Section V: Inference

Investigation 15

How Many Can You Expect to Have a Job?

Sampling Distribution

Overview

This investigation explores the concept of a sampling distribution. Students will use simulation to model the sampling distribution of the sample number of successes by drawing slips of paper from a bag where 60% of the slips represent 16- to 24-year-olds with a job (success). Investigating the sampling distribution leads to the key idea that the mean of the sample number of successes will approximately equal the population expected number of successes.

This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report.* The four components are formulate a statistical question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level C activity.

This investigation is based on lessons from *Probability Models* by Patrick Hopfensperger, Henry Kranendonk, and Richard Scheaffer, available as a free download at *www.amstat.org/ASA/Education/K-12-Educators*.

Instructional Plan

Brief Overview

» Read the scenario about the percent of teens who have a job.

» Discuss possible results of taking a random sample from a population with 60% of teens who have a job.

» Take a random sample of size 20 from a bag of slips of paper with 60% of the slips indicating teens have a job.

» Continue to take random samples to build a sampling distribution of the sample number of successes with sample sizes of 20.

» Describe the distribution and conclude the mean of the sampling distribution equals the expected number of successes (have a job).

Scenario

Many high-school and college students have a job after school or on weekends. Many work in a fast-food restaurant or as a clerk in a store.

Do you have a job? What kind of work do you do? Do you like your job?

If you don't have a job, do you think this is unusual?

The youth labor force of 16- to 24-year-olds working or actively looking for work increases sharply between April and July each year. During these months, large numbers of high-school and college students search for or take summer jobs, and many graduates enter the

labor market to look for or begin permanent employment.

According to the Bureau of Labor Statistics, the labor force participation rate for all youth was 60.6% in July of 2017. (The labor force participation rate is the proportion of the civilian noninstitutional population that is working or looking and available for work. The civilian noninstitutional is made up of people 16 years and over residing in the US and not inmates or on active military duty.) *Source: www.bls.gov/news.release/youth.nr0.htm*

Learning Goal

Investigate the sampling distribution of the sample number of successes through simulation.

Mathematical Practices Through a Statistical Lens

MP8. Look for and express regularity in repeated reasoning.

Statistically proficient students maintain oversight of the process, attend to the details, and continually evaluate the reasonableness of their results as they are carrying out the statistical problem-solving process.

Materials

Student worksheets are available at *www.statisticsteacher.org/statistics-teacher-publications/focus.*

» Cloth or paper bag

» Copies of the template run on stock paper; both sheets of the template for each group

» Student Worksheet 15.1 Data Collecting

» Student Worksheet 15.2 Template

» Exit Ticket

Estimated Time

Two 50-minute class periods. One period to collect the data and build a sampling distribution. A second period to analyze and draw conclusions pertaining to the sampling distribution.

Pre-Knowledge

» Students should be able to:

» Design and conduct a simulation

» Construct a dot plot

» Find the mean and standard deviation of a distribution using technology

Formulate a Statistical Question

Hand out Student Worksheet 15.1 Data Collecting. Have students work with a partner to answer questions 1 to 4.

Assume the labor force participation rate for all 16- to 24-year-olds in your city is 60%.

1. If your class is to select a random sample of 20 16- to 24-year-olds from the community, how many of the randomly selected 16- to 24-year-olds would you expect to be part of the labor force?

Answer: *Approximately 0.6(20), or 12 16- to 24-year-olds.*

2. If all the students in your class would each take a random sample of 20 16- to 24-year-olds from the city and ask each selected 16- to 24-year-old if they were working, do you think each class member would get the same number of 16- to 24-year-olds in the sample who are part of the labor force?

Answer: *It is unlikely they would all get exactly same result.*

3. Do you think it would be likely to find survey results of 13 out of the 20 16- to 24-year-olds, or 65%, reporting they have a job? Explain your answer.

Possible answer: *This result is likely; 13 is not much larger than the expected value of 12.*

4. Do you think it would be likely to find survey results of fewer than seven of the 20 16- to 24-year-olds, or 35% or less, reporting they have a job? Explain your answer.

Possible answer: *It is unlikely to get this result; seven is much smaller than the expected value of 12.*

Formulate a Statistical Question

Discuss the answers to questions 1 to 4. When discussing Question 2, encourage your students to begin to think about what they would expect for results. This leads into questions 3 and 4, in which the objective is to help students get an idea of what they think are likely and unlikely results. Consider having students give an interval of likely results.

Explain that the next step is to take many random samples of 20 slips of paper from a bag. Next, the students will investigate what results are likely and whether any patterns develop as they model the behavior of the results by building a sampling distribution of the sample number of successes (have a job).

Ask your students to consider the statistical question: "How many 16- to 24-year-olds out of 20 could have jobs if random samples of size 20 are taken from a population in which 60% of 16- to 24-year-olds have jobs?"

Collect Appropriate Data

Examine the template (Student Worksheet 15.2—two sheets each with 100 squares numbered 1 to 200) provided for this investigation. Note that the slips numbered from 1 to 120 are labeled with the word Job and those numbered 121 to 200 are labeled No Job. Cut out the slips and place the slips in a bag or container. Thoroughly mix the slips.

Tell your students the 200 slips of paper represent a population of 16- to 24-year-olds.

Explain that we are going to take a random sample of 20 slips of paper and count the number of slips that have the word Job written on them.

One way to collect this sample is to go around the room and have your students reach in the bag and select a slip. Note what the slip says. Continue selecting a slip until a sample of 20 slips has been selected. Count the number of slips that have the word Job written on them.

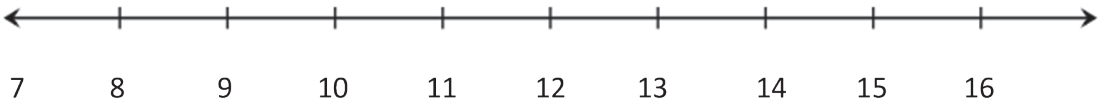

Figure 15.1: Number with a job from sample of 20

Ask your students what the results of the random sample of 20 slips represent.

Answer: *The results represent the number of 16- to 24-year-olds who said they had a job based on a random sample of 20 from a population of 16- to 24-year-olds in which 60% have a job.*

Ask your students if they will get the same number of slips with Job written on them if another random sample of 20 slips is selected?

Answer: *It is unlikely they would get the same result.*

Draw another random sample of 20 from the bag and find the number of 16- to 24-year-olds with a job. Again, ask your students what the results represent.

On the board or on poster paper, draw a number line like in Figure 15.1.

Give each group of students a bag and the template (Student Worksheet 15.2). Ask your students to cut out the slips of paper and place the slips in the bag. Have them thoroughly mix the slips.

Ask each group to take a random sample of 20 slips and record the number of successes (number that said Job) on the class dot plot.

Ask your students to repeat the simulation until the class has completed at least 50 trials.

Analyze the Data

Once students have completed their trials and reported their sample results on the class dot plot, point out to your students that the dot plot is an example of an approximate sampling distribution. Define a sampling distribution of the sample number of successes (students with a job) as a distribution of the sample number of successes from all possible samples of the same size from a population with a known proportion of successes (in this case, 60%).

Note: The total number of all possible samples of size 20 from 200 slips is ${}_{200}C_{20} = 1.61 \times 10^{27}$. Since it is not practical to create the sampling distribution of all possible samples, we will create an approximate sampling distribution.

Ask your students to answer questions 5 to 7. Discuss the student answers.

5. What should the title of our graph be? What did we graph?

Possible answer: *Simulated sampling distribution of the sample number of successes based on sample size = 20.*

6. Estimate the mean and standard deviation of the sampling distribution.

Possible answer: *Mean is approximately 11.5 and the standard deviation is approximately 1.8. (These are values for the example in Figure 15.2; the class data will vary from this example but be approximately the same.)*

7. Describe the shape of the sampling distribution.

Possible answers: *For this example in Figure 15.2, the shape is approximately symmetric and mound shaped.*

The sample results (Figure 15.2) are based on 68 student responses.

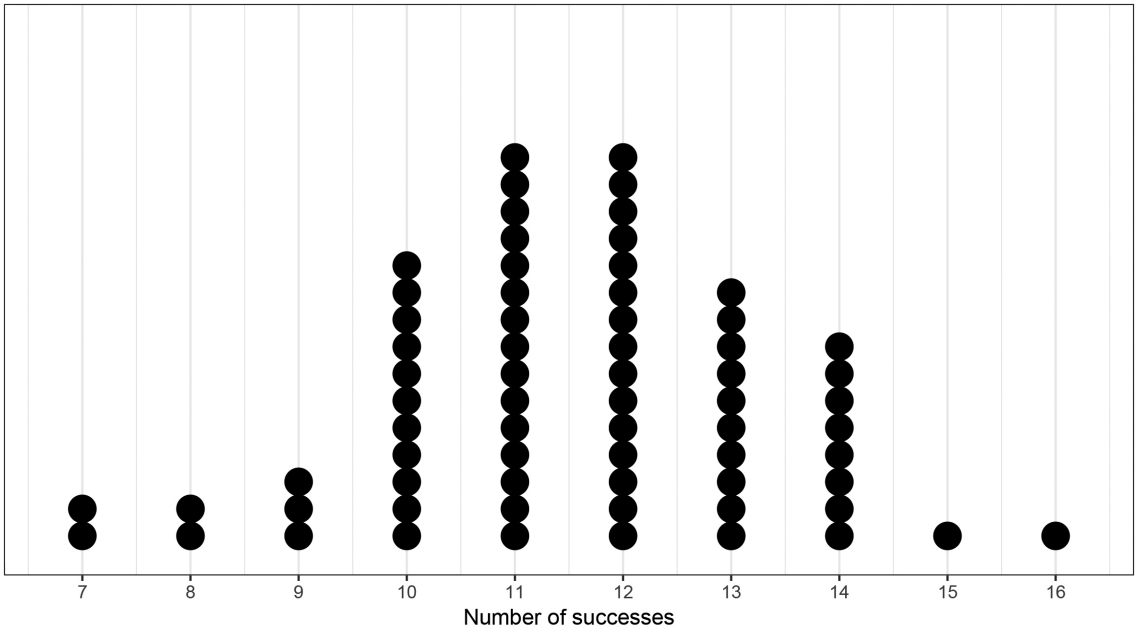

Figure 15.2: Dot plot of the results of 68 student responses

Ask your students to answer questions 8 to 11 on the worksheet.

8. Are you surprised the center of the distribution is close to 12?

Possible answer: *No, since the population proportion is 60%, this distribution should center around 0.60(20), or 12.*

9. If you took another random sample of 20 and found 10 said Job, would you call this a likely result? Explain.

Answer: *Ten is a likely outcome because 10 is close to the expected number of 12.*

10. If you took another random sample of 20 and found 15 said Job, would you call this a likely result? Explain.

Answer: *Fifteen is not a likely result. In our simulation, 15 or more was reported only twice out of 68 trials.*

11. If you took one more random sample of 20, give an interval you think would constitute a likely result? Explain.

Answer: *Between 9 and 14. This interval contained all the outcomes, except for six.*

Note: This answer is based on the dot plot shown in Figure 15.2.

Now explain that we want to change the *number* of successes to the *proportion* of students who have a job.

Add a second number line underneath the class dot plot, as shown in Figure 15.3.

Ask your students to change the number of successes to the proportion of students with a job from a sample of 20.

Answer: *Figure 15.3*

Ask your students to answer questions 12 and 13 on the worksheet.

12. Estimate the mean of the sample proportions.

Answer: *The mean will be approximately 0.57, or close to 0.60.*

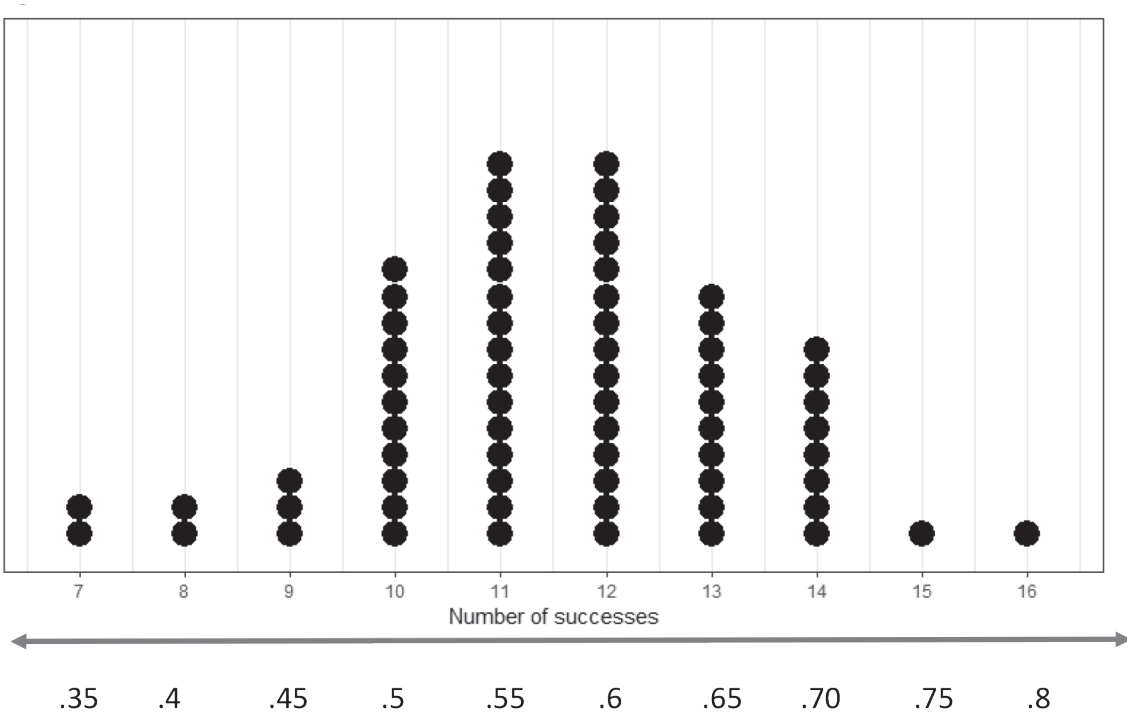

Figure 15.3: Dot plot of the proportion of students with a job from a sample of 20

13. What proportion would you expect for the mean? Explain.

Answer: *The expected mean would be 0.60, the population proportion.*

Interpret the Results in the Context of the Original Question

Based on the class simulation results, answer questions 14 to 16.

14. How many 16- to 24-year-olds out of 20 could have jobs if random samples of 20 are taken from a population in which 60% of 16- to 24-year-olds have jobs?

Answer: *Between 9 and 14*

15. What would be an unusual number to find through random sampling of 20 from a population in which 60% of 16- to 24-year-olds have a job?

Answer: *Eight and fewer and 15 or more*

16. Complete the following sentence:

The mean of the sample proportions will be equal to the value of the _____.

Answer: *population proportion*

Additional Ideas

Search the Census Bureau website: (*www.census.gov*) for one of the following population proportions:

» Proportion of US households that subscribes to cable TV

» Proportion of US households that have a telephone landline as their only phone service

» Proportion of US residents over the age of 25 who are high-school graduates

Use this population proportion to develop a sampling distribution of the number of sample successes or sample proportions based on samples of size 20.

Exit Ticket

A random sample of 20 from 200 US households was taken and the number of cat owners was calculated.

This was repeated 75 times, and the dot plot of the results is shown in Figure 15.4.

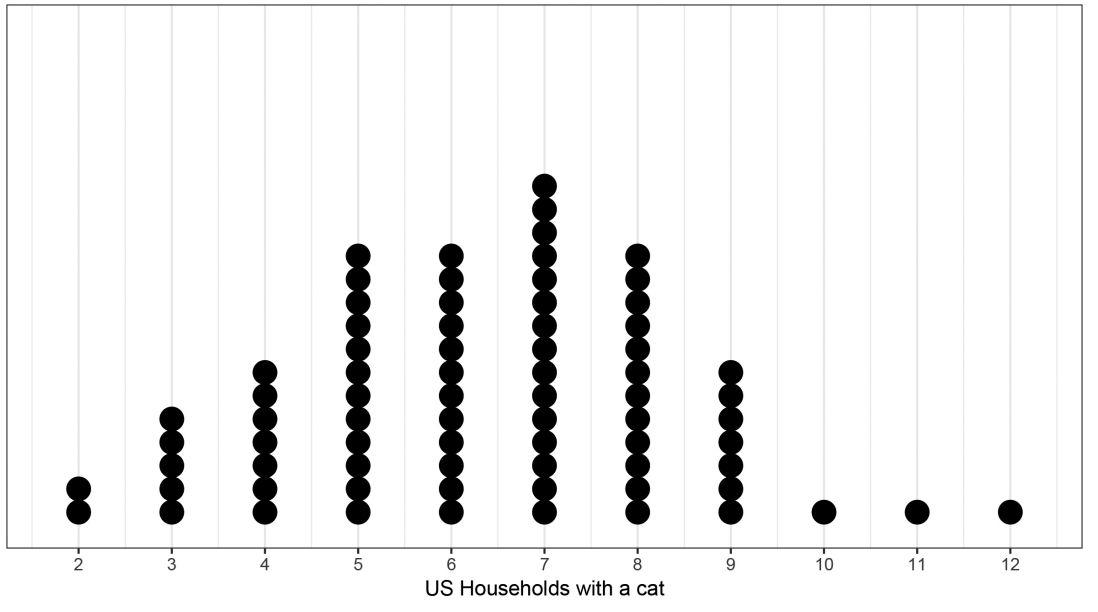

Figure 15.4: Dot plot of number of cat owners from a random sample of 20

1. Give a title to this graph.

Answer: *Simulated Sampling Distribution of the number of Cat Owners from a Sample Size of 20.*

2. Describe the shape and estimate the center and spread of the distribution.

Answer: *The distribution is mound shaped, centered at about 7, and has a spread (standard deviation) of about two.*

3. What is your estimate for the proportion of all US households that own a cat? Explain.

Answer: *The estimate is 7/20, or 0.35, the mean of the sampling distribution.*

Further Exploration and Extensions

1. Ask your students to find the standard deviation for the class results of the simulated distributions of the sample number of successes based on sample size of 20.

Answer: *In this example, the standard deviation is approximately 1.8.*

2. Have your students refer to the sampling distribution based on a sample size of 20. Find the percent of sample proportions within one standard deviation of the mean.

Answer: *In this example, with a mean of 11.5 and standard deviation of 1.8, the interval 9.7 to 13.3 would represent one standard deviation from the mean. Counting the dots between 10 and 13, inclusive, results in 51 out of 68, or about 75%.*

3. Have your students refer to the sampling distribution based on a sample size of 20. Find the percent of sample proportions within two standard deviations of the mean.

Answer: *In this example, with a mean of 11.5 and standard deviation of 1.8, the interval 7.9 to 15.1 would represent two standard deviations from the mean. Counting the dots between 8 and 1.5, inclusive, results in 65 out of 68, or about 95%.*

4. Could the Normal distribution be used to model the sampling distribution of the sample number of successes?

Answer: *The Normal distribution could be used to model the sampling distribution because the distribution is mound shaped and symmetrical. The proportion of data within one and two standard deviations is close to the theoretical results from the empirical rule.*

Investigation 16

Too Many Peanuts?

Investigating a Claim

Overview

This investigation introduces the concept of informal statistical inference. Using technology, students construct a sampling distribution of sample proportions to determine if an observed sample proportion would be considered unusual for a given population proportion. Students will be testing the claim that cans of mixed nuts contain approximately 50% peanuts. This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are formulate a statistical question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level C activity.

Note: If any students have an allergy to peanuts, rather than open a can of nuts, use the data presented in the lesson.

Instructional Plan

Brief Overview

» Develop a statistical question about the proportion of peanuts in a can of mixed nuts that is claimed to contain approximately 50% peanuts.

» Prior to class, count the number of nuts in the can and consider that number to be the random sample size of mixed nuts taken from the population of all mixed nuts processed by the manufacturer. In this investigation, the number of mixed nuts is 258.

» Calculate the proportion of peanuts in the sample. In this investigation, the proportion of peanuts in the sample of 258 mixed nuts is 55% peanuts.

» Construct a sampling distribution of sample proportions of size (number of nuts in the can). In this investigation, 258 mixed nuts from a population with a proportion of 50% peanuts is used as an example.

» Based on the sampling distribution, find the probability of randomly obtaining a sample proportion of peanuts of at least 55% peanuts, assuming the population from which the sample is taken contains 50% peanuts.

Hand out Student Worksheet 16.1 Peanut Investigation.

Ask your students to read the scenario.

Scenario

Did you ever buy a can of mixed nuts and it seemed all you got in the can was peanuts and you were hoping for a lot of cashews and almonds?

A 1964 *Consumer Reports* investigation of 124 cans of mixed nuts, representing 31 brands bought in 17 American cities, determined that most mixed nuts at that time were mostly peanuts, often 75%. As of 1993, the Food and

Drug Administration (FDA) has required a container of mixed nuts to contain at least four varieties of tree nuts or peanuts. Each kind of nut must be present not less than 2% and not more than 80% of the number of nuts.

A major manufacturer of cans of mixed nuts makes the claim that their 10.3 oz. cans containing a mixture of peanuts, almonds, cashews, pecans, and Brazil nuts have approximately 50% peanuts.

As part of a statistics project, an 11th grader purchased a 10.3 oz. can of mixed nuts and found 142 peanuts in the can that contained 258 mixed nuts or approximately 55% peanuts.

Learning Goal

Use the sampling distribution of sample proportions and informally decide if a single sample proportion is unusual.

Mathematical Practices Through a Statistical Lens

MP3. Construct viable arguments and critique the reasoning of others.

Statistically proficient students use appropriate data and statistical methods to draw conclusions about a statistical question. They reason inductively about data, making inferences that take into account the context from which the data arose. They justify their conclusions and communicate them to others.

Materials

Student worksheets are available at *www.statisticsteacher.org/statistics-teacher-publications/focus.*

- 10.3 oz. can of mixed nuts that the manufacturer claims contains approximately 50% peanuts
- Statistical software or application to generate a sampling distribution of sample proportions. Possible applications: Graphing calculator with ProbSim app or computer software like GeoGebra or StatKey
- Student Worksheet 16.1 Peanut Investigation
- Exit Ticket
- Optional: Student Worksheet 16.2 StatKey Directions

Estimated Time

One 50-minute class

Pre-Knowledge

Students should be able to find the mean and standard deviation of a distribution using technology.

Does this mean the manufacturer's claim of approximately 50% peanuts is not correct? Does this provide convincing evidence that cans of mixed nuts from this manufacturer contain more than 50% peanuts?

Note: If appropriate, open the can of mixed nuts and count the total number of nuts and the number of peanuts. Determine the proportion of peanuts in the can. Use these results to complete this investigation.

Note: If there are students with peanut allergies, then use the 55% result computed from a can containing 258 nuts. This investigation will use the 55% results from a can containing 258 nuts as the example.

Formulate a Statistical Question

Start the discussion by explaining that we are going to assume the claim of approximately 50% of mixed nuts are peanuts is true. Another way to say this is the population proportion of peanuts in all the mixed nuts is approximately 50%. Explain that we are also going to assume the number of nuts in the can of mixed nuts represents a random sample from a population of all mixed nuts produced by this manufacturer.

Next, ask your students, "If the manufacturer's claim of 50% peanuts is true, how likely is it that we get a can of mixed nuts that contains 55% peanuts? Is this an unusual result? What is the probability we could get a sample containing 55% peanuts by chance from a population containing 50% peanuts?"

Ask your students to consider the statistical question: "Assuming the manufacturer's claim that a can of mixed nuts contains 50% peanuts is true or the population proportion is 0.5, is the proportion of peanuts of 0.55 found in a can (sample) of mixed nuts an unusual result?"

Collect Appropriate Data

Ask your students to answer questions 1 to 3.

We are going to assume the population proportion of peanuts in all the mixed nuts processed by the manufacturer is 50% and the sample of 258 nuts (one can) was a random sample of all the mixed nuts produced by the manufacturer.

1. Assuming the claim that 50% of a can is peanuts, how many peanuts would you expect to be in a can of 258 nuts?

Answer: *Fifty percent of 258 is 129.*

2. If a random sample of 258 mixed nuts yielded 134 peanuts, would you think it an unusual result? Why or why not?

Answer: *This would not be considered unusual, since 134/258 is approximately 52%, which is close to 50%.*

3. If a random sample of 258 mixed nuts yielded 155 peanuts, would you think it an unusual result? Why or why not?

Answer: *This is a very unusual result. 155/258 is about 60%, which is much higher than the claim of 50%.*

Ask your students to complete questions 4 and 5.

Note: The following results were constructed using the statistical computer application-StatKey (*www.lock5stat.com/StatKey*).

Steps for using StatKey are on Student Worksheet 16.2 StatKey Directions.

4. As directed by your teacher, use statistical software to construct a simulated sampling distribution of at least 200 sample proportions based on a sample size of 258—number of nuts in the can—and assuming a population proportion of 50%.

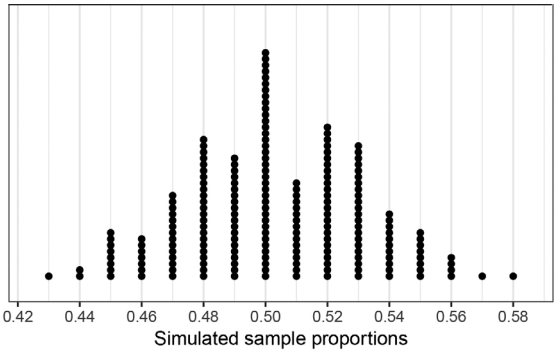

Figure 16.1: Sampling distribution of 200 sample proportions

Sample answer: *Figure 16.1*

Analyze the Data

Ask your students to answer questions 5 to 9.

5. What do you expect the mean of the simulated sampling distribution of sample proportions to equal?

Answer: *Approximately 0.50*

6. Using statistical software, find the mean and standard deviation of the simulated sampling distribution.

Sample answer: *Mean = 0.501, or 0.5; standard deviation = 0.030*

Note: The standard deviation of the sampling distribution is called the standard error of the sample proportion.

7. Describe the simulated sampling distribution of the sample proportions.

Sample answer: *Mound shaped or approximately Normal with a mean of approximately 0.5 and a standard deviation of approximately 0.03. Most sample proportions are between 0.44 and 0.56.*

8. Count the sample proportions on the plot that are greater than or equal to the proportion of peanuts in the class can (0.55 in this example). How many sample proportions were greater than or equal to the class proportion of peanuts?

Sample answer: *14 out of 200 (based on this example)*

9. Estimate the probability of the class getting a can of mixed nuts and obtaining a sample proportion of __% (in this example, 55%) peanuts or greater from a population with the population proportion equal to 0.50 peanuts.

Sample answer: *14/200 or 7% chance (based on this example)*

Interpret the Results in the Context of the Original Question

Ask the students to answer questions 10 to 12.

10. Do you think the proportion of peanuts in the class can of mixed nuts was an unusual result assuming the manufacturer's claim of 50% is correct?

Answer: *Since an estimate for the probability of obtaining the class sample proportion is 7% (answer and interpretation will vary based on the class results), the sample proportion is not that unusual. There is "some" evidence that the number of peanuts is higher than the usual amount, but there is not enough evidence to say the company's claim of 50% peanuts is not true.*

11. What proportion of peanuts would you consider to be an unusual result? Based on the simulated sampling distribution, what is an estimate for the probability of obtaining that proportion or more by chance?

Possible answers: *Answers will vary, but students may respond with a sample proportion of around 57% or higher. In this example, the probability is approximately 2/200, or 0.01.*

12. If you got such a can (high proportion of peanuts), would you have reason to believe the manufacturer's claim is not correct?

Possible answer: *Even though it can happen, the probability is very low, and I would not believe the manufacturer claim of 50% peanuts.*

Additional Ideas

» Use survey results from your class or classes and test the claim that 75% of teens use Snapchat.

» Use survey results from your class or classes and test the claim that 50% of teens use Twitter.

» Use survey results from your class or classes and test the claim that fewer than 30% of teens use Tumblr, Twitch, or Linkedln.

Exit Ticket

The American Society for the Prevention of Cruelty to Animals (ASPCA) claims that approximately 35% of US households have at least one cat. Assuming the ASPCA's claim is correct, a sampling distribution of 100 sample proportions based on a sample size of 50 and population proportion of 0.35 is shown in Figure 16.2.

Mean of simulated sampling distribution = 0.35 and a standard deviation = 0.06

1. Describe the simulated sampling distribution of the sample proportion.

Answer: *The distribution is mound shaped or approximately Normal with a mean of 0.35 and standard deviation of 0.06.*

2. A random sample of 50 US households found the proportion of households with at least one cat was 0.22. Mark the sample result of 0.22 on the dot plot. How many sample proportions are less than or equal to the sample result of 0.22?

Answer: *Approximately 3 out of 100, as shown in Figure 16.3*

3. What is an estimate for the probability of obtaining a sample proportion of 0.22 or less from a population with 0.35 households with a cat?

Exit Ticket Cont.

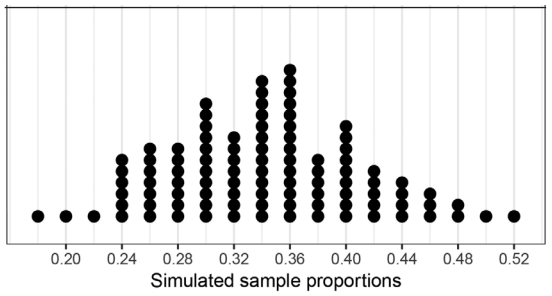

Figure 16.2: A sampling distribution of 100 sample proportions based on a sample size of 50 and population proportion of 0.35

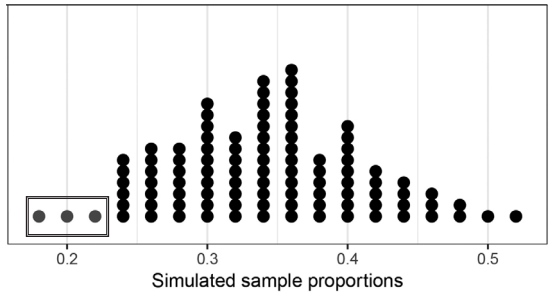

Figure 16.3: Sample result

Answer: *Approximately 3/100 or 0.03*

4. Do you think the proportion of households with a cat (22%) was an unusual result, assuming the ASPCA's claim of 35% is correct? Explain your answer.

Possible answer: *The probability of obtaining a sample result of 0.22 households with a cat from a population with a proportion of 0.35 households with a cat is about 3%. This is a low probability, so I think this is an unusual result.*

Further Exploration and Extensions

1. Introduce the *p*-value.

   The 7% (peanut example) is called a *p*-value. Assuming the company's claim of 50% peanuts in the can is correct, the *p*-value is the probability of getting the results you did (or more extreme) purely by chance.

   A *p*-value less than or equal to 5% is considered statistically significant, to where the researcher would reject the assumption that the observed results were due to random variation and conclude there is strong evidence to support that the results indicate the claim is not true.

   The concept of a *p*-value was formally introduced by Karl Pearson, in his Pearson's Chi-Squared test. The use of the *p*-value in statistics was popularized by Sir Ronald Fisher. In his book *Statistical Methods for Research Workers* (1925), Fisher proposed the level $p = 0.05$ as a possible limit for statistical significance and applied this to a Normal distribution, thus yielding the rule of two standard deviations (on a Normal distribution) for statistical significance using the empirical rule, or 68–95–99.7 rule.

2. Activity to illustrate the general rule that a *p*-value of less than or equal to 5% is considered statistically significant. *Note:* This activity usually takes a few minutes to complete.
   - » You will need a deck of all red cards. All the cards need to have the same design on the front so they look like a regular deck of cards. Have the cards in the box so the students assume the box contains the normal arrangement of 26 red and 26 black cards.
   - » Tell the students you are going to randomly divide them into two groups based on the color of the card. Red card they are in Group 1 and black card in Group 2.
   - » Remove the cards from the box and carefully shuffle them without the students seeing any of the red color on the back of the cards.
   - » Go to the first student and turn over a card. Since it will be red, tell the student s/he is in Group 1.
   - » Go to the second student and turn over a card. That student will also be in Group 1.
   - » Continue until the students get suspicious, usually around the fourth or fifth student.

Further Exploration and Extensions Cont.

» Once they get suspicious, stop and discuss the results:

» Why did you get suspicious?

Answer: *They might say too many reds in a row.*

» What did they expect to happen? Why did they expect this?

Answer: *Expect about half the cards to be red and the others black. If this were a regular deck of cards, we would expect half to be red and half to be black.*

» What is the probability that we would get this many red cards in a row, assuming this was a regular deck of cards?

Answer: *4th student – $(1/2)^4$, about 0.06; 5th student – $(1/2)^5$, about 0.03*

Point out that they got suspicious when the probability of observing five red cards in a row was about 3%. They assumed the deck was a regular deck of cards—that is, the probability of turning over a red card was 50% and they reacted (the results they saw were unusual) between 6% and 3%.

Note: The process followed is comparable to what is done in traditional hypothesis testing. Hypothesis testing refers to the formal procedures used by statisticians to reject or fail to reject the null hypothesis. In the flipping card example, the null hypothesis is the probability of turning over a red card is 50%. We assumed the null hypothesis (probability of red 50%) is true. Given the results of the card-turning simulation, we decided to reject the null hypothesis and conclude the probability of turning over a red card is not 50%.

Investigation 17

How Many Hours of Volunteer Time?

Bootstrapping

Overview

This investigation develops an interval estimate for a population mean through a resampling method called bootstrapping. Bootstrapping is a technique in which a large number of random samples of the same size are repeatedly drawn, with replacement, from a single original sample. The interval that includes the middle 95% of the resampled sample means forms a bootstrap interval.

This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are formulate a statistical question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level C activity.

Note: Before having your students read the scenario, discuss with them that some high schools require students to perform a minimum number of community service hours to graduate. Some high schools also allow students to earn credit toward their high-school graduation through community service.

Hand out Student Worksheet 17.1 Volunteer Work. Have your students read the scenario.

Scenario

Does your high school have a requirement of students to perform community service hours? If there is a requirement, how many hours are required to fulfill the responsibility? Do you already volunteer your time? What type of volunteer work do you do? How many hours do you volunteer? What are the benefits of volunteering?

There are many volunteer opportunities available for high-school students to take part in.

Some places one might volunteer include a hospital, nursing home, animal shelter, food bank, library, tutoring center, museum, beach or park, or church. The excerpt below discuses benefits for high-school students who volunteer.

> Volunteering has many benefits. Through volunteering, you'll get to explore a passion you have (such as literature or medicine). Also, by volunteering, you can support a cause you love such as helping the homeless. You can also meet like-minded students, who share your passion or want to support that cause.
>
> Volunteering is a great opportunity to test out whether you'd like to pursue a specific career (such as medicine, education, etc.). It's great to try and find your passion in high school, so you don't waste time and money during college trying to figure out what you want to major in. If you don't enjoy volunteering at a hospital, maybe pre-med isn't for you. If you love volunteering at an animal shelter, maybe you should pursue a career as a veterinarian.

> Volunteering is also a great extracurricular for your college application. It shows you selflessly dedicated your time and effort to helping others! Additionally, volunteering is a free experience that won't cost you anything other than time.

Source: https://blog.prepscholar.com/volunteer-opportunities-for-teens

Formulate a Statistical Question

A local school board is considering adding a community service graduation requirement for all district high-school students. To help the school board make an informed decision, a small group of statistics students decided to select a random sample of district high-school students who are already volunteering to determine the type of volunteer service and how many

Learning Goal

Understand the concept of bootstrapping and use bootstrapping to construct an interval estimate for a population mean.

Mathematical Practices Through a Statistical Lens

MP5. Use Appropriate Tools Strategically

Statistically proficient students can use technological tools to carry out simulations for exploring and deepening their understanding of statistical and probabilistic concepts.

Materials

Student worksheets are available at *www.statisticsteacher.org/statistics-teacher-publications/focus.*

- Student Worksheet 17.1 Random Sample of 50 Hours (copies run on stock paper)
- Student Worksheet 17.2 Volunteer Work
- Cloth or paper bag
- Statistical software or application that can generate a sampling distribution of sample means using a bootstrap method (possible application: StatKey, *www.lock5stat.com/StatKey*)
- Optional: Student Worksheet 17.3 StatKey Directions
- Exit Ticket

Estimated Time

One to two 50-minute class periods. One period to read the scenario and collect data. A second period to discuss bootstrapping and analyze the collected data and interpret the results.

Pre-Knowledge

Students should be able to use technology to construct a dot plot and find the mean and standard deviation of a sampling distribution.

Table 17.1

10	13	13	7	10	46	23	21	30	41	18	23
17	27	44	31	83	81	59	111	12	12	23	118
182	124	101	262	349	89	68	50	63	350	249	271
311	10	27	45	19	36	311	486	503	33	42	20
29	31										

hours the students are volunteering. They decided to investigate the statistical question: "For district high-school students who volunteer, what is an interval estimate for the mean number of hours they volunteer per year?"

Collect Appropriate Data

Explain that the 50 values on Student Worksheet 17.2 Random Sample of 50 hours represent the number of hours per year that 50 randomly selected district high-school students reported they volunteer (see Table 17.1). The 50 students were randomly selected from a large group of district students who reported they volunteered during the past year.

Explain that the school board would like to use these sample data to make an inference about the number of hours all district high-school students volunteer.

Ask your students for any observations or questions they have about the data.

Possible answer: *There is a large spread in the data going from 7 to more than 500.*

Ask your students to answer questions 1 to 4.

1. Construct a dot plot of the 50 times.

Answer: *Figure 17.1*

2. Find the mean of the distribution of times and describe the distribution.

Answer: *98.68 hours. The distribution is skewed to the right, with much of the data clustered between 0 and 50 hours. There are six times that are greater than 300 hours. Note: Standard deviation is 126.7 hours.*

3. What would happen if another random sample of 50 district students were taken?

Answer: *We would get different results but likely a similar looking distribution.*

4. What patterns would emerge if a large number of random samples of 50 students were taken and sample means were used to build a sampling distribution?

Answer: *The distribution of sample means would be approximately mound shape with a mean approximately equal to the population mean.*

Explain that the issue is that the school board only has the one sample of 50 times listed above. It is time consuming and usual-

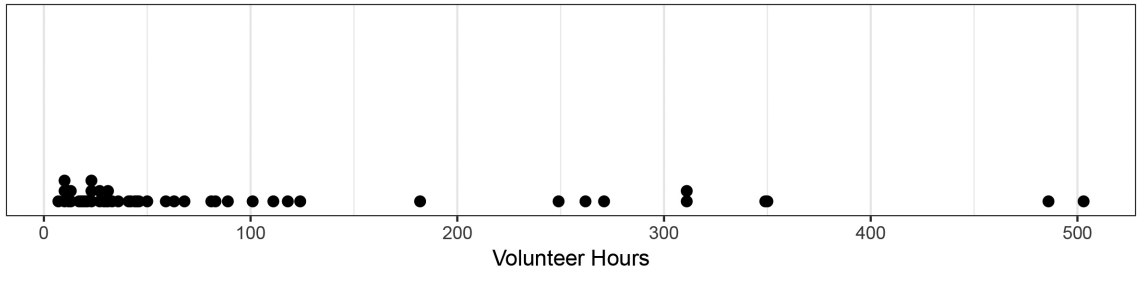

Figure 17.1: Dot plot of the 50 times high-school students reported they volunteer

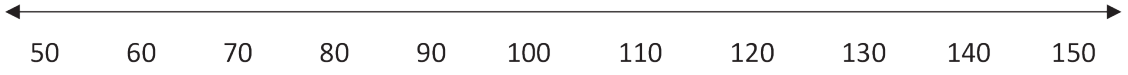

Figure 17.2: Number line of sample means

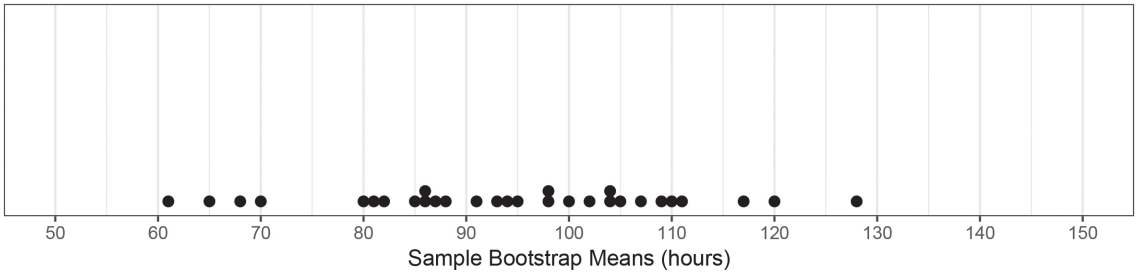

Figure 17.3: Number line of sample means

ly difficult to take a large number of random samples. In this investigation, we are going to use the single random sample result as if it is a population where the mean is approximately 99 hours. To help the school board draw conclusions about the number of volunteer hours, we are going to use a technique called bootstrapping. This technique uses the random sample of 50 in place of the actual population distribution of volunteer hours. That is, we will think of the distribution we have as being an estimate of the population distribution of volunteer hours. The bootstrapping method takes repeated samples of the same sample size with replacement and creates a sampling distribution of the sample bootstrap means.

Note: The use of the term bootstrap derives from the phrase to pull oneself up by one's bootstraps. You're trying to pull yourself up from what you've got. In a data sense, you're going to use the sample data itself to try to get more information about a population mean—the mean number of hours district high-school students volunteer.

Hand out Student Worksheet 17.1 Random Sample of 50 Hours.

Cut out the slips of paper with the 50 times and place the slips in a bag or container. Thoroughly mix the slips.

Explain that we are going to take a random sample (with replacement) of 50 slips. Go around the room and have a student draw out a slip, record the outcome, and return the slip to the bag. Continue until you have a total of 50 observations. This is called a bootstrap sample.

Note: It is possible you will draw each of the 50 slips exactly once, but this is highly unlikely. It is more likely you will draw some slips more than once and some slips not at all.

Find the mean of this bootstrap sample. This is a bootstrap estimate of the population mean.

Now have your students cut out the slips with the 50 numbers and place the slips in a bag or container. Have them thoroughly mix the slips.

5. Ask students to take a random sample of 50 slips of times with replacement and find the mean of their sample.

Place a number line on the board or poster paper similar to the one shown in Figure 17.2.

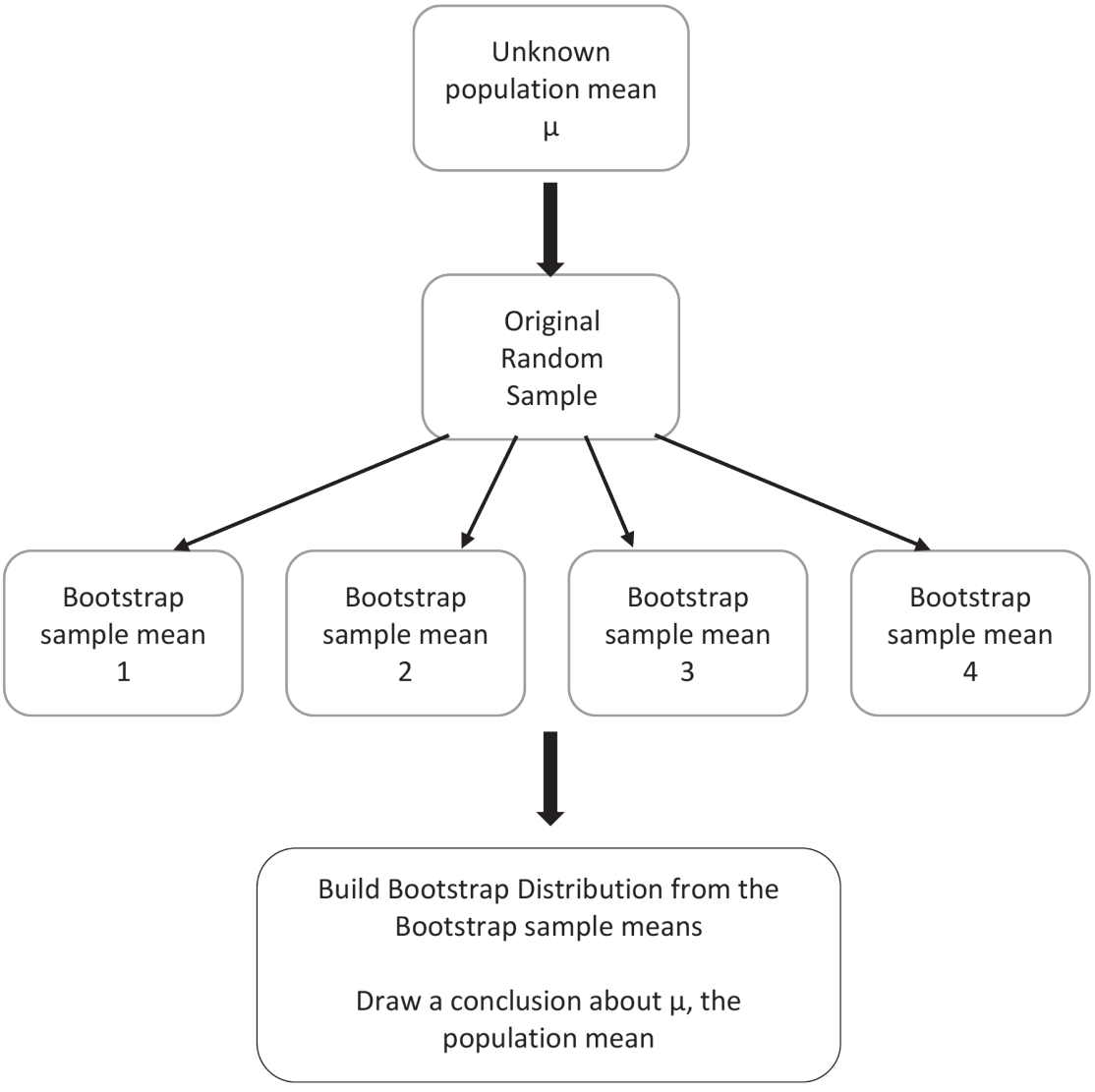

Figure 17.4: Bootstrapping method

Have your students record their sample means on the number line.

Answer: *Sample dot plot (Figure 17.3) of 30 bootstrap sample means.*

Note: Here is a diagram (Figure 17.4) that could be used to help explain the bootstrapping method. The symbol μ represents the population mean or the actual mean number of hours for all the student volunteers.

Analyze the Data

6. Using the class distribution of sample bootstrap means, find the mean of the distribution.

Answer: *The mean will be approximately 98 or 99 hours.*

Explain that it would be helpful to have more bootstrap sample means so we could get a clearer picture of the sampling distribution.

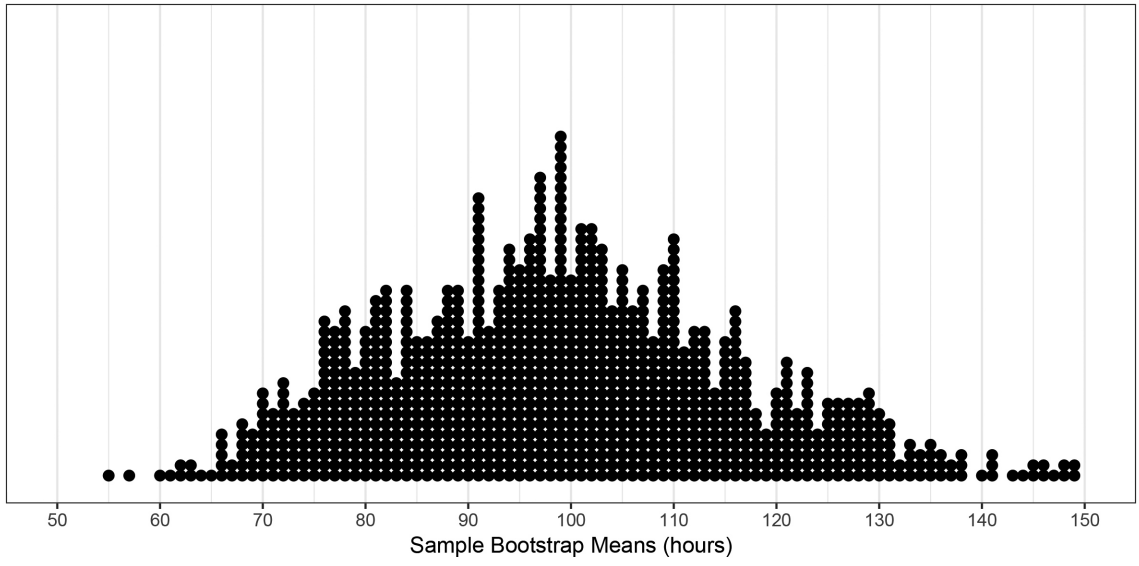

Figure 17.5: 1000 bootstrap sample means

So, rather than continuing to take random samples of 50 from the bag, use statistical software to generate a large number of bootstrap samples and display the sampling distribution of the sample means from the bootstrap samples.

Note: Worksheet 17.3 StatKey Directions contains the directions for using StatKey (*www.lock5stat.com/StatKey*) to generate the bootstrap samples. If students have access to computers, it is recommended they use software and construct a bootstrap confidence interval.

7. Use the statistical software and generate 1000 bootstrap sample means.

Possible answer: *Figure 17.5*

Rather than give just the mean, it would be helpful for the school board to have an interval where the population mean (volunteer hours of all the district students) would likely fall.

8. Using the sample distribution of the 1000 bootstrap sample means, between which two sample means is approximately the middle 95% of the distribution?

Possible answer: *Since there are 1000 sample means, we could eliminate the bottom 2.5%, or the lowest 25 sample means, and eliminate the top 2.5%, or the highest 25 sample means. See Figure 17.6.*

The two sample means 69 and 135 form an interval between which 95% of the bootstrap sample means are located. This interval gives an interval where the population mean is likely to be.

Interpret the Results in the Context of the Original Question

Ask the students to complete questions 9 to 12.

9. Using the results from the bootstrapping resample method, answer the original statistical question, "For district high-school students who volunteer, what is an interval estimate for the mean number of hours they volunteer per year?"

Possible answer: *The true mean number of hours of all the district high-school students who perform volunteer work is approximately between 68 and 136 hours, exclusive.*

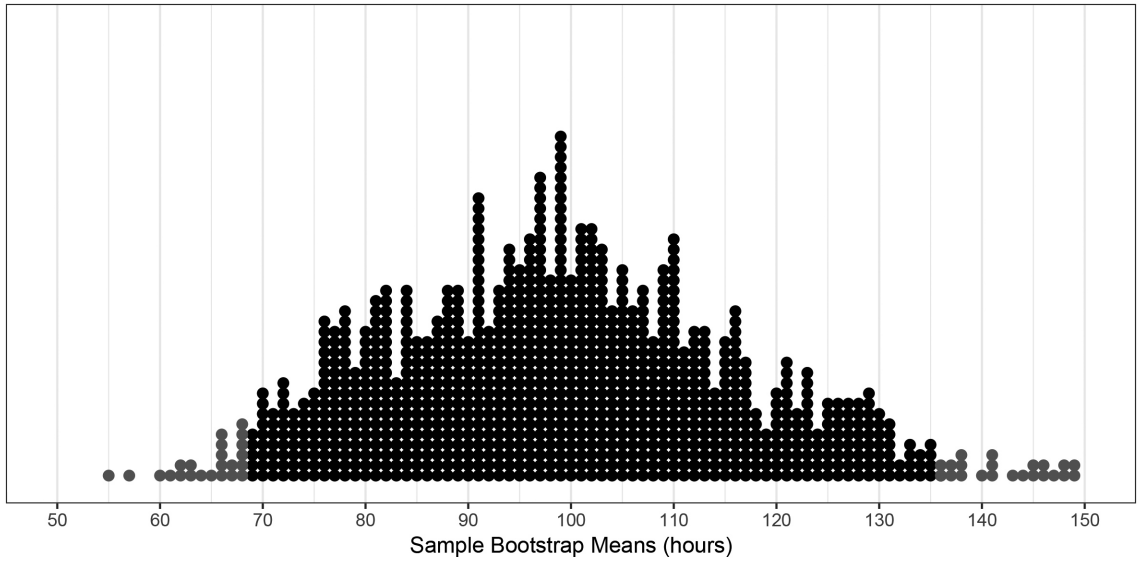

Figure 17.6: 1000 bootstrap sample means with lowest and highest sample means shaded lighter

10. Share your interval with others in class. Compare the intervals and discuss the similarities and differences.

Possible answer: *Most of the intervals will be close to the same.*

11. Write a brief summary of the bootstrap method and how it works.

Possible answer: *First, a random sample is taken from a large population. This random sample is used as representing the whole population. Many random samples are taken with replacement from the original sample, and the distribution of the sample means is created. This distribution is used to construct an interval estimate of the middle 95% of the sample means, which represents the actual value of the population mean.*

12. Explain how the bootstrap method could be used to construct an interval estimate for the middle 90% of the distribution.

Possible answer: *Take the bottom 5% off and the top 5% off the distribution of bootstrap sample means.*

As a final discussion, emphasize to your students that the interval we created is really dependent on how well the original random sample of 50 represents the population. If the original sample was biased or too small to truly represent the population distribution of volunteer hours, then the bootstrap method won't produce a valid result.

Additional Ideas

In Investigation 2: Are Baseball Games Taking Longer?, random samples of the length of Major League (MLB) Baseball games in three years (1957, 1987, and 2017) were given. Use each of the years and generate a sampling distribution of bootstrap sample means and construct a 95% interval estimate for the actual population mean (true mean length of all the games played during that year). Compare the three intervals and investigate whether the average length of MLB games is getting longer.

Exit Ticket

Chicago has been keeping records for the number of inches of snow for many years. Forty years were randomly selected from all the years snowfall has been recorded, and the amount of snowfall for those years formed a random sample.

This random sample of 40 snowfall amounts was used to take 100 bootstrap samples, and the sample bootstrap means are shown in the dot plot in Figure 17.7.

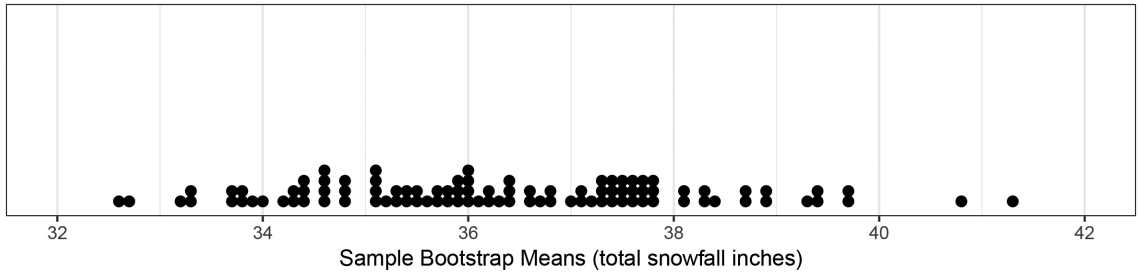

Figure 17.7: Sample bootstrap means of 40 snowfall amounts

1. What is an estimate for the mean number of inches of snow in Chicago over all the years snowfall has been recorded?

Answer: *Approximately 36 inches—the mean of the sampling distribution.*

2. What is a 95% bootstrap interval estimate for the mean number of inches of snow in Chicago for all the years snowfall has been recorded? Explain how you got your answer.

Answer: *Approximately 33 to 41 inches. Since there are 100 sample means, eliminate the lowest 2.5%, or the lowest two or three, and eliminate the highest 2.5%, or the top two or three.*

Investigation 18

How Stressed Are You?

Exploratory Lesson: Comparing the Differences in Proportions

Overview

This investigation offers two options for students. One option provides students an opportunity to use the four components of statistical problem solving by designing their own investigation around a topic of interest that involves exploring whether two proportions are significantly different. Several suggestions are included in this investigation, and students could be encouraged to come up with their own questions of interest. Encourage students to work in pairs or small groups.

The results could include written and oral presentations and/or construction of a poster to display the data and answer the statistical question. Information about creating a statistical poster, a rubric, and competition information can be found at *www.amstat.org/asa/education/ASAStatistics-Poster-Competition-for-Grades-K-12.aspx*.

A second option provides the set of directions for an investigation titled, "American Teens Compared to New Zealand Teens." This option is suggested for students who may need more scaffolding and direction when designing a simulation and analyzing the simulation results. This option is based on Lesson 14 from *Making Sense of Statistical Studies,* published by the American Statistical Association and available at *ww2.amstat.org/education/msss/index.cfm*.

This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are formulate a question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level B or C activity, depending on the amount of scaffolding provided.

Instructional Plan

Instructions for Designing Your Own Investigation

Explain that students will follow the four steps of the statistical problem-solving process. Have students work in pairs or small groups. Distribute Student Worksheet 18.1 Directions for Student-Designed Investigation.

Formulate a Statistical Question

Students will brainstorm topics that might interest their group that include a categorical variable. Your students could design a question and take a random sample of two groups in their high school, such as freshmen and seniors or students and teachers. Students could model their question based on the questionnaire on the Census at School website or design questions based on a topic of student interest.

Some possible questions to explore based on ideas from the Census at School website (*ww2.amstat.org/censusatschool/students.cfm*) include the following:

» Is the proportion of students who are vegetarians significantly different from

the proportion of teachers who are vegetarians?

22. Are you vegetarian?
□Yes □No

» Is the proportion of students who drink energy or sports drinks significantly different from the proportion of teachers who drink energy or sports drinks?

24. What type of beverage do you drink most often during the day?
□Water □Soft drink (caffeinated) □Tea □Milk □Soft drink (non-caffeinated) □Coffee □Juice □Energy drink □Sports drink □Powdered drink (e.g., Kool-Aid, Tang)

» Is the proportion of teachers who like country music significantly different

Learning Goal

Understand what it means for two proportions to be significantly different.

Mathematical Practices Through a Statistical Lens

MP1. Make sense of problems and persevere in solving them.

Statistically proficient students understand how to carry out the four steps of the statistical problem-solving process: formulating a statistical question, designing a plan for collecting data and carrying out that plan, analyzing the data, and interpreting the results.

Materials

Student worksheets are available at *www.statisticsteacher.org/statistics-teacher-publications/focus.*

» Student Worksheet 18.1 Directions for Student-Designed Investigation

» Student Worksheet 18.2 Difference Between Proportions

» Student Worksheet 18.3 Template

Estimated Time

One to three 50-minute class periods, depending on final report and amount of work required outside of class.

Pre-Knowledge

Students should be able to:

» Summarize data using a dot plot

» Find the mean and standard deviation

» Conduct a simulation to construct a sampling distribution

» Use a simulated sampling distribution to determine what values are unlikely outcomes

than the proportion of students who like country music?

36. What is your favorite type of music? Select one.
□Classical □Pop □Rhythm and blues (R&B) □Other □Country □Punk rock □Rock and roll □Heavy metal □Rap/Hip hop □Techno/Electronic □Jazz □Reggae □Gospel

» Is the proportion of females who would like to be able to fly significantly different than the proportion of males who would like to fly?

37. Which of the following superpowers would you most like to have? Select one.
□Invisibility □Telepathy (read minds) □Freeze time □Super strength □Fly

Is the proportion of 9th graders who would give $1000 to environmental causes significantly different than the proportion of 12th graders who would give $1000 to environmental causes?

40. If you had $1000 to donate to a charity of your choice, what type of organization would you choose?
□Arts, culture, sports (e.g., community centers, museums, sports teams, music programs)

□Health (e.g., cancer, AIDS, diabetes research)

□Religious (e.g., church or activities related to worship)

□Environmental (e.g., saving forests, clean air, clean water)

□Wildlife, animals (e.g., endangered species, prevention of cruelty to animals)

□Education/Youth development (e.g., reading, literacy and skills training, after-school programs)

□International aid (e.g., disaster relief, health, education and food aid in poor countries)

Students should then develop a statistical question. Students are expected to check in for approval at this point before moving on to collecting data.

Collect Appropriate Data

Students can either take a random sample from the database at the Census at School website or take a random sample of the appropriate groups in their school. Direct students to outline the data-collection process, including possible complications and how these might be handled.

Once the data are collected, direct students to design a two-way table and find the difference between the proportions of interest. They will then design a simulation similar to the simulation outlined in Investigation 11 or the second part of this investigation based on the assumption that there is no difference between the two proportions. Students should run a large number of trials. For each trial, they should record the simulated difference between the two proportions.

Analyze the Data

Data analysis should include a dot plot and summary of the simulation.

Interpret the Results in the Context of the Original Question

Interpret the analysis of the data in the context of the situation. Be sure to answer the statistical question and support the answer with the data analysis.

Option 1: Write and orally present a report summarizing your results. Your report and presentation should include the following:

» The statistical question investigated and why it was chosen

» A description of the population sampled

» A summary of the data collection

» The collected data, organized as appropriate

» Analysis and descriptions of the data, using calculations, tables, graphs, and plots. Note any unusual results.

» Conclusions about the statistical question

» Recommendations for any follow-up studies or questions that may be investigated

Option 2: Create a data visualization poster and orally present the poster summarizing your results.

» The poster should include the following:

» The statistical question as the title of the poster

» The organized collected data—tables and graphs

» Conclusions about the statistical question

The oral report should include the following:

» The reason the statistical question was chosen

» A description of the populations sampled

» A summary of the data collection

» Analysis and descriptions of the data, using calculations, tables, graphs, and plots. Note any unusual results.

» Recommendations for any follow-up studies or questions that may be investigated

Instructions for "American Teens Compared to New Zealand Teens"

Scenario

The World Happiness Report is a survey of the state of global happiness. The World Happiness Report 2018 ranks 156 countries by their happiness levels. The report named Finland as the top-ranked—happiest—country in the world. New Zealand ranked eighth, and the United States ranked 18th. All the top countries tend to have high values for all six of the key variables that have been found to support well-being: income, healthy life expectancy, social support, freedom, trust, and generosity.

As a student, how happy are you? High school can be challenging—students are under a lot of stress from the amount of homework, studying for exams, preparing college applications, social anxieties, and athletic competitiveness. Teens routinely say their school-year stress levels are far higher than they think is healthy and their average reported stress exceeds that of adults, according to an annual survey published by the American Psychological Association.

Do you feel stressed because of the amount of homework you have?

The Census at School website contains a large database of responses to questions and responses from students in the United States and other countries.

Note: See the appendix for more information concerning the Census at School website (*ww2.amstat.org/censusatschool/students.cfm*).

Table 18.1

	Some or A Lot	Little or None	Total
US Teens			100
New Zealand Teens			100
Total	102	98	200

Sample result:

	Some or A Lot	Little or None	Total
US Teens	52	48	100
New Zealand Teens	50	50	100
Total	102	98	200

One question from the Census at School questionnaire students gave responses to is:

"How much pressure do you feel because of the schoolwork you have to do?"

□None □Very little □Some □A lot

A random sample from the Census at School database of 100 US 16- 17-year-old students who answered the question about stress found that approximately 54% of the students responded "some" or "a lot" of pressure.

A random sample from the New Zealand Census at School database of 100 New Zealand 16- 17-year-old students who answered the question about stress found approximately 48% of the students responded "some" or "a lot" of pressure.

Formulate a Statistical Question

The results of the two surveys indicate the two proportions of students who responded and feel "some" or "a lot" of pressure are different: 0.54 for U.S. 16- 17-year-old students and 0.48 for New Zealand 16- 17-year-old students. The difference between the two proportions is 0.06.

Are the two groups of students really not that far apart, or is the difference of 0.06 a significant difference? By significant difference, we mean the difference in the response proportions for the two samples is larger than what we would expect to see due to sampling variability.

The statistical question we want to answer is: "Is there a significant difference between the proportion of US 16- 17-year-old students and the proportion of New Zealand 16- 17-year-old students who responded they feel 'some' or 'a lot' of pressure because of schoolwork?"

Collect Appropriate Data

A simulation (similar to the one designed in Investigation 11) will be used to help answer the statistical question. In this simulation, a population is created of 200 students representing the combined number of US and New Zealand students who answered the question.

Create 200 slips of paper (Student Worksheet 18.3 Template) of the same size and mark 102 of the slips with an **L** to represent the 102 students (54 US and 48 New Zealand) who responded they feel "some" or "a lot" of pressure. Cut out the slips and thoroughly mix them. Then, randomly select 100 slips from the bag. These 100 slips represent the number of US teens. Count the number of slips with an **L** and record in the cell labeled US teens and "some" or "a lot." Based on this count, complete Table 18.1.

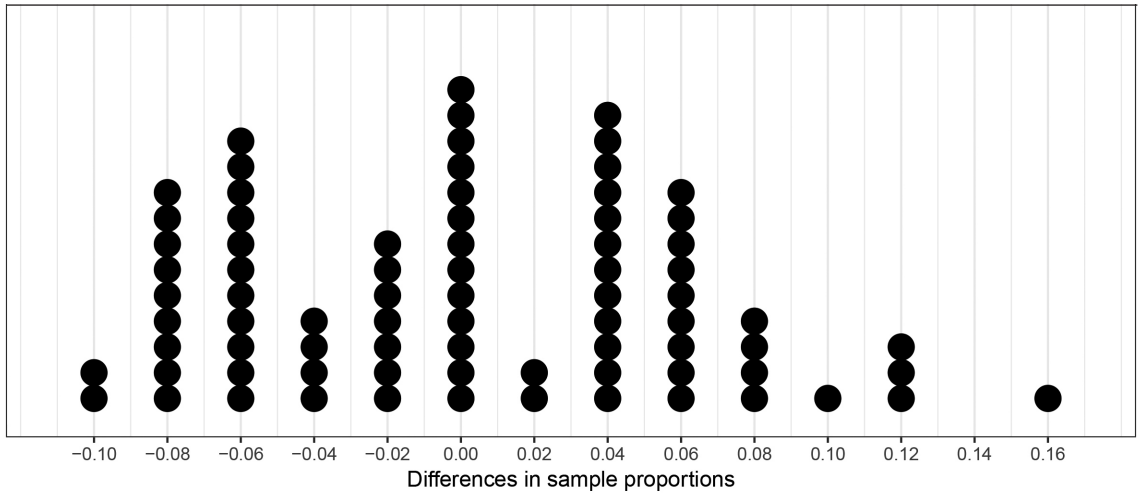

Figure 18.1: Dot plot showing sample result of 80 trials

Find the proportion of US teens who responded with "some" or "a lot" of pressure.

Sample answer: *52/100 or 0.52*

Find the proportion of New Zealand teens who responded with "some" or "a lot" of pressure.

Sample answer: *50/100 or 0.50*

Find the difference between these two proportions and record this difference.

Sample answer: *0.52-0.50 = 0.02*

Repeat this simulation a large number of times. Each time, record the difference between the two simulated proportions.

Analyze the Data

Construct a dot plot of the simulated differences between the two proportions.

A sample result (Figure 18.1) is based on 80 trials.

Interpret the Results in the Context of the Original Question

The original surveys showed 54% of US 16-17-year-old students responded that they felt "some" or "a lot" of pressure and 48% of the New Zealand 16- 17-year-old students responded they felt "some" or "a lot" of pressure. This gave an observed difference of 0.06.

Based on the simulated differences, which were based on assuming there was no difference between the proportions, write a few sentences that address the statistical question: "Is there a significant difference between the proportion of US 16- 17-year-old students and the proportion of New Zealand 16- 17-year-old students who responded they feel "some" or "a lot" of pressure because of schoolwork?"

Possible answer: *Based on the dot plot of the simulated differences, an observed difference of 0.06 would not be that unusual. Eighteen of the 80 (22.5%) simulated differences were greater than or equal to the observed difference of 0.06. This difference could be due to sampling variability and therefore there is no significant difference between the proportion of US 16-17year-old students and the proportion of New Zealand 16- 17-year-old students who responded they feel "some" or "a lot" of pressure because of schoolwork. See Figure 18.2.*

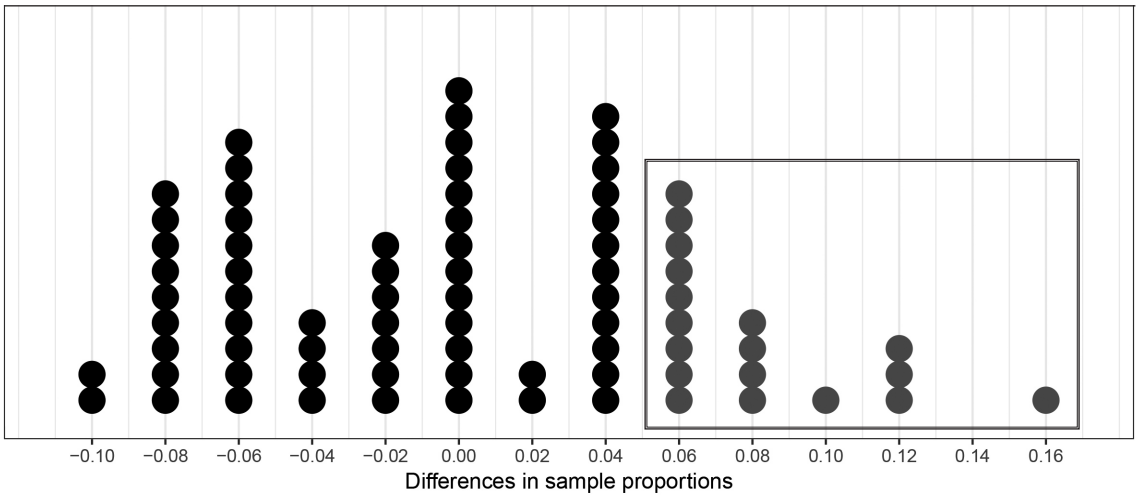

Figure 18.2: Dot plot of simulated differences