Making Cents of Your Income

What's the average income in your state? If a researcher claimed that the mean is $30,000 after surveying 100 residents, would you find this claim believable? In this module, you investigate how to use sampling to test claims about populations.

Doug Mack • David Thiel • Deanna Turley
Danny Jones • John Knudson-Martin

Making Cents of Your Income

Introduction

How could you determine the average income of a high school graduate in your state? Since it is not practical to contact every high school graduate, you might survey a sample of this population. In this case, you would want to contact enough graduates to obtain an accurate estimate, but not more than is necessary. How many are enough?

Using a sample to obtain accurate information about a population can be a complicated process. In this module, you investigate how sample size affects the reliability of an estimate, determine the confidence you should have in an estimate, and use samples to test claims about a population.

Activity 1

In order to decide how large a sample you need to make reasonable predictions about a population, you must first understand how information gained from samples of different sizes can vary.

Exploration 1

In this exploration, you draw samples of different sizes from a population, then compare the means and standard deviations of each sample.

Mathematics Note

Standard deviation is a measure of the spread in a data set. The standard deviation of a population, often denoted by σ, can be calculated using the formula below:

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N}}$$

where $x_1, x_2, \cdots, x_N$ represent all the individual values in the population, μ represents their mean, and *N* represents the population size.

The standard deviation of a sample is the **sample standard deviation**, denoted by *s*. Statisticians often use *s* to approximate σ. The formula for calculating *s* is:

$$s = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}}$$

where $x_1, x_2, \cdots, x_n$ represent the individual values in the sample, $\bar{x}$ represents the sample mean, and *n* represents the sample size.

For example, suppose that you selected a sample of five cars during a study of the ages of motor vehicles. The ages of the cars, in years, were 1, 4, 6, 8, and 11. In this case, the sample mean $\bar{x}$ can be found as follows:

$$\bar{x} = \frac{1+4+6+8+11}{5} = 6 \text{ years.}$$

The sample standard deviation *s* can be found as shown below:

$$s = \sqrt{\frac{(1-6)^2+(4-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5-1}} \approx 3.8 \text{ years}$$

When *n* is large, $\bar{x}$ approximates μ and the effect caused by subtracting 1 from *n* in the formula for *s* is very small. In such cases, the value of *s* approximates σ.

a. Obtain a population of pennies from your teacher and place them in a container. To obtain a sample of size 5 from this population, complete the following steps:

1. Draw one member of the population at random from the container.

2. Record its age.

3. Return it to the container and mix the population thoroughly.

4. Repeat Steps **1–3** until the desired sample size has been reached.

b. Determine the mean age $\bar{x}$ and standard deviation *s* of this sample. Record these values in a table with headings like those in Table **1**.

Table 1: Statistics for samples of five pennies

Sample Number	$\bar{x}$	*s*
1		
2		
3		
⋮		
10		

c. Repeat Parts **a** and **b** nine more times. **Note:** Save a copy of Table **1** for use in Activity **2**.

d. Create a frequency histogram of the means of your 10 samples.

e. Collect all the sample means from each group in your class. Create a frequency histogram of these sample means. **Note:** Save the class data for use in Exploration **2**.

Discussion 1

a. In Part **b** of Exploration **1**, why were you instructed to use s for the standard deviation rather than σ?

b. Compare the frequency histogram of your 10 sample means with the one for the class data.

c. Describe how you could use a frequency histogram to estimate the mean age of the population in Exploration **1**.

d. Which histogram should provide a better estimate of the population mean: your histogram for 10 sample means or the one for the class data?

Exploration 2

In this exploration, you use a simulation to investigate how sample size can affect the results of a sampling.

a. With the help of technology, you can examine many samples from the population in Exploration **1** in a relatively short time.

1. Use the simulation provided by your teacher to obtain 10 samples of size 20 from the population.

2. Determine the mean age and standard deviation of each sample and record these values in a table with headings like those in Table **1**. **Note:** Save this data for use in Activity **2**.

3. Create a frequency histogram of the means of your 10 samples.

4. Collect all the sample means from each group in your class. Create a frequency histogram of these sample means.

b. Repeat Part **a** for samples of size 40.

c. Using the class data for samples of size 5 from Exploration **1**, determine the mean and standard deviation for the sample means. Record these values in a table with headings like those in Table **2**.

Table 2: Sample sizes, mean, and standard deviation

Sample Size (n)	Mean of Sample Means	Standard Deviation of Sample Means
5		
20		
40		

d. Repeat Part **c** for sample sizes of 20 and 40.

e. Estimate the mean age of the population of pennies.

Discussion 2

a. Describe how the frequency histograms in Exploration **2** change as the sample size increases.

b. Describe how the means and standard deviations recorded in Table **2** change as the sample size increases.

c. Why would you expect an increase in sample size n to produce a decrease in the standard deviation of the sample means?

d.
1. How did you estimate the mean age μ for the population of pennies?

2. How accurate do you think your estimate is? Explain your response.

e.
1. Why should the mean of one large sample from a population be approximately the same as the mean of a large number of sample means from the same population?

2. Why should the standard deviation s of one large sample from a population be approximately the same as the standard deviation μ of the entire population?

Mathematics Note

The **sampling distribution of sample means** contains the means ($\bar{x}$) of *all* possible samples of size n from a population.

The mean of the sampling distribution of sample means, denoted by $\mu_{\bar{x}}$, equals the population mean μ.

The standard deviation of the sampling distribution, denoted by $\sigma_{\bar{x}}$, equals $\sigma/\sqrt{n}$, where σ is the population standard deviation and n is the sample size. When σ is unknown, the standard deviation of the sample (s) may be used as an estimate of σ.

For example, consider a jar containing three pennies, A, B, and C with ages 2 yr, 6 yr, and 6 yr, respectively. Table **3** shows the mean ages, in years, of all possible samples of size 2 that can be taken from this population, drawn one at a time with replacement.

Table 3: A sampling distribution of sample means

Sample	AA	AB	AC	BA	BB	BC	CA	CB	CC
$\bar{x}$	2	4	4	4	6	6	4	6	6

In this case,

$$\mu_{\bar{x}} = \frac{2+4+4+4+6+6+4+6+6}{9} \approx 4.66 \text{ years.}$$

Since $\mu_{\bar{x}} = \mu$, the mean also can calculated as follows: $(2 + 6 + 6)/3 \approx 4.66$ yr.

Using the formula for the standard deviation of a population, the standard deviation of all possible sample means is approximately 1.33 yr. This also can be calculated as shown below:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{1.89}{\sqrt{2}} \approx 1.33 \text{ yr.}$$

f. Obtain the actual mean and standard deviation of the population of pennies used in Explorations **1** and **2**. Use the formulas $\mu_{\bar{x}} = \breve{u}$ and $\sigma_{\bar{x}} = \underline{L}/\sqrt{n}$ to determine $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ for sample sizes of 5, 20, and 40. Compare your results with the values you recorded in Table **2**.

g. How does sample size affect the spread of all the possible sample means about the population mean?

h. Taking large samples from a population can be expensive. In what situations might it be worth the cost of collecting a very large sample?

Assignment

1.1 The following table shows the ages, in years, of a sample of pennies found in a piggy bank:

28	8	6	5	1	1	3	4	2	2
14	1	2	3	6	6	12	7	5	1
12	7	1	6	1	2	1	8	1	8
16	10	5	2	3	10	16	3	4	7
3	5	2	4	8	3	4	1	4	5

a. Determine the mean age $\bar{x}$ of these pennies.

b. Determine the sample standard deviation *s*.

c. Estimate the mean age μ for the population of pennies from which this sample came.

1.2 To determine the mean annual income in a large city, a research group analyzed 1000 random samples of 40 adults from this population. The results of the study are shown in the following table.

Sample Mean (to nearest $1000)	Frequency
18,000	1
19,000	11
20,000	45
21,000	111
22,000	184
23,000	220
24,000	197
25,000	131
26,000	66
27,000	24
28,000	6
29,000	2
30,000	1
31,000	1

a. Use the mean of these sample means to estimate the mean $\mu_{\bar{x}}$ of all possible sample means (to the nearest thousand dollars) using samples of size 40.

b. Use the standard deviation of these sample means to estimate the standard deviation $\sigma_{\bar{x}}$ of all possible sample means using samples of size 40.

c. Estimate the mean μ and standard deviation σ for the population from which these samples came.

d. How accurate do you think your estimate is for μ? Explain your response.

e. Using the given data, estimate the probability that a random sample of 40 adults from this city has an average income that is:

1. more than $29,000

2. less than $21,000

3. between $21,000 and $29,000, inclusive.

1.3 When trying to estimate the average income of high school graduates in your state, Andreas claims that using a sample size of 240 instead of 120 would reduce the standard deviation of all possible sample means by half. Likewise, using a sample size of 480 instead of 240 would reduce $\sigma_{\bar{x}}$ again by half. Defend or refute Andreas' claim.

* * * * *

1.4 At Lincoln High School, many students take the Scholastic Aptitude Test (SAT). The following table shows the scores on the mathematics portion of the SAT for the students in one classroom.

410	770	430	420	400	780	440	420
610	630	400	440	430	500	430	450
680	500	720	450	520	470	440	740
440	580	550	590	770	400	500	610

a. Determine the mean score $\bar{x}$ (to the nearest whole number).

b. Determine the sample standard deviation s.

c. Estimate the mean score μ for the entire population of Lincoln High School students who took the exam.

d. Why might you hesitate to use the results of this sample to estimate the mean for the entire population?

1.5 The table below shows the ages of a sample of high school students in a community.

Age	**Frequency**
13	100
14	150
15	120
16	93
17	157
18	82
19	51

a. Determine the mean age, to the nearest year, of students in this sample.

b. Determine the standard deviation of this sample.

c. Estimate the mean μ for the population from which this sample came.

d. Estimate the proportion of students in the community that are:

1. older than 16

2. younger than 17

3. between 15 and 17, inclusive.

e. How confident are you that the estimates made from this sample are accurate? Explain your response.

1.6 The following table shows the numbers of persons per household for a sample of households taken in a large city:

Number in Household	Frequency
1	15
2	20
3	37
4	23
5	14
6	4
7	2

a. Determine the mean number of persons per household for this sample.

b. Determine the standard deviation of this sample.

c. Estimate the mean number of persons per household for the city from which this sample came.

d. Estimate the proportion of households in this city that have:

1. more than 4 persons

2. less than 4 persons.

e. How confident are you that the estimates made from this sample are accurate? Explain your response.

* * * * * * * * * *

Activity 2

In Activity **1**, you examined how sample means can vary by taking many samples of different sizes. In the real world, however, researchers typically do not collect a large number of samples. In fact, they often take only one.

In this activity, you investigate the probability that the mean from a single sample accurately estimates the population mean.

Exploration

a. Use the data you compiled in Activity **1** for samples of size 5 to complete a table with headings like those in Table **4**. To estimate $\sigma_{\bar{x}}$, use the value of *s* for each sample to approximate σ, then use the formula $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Table 4: Statistics for samples of five pennies

Sample Number	$\bar{x}$	*s*	Estimate of $\sigma_{\bar{x}}$
1			
2			
3			
⋮			
10			

b. **1.** For each sample in Table **4**, find the interval $[\bar{x} - \sigma_{\bar{x}}, \bar{x} + \sigma_{\bar{x}}]$.

2. Graph each interval above a number line as a line segment with the midpoint $\bar{x}$ indicated as shown in Figure **1** below.

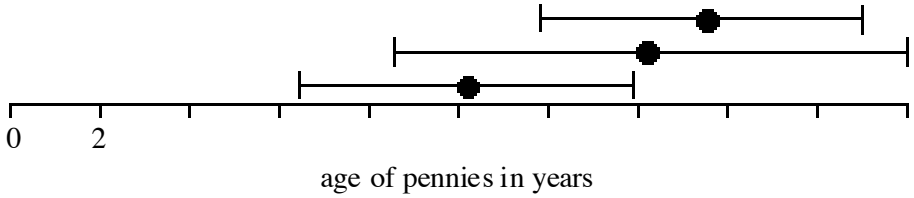

Figure 1: Intervals for three samples of pennies

c. **1.** Obtain the mean age μ of all the pennies in the population used in Activity **1**.

2. Draw a line on your graph to represent the population mean, as shown in Figure **2**:

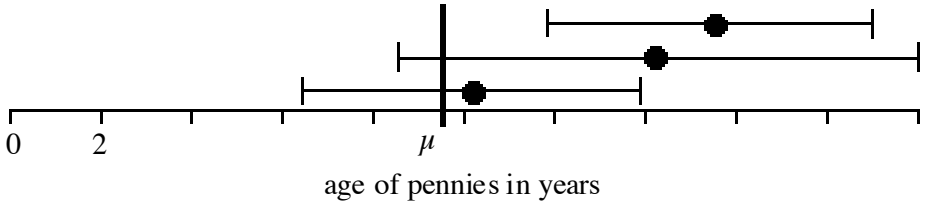

Figure 2: Graph of intervals with line at μ

d. Determine the percentage of intervals that contains the population mean. In Figure **2**, for example, approximately 67% of the intervals contain the population mean. Record the result in a table with headings like those in Table **5**.

Table 5: Percentage of intervals containing μ 7

Samples sizes n	Interval	Percentage that Contained μ
5	$[\bar{x} - \sigma_{\bar{x}}, \bar{x} + \sigma_{\bar{x}}]$	
20	$[\bar{x} - \sigma_{\bar{x}}, \bar{x} + \sigma_{\bar{x}}]$	
40	$[\bar{x} - \sigma_{\bar{x}}, \bar{x} + \sigma_{\bar{x}}]$	
5	$[\bar{x} - 2\sigma_{\bar{x}}, \bar{x} + 2\sigma_{\bar{x}}]$	
20	$[\bar{x} - 2\sigma_{\bar{x}}, \bar{x} + 2\sigma_{\bar{x}}]$	
40	$[\bar{x} - 2\sigma_{\bar{x}}, \bar{x} + 2\sigma_{\bar{x}}]$	
5	$[\bar{x} - 3\sigma_{\bar{x}}, \bar{x} + 3\sigma_{\bar{x}}]$	
20	$[\bar{x} - 3\sigma_{\bar{x}}, \bar{x} + 3\sigma_{\bar{x}}]$	
40	$[\bar{x} - 3\sigma_{\bar{x}}, \bar{x} + 3\sigma_{\bar{x}}]$	

e. Repeat Parts **a–d** for sample sizes of 20 and 40 pennies.

f. For each sample size, determine the percentage of intervals of $[\bar{x} - 2\sigma_{\bar{x}}, \bar{x} + 2\sigma_{\bar{x}}]$ that contain the population mean. Record your results in your copy of Table **5**.

g. For each sample size, determine the percentage of intervals of $[\bar{x} - 3\sigma_{\bar{x}}, \bar{x} + 3\sigma_{\bar{x}}]$ that contain the population mean. Record your results in your copy of Table **5**.

Discussion

a. What appears to be the relationship between the number of standard deviations used to create the interval and the percentage of intervals that contain the population mean? Explain your response.

Mathematics Note

The **central limit theorem** states that, regardless of the population, as the sample size increases, the sampling distribution of sample means approaches a **normal distribution**.

Figure **3** below shows a graph of a normal distribution. The curve that describes the shape of the graph is the **normal curve**. As in all continuous probability distributions, the total area between the *x*-axis and a normal curve is 1. Approximately 68% of this area falls within 1 standard deviation of the mean, 95% within 2 standard deviations of the mean, and 99.7% within 3 standard deviations of the mean. This is the **68–95–99.7 rule**.

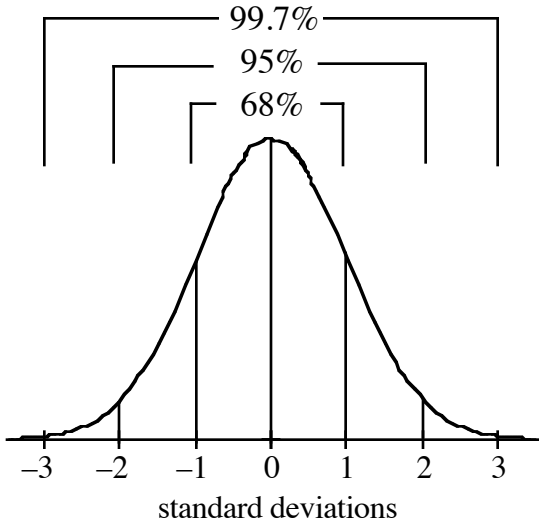

Figure 3: A normal curve and the 68–95–99.7 rule

As a rule of thumb, samples of size $n \geq 30$ are large enough to assume that the sampling distribution of sample means approaches a normal distribution.

b. How do your results in the exploration compare with the percentages predicted by the 68–95–99.7 rule?

c. As the sample size changes, what happens to the sizes of the corresponding intervals? Explain your response.

Mathematics Note

A **confidence interval** for a population mean μ is an interval of numbers in which you would expect to find the value of μ. The 68–95–99.7 rule implies the following:

- For approximately 68% of all sample means $\bar{x}$, the confidence interval $[\bar{x} - \sigma_{\bar{x}}, \bar{x} + \sigma_{\bar{x}}]$ contains the population mean μ.
- For approximately 95% of all sample means $\bar{x}$, the confidence interval $[\bar{x} - 2\sigma_{\bar{x}}, \bar{x} + 2\sigma_{\bar{x}}]$ contains the population mean μ.
- For approximately 99.7% of all sample means $\bar{x}$, the confidence interval $[\bar{x} - 3\sigma_{\bar{x}}, \bar{x} + 3\sigma_{\bar{x}}]$ contains the population mean μ.

For example, consider a sample of 40 pennies with a mean age $\bar{x}$ of 9 yr and a standard deviation *s* of 3 yr. In this case, $\sigma_{\bar{x}}$ can be estimated by $3/\sqrt{40}$=0.47, Using the 68–95–99.7 rule, you can be 68% confident that the population mean μ is in the interval $[9 - 0.47, 9 + 0.47]$, 95% confident μ is in the interval $[9 - 2(0.47), 9 + 2(0.47)]$, and 99.7% confident μ is in the interval $[9 - 3(0.47), 9 + 3(0.47)]$.

d. What is meant by a "95% confidence interval?"

e. How often would you expect a 95% confidence interval to not contain the population mean? Explain your response.

f. Suppose that a researcher surveys a random sample of 100 high school graduates and calculates their mean annual income. What other information is necessary to give a reasonable estimate of how close the sample mean is to the population mean?

Assignment

2.1 Imagine that you have taken 20 samples of size 30 from a population of pennies. For each sample, you determine the mean age and estimate the corresponding 95% confidence interval. Approximately how many of these intervals do you think will include the population mean?

2.2 For a random sample of 100 eggs, the mean mass was 67 g, with a standard deviation of 45 g.

a. Estimate the standard deviation $\sigma_{\bar{x}}$ of all sample means for $n = 100$.

b. Construct the interval $[\bar{x} - \sigma_{\bar{x}}, \bar{x} + \sigma_{\bar{x}}]$.

c. How confident should you be that the population mean falls in the interval in Part **b**? Justify your response.

2.3 The Shiny Bright Company manufactures light bulbs. To determine the mean life expectancy of their 60-watt bulbs, the company sampled 1000 bulbs. They found that the mean life expectancy of the sample was 827 hr, with a sample standard deviation of 424 hr.

a. Estimate $\sigma_{\bar{x}}$ for these 60-watt bulbs.

b. Write a conclusion about the mean life expectancy at each of the following:

1. a 68% confidence level

2. a 95% confidence level

3. a 99.7% confidence level.

2.4 Ken sampled 20 bags of a certain brand of candy. He discovered that the mean mass was 52 g, with a standard deviation of 4 g.

a. Write a conclusion about the mean mass of all bags of this candy.

b. Considering the sample size, how sure should you be of your conclusion in Part **a**?

c. How could you make the results of this experiment more conclusive?

2.5 As a quality control specialist, you have been asked to determine the mean volume of soft drinks packaged in 2-L bottles. The table below shows the data (in liters) collected from a sample of 30 bottles. Use this information to write a conclusion about the mean volume of soft drinks in 2-L bottles.

1.90	1.95	1.97	1.98	2.00	1.94
1.91	2.04	2.02	1.99	1.97	1.96
1.92	2.00	1.97	1.96	1.94	1.94
1.93	2.05	2.04	2.00	1.98	1.97
2.00	1.98	1.97	1.94	2.01	1.99

2.6 A random sample of the savings account balances of 500 people banking at Third National Bank resulted in $\bar{x} = \$720$ and $s = \$217$.

a. Estimate the standard deviation $\sigma_{\bar{x}}$ of all sample means for this sample size.

b. Construct the interval $[\bar{x} - 2\sigma_{\bar{x}}, \bar{x} + 2\sigma_{\bar{x}}]$.

c. How confident are you that the population mean falls in the interval from Part **b**?

* * * * *

2.7 The Rolling-On Company manufactures automobile tires. To determine the mean life expectancy of their four-ply tires, the company sampled 800 tires. The mean life expectancy of this sample was 197,124 km, with a standard deviation of 79,678 km. Assuming that the lives of these tires are normally distributed, write a conclusion about the mean life expectancy at:

a. a 68% confidence level

b. a 95% confidence level

c. a 99.7% confidence level.

2.8 To determine the mean life expectancy of their tennis shoes, the Brand 10 company surveyed a random sample of 32 people. The results of the survey are shown in the table below.

Life of Tennis Shoes (in months)							
35	39	6	41	23	61	45	18
16	3	13	8	21	13	15	27
27	40	26	18	22	23	6	40
25	42	17	41	38	29	13	12

Assuming that the lives of these tires are normally distributed, write a conclusion about the mean life of Brand 10 shoes**,** using:

a. a 68% confidence level

b. a 95% confidence level

c. a 99.7% confidence level.

2.9 After obtaining a sample of size 1000, a researcher constructed the following 95% confidence interval: [1250, 1420]. Use this information to identify $\bar{x}$ and $\sigma_{\bar{x}}$.

* * * * * * * * * *

Activity 3

Researchers, pollsters, and statisticians sample populations in order to gain information about the population. Since statistics gathered by sampling only provide estimates of the population parameters, errors are possible. In this activity, you examine how to quantify and minimize the risk of making mistakes.

Exploration 1

Since finding the actual mean of a large population can be difficult, it is often necessary to use statistics to make predictions about a population. In this exploration, you investigate the use of confidence intervals in estimating a population mean.

Imagine that your school business club is completing a survey of the average incomes of high school graduates in your state. The club has budgeted enough money to conduct a telephone survey of 30 former students. The information gathered in the survey is shown in Table **6**:

Table 6: A sample of 30 incomes (in dollars)

4000	17,400	10,900	15,600	8000
9800	10,000	19,400	14,400	42,000
32,000	4500	46,000	90,400	24,400
112,000	19,600	25,400	7000	16,000
16,800	23,600	26,200	19,600	28,400
14,800	19,400	13,500	18,200	6600

a. Construct 68%, 95%, and 99.7% confidence intervals of the sample mean of the data. (Estimate $\sigma_{\bar{x}}$ using the sample standard deviation s.)

b. Graph the confidence intervals above a number line as in Activity **2**. **Note**: Save this graph for use in Exploration **2**.

c. Make a statement about where you would expect to find the actual mean income of high school graduates in your state at a:

1. 68% confidence level

2. 95% confidence level

3. 99% confidence level.

Discussion 1

a. Which confidence interval is most likely to include the actual average income of high school graduates in your state? Explain your response.

b. Which confidence level has the least range of estimates of the actual average income of the graduates? Explain your response.

c. What is the relationship between interval size and the probability of making an error?

d. How could the confidence intervals be narrowed without increasing the chances of making an error? Explain your response.

e. Describe some of the advantages and disadvantages of using:

1. a 99.7% confidence level

2. a 95% confidence level

3. a 68% confidence level.

f. Explain why it is just as important to use a random sample to estimate a confidence interval as it is to use a random sample to estimate the population mean.

Exploration 2

Roberto is writing an article for the school newspaper. Based on the study by the school business club, he wants to claim that the average income for high school graduates in your state is $36,000. However, the editor of the newspaper disagrees. She believes that a better value for the average income is $16,000. Who is right? In this activity, you use confidence intervals to test the claims of Roberto and his editor.

Mathematics Note

Statisticians often make hypotheses or claims about the parameters of a population, then use sampling techniques to test their claims. If a researcher assumes that a population parameter has a specific value, then a hypothesis can be formed about the consequences of that assumption. In statistical analysis, there are two types of hypotheses.

A **null hypothesis (H_0)** is a statement about one or more parameters. The **alternative hypothesis (H_a)** is the statement that must be true if the null hypothesis is false. The null hypothesis usually involves a claim of no difference or no relationship. In many situations, the null hypothesis and alternative hypothesis are negations of each other, but this is not necessarily the case.

For example, if a researcher wants to test the claim that the mean income of the population is $25,000, then the null hypothesis is that the mean income of the population equals $25,000. The alternative hypothesis is that the mean income of the population does not equal $25,000. Symbolically, this can be represented as shown below:

$$H_0: \mu = \$25{,}000$$

$$H_a: \mu \neq \$25{,}000$$

a. State the null and alternative hypotheses for Roberto's claims about the average income for high school graduates.

b. On your graph of the 68%, 95%, and 99.7% confidence intervals of the income data from Table **6** (from Part **b** of Exploration **1**), draw a vertical line to represent Roberto's claim about the average income.

c. Record the confidence intervals that include Roberto's predicted mean.

Mathematics Note

A **hypothesis test** may consist of the following steps.

- State null and alternative hypotheses about a parameter of a population.
- If the null hypothesis is true, predict what this implies about a sample of the population.
- Take a sample of the population and compare the results with your prediction.
- If the results are inconsistent with the prediction, then you can conclude, with some level of certainty, that the null hypothesis is false and, therefore, reject it.
- If the results are consistent with the prediction, you fail to reject the null hypothesis. The failure to reject the null hypothesis does not guarantee that the null hypothesis is true, only suggests that it might be true.

For example, to test the claim that the mean income of high school graduates is \$25,000 at the 95% confidence level, you would:

- State the null hypothesis: $H_0: \mu = \$25{,}000$.
- Sample the population, calculate the mean H_0 of the sample, and construct the 95% confidence interval around $\bar{x}$.
- If \$25,000 is not in the 95% confidence interval, then you would reject the null hypothesis and accept the alternative hypothesis $H_a: \mu \neq \$25{,}000$

. If \$25,000 is in the 95% confidence interval, then you would fail to reject the null hypothesis. You can only conclude that H_0 may or may not be true.

d. Determine whether you would reject or fail to reject the hypotheses from Part **a** at the 68%, 95%, and 99.7% confidence levels.

e. Repeat Parts **a–d** for the editor's claim about the average income for high school graduates.

Discussion 2

a. Given your results in Exploration **2**, what can you conclude about the claims of Roberto and his editor? Explain your response.

b. What is the difference between "failing to reject" a null hypothesis and "accepting" a null hypothesis?

c. Why does the rejection of the null hypothesis result in the acceptance of the alternative hypothesis?

d. Does failing to reject a null hypothesis prove that it is true? Explain your response.

e. Does rejecting a null hypothesis prove that it is false?

f. Scientists usually require at least a 95% confidence level to reject a hypothesis. Using this standard, state conclusions about the editor's and Roberto's hypotheses.

g. After Roberto discovers that his hypothesis cannot be rejected at a 99.7% confidence level, he exclaims, "This shows that there is a 99% chance that my hypothesis is right." Explain what is wrong with Roberto's reasoning.

Assignment

3.1 To test her null hypothesis, Roberto's editor conducted a survey of 30 acquaintances who recently graduated from high school. Their incomes (in dollars) are shown in the table below.

6200	9400	22,300	6200	18,000	38,000
17,100	21,500	17,300	27,500	11,900	23,200
19,000	11,000	13,200	16,500	13,800	13,400
15,500	8700	33,000	34,000	16,200	9500
14,000	11,000	32,000	14,700	6400	3400

a. The editor claimed that the mean income is $16,000. Use the data in the table above to test her null hypothesis at a confidence level of your choice.

b. Based on your test in Part **a**, state your conclusions.

c. Does this test prove that the editor's claim is correct? Explain your response.

d. Roberto complains that the editor's sample is not representative of the high school graduates in their state. Describe some possible sources of bias in the sample.

3.2 Roberto decides to survey his own sample of 30 graduates. He telephones the class presidents from each of the past 30 yr and records their incomes in the table below.

74,200	29,500	17,000	74,000	44,500	31,000
26,000	19,000	36,000	12,500	32,000	21,000
6100	27,700	34,000	29,000	72,100	30,500
93,000	12,000	19,000	31,000	26,200	33,200
34,000	31,000	23,400	46,300	35,400	6200

a. Recall that Roberto claimed that the mean income is $36,000. Use the data in the table above to test his claim at a 95% confidence level.

b. Based on your test in Part **a**, state your conclusions.

c. What possible sources of bias are there in Roberto's sample?

3.3 **a.** Make a claim about the mean income of graduates in your state.

b. Write null and alternative hypotheses for your claim.

c. Describe what group you would sample to test your hypothesis and how you would collect your information.

d. What level of confidence would you choose to test your hypothesis? Defend your choice.

3.4 In its advertisements, the Shiny Bright Company claims that its 60-watt bulbs have an average life expectancy of 1250 hr. They based this conclusion on a sample of 1000 light bulbs in which the mean life span was 827 hr, with a standard deviation of 424 hr.

According to the company, their advertised value is within 1 standard deviation of the sample mean. Therefore, their claim cannot be rejected. What is wrong with their logic?

3.5 In tests conducted by outside experts, Shiny Bright's "Best Bulb" had an average life expectancy of 2000 hr, with a standard deviation of 300 hr. The Hi-Glow Company claims their "Long-Life" bulbs are better because they last even longer. Both companies decide to test the hypothesis that the mean life expectancy of Long-Life bulbs is 2000 hr.

Using a random sample of 100 Long-Life bulbs, Shiny Bright found a mean life span of 2040 hr, with a standard deviation of 470 hr. The firm concluded that the mean life of Long-Life bulbs is the same as that of their Best Bulbs.

Hi-Glow tested 10,000 Long-Life bulbs and found a mean life span of 2010 hr, with a standard deviation of 400 hr. They concluded that the mean life span of their bulbs is not the same as that of Shiny Bright's bulbs. Perform a hypothesis test at the 95% confidence level to determine which company you think is right.

* * * * *

3.6 Cereal boxes usually display the disclaimer that the boxes are filled by weight, not by volume, and that some settling may occur during shipping. A high school statistics class wants to see if this is true or if the actual mass of cereal is significantly less than is claimed. To test a manufacturer's claims, students randomly sample forty 397-g boxes of the same brand of cereal. The table below shows the mass of each box of cereal, rounded to the nearest gram:

402	397	404	384	390	395	397	385	392	399
380	390	408	403	389	389	393	381	402	401
383	403	383	392	400	392	395	395	406	396
408	383	381	390	401	385	382	404	409	387

a. State the null and alternative hypotheses for this experiment.

b. Test the hypothesis at the 95% confidence level.

c. Decide whether to reject or fail to reject the null hypothesis. Justify your reasoning.

d. Explain what conclusion the class should reach about the net mass of the boxes of cereal.

3.7 Loretta has bowled in a league for years. Last season, her average score per game was 152. This season, she bought a new ball. Her weekly average scores for this season with the new ball are shown below.

194	153	171	199	141	146
151	190	171	168	150	128
166	166	160	183	141	210
169	172	132	126	195	191
127	155	204	191	129	170
180	150	155	167	192	168

Has Loretta's average score changed significantly?

a. Use the information in the table to construct a 68% confidence interval.

b. Test the claim that Loretta's average has changed significantly from last year's average of 152 at the 68% confidence level.

c. Based on your hypothesis test, determine whether or not Loretta's bowling average has changed significantly since she started using the new ball.

* * * * * * * * * *

Research Project

Design and implement a plan to determine the average age of a home in your community. Your report should include at least the following information:

a. A claim about the mean age of homes in your community.

b. The null and alternative hypotheses for your claim.

c. A description of your sample and how you collected the data.

d. The level of confidence you chose to test your hypothesis.

Summary Assessment

Whether they recognize this fact or not, anglers often use sampling to judge the waters in which they fish. For example, after catching several big fish, an angler may conclude that a lake contains a healthy population of large fish. On the other hand, anglers who have little success may grumble that, "There aren't many big fish in this lake."

In such informal evaluations, anglers rarely use confidence intervals or take the variability of their samples into account. Wildlife managers, however, need a more reliable method to describe a fish population.

Suppose you are a wildlife manager responsible for estimating the mean size of the fish in a "lake" provided by your teacher. You must evaluate your estimate by taking a sample of the fish and performing a hypothesis test.

1. a. After examining the fish in the lake, make null and alternative hypotheses about their average length.

   b. Select a confidence level with which to test your hypothesis.

   c. Sample the population using an appropriate sample size.

2. **a.** Construct the appropriate confidence interval and use it to test your hypothesis.

   b. Write a conclusion based on your test.

3. If you had the opportunity to sample the fish population again, what would you change about your technique?

Module Summary

- The **standard deviation** of a population, denoted by σ, is calculated using the formula below, where Type equation here. represent all the individual values in the population, μ represents their mean, and *N* represents the population size:

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N}}$$

- The standard deviation of a sample is the **sample standard deviation**, denoted by *s*. The formula for calculating *s* is:

$$s = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}}$$

  where $x_1, x_2, \cdots, x_n$ represent the individual values in the sample, $\bar{x}$ represents the sample mean, and *n* represents the sample size.

- The **sampling distribution of sample means** contains the means ($\bar{x}$) of *all* possible samples of size *n* from a population.
- The mean of the sampling distribution of sample means, denoted by $\mu_{\bar{x}}$, equals the population mean μ.
- The standard deviation of the sampling distribution, denoted by $\sigma_{\bar{x}}$, equals $\sigma/\sqrt{n}$, where σ is the population standard deviation and *n* is the sample size. When σ is unknown, the standard deviation of the sample (*s*) may be used as an estimate of σ.
- The **central limit theorem** states that, regardless of the population, as the sample size increases, the sampling distribution of sample means approaches a normal distribution. As a rule of thumb, samples of size $n \geq 30$ are large enough to assume that the sampling distribution of sample means approaches a normal distribution.
- The curve that describes the shape of a normal distribution is the **normal curve**. As in all continuous probability distributions, the total area between the *x*-axis and a normal curve is 1. Approximately 68% of this area falls within 1 standard deviation of the mean, 95% within 2 standard deviations of the mean, and 99.7% within 3 standard deviations of the mean. This is the **68–95–99.7 rule**.

- A **confidence interval** for a sample mean is an interval of numbers in which you would expect to find the population mean. The 68–95–99.7 rule for normal distributions implies the following:
- For approximately 68% of all sample means $\bar{x}$, the confidence interval $[\bar{x} - \sigma_{\bar{x}}, \bar{x} + \sigma_{\bar{x}}]$ contains the population mean μ.
- For approximately 95% of all sample means $\bar{x}$, the confidence interval $[\bar{x} - 2\sigma_{\bar{x}}, \bar{x} + 2\sigma_{\bar{x}}]$ contains the population mean μ.
- For approximately 99.7% of all sample means $\bar{x}$, the confidence interval $[\bar{x} - 3\sigma_{\bar{x}}, \bar{x} + 3\sigma_{\bar{x}}]$ contains the population mean μ.
- In statistical analysis, there are two types of hypotheses. A **null hypothesis (H_0)** is a statement about one or more parameters. The **alternative hypothesis (H_a)** is the statement that must be true if the null hypothesis is false. The null hypothesis usually involves a claim of no difference or no relationship. In many situations, the null hypothesis and alternative hypothesis are negations of each other, but this is not necessarily the case.
- A **hypothesis test** may consist of the following steps.
- State null and alternative hypotheses about a parameter of a population.
- If the null hypothesis is true, predict what this implies about a sample of the population.
- Take a sample of the population and compare the results with your prediction.
- If the results are inconsistent with the prediction, then you can conclude, with some level of certainty, that the null hypothesis is false and, therefore, reject it.
- If the results are consistent with the prediction, you fail to reject the null hypothesis. The failure to reject the null hypothesis does not guarantee that the null hypothesis is true, only suggests that it might be true.

Selected References

Neter, J., W. Wasserman, and G. A. Whitmore. *Applied Statistics*. Boston: Allyn and Bacon, 1993.

Billingsley, P., and D. V. Huntsberger. *Elements of Statistical Inference*. Boston: Allyn and Bacon, 1981.