution, one page for the exponential distribution, and one page for the normal distribution.

____ State the distribution: $X \sim$ _____
____ Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.
____ Find the following theoretical probabilities (rounded to four decimal places).

a. $P(X < k) =$ _____
b. $P(X > k) =$ _____
c. $P(X = k) =$ _____

____ Compare the relative frequencies to the corresponding probabilities. Are the values close?
____ Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

Part III: CLT Experiments

_____ From your original data (before ordering), use a random number generator to pick 40 samples of size five. For each sample, calculate the average.
_____ On a separate page, attached to the summary, include the 40 samples of size five, along with the 40 sample averages.

_____ List the 40 averages in order from smallest to largest.
_____ Define the random variable, $\bar{X}$, in words. $\bar{X}$ = _____
_____ State the approximate theoretical distribution of $\bar{X}$. $\bar{X} \sim$ _____
_____ Base this on the mean and standard deviation from your original data.
_____ Construct a histogram displaying your data. Use five to six intervals of equal width. Label and scale it.
Calculate the value $\bar{k}$ (an $\bar{X}$ value) that is 1.75 standard deviations above the sample mean. $\bar{k}$ = _____ (rounded to two decimal places)
Determine the relative frequencies (*RF*) rounded to four decimal places.

a. $RF(\bar{X} < \bar{k}) =$ _____
b. $RF(\bar{X} > \bar{k}) =$ _____
c. $RF(\bar{X} = \bar{k}) =$ _____

Find the following theoretical probabilities (rounded to four decimal places).

a. $P(\bar{X} < \bar{k}) =$ _____
b. $P(\bar{X} > \bar{k}) =$ _____
c. $P(\bar{X} = \bar{k}) =$ _____

_____ Draw the graph of the theoretical distribution of X.
_____ Compare the relative frequencies to the probabilities. Are the values close?

_____ Does it appear that the data of averages fit the distribution of $\bar{X}$ well? Justify your answer by comparing the probabilities to the relative frequencies, and the histogram to the theoretical graph.
In three to five complete sentences for each, answer the following questions. Give thoughtful explanations.
_____ In summary, do your original data seem to fit the uniform, exponential, or normal distributions? Answer why or why not for each distribution. If the data do not fit any of those distributions, explain why.

_____ What happened to the shape and distribution when you averaged your data? **In theory,** what should have happened? In theory, would â€œitâ€? always happen? Why or why not?

_____ Were the relative frequencies compared to the theoretical probabilities closer when comparing the X or $\bar{X}$ distributions? Explain your answer.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:
____ **Cover sheet**: name, class time, and name of your study
____ **Summary pages**: These should contain several paragraphs written with complete sentences that describe the experiment, including what you studied and your sampling technique, as well as answers to all of the questions previously asked questions
____ **URL** for data, if your data are from the World Wide Web
____ **Pages, one for each theoretical distribution**, with the distribution stated, the graph, and the probability questions answered
____ **Pages of the data requested**
____ **All graphs required**

Hypothesis Testing-Article

Student Learning Objectives

- The student will identify a hypothesis testing problem in print.
- The student will conduct a survey to verify or dispute the results of the hypothesis test.
- The student will summarize the article, analysis, and conclusions in a report.

Instructions

As you complete each task, check it off. Answer all questions in your summary.
____**Find an article** in a newspaper, magazine, or on the internet which makes a claim about **ONE** population mean or **ONE** population proportion. The claim may be based upon a survey that the article was reporting on. Decide whether this claim is the null or alternate hypothesis.
____**Copy or print out the article** and include a copy in your project, along with the source.
____**State how you will collect your data.** (Convenience sampling is not acceptable.)
____**Conduct your survey. You must have more than 50 responses in your sample.** When you hand in your final project, attach the tally sheet or the packet of questionnaires that you used to collect data. Your data must be real.
____**State the statistics** that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.
____**Make two copies of the appropriate solution sheet.**
____**Record the hypothesis test** on the solution sheet, based on your experiment. **Do a DRAFT solution** first on one of the solution sheets and check it over carefully. Have a classmate check your solution to see if it is done correctly. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution sheet.
____**Create a graph that illustrates your data.** This may be a pie or bar graph or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for the type of data in your project.
____**Write your summary** (in complete sentences and paragraphs, with proper grammar and correct spelling) that describes the project. The summary **MUST** include:

a. Brief discussion of the article, including the source
b. Statement of the claim made in the article (one of the hypotheses).
c. Detailed description of how, where, and when you collected the data, including the sampling technique; did you use cluster, stratified, systematic, or simple random sampling (using a random number generator)? As previously mentioned, convenience sampling is not acceptable.
d. Conclusion about the article claim in light of your hypothesis test; this is the conclusion of your hypothesis test, stated in words, in the context of the situation in your project in sentence form, as if you were writing this conclusion for a non-statistician.
e. Sentence interpreting your confidence interval in the context of the situation in your project

Assignment Checklist

Turn in the following typed (12 point) and stapled packet for your final project:
____**Cover sheet** containing your name(s), class time, and the name of your study
____**Summary**, which includes all items listed on summary checklist
____**Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.
____**Graphic representation of your data**, created following the guidelines previously discussed; include only graphs which are appropriate and useful.

____**Raw data collected AND a table summarizing the sample data** (*n*, $\hat{A}\bar{x}$ and *s*; or *x*, *n*, and *p*â€™, as appropriate for your hypotheses); the raw data does not need to be typed, but the summary does. Hand in the data as you collected it. (Either attach your tally sheet or an envelope containing your questionnaires.)

Bivariate Data, Linear Regression, and Univariate Data

Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.
- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

Part I: Bivariate Data

Introduction

____State the bivariate data your group is going to study.

Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

____Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.
____Conduct your survey. Your number of pairs must be at least 30.
____Print out a copy of your data.

Analysis

____On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.
____State the least squares line and the correlation coefficient.
____On your scatter plot, in a different color, construct the least squares line.
____Is the correlation coefficient significant? Explain and show how you determined this.
____Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.
____Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.
____Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your independent variable is sequential data such as year with 30 years and one piece of data per year, your *x*-values might be 1971, 1972, 1973, 1974, â€¦, 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.

_____Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.

_____Answer the following question, rounded to two decimal places:

a. Sample mean = _____
b. Sample standard deviation = _____
c. First quartile = _____
d. Third quartile = _____
e. Median = _____
f. 70th percentile = _____
g. Value that is 2 standard deviations above the mean = _____
h. Value that is 1.5 standard deviations below the mean = _____

_____Construct a histogram displaying your data. Group your data into six to ten intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26,27-33,34-40,41-47,48-54,55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .

_____In complete sentences, describe the shape of your histogram.

_____Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in **Descriptive Statistics** (since you are now using univariate data) to determine which values might be outliers.

_____Construct a box plot of your data.

_____Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.

_____Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

Due Dates

- Part I, Intro: _____ (keep a copy for your records)
- Part I, Analysis: _____ (keep a copy for your records)
- Entire Project, typed and stapled: _____

  ____ Cover sheet: names, class time, and name of your study

  ____ Part I: label the sections â€œIntroâ€? and â€œAnalysis.â€?
  ____ Part II:
  ____ Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.
  ____ All graphs requested in the project
  ____ All calculations requested to support questions in data
  ____ Description: what you learned by doing this project, what challenges you had, how you overcame the challenges

NOTE

Include answers to ALL questions asked, even if not explicitly repeated in the items above.

APPENDIX E: SOLUTION SHEETS

Hypothesis Testing With One Sample

Class Time: _____
Name: _____

a. H_0: _____
b. H_a: _____
c. In words, clearly state what your random variable $\bar{X}$ or P' represents.
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one or two *complete sentences*, explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. *Clearly*, label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

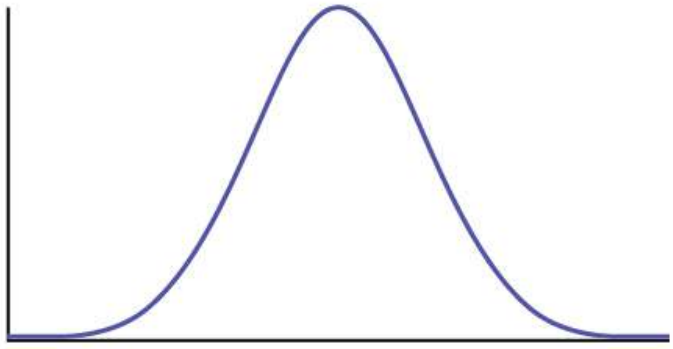


Figure E1

h. Indicate the correct decision (*reject* or *do not reject* the null hypothesis), the reason for it and write appropriate conclusions using *complete sentences."*
   i. Alpha: _____
   ii. Decision: _____
   iii. Reason for decision: _____
   iv. Conclusion: _____
i. Construct a 95 percent confidence interval for the true mean or proportion. Sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.

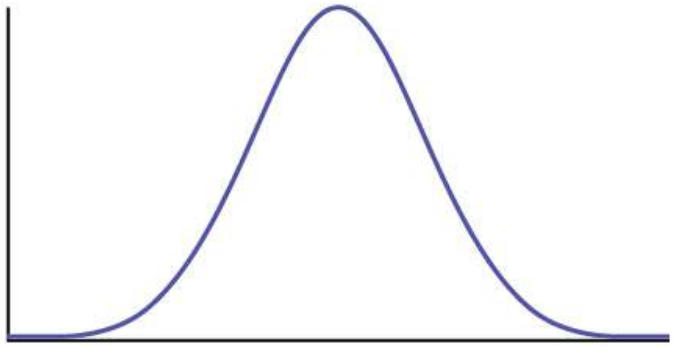


Figure E2

Hypothesis Testing With Two Samples

Class Time: _____
Name: _____

a. H_0: _____

b. H_a: _____

c. In words, *clearly* state what your random variable $\bar{X}_1 - \bar{X}_2$, $P'_1 - P'_2$ or $\bar{X}_d$ represents.

d. State the distribution to use for the test.

e. What is the test statistic?

f. What is the *p*-value? In one to two *complete sentences*, explain what the *p*-value means for this problem.

g. Use the previous information to sketch a picture of this situation. *Clearly* label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

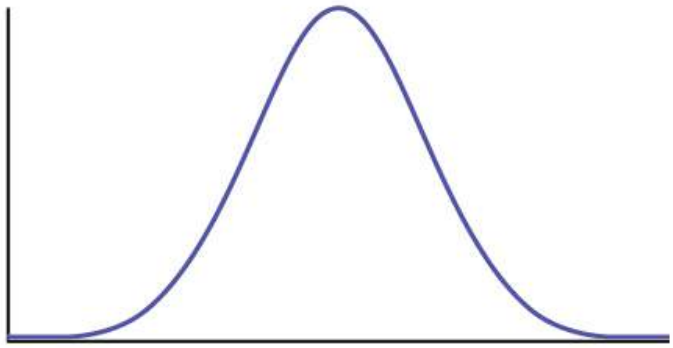


Figure E3

h. Indicate the correct decision (*reject* or *do not reject* the null hypothesis), and write appropriate conclusions using *complete sentences.*

   i. Alpha: _____

   ii. Decision: _____

   iii. Reason for decision: _____

   iv. Conclusion: _____

i. In *complete sentences*, explain how you determined which distribution to use.

The Chi-Square Distribution

Class Time: _____
Name: _____

a. H_0: _____
b. H_a: _____
c. What are the degrees of freedom?
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one to two *complete sentences*, explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. *Clearly* label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

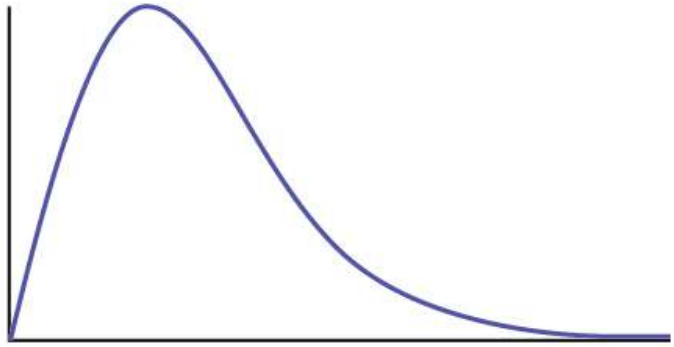


Figure E4

h. Indicate the correct decision (*reject* or *do not reject* the null hypothesis) and write appropriate conclusions, using *complete sentences.*
   i. Alpha: _____
   ii. Decision: _____
   iii. Reason for decision: _____
   iv. Conclusion: _____

F Distribution and One-Way ANOVA

Class Time: _____
Name: _____

a. H_0: _____
b. H_a: _____
c. $df(n)$ = _____ $df(d)$ = _____
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value?
g. Use the previous information to sketch a picture of this situation. *Clearly* label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

Figure E5

h. Indicate the correct decision (*reject* or *do not reject* the null hypothesis) and write appropriate conclusions, using *complete sentences*.
   a. Alpha: _____
   b. Decision: _____
   c. Reason for decision: _____
   d. Conclusion: _____

APPENDIX F: MATHEMATICAL PHRASES, SYMBOLS, AND FORMULAS

English Phrases Written Mathematically

When the English says:	Interpret this as:
X is at least 4.	$X \geq 4$
The minimum of *X* is 4.	$X \geq 4$
X is no less than 4.	$X \geq 4$
X is greater than or equal to 4.	$X \geq 4$
X is at most 4.	$X \leq 4$
The maximum of *X* is 4.	$X \leq 4$
X is no more than 4.	$X \leq 4$
X is less than or equal to 4.	$X \leq 4$
X does not exceed 4.	$X \leq 4$
X is greater than 4.	$X > 4$
X is more than 4.	$X > 4$
X exceeds 4.	$X > 4$
X is less than 4.	$X < 4$
There are fewer *X* than 4.	$X < 4$
X is 4.	$X = 4$
X is equal to 4.	$X = 4$
X is the same as 4.	$X = 4$
X is not 4.	$X \neq 4$
X is not equal to 4.	$X \neq 4$
X is not the same as 4.	$X \neq 4$
X is different than 4.	$X \neq 4$

Table F1

Formulas

Formula 1: Factorial

$$n! = n(n-1)(n-2)...(1)$$

$$0! = 1$$

Formula 2: Combinations

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

Formula 3: Binomial Distribution

$$X \sim B(n,\ p)$$

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \text{ for } x = 0,\ 1,\ 2,\ ...,\ n$$

Formula 4: Geometric Distribution

$$X \sim G(p)$$

$$P(X = x) = q^{x-1} p, \text{ for } x = 1,\ 2,\ 3,\ ...$$

Formula 5: Hypergeometric Distribution

$$X \sim H(r,\ b,\ n)$$

$$P(X = x) = \left(\frac{\binom{r}{x}\binom{b}{n-x}}{\binom{r+b}{n}}\right)$$

Formula 6: Poisson Distribution

$$X \sim P(\mu)$$

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$$

Formula 7: Uniform Distribution

$$X \sim U(a,\ b)$$

$$f(X) = \frac{1}{b-a},\ a < x < b$$

Formula 8: Exponential Distribution

$$X \sim Exp(m)$$

$$f(x) = me^{-mx} m > 0,\ x \geq 0$$

Formula 9: Normal Distribution

$$X \sim N(\mu,\ \sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}},\ -\infty < x < \infty$$

Formula 10: Gamma Function

$$\Gamma(z) = \int_{\infty}^{0} x^{z-1} e^{-x} dx \quad z > 0$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$\Gamma(m+1) = m!$ for m, a nonnegative integer

otherwise: $\Gamma(a+1) = a\Gamma(a)$

Formula 11: Student's *t*-distribution

$X \sim t_{df}$

$$f(x) = \frac{\left(1 + \frac{x^2}{n}\right)^{\frac{-(n+1)}{2}} \Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\mathrm{n}\pi}\Gamma\left(\frac{n}{2}\right)}$$

$$X = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

$Z \sim N(0, 1), Y \sim X^2_{df}$, n = degrees of freedom

Formula 12: Chi-Square Distribution

$X \sim X^2_{df}$

$f(x) = \frac{x^{\frac{n-2}{2}} e^{\frac{-x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}$, $x > 0$, n = positive integer and degrees of freedom

Formula 13: F Distribution

$X \sim F_{df(n),\, df(d)}$

$df(n)$ = degrees of freedom for the numerator

$df(d)$ = degrees of freedom for the denominator

$$f(x) = \frac{\Gamma(\frac{u+v}{2})}{\Gamma(\frac{u}{2})\Gamma(\frac{v}{2})}(\frac{u}{v})^{\frac{u}{2}} x^{(\frac{u}{2}-1)}[1 + (\frac{u}{v})x^{-0.5(u+v)}]$$

$X = \frac{Y_u}{W_v}$, Y, W are chi-square

Symbols and Their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Sampling and Data	$\sqrt{\quad}$	The square root of	same
Sampling and Data	π	Pi	3.14159… (a specific number)
Descriptive Statistics	Q_1	Quartile one	the first quartile

Table F2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Descriptive Statistics	Q_2	Quartile two	the second quartile
Descriptive Statistics	Q_3	Quartile three	the third quartile
Descriptive Statistics	*IQR*	interquartile range	$Q_3 - Q_1 = IQR$
Descriptive Statistics	$\bar{x}$	x-bar	sample mean
Descriptive Statistics	μ	mu	population mean
Descriptive Statistics	s s_x *sx*	s	sample standard deviation
Descriptive Statistics	s^2 s_x^2	s squared	sample variance
Descriptive Statistics	σ σ_x *σx*	sigma	population standard deviation
Descriptive Statistics	σ^2 σ_x^2	sigma squared	population variance
Descriptive Statistics	Σ	capital sigma	sum
Probability Topics	$\{\}$	brackets	set notation
Probability Topics	S	S	sample space
Probability Topics	A	Event A	event A
Probability Topics	$P(A)$	probability of A	probability of A occurring
Probability Topics	$P(A\|B)$	probability of A given B	prob. of A occurring given B has occurred
Probability Topics	$P(A \text{ OR } B)$	prob. of A or B	prob. of A or B or both occurring
Probability Topics	$P(A \text{ AND } B)$	prob. of A and B	prob. of both A and B occurring (same time)
Probability Topics	*A′*	A-prime, complement of A	complement of A, not A
Probability Topics	*P*(*A*')	prob. of complement of A	same
Probability Topics	$P(G_1)$	green on first pick	same
Probability Topics	$P(G_1)$	prob. of green on first pick	same
Discrete Random Variables	*PDF*	prob. distribution function	same
Discrete Random Variables	*X*	X	the random variable X
Discrete Random Variables	*X* ~	the distribution of X	same
Discrete Random Variables	*B*	binomial distribution	same
Discrete Random Variables	*G*	geometric distribution	same
Discrete Random Variables	*H*	hypergeometric dist.	same
Discrete Random Variables	*P*	Poisson dist.	same
Discrete Random Variables	λ	Lambda	average of Poisson distribution
Discrete Random Variables	$\geq$	greater than or equal to	same
Discrete Random Variables	$\leq$	less than or equal to	same
Discrete Random Variables	=	equal to	same

Table F2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Discrete Random Variables	≠	not equal to	same
Continuous Random Variables	*f*(*x*)	*f* of *x*	function of *x*
Continuous Random Variables	*pdf*	prob. density function	same
Continuous Random Variables	*U*	uniform distribution	same
Continuous Random Variables	*Exp*	exponential distribution	same
Continuous Random Variables	*k*	*k*	critical value
Continuous Random Variables	*f*(*x*) =	*f* of *x* equals	same
Continuous Random Variables	*m*	*m*	decay rate (for exp. dist.)
The Normal Distribution	*N*	normal distribution	same
The Normal Distribution	*z*	*z*-score	same
The Normal Distribution	*Z*	standard normal dist.	same
The Central Limit Theorem	*CLT*	Central Limit Theorem	same
The Central Limit Theorem	$\bar{X}$	*X*-bar	the random variable *X*-bar
The Central Limit Theorem	μ_x	mean of *X*	the average of *X*
The Central Limit Theorem	$\mu_{\bar{x}}$	mean of *X*-bar	the average of *X*-bar
The Central Limit Theorem	σ_x	standard deviation of *X*	same
The Central Limit Theorem	$\sigma_{\bar{x}}$	standard deviation of *X*-bar	same
The Central Limit Theorem	ΣX	sum of *X*	same
The Central Limit Theorem	Σx	sum of *x*	same
Confidence Intervals	*CL*	confidence level	same
Confidence Intervals	*CI*	confidence interval	same
Confidence Intervals	*EBM*	error bound for a mean	same
Confidence Intervals	*EBP*	error bound for a proportion	same
Confidence Intervals	*t*	Student's *t*-distribution	same
Confidence Intervals	*df*	degrees of freedom	same
Confidence Intervals	$t_{\frac{\alpha}{2}}$	student t with *a*/2 area in right tail	same
Confidence Intervals	p' ; $\hat{p}$	*p*-prime; *p*-hat	sample proportion of success
Confidence Intervals	q' ; $\hat{q}$	*q*-prime; *q*-hat	sample proportion of failure

Table F2 Symbols and their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Hypothesis Testing	H_0	*H*-naught, *H*-sub 0	null hypothesis
Hypothesis Testing	H_a	*H*-*a*, *H*-sub *a*	alternate hypothesis
Hypothesis Testing	H_1	*H*-1, *H*-sub 1	alternate hypothesis
Hypothesis Testing	α	alpha	probability of Type I error
Hypothesis Testing	β	beta	probability of Type II error
Hypothesis Testing	$\overline{X1} - \overline{X2}$	*X*1-bar minus *X*2-bar	difference in sample means
Hypothesis Testing	$\mu_1 - \mu_2$	*mu*-1 minus *mu*-2	difference in population means
Hypothesis Testing	$P'_1 - P'_2$	*P*1-prime minus *P*2-prime	difference in sample proportions
Hypothesis Testing	$p_1 - p_2$	*p*1 minus *p*2	difference in population proportions
Chi-Square Distribution	X^2	*Ky*-square	Chi-square
Chi-Square Distribution	O	Observed	Observed frequency
Chi-Square Distribution	E	Expected	Expected frequency
Linear Regression and Correlation	$y = a + bx$	*y* equals a plus *b*-*x*	equation of a line
Linear Regression and Correlation	$\hat{y}$	*y*-hat	estimated value of *y*
Linear Regression and Correlation	r	correlation coefficient	same
Linear Regression and Correlation	ε	error	same
Linear Regression and Correlation	*SSE*	Sum of Squared Errors	same
Linear Regression and Correlation	1.9*s*	1.9 times *s*	cut-off value for outliers
F-Distribution and ANOVA	*F*	*F*-ratio	*F*-ratio

Table F2 Symbols and their Meanings

APPENDIX G: NOTES FOR THE TI-83, 83+, 84, 84+ CALCULATORS

Quick Tips

Legend

-  represents a button press
- [ ] represents yellow command or green letter behind a key
- < > represents items on the screen

To adjust the contrast

Press  , then hold  to increase the contrast or  to decrease the contrast.

To capitalize letters and words

Press  to get one capital letter, or press  , then  to set all button presses to capital letters. You can return to the top-level button values by pressing  again.

To correct a mistake

If you hit a wrong button, press  and start again.

To write in scientific notation

Numbers in scientific notation are expressed on the TI-83, 83+, 84, and 84+ using E notation, such that...

- 4.321 E 4 = 4.321×10^{4}
- 4.321 E –4 = 4.321×10^{-4}

To transfer programs or equations from one calculator to another

Both calculators: Insert your respective end of the link cable cable and press  , then [LINK].

Calculator receiving information

1. Use the arrows to navigate to and select <RECEIVE>.
2. Press  .

Calculator sending information

1. Press the appropriate number or letter.

2. Use the up and down arrows to access the appropriate item.
3. Press ENTER to select the item to transfer.
4. Press the right arrow to navigate to and select <TRANSMIT>.
5. Press ENTER.

NOTE

ERROR 35 LINK generally means that the cables have not been inserted far enough.

Both calculators—Insert your respective end of the link cable, press 2nd, then [QUIT] to exit when done.

Manipulating One-Variable Statistics

NOTE

These directions are for entering data using the built-in statistical program.

Data	Frequency
–2	10
–1	3
0	4
1	5
3	8

Table G1 Sample Data We are manipulating one-variable statistics.

To begin

1. Turn on the calculator.

   ON

2. Access statistics mode.

   STAT

3. Select <4:ClrList> to clear data from lists, if desired.

   4 ENTER, then .

4. Enter the list [L1] to be cleared.

   2nd , [L1], ENTER.

5. Display the last instruction.

   2nd , [ENTRY].

6. Continue clearing any remaining lists in the same fashion, if desired.

   ◀ , 2nd , [L2] , ENTER

7. Access statistics mode.

   STAT

8. Select <1:Edit . . .>.

   ENTER

9. Enter data. Data values go into [L1]. (You may need to arrow over to [L1]).
   - Type in a data value and enter it. For negative numbers, use the negate – key at the bottom of the keypad.

     (–) , 9 , ENTER .
   - Continue in the same manner until all data values are entered.
10. In [L2], enter the frequencies for each data value in [L1].
    - Type in a frequency and enter it. If a data value appears only once, the frequency is *1*.

      4 , ENTER .
    - Continue in the same manner until all data values are entered.
11. Access statistics mode.

    STAT

12. Navigate to <CALC>.
13. Access <1:1-var Stats>.

    ENTER

14. Indicate that the data is in [L1]...

    2nd , [L1] , , ,

15. ...and indicate that the frequencies are in [L2].

    2nd , [L2] , ENTER .

16. The statistics should be displayed. You may arrow down to get remaining statistics. Repeat as necessary.

Drawing Histograms

NOTE

We will assume that the data are already entered.

We will construct two histograms with the built-in [STAT PLOT] application. In the first method, we will use the default ZOOM. The second method will involve customizing a new graph.

1. Access graphing mode.

   2nd , [STAT PLOT].

2. Select <1:plot 1> to access plotting - first graph.

   ENTER

3. Use the arrows to navigate to <ON> to turn on Plot 1.

   ENTER
   <ON> , .

4. Use the arrows to go to the histogram picture and select the histogram.

   ENTER

5. Use the arrows to navigate to <Xlist>.
6. If [L1] is not selected, select it.

   2nd ENTER
   , [L1] , .

7. Use the arrows to navigate to <Freq>.
8. Assign the frequencies to [L2].

   2nd ENTER
   , [L2] , .

9. Go back to access other graphs.

   2nd , [STAT PLOT].

10. Use the arrows to turn off the remaining plots.
11. **Be sure to deselect or clear all equations before graphing.**

To deselect equations

1. Access the list of equations.

   Y=

2. Select each equal sign (=).

   ▼ ▶ ENTER
   .

3. Continue until all equations are deselected.

To clear equations

1. Access the list of equations.

   Y=

2. Use the arrow keys to navigate to the right of each equal sign (=) and clear them.

.

3. Repeat until all equations are deleted.

To draw default histogram

1. Access the ZOOM menu.

2. Select <9:ZoomStat>.

3. The histogram will display with a window automatically set.

To draw a custom histogram

1. Access window mode to set the graph parameters.

2.
   - $X_{\min} = -2.5$
   - $X_{\max} = 3.5$
   - $X_{scl} = 1$ (width of bars)
   - $Y_{\min} = 0$
   - $Y_{\max} = 10$
   - $Y_{scl} = 1$ (spacing of tick marks on *y*-axis)
   - $X_{res} = 1$

3. Access graphing mode to see the histogram.

To draw box plots

1. Access graphing mode.

, [STAT PLOT].

2. Select <1:Plot 1> to access the first graph.

3. Use the arrows to select <ON> and turn on Plot 1.

4. Use the arrows to select the box plot picture and enable it.

5. Use the arrows to navigate to <Xlist>.

6. If [L1] is not selected, select it.

   2nd ENTER , [L1] , .
7. Use the arrows to navigate to <Freq>.
8. Indicate that the frequencies are in [L2].

   2nd ENTER , [L2] , .
9. Go back to access other graphs.

   2nd , [STAT PLOT].
10. **Be sure to deselect or clear all equations before graphing** using the method mentioned above.
11. View the box plot.

    GRAPH , [STAT PLOT].

Linear Regression

Sample Data

The following data are real. The percent of declared ethnic minority students at De Anza College for selected years from 1970–1995 is indicated in the following table.

Year	Student Ethnic Minority Percentage
1970	14.13%
1973	12.27%
1976	14.08%
1979	18.16%
1982	27.64%
1983	28.72%
1986	31.86%
1989	33.14%
1992	45.37%
1995	53.1%

Table G2 The independent variable is *Year,* while the independent variable is *Student Ethnic Minority Percentage*.

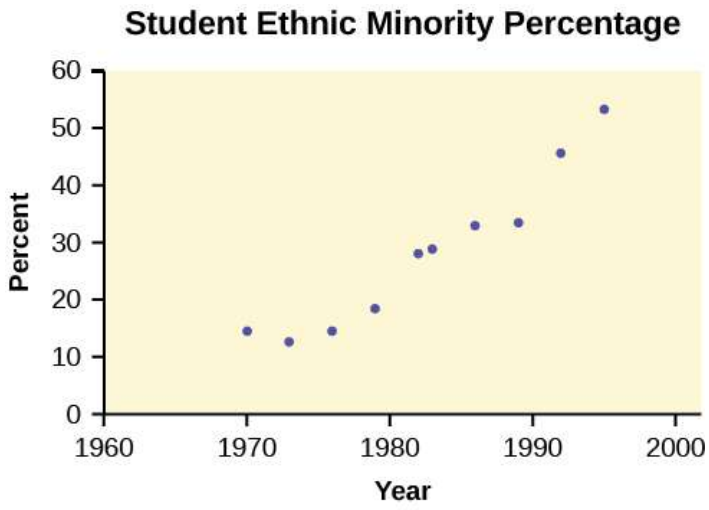


Figure G1 Student Ethnic Minority Percentage By hand, verify the scatterplot above.

NOTE

The TI-83 has a built-in linear regression feature, which allows the data to be edited. The *x*-values will be in [L1]; the *y*-values in [L2].

To enter data and perform linear regression

1. ON Turns calculator on.

   ON

2. Before accessing this program, be sure to turn off all plots.
   - Access graphing mode.

     2nd , [STAT PLOT].
   - Turn off all plots.

     4 ENTER , .
3. Round to three decimal places.
   - Access the mode menu.

     MODE , [STAT PLOT].
   - Navigate to <Float> and then to the right until you reach <3>.

     ▼ ▶ .
   - All numbers will be rounded to three decimal places until changed.

     ENTER
4. Enter statistics mode and clear lists [L1] and [L2], as described previously.

STAT 4

, .

5. Enter editing mode to insert values for *x* and *y*.

STAT ENTER

, .

ENTER

6. Enter each value. Press to continue.

To display the correlation coefficient

1. Access the catalog.

2nd

, [CATALOG].

2. Arrow down and select <DiagnosticOn>.

▼ ENTER ENTER

... , , .

3. r and r^2 will be displayed during regression calculations.
4. Access linear regression.

STAT ▶

.

5. Select the form of $y = a + bx$.

8 ENTER

, .

The display will show the following information

LinReg

- $y = a + bx$
- $a = -3176.909$
- $b = 1.617$
- $r^2 = 0.924$
- $r = 0.961$

This means the Line of Best Fit (Least Squares Line) is:

- $y = -3176.909 + 1.617x$
- % = –3176.909 + 1.617 (year #)

The correlation coefficient is $r = 0.961$.

To see the scatter plot

1. Access graphing mode.

2nd

, [STAT PLOT].

2. Select <1:Plot 1> To access plotting - first graph.

ENTER

3. Navigate and select <ON> to turn on <1:Plot 1>.

   ENTER
   <ON> .
4. Navigate to the first picture.
5. Select the scatter plot.

   ENTER
6. Navigate to <Xlist>.

   2nd
7. If [L1] is not selected, press , then [L1] to select it.
8. Confirm that the data values are in [L1].

   ENTER
   <ON>, .
9. Navigate to <Ylist>.
10. Select that the frequencies are in [L2].

    2nd ENTER
    , [L2] ,
11. Go back to access other graphs.

    2nd
    , [STAT PLOT]
12. Use the arrows to turn off the remaining plots.
13. Access window mode to set the graph parameters.

    WINDOW

    - $X_{\min} = 1970$
    - $X_{\max} = 2000$
    - $X_{scl} = 10$ (spacing of tick marks on x-axis)
    - $Y_{\min} = -0.05$
    - $Y_{\max} = 60$
    - $Y_{scl} = 10$ (spacing of tick marks on y-axis)
    - $X_{res} = 1$
14. Be sure to deselect or clear all equations before graphing, using the instructions above.

    GRAPH
15. Press the graph button to see the scatter plot.

To see the regression graph

1. Access the equation menu. The regression equation will be put into Y1.

   Y=

2. Access the vars menu and navigate to <5: Statistics>.

   VARS 5

   , .

3. Navigate to <EQ>.
4. <1: RegEQ> contains the regression equation which will be entered in Y1.

   ENTER

5. Press the graphing mode button. The regression line will be superimposed over the scatter plot.

   GRAPH

To see the residuals and use them to calculate the critical point for an outlier

1. Access the list. <RESID> will be an item on the menu. Navigate to it.

   2nd , [LIST], then <RESID>.

2. Press enter twice to view the list of residuals. Use the arrows to select them.

   ENTER ENTER

   , .

3. The critical point for an outlier is $1.9V\frac{\text{SSE}}{n-2}$, where
   - n = number of pairs of data
   - SSE = sum of the squared errors
   - $\sum\left(\text{residual}^2\right)$
4. Store the residuals in [L3].

   STO► 2nd ENTER

   , , [L3], .

5. Calculate the $\frac{(\text{residual})^2}{n-2}$. Note that $n-2=8$.

   2nd x^2 ÷ 8

   , [L3], , , then .

6. Store this value in [L4].

   STO► 2nd ENTER

   , , [L4], .

7. Calculate the critical value using the equation above.

   1 . 9 × 2nd 2nd ►

   , , , , , [V], , [LIST]

   ► 5 2nd)) ENTER

   , , , , [L4], , , then .

8. Verify that the calculator displays 7.642669563. This is the critical value.
9. Compare the absolute value of each residual value in [L3] to 7.64. If the absolute value is greater than 7.64, then the (x, y) corresponding point is an outlier. In this case, none of the points is an outlier.

To obtain estimates of *y* for various *x*-values

There are various ways to determine estimates for "*y*." One way is to substitute values for "*x*" in the equation. Another way is to use the TRACE on the graph of the regression line.

TI-83, 83+, 84, 84+ instructions for distributions and tests

Distributions

Access DISTR for *Distributions*.

For technical assistance, visit the Texas Instruments website at **http://www.ti.com (http://www.ti.com)** and enter your calculator model into the *search* box.

Binomial Distribution

- `binompdf(n,p,x)` corresponds to $P(X = x)$
- `binomcdf(n,p,x)` corresponds to $P(X \le x)$
- To see a list of all probabilities for x: 0, 1, . . . , n, leave off the "x" parameter.

Poisson Distribution

- `poissonpdf(λ,x)` corresponds to $P(X = x)$
- `poissoncdf(λ,x)` corresponds to $P(X \le x)$

Continuous Distributions (general)

- $-\infty$ uses the value –1EE99 for left bound
- $+\infty$ uses the value 1EE99 for right bound

Normal Distribution

- `normalpdf(x,μ,σ)` yields a probability density function value, only useful to plot the normal curve, in which case "x" is the variable
- `normalcdf(left bound, right bound, μ, σ)` corresponds to P(left bound < X < right bound)
- `normalcdf(left bound, right bound)` corresponds to P(left bound < Z < right bound) – standard normal
- `invNorm(p,μ,σ)` yields the critical value, k: $P(X < k) = p$
- `invNorm(p)` yields the critical value, k: $P(Z < k) = p$ for the standard normal

Student's *t*-Distribution

- `tpdf(x,df)` yields the probability density function value, only useful to plot the student-*t* curve, in which case "x" is the variable)
- `tcdf(left bound, right bound, df)` corresponds to P(left bound < t < right bound)

Chi-square Distribution

- X^2`pdf(x,df)` yields the probability density function value, only useful to plot the chi^2 curve, in which case "x" is the variable
- X^2`cdf(left bound, right bound, df)` corresponds to P(left bound < X^2 < right bound)

F Distribution

- `Fpdf(x,dfnum,dfdenom)` yields the probability density function value, only useful to plot the F curve, in which case "x" is the variable
- `Fcdf(left bound,right bound,dfnum,dfdenom)` corresponds to P(left bound < F < right bound)

Tests and Confidence Intervals

Access STAT and TESTS.

For the confidence intervals and hypothesis tests, you may enter the data into the appropriate lists and press DATA to have the calculator find the sample means and standard deviations. Or, you may enter the sample means and sample standard deviations directly by pressing STAT once in the appropriate tests.

Confidence Intervals

- `ZInterval` is the confidence interval for mean when σ is known.
- `TInterval` is the confidence interval for mean when σ is unknown; *s* estimates σ.
- `1-PropZInt` is the confidence interval for proportion.

NOTE

The confidence levels should be given as percents (e.g., enter "`95`" or "`.95`" for a 95 percent confidence level).

Hypothesis Tests

- `Z-Test` is the hypothesis test for single mean when σ is known.
- `T-Test` is the hypothesis test for single mean when σ is unknown; *s* estimates σ.
- `2-SampZTest` is the hypothesis test for two independent means when both σs are known.
- `2-SampTTest` is the hypothesis test for two independent means when both σs are unknown.
- `1-PropZTest` is the hypothesis test for a single proportion.
- `2-PropZTest` is the hypothesis test for two proportions.
- X^2`-Test` is the hypothesis test for independence.
- X^2`GOF-Test` is the hypothesis test for goodness-of-fit (TI-84+ only).
- `LinRegTTEST` is the hypothesis test for Linear Regression (TI-84+ only).

NOTE

Input the null hypothesis value in the row below "`Inpt`." For a test of a single mean, "`μ⌀`" represents the null hypothesis. For a test of a single proportion, "`p⌀`" represents the null hypothesis. Enter the alternate hypothesis on the bottom row.

APPENDIX H: TABLES

The module contains links to government site tables used in statistics.

NOTE

When you are finished with the table link, use the back button on your browser to return here.

Tables (NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 3, 2009)

- **Student t table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm)**
- **Normal table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm)**
- **Chi-Square table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm)**
- **F-table (http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm)**
- All **four tables (http://www.itl.nist.gov/div898/handbook/eda/section3/eda367.htm)** can be accessed by going to http://www.itl.nist.gov/div898/handbook/eda/section3/eda367.htm

95% Critical Values of the Sample Correlation Coefficient Table

- **95% Critical Values of the Sample Correlation Coefficient**

INDEX