Section III: Two-Variable Data Analysis

Investigation 5

How Many Calories?

Scatterplots

Overview

This investigation begins the exploration of relationships within bivariate data by investigating errors made by students between guesses and actual values, setting the stage for the concept of a residual.

The concept of a residual and a residual plot will be used in Investigation 7 as a tool for exploring variability about the least squares regression line.

This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Pre-K-12 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 B activity.

This activity is based on lessons from *Exploring Linear Relations* (Lesson 7), published by the American Statistical Association (original copyright by Dale Seymour Publications 1999). *Exploring Linear Relations* is a module in the ASA Data-Driven Mathematics Project. It is available as a free download at *www.amstat.org/asa/files/pdfs/ddmseries/ExploringLinearRelations--TeachersEdition.pdf*.

Instructional Plan

Brief Overview

» Read and discuss the scenario about obesity and caloric information.

» Formulate the statistical question: What is the typical error made by students in estimating the number of calories in bite-sized candies?

» Have students estimate the number of calories in each item and find their errors.

» Use the squares of the errors to determine who is the "best" guesser.

Hand out Student Worksheet 5.1 Guess the Calories. Have your students read the Scenario.

Scenario

The following excerpt comes from *Attacking the Obesity Epidemic: The Potential Health Benefits of Providing Nutrition Information in Restaurants* by Scot Burton, Elizabeth H. Creyer, Jeremy Kees, and Kyle Huggins. The entire article can be found at *www.ncbi.nlm.nih.gov/pmc/articles/PMC1551968*.

> Sixty-four percent of American adults are either overweight or obese, and the obesity epidemic shows few signs of weakening. Although the precise number of deaths attributable to obesity is difficult to estimate, obesity is clearly a major cause of preventable death. Not surprisingly, improving the healthfulness of the American diet has become a national health priority. The increasing prevalence of obesity-related diseases has been blamed, in part, on the increased consumption of foods prepared outside the home. Restaurant expenditures have increased consistently in

Learning Goals

» Represent data on two quantitative variables on a scatterplot and describe how the variables are related

» Develop understanding of an error

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 follow the logical progression of the statistical problem-solving process to investigate answers to a statistical question and provide insights into the research topic. They reason inductively about data, making inferences that consider the context from which the data arose. They justify their conclusions, communicate them to others (orally and in writing), and critique the conclusions of others.

Materials

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

» Student Worksheet 5.1 Guess the Calories

» Graph paper (.5 cm or .25 inch)

» "Fun" size Milky Way candy bar

Estimated Time

One 50-minute class period

Pre-Knowledge

Students should already be able to:

» Construct scatterplots

» Draw the y=x line on a scatterplot

recent decades; consumers now spend more than $400 billion annually.

Results: Survey results showed that levels of calories, fat, and saturated fat in less-healthful restaurant items were significantly underestimated by consumers. Actual fat and saturated fat levels were twice consumers' estimates and calories approached two times more than what consumers expected. In the subsequent experiment, for items for which levels of calories, fat, and saturated fat substantially exceeded consumers' expectations, the provision of nutrition information had a significant influence on product attitude, purchase intention, and choice.

Conclusions: Most consumers are unaware of the high levels of calories, fat, saturated fat, and sodium found in many menu items. Provision of nutrition information on restaurant menus could potentially have a positive impact on public health by reducing the consumption of less-healthful foods.

Formulate a Statistical Question

After students have read the scenario ask:

How well do you think you might estimate the calories in restaurant items? Have you noticed restaurants providing this information more readily? For example, some restaurants list this on their menus, such as Panera and Noodles and Co.

Explain that they will be asked to estimate the number of calories in "fun" size candy bars. Hold up a fun size Milky Way candy bar and ask students to guess the number of calories. Have them write down their guesses, and then ask a few students to share. Then share that the fun size Milky Way bar contains 80 calories.

Ask your students how close their estimates were for the number of calories.

Ask students to create a possible statistical question for this activity.

Example: *What is the typical error made by students in estimating the number of calories in bite-sized candies?*

Note: If it appears students may interpret this investigation as the worst guessers will become obese, it may be worth making the point that the study was looking for a positive relationship between knowing the calories of a food prior to consumption and making healthier choices. This does not indicate not knowing the calories in food means making unhealthy choices.

Table 5.1 Candy

Candy Item – Fun Size	Actual
Snickers	80
Skittles	80
Butterfinger	100
Kit Kat	70
M&M's Plain	73
M&M's Peanut	90
Reese's Peanut Butter Cup	110
Starburst	40
Whoppers	100
Twizzlers	50
Jolly Ranchers (3 Pieces)	70

Collect Appropriate Data

Ask students to complete problem 1 on Student Worksheet 5.1 Guess the Calories.

1. Fill in the "Guess?" column with your guesses for the number of calories in each fun size candy item.

When students have completed their Guess? column, reveal the actual number of calories (Table 5.1). Have the students complete Problem 2.

2. Fill in the "Actual" column with the actual number of calories in each fun size candy item.

Analyze the Data

Ask your students to complete Question 3.

3. How might you decide who is the best guesser in the class? Justify your answer.

After giving your students time to individually write about their methods of determining the best guesser, have them share with a partner or in small groups. Then, as a whole group, ask students to share their ideas about how they might determine who is the best guesser. Does a person need to be exactly

correct? Or just close? Is underestimating different from overestimating?

Possible answers: *Students might suggest a variety of ways to calculate who is the best guesser. Encourage discussion and collect several ideas. For example, the greatest number of guesses that match actual calories or the greatest number of guesses that were within 10 calories of the actual calories. If it doesn't come up, guide discussion around whether it matters how far off someone was.*

Explain that we are going to create a visual display of their data to help determine who is the best guesser. Distribute graph paper to each student. Give students time to answer questions 4 and 5.

4. Create a scatterplot of your data on graph paper, plotting the actual calories on the x-axis and your guess on the y-axis.

Sample answer: *Figure 5.1*

5. Describe the relationship between your guesses and the actual calories in each candy item.

Possible answer: *As the actual calories increase, my guessed calories increased. I tended to underestimate the actual calories. For the Butterfinger, my guess for number of calories was much higher than the actual number of calories.*

Table 5.2 Sample Data

Candy Item – Fun Size	Actual	Guess?
Snickers	80	90
Skittles	80	75
Butterfinger	100	150
Kit Kat	70	50
M&M's Plain	73	50
M&M's Peanut	90	85
Reese's Peanut Butter Cup	110	90
Starburst	40	25
Whoppers	100	50
Twizzlers	50	30
Jolly Ranchers (3 Pieces)	70	50

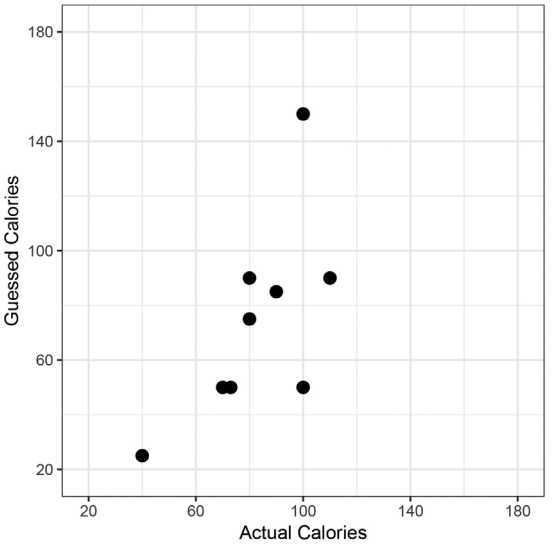


Figure 5.1: Scatterplot of guessed and actual calories

Use one of the student's scatterplots as an example or show students a scatterplot with the x-axis labeled Actual Calories and the y-axis labeled Guessed Calories. Ask what a point on this graph represents. For example, the point (100, 75) represents a candy item that has 100 calories and was guessed to have 75 calories. Refer to the study from the beginning of the lesson—if someone often underestimated the number of calories in the candy items, how would that appear in a scatterplot of the data?

The points would be closer to the x-axis since the guessed calories would be lower than the actual calories.

Ask your students to answer Question 6,

6. What would the scatterplot look like if someone had guessed the correct actual calories in each candy item?

Possible answer: *The points would make a straight line. If it doesn't come up, ask for the equation of the line, which is y = x, the line*

where all the y-values (guessed calories) are the same as the x-values (actual calories).

Have students draw the y = x line on their scatterplots. Ask what it means for a point to be on the line, below the line, and above the line.

Possible answers: *If a point is on the line, the guess matches the actual calories. If a point is below the line, the guess is lower than the actual calories. If a point is above the line, the guess is higher than the actual calories.*

Ask students to answer Question 7, and then have some students share their scatterplots and explain the type of guesser their graph shows.

7. Describe the type of guesser your scatterplot shows. Explain.

Possible answer: *Based on my scatterplot (Figure 5.2), I tended to underestimate the calories in the fun size items as shown by most of my dots being below the line y = x.*

Ask students how this line might help us determine who is a better guesser.

Possible answer: *Students might suggest a person whose points are closer to the line is a better guesser.*

How might we measure the distance from the line?

Possible answer: *Students might suggest measuring the perpendicular distances, horizontal distances, vertical distances, or other methods.*

Explain that we are going to use the vertical distances, as this is a convention used in statistics. What do the vertical distances represent?

Possible answer: *The difference between the guessed calories (y-value of the point) and the actual calories (y-value on the line).*

What could we call these differences?

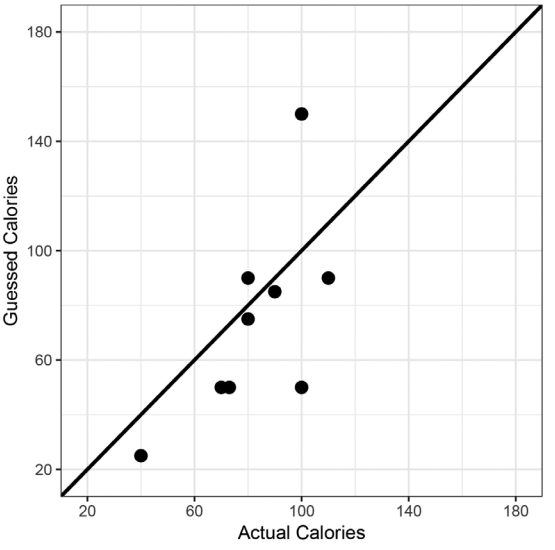


Figure 5.2: Scatterplot of someone who underestimated the calories in the candy.

Possible answer: *These differences between the guessed calories and actual calories for each candy item are the amounts of error for each candy item.*

Note: You should not use the term "residual." This activity is setting the stage for understanding the concept of the residual in Investigation 7.

The values for the vertical distances represent the error for each guess. Using a student's scatterplot, choose a point above the line and draw the vertical distance between the point and the line y = x. The example here shows the point (100, 150), which means the difference between the guessed calories (150) minus the actual calories (100) is 50, thus the error is 50 calories.

If an error is 10, how could we determine if the person overestimated or underestimated? If no students make a suggestion, explain that positives and negatives are used to determine this. The error is positive if the point is above the line, and the error is negative if the point is below the line. In this situation, points above

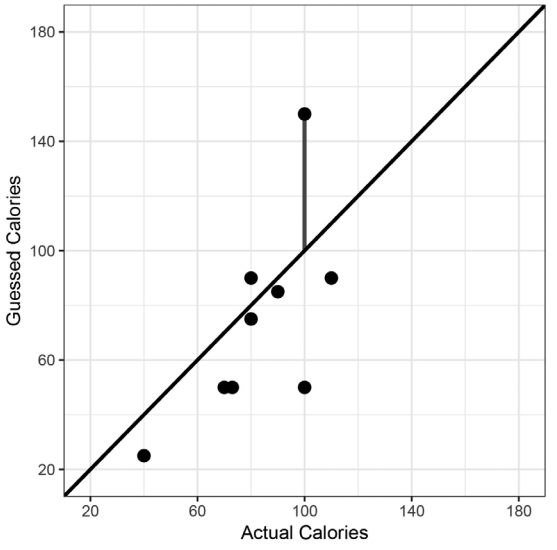


Figure 5.3: Scatterplot showing the difference between the guessed calories and actual calories for each candy

Table 5.3 Errors Based on Sample Data

Candy Item – Fun Size	Actual	Guess?	Errors
Snickers	80	90	10
Skittles	80	75	-5
Butterfinger	100	150	50
Kit Kat	70	50	-20
M&M's Plain	73	50	-23
M&M's Peanut	90	85	-5
Reese's Peanut Butter Cup	110	90	-20
Starburst	40	25	-15
Whoppers	100	50	-50
Twizzlers	50	30	-20
Jolly Ranchers (3 Pieces)	70	50	-20

the line are overestimates and positive errors; points below the line are underestimates and negative errors. It might be helpful to show another example on a student's graph.

Ask what an error of 0 represents.

Answer: *The guess and the actual value were the same.*

Have your students answer questions 8 and 9.

8. On your scatterplot, draw the y=x line. Then draw the vertical distances representing the "errors" on your scatterplot.
9. On the table of guesses and actual number of calories, add a third column labeled "Errors" and calculate the errors (guess minus actual) for each candy item. Find the sum of the errors.

Answer based on the given example (Table 5.3).

Ask students how we might use the errors to help us determine who is the best guesser. Encourage discussion. Students might suggest how many are close to 0 or +/- 10 calories or adding up the errors. Propose that the best guesser is one who has a sum of errors close to zero. After students have found their sum, ask who the best guesser was using this method. Ask students if they have any concerns about using the sum of the errors.

Possible answer: *This would not be a good method since someone could have a very high error (extreme overestimate) balanced by a very low error (extreme underestimate).*

Ask students how we might handle values that are negative when we might like them to be positive.

Possible answer: *Students will most likely say take the absolute value.*

Have your students answer Question 10.

10. On the table of guesses and actual number of calories, add a fourth column labeled "Absolute Value" and complete the column. Find the sum of the absolute values.

Note: Finding the absolute values and the sum of the absolute values could be connected to earlier learning of the Mean Absolute Deviation (MAD) from middle school.

Now ask again who is the best guesser. The best guesser would have the lowest sum of the absolute value of errors.

Ask students for another way we might handle values that are negative when we might like them to be positive.

Possible answer: *We could square the values.*

Ask students where else squaring has been used in statistics to "handle" negatives.

Answer: *Squaring was used when calculating deviations from the mean for standard deviation.*

Ask how this would help determine who the best guesser is.

Possible answer: *The best guesser would have the lowest sum of the squares of the errors, though this will usually be much higher than the sum of the absolute value of the errors.*

Note: If all the errors are less than 1, then the sum of squares will be less.

Have your students answer Question 11.

11. On the table of guesses and actual number of calories, add a fifth column labeled "Squares" and complete the column. Find the sum of the squares.

Ask again who the best guesser is. Did this answer change from the best guesser based on absolute values of the errors?

Interpret the Results in the Context of the Original Question

In groups of four, ask your students to complete problems 12 to 15. Then discuss.

12. Compare your results from Question 11 with the other students in your group. Who in your group was the best guesser of calories? Justify your answer.

Possible answer: *I know ____ is the best guesser because his/her sum of the squared errors is the lowest. This means his/her guesses were closest to the line y=x, which would be the line created if all the guesses were correct.*

13. Using the scatterplots and analysis of the errors, answer the statistical question: What is the typical error made by students in estimating the number of calories in bite-sized candies?

Answers will vary depending on the class results. Reference whether the students tended to overestimate or underestimate. Also refer to the usual number of calories their estimate was off.

14. How do these results relate to the study results?

Possible answer: *Based on the results from the class, either support or do not support the claim in the study that people tend to underestimate the number of calories in restaurant items.*

15. Why might finding errors be important when looking at data?

Possible answer: *Errors can help determine how close guesses are to the actual data values and who is the best guesser from many guessers.*

Additional Ideas

Do the same investigation steps, but instead of using the candy items, guess the calories (or fat grams) of fast food items from a particular restaurant or several restaurants. Students can find this information online.

Do the same investigation steps, but instead of using the candy items, guess the ages of various celebrities such as actors, sports figures, or politicians (international, national, and/or local) students would know. It can be helpful to have recent pictures of the celebrities to show. Make sure to have variety in ages.

Exit Ticket

1. The Price Is Right - Cliffhanger Game: "The contestant bids on three small prizes. For every dollar the contestant is away from the actual prices, a mountain climber takes one step up a mountain. If the mountain climber does not exceed 25 steps after the contestant has bid on all three prizes, then the contestant wins a bonus prize."
   Source: www.priceisright.com/games

Here are some items and the prices the contestants bid. Who won the bonus prize? Who was the best price guesser? Who was the worst guesser? Justify your answer.

Jenna		
Item	**Bid**	**Actual**
Passport Holder	12	16
Toaster	35	22
Coffee Pot	35	40
Lindsey		
Item	**Bid**	**Actual**
Inflatable Pool Lounge Chair	15	20
Electronic Piggy Bank	37	30
Pet Brush and Accessories	27	40
Debra		
Item	**Bid**	**Actual**
Liquid Measuring Cup	5	15
Electric Egg Cooker	7	22
Whipped Cream Dispenser	6	26
Kristen		
Item	**Bid**	**Actual**
Steam Iron	25	22
Electric Heater Fan	35	35
Milkshake Drink Mixer	50	49

Exit Ticket Cont.

Exit Ticket Answer:

Jenna					
Item	**Bid**	**Actual**	**Error**	**Absolute Value of Error**	**Squared Error**
Passport Holder	12	16	-4	4	16
Toaster	35	22	13	13	169
Coffee Pot	35	40	-5	5	25
Sum	22	210			
Lindsey					
Item	**Bid**	**Actual**	**Error**	**Absolute Value of Error**	**Squared Error**
Inflatable Pool Lounge Chair	15	20	-5	5	25
Electronic Piggy Bank	37	30	7	7	49
Pet Brush and Accessories	27	40	-13	13	169
Sum	25	243			
Debra					
Item	**Bid**	**Actual**	**Error**	**Absolute Value of Error**	**Squared Error**
Liquid Measuring Cup	5	15	-10	10	100
Electric Egg Cooker	7	22	-15	15	225
Whipped Cream Dispenser	6	26	-20	20	400
Sum	45	725			
Kristen					
Item	**Bid**	**Actual**	**Error**	**Absolute Value of Error**	**Squared Error**
Steam Iron	25	22	3	3	9
Electric Heater Fan	35	35	0	0	0
Milkshake Drink Mixer	50	49	1	1	1
Sum	4	10			

Jenna, Lindsey, and Kristen all won the bonus prize because their sum of the absolute values of the errors did not exceed 25. The best guesser is Kristen, since her sum of squared errors is only 10. The worst guesser is Debra, as her sum of squared errors is 725. Debra underestimated all the prices by quite a bit.

Further Explorations and Extensions

1. The Price Is Right - Bargain Game: "Two prizes are shown to the contestant. Each prize displays a bargain price that is below its actual retail price. If the contestant selects the bigger bargain (the prize with the bargain price that is farther from the actual retail price), then the contestant wins both prizes."
   Source: www.priceisright.com/games

Sketch a scatterplot that would depict what several prizes and their bargain prices might look like, as well as the line y = x.

Example video: *www.youtube.com/watch?v=crzYmKNvISg*

Example from video: Which price is the bigger bargain? Billiards table at $1600 or 55" HDTV at $2649?

Billiards table: bargain price: $1600; actual price: $2600

55" HDTV: bargain price: $2649; actual price: $3149

Answers: *Figure 5.4 and Figure 5.5*

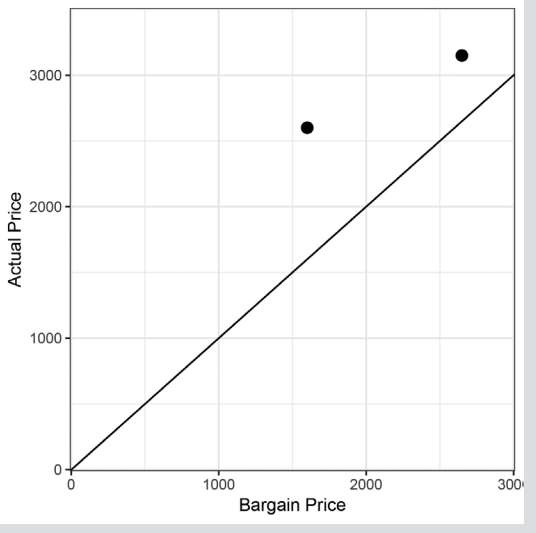


Figure 5.4

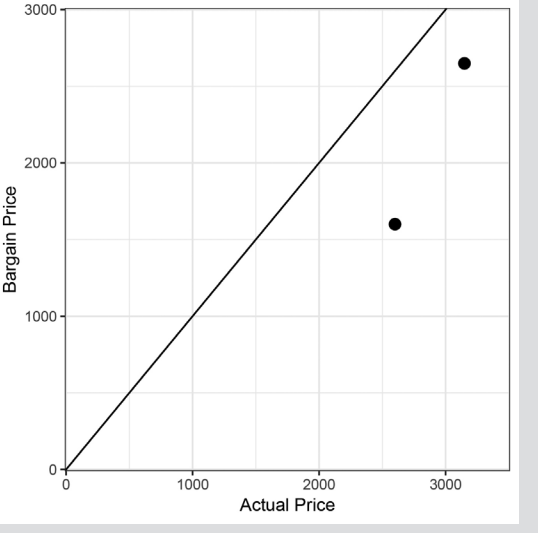


Figure 5.5

Investigation 6

Are Gender and Pay Related?

Correlation

Overview

This investigation continues to explore relationships between two quantitative variables. It focuses on determining whether there is an association between two quantitative variables by looking for patterns in scatterplots and describing the relationship between two quantitative variables by finding and interpreting the correlation coefficient.

The next two investigations focus on these relationships by exploring the variability about the least squares regression line using two tools: correlation coefficient and residual plots.

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 activity.

Instructional Plan

Brief Overview

Part I: Introduce the three types of association.

Part II: Develop the concept of the correlation coefficient.

Part III: Interpret the correlation coefficient in context.

Part I: Association

Hand out Student Worksheet 6.1 Patterns in Scatterplots. Ask your students to answer questions 1 and 2.

Do you own an MP3 player? Or do you have a lot of songs stored on your cell phone? How much memory are the songs using on the device?

Do you think there is a relationship between the length of a song in minutes and the size of the file on an MP3 player?

The scatterplot in Figure 6.1 shows the file sizes (in MB) for songs of different lengths (in seconds).

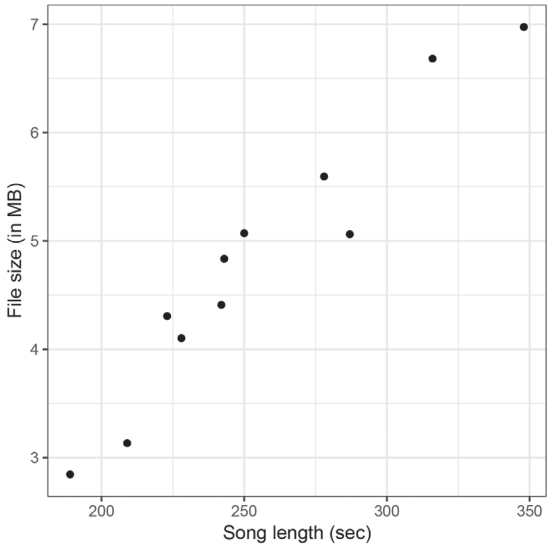


Figure 6.1: File sizes for songs of different lengths

1. What trends do you observe in the scatterplot?

Possible answer: *Longer songs (seconds) tend to use more space (MB).*

2. As the song increases in length, what is happening to the size of the file?

Possible answer: *The file size is increasing.*

Explain that, on this scatterplot, the length of the song is called the *independent variable* or *explanatory variable* and the size of the file is called the *dependent variable* or *response variable*.

Independent Variable or Explanatory Variable The explanatory variables are those variables that influence changes in the response variable.

Learning Goal

Represent data of two quantitative variables on a scatterplot and discuss whether the two variables are related. Compute (using technology) and interpret the correlation coefficient of a linear fit.

Mathematical Practices Through a Statistical Lens

MP7. Look for and make use of structure.

Students use structure to separate the 'signal' from the 'noise' in a set of data—the 'signal' being the structure, the 'noise' being the variability. They look for patterns in the variability around the structure and recognize that these patterns can often be quantified.

Materials

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

» Statistical technology that can do the following:

  - Create a scatterplot
  - Calculate Pearson's correlation coefficient

» Student Worksheet 6.1 Patterns in Scatterplots

» Student Worksheet 6.2 Gender and Pay

» Student Worksheet 6.3 Graphs to Illustrate Association

» Exit Ticket

Estimated Time

Two 50-minute class periods: One period for introducing correlation (parts I and II) and another period to apply the concept of correlation (Part III)

Pre-Knowledge

Students should be able to construct scatterplots using technology.

Dependent Variable or Response Variable
The response variable is the observed result of the explanatory variable being manipulated.

When describing the scatterplot, statisticians generally list the dependent variable (located on the y-axis) vs. the independent variable (located on the x-axis). For example, statisticians would describe the previous scatterplot as file size vs. length of song.

There are three types of association when describing a scatterplot: positive association, negative association, and no association.

Note: Figure 6.2 (Flight time vs. Distance), Figure 6.3 (Olympic 100 m freestyle time vs. Years since 1900), and Figure 6.4 (Gross earnings vs. movie running time) can be found on Student Worksheet 6.3 Graphs to Illustrate Association.

Display Figure 6.2, an example of a scatterplot showing a **positive association**.

The data show the distance in miles and time in minutes for a sample of United Airline nonstop flights from Chicago to various cities west of Chicago.

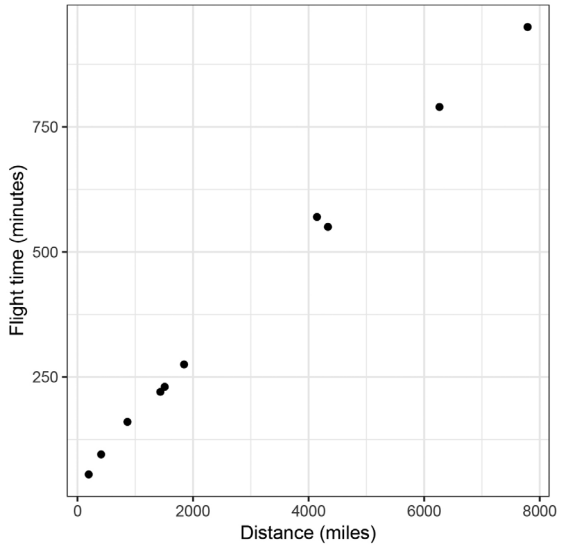


Figure 6.2: A scatterplot showing a positive association

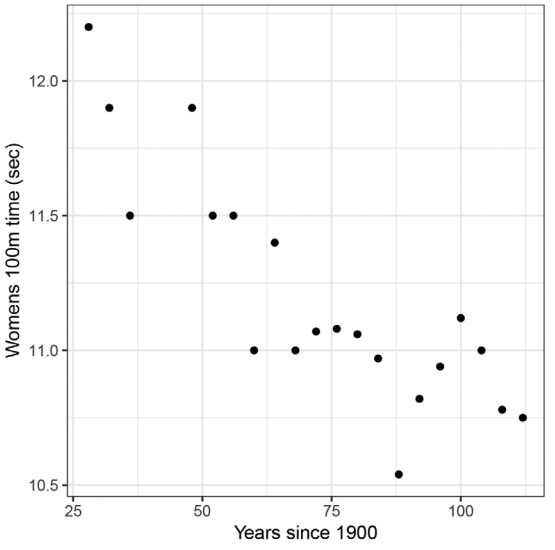


Figure 6.3: A scatterplot showing a negative association

A positive association occurs when large values of one variable are associated with large values of the other and small with small.

Or

As the distance from Chicago increases, the flight time increases.

Display Figure 6.3, an example of a scatterplot showing a **negative association**.

The data show Olympic Women's 100 m freestyle times since 1912.

A negative association occurs when the values of one variable tend to decrease as the values of the other variable increase.

Or

As the years have increased from 1912, the Olympic times in the women's 100 m freestyle have decreased.

Display Figure 6.4, an example of a scatterplot showing **no association**.

Data are the running times (minutes) and gross receipts for a selected group of movies.

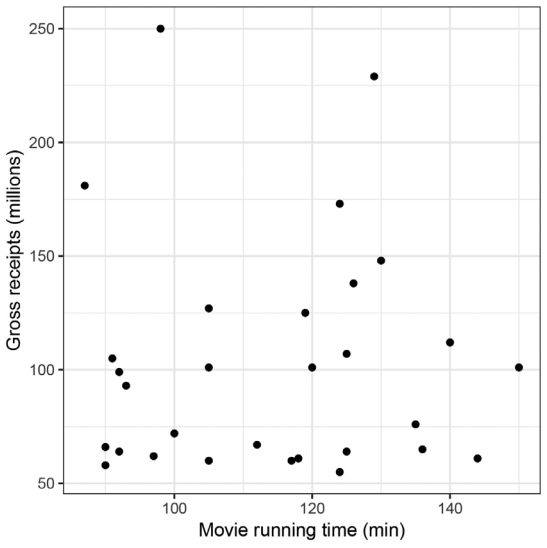


Figure 6.4: A scatterplot showing no association

No association occurs when one variable increases and there is no pattern for how the other variable reacts.

Or

As the running time of a movie increased, the amount of money the movie grossed was hard to predict—sometimes it was large, sometimes it remained fairly constant.

Ask your students to answer Question 3.

3. How would you describe the relationship between the length of a song and the size of the file?

Answer: *There is a positive association. As the song length increases, the file size increases.*

This is a short activity to help solidify student understanding of association. Ask students to stand up. Explain that sets of variables will be shared and students should raise both arms if they think there is a positive correlation between the variables, put their arms straight out if they think there is no correlation, and put their arms down at their sides if they think there is a negative correlation. (These variables could be altered to reflect the interests of your students.)

» Reading achievement vs. IQ
Positive Association

» Test scores vs. level of anxiety
Negative Association

» Weight gain vs. amount of exercise
Negative Association

» Math achievement vs. reading achievement
Positive Association

» Math achievement vs. athletic achievement
No Association

» Lifetime earnings vs. years of schooling
Positive Association

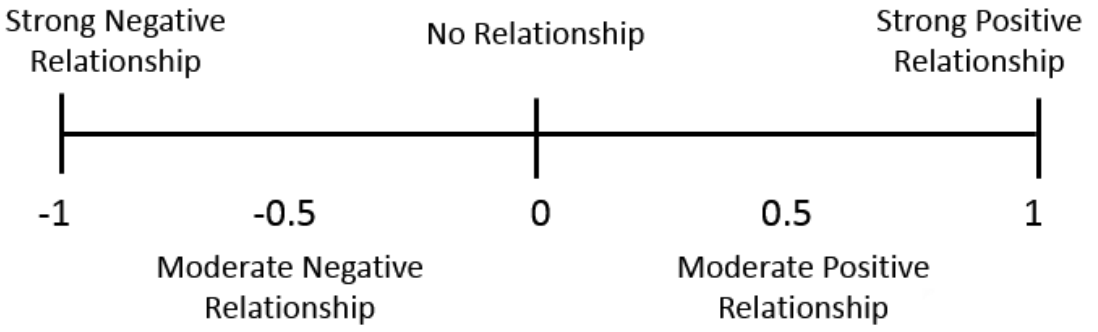
» Number of texts sent/received in a day vs. number of Facebook friends
Could be arguments for all three answers—most likely No Association or Positive Association

Have a brief discussion about the amount of time it took students to position their arms. Some of the sets of variables may have been easier than others for students to make decisions. The amount of time to decide about correlation is related to the strength of the correlation.

Note: It is important to point out to your students that association does not mean causation. A positive association between reading achievement and math achievement does not mean high reading achievement causes high math achievement.

Part II: Correlation Coefficient

Explain that our goal is to find a mathematical model to describe the relationship between two quantitative variables. For example, what mathematical model could be used



to describe the association between the length of a song and its file size?

Answer: *A linear model makes sense because the data appear to fit close to a line.*

If the association appears linear, then the "tightness" of the data points to the line fitted to the data can be measured. This value is called the correlation coefficient.

A correlation coefficient is a number that measures the direction and strength of a linear relationship between two quantitative variables. One such measure is called Pearson's correlation coefficient, represented by the letter *r*.

Note: Explain that the convention is to call this value "r." The letter r was used because Sir Francis Galton, who developed the concept of regression, originally used r to describe the slope of the regression line. While Pearson developed the formula we use today, the use of r stuck with correlation coefficient.
Source: www.buttelake.com/corr.htm

Explain that the closer Pearson's correlation coefficient is to +1, the stronger the positive linear relationship. The closer Pearson's correlation coefficient is to -1, the stronger the negative linear relationship.

Display the diagram at the top of the page.

Note: For more information about the development of the formula for Pearson's correlation coefficient, see the Further Explorations and Extensions section at the end of this investigation.

The next step is to have students estimate the correlation coefficient for the following four scatterplots. The plots are the same graphs used to introduce association. They are found on Worksheet 6.3 Graphs to Illustrate Association.

Note: Students could be asked to place themselves along an imaginary number line in the room to estimate the correlation coefficient.

Display the scatterplot of flight time versus distance from Chicago again and ask students to estimate what the correlation coefficient might be.

Answer: *The correlation coefficient is 0.999.*

Display the scatterplot of the Olympic times versus years since 1900 again and ask students to estimate what the correlation coefficient might be.

Answer: *The correlation coefficient is -0.95.*

Display the scatterplot of the gross receipts versus movie run times again and ask students to estimate what the correlation coefficient might be.

Answer: *The correlation coefficient is 0.01.*

Display the scatterplot of the length of song and file size versus length of song again and

ask students to estimate what the correlation coefficient might be.

Answer: *The correlation coefficient is 0.96.*

Optional: The websites shown below can provide extra practice for students to get a good feel for what different values of r look like.

www.istics.net/Correlations: Students are presented four scatterplots and asked to match each correlation coefficient to the correct graph.

http://guessthecorrelation.com: A game in which students guess the correlation and compete for high scores. Students can also play against each other by entering the competitor's user name.

Part III: Investigating a Scenario

Note: This part of the lesson is based on data from *www.census.gov*, obtained through Core Math Tools at *www.nctm.org/Classroom-Resources/Core-Math-Tools/Core-Math-Tools*. Another website that provides quite a bit of information, including a breakdown by state, is the American Association of University Women at *www.aauw.org/research/the-simple-truth-about-the-gender-pay-gap*. Students might be interested in doing more research on this topic.

Hand out Student Worksheet 6.2 Gender and Pay. Have students read the scenario.

Scenario

> Did you know that in 2016, women working full time in the United States typically were paid just 80 percent of what men were paid, a gap of 20 percent? The gap has narrowed since the 1970s, due largely to women's progress in education and workforce participation and to men's wages rising at a slower rate. Still, the pay gap does not appear likely to go away on its own. At the rate of change between 1960 and 2016, women are expected to reach pay equity with men in 2059. But even that slow progress has stalled in recent years. If change continues at the slower rate seen since 2001, women will not reach pay equity with men until 2119. *Source: www.aauw.org/research/the-simple-truth-about-the-gender-pay-gap*

Note: Another interesting article for students to explore the gender pay gap is by Amanda Golbeck and titled "How One Woman Used Regression to Influence the Salaries of Many." It tells the story of Elizabeth Scott, the Berkeley statistics professor who spent two decades analyzing inequities in academic salaries and advocating for change. Find it at *http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2017.01092.x/epdf*.

After students have read the scenario, ask them to brainstorm about why this gap might exist. Ideas that might surface include type of job, education level, experience, age, and race. Some students might debate whether the pay gap between gender exists, as some claim the 20% is inflated. Some sites to explore are the following:

» *www.huffingtonpost.com/christina-hoff-sommers/wage-gap_b_2073804.html*

» *www.bls.gov/opub/reports/womens-earnings/archive/womensearnings_2009.pdf*

Formulate a Statistical Question

Point out to students that, in the scenario, the gap was measured as a percent, calculating the percent women were paid compared to men. This is often referred to as women's to men's earnings ratio. For this investigation, this value will simply be referred to as the "earnings ratio."

After the discussion concerning the pay gap, discuss possible statistical questions students could investigate. To help the discussion, remind them they need to identify the populations of interest and how these populations could be compared.

Table 6.1 Answers to Question 1

Years since 1970	Median Income Men ($)	Median Income Women ($)	Earnings Ratio
0	9,184	5,440	0.592
5	12,934	7,719	0.597
10	19,173	11,591	0.605
15	24,999	16,252	0.650
20	28,979	20,591	0.711
25	32,199	23,777	0.738
30	38,891	29,123	0.749
35	42,188	33,256	0.788
39	49,164	37,234	0.757
45	50,119	40,022	0.799

On Student Worksheet 6.2 Gender and Pay, ask your students to write a possible statistical question. Have them compare with others in their group.

In this investigation, the statistical question we will be investigating is: "To what extent are the median income of US males and the median income of US females related?"

Collection of Data

Explain how the earnings ratio is calculated for 1970 (0 years since 1970).

Ask your students to interpret the earnings ratio in 1970.

Possible answer: *The earnings ratio is 5440/9184, also written as approximately 0.592, which means women earned about $0.59 for every $1 men earned in 1970.*

Analyze the Data

Ask your students to complete questions 1 to 4 on the student worksheet.

1. Find the remaining earnings ratios.

See answers in Table 6.1.

2. Interpret the earnings ratio for 45 years since 1970.

Possible answer: *Women earned about $0.80 for every $1 men earned in 2015.*

3. What trends do you observe in the earnings ratio over time?

Possible answer: *The earnings ratio has increased since 1970.*

4. Could the earnings ratio exceed 1? What would that mean?

Possible answer: *The earnings ratio would exceed 1 when the median income for women exceeds the median income for men.*

Explain that the response variable will be the earnings ratio, as this amount is changing over time. Time (years since 1970) is the explanatory variable, as it is possibly influencing the change in the earnings ratio.

Ask your students to complete Question 5 on the student worksheet.

5. Use technology to create a scatterplot of time and earnings ratio. Sketch a copy of the scatterplot.

Answer: *Figure 6.5.*

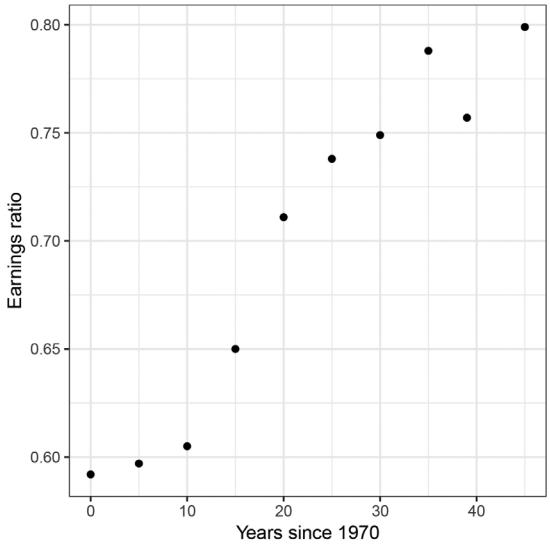


Figure 6.5: A scatterplot of time and earnings ratio

Ask students to complete questions 6 to 8 on the student worksheet.

6. Describe the relationship between the earnings ratio and time since 1970.

Possible answer: *There is a positive association. As the years since 1970 increases, the earnings ratio increases. In other words, the ratio of the median earnings for women to the median earnings for men is getting larger, thus the gap between pay based on gender is diminishing over time.*

7. Estimate r, the correlation coefficient.

Possible answer: *Approximately 0.95.*

8. Do you think it would be appropriate to draw a line through the data?

Possible answer: *Although there is some curvature, it appears a linear function could be used to model the relationship between earnings ratio and time since 1970. The fit should be relatively strong.*

Explain that the correlation coefficient can easily be found using technology. Graphing calculators, spreadsheets, and statistical software/applications all have the capability of finding the value of r. Demonstrate how to find the value of r. Ask students to complete Question 9.

9. Use technology to find the value of r and interpret this value in terms of the data.

Possible answer: *r is approximately 0.96. The data fit very tightly to a line. The earnings ratio and time have a strong positive correlation.*

Interpret the Results in the Context of the Original Question

Ask students to revisit or restate the statistical question: "To what extent are the median income of males and the median income of females related?"

Ask students to discuss with a partner or small group how they would answer the question based on the scatterplot and correlation coefficient, and then write a few sentences to answer the statistical question based on the analysis.

Optional: Ask students to comment on possible social implications of the results or other data they might be interested in collecting that might influence the answer to the statistical question.

Example: A scatterplot of the data shows a positive linear pattern. The correlation coefficient is approximately 0.963, indicating a strong positive correlation between the years since 1970 and the earnings ratio of median incomes of women to men. If this pattern continues, the earnings ratio will eventually be 1, indicating no gender gap in income.

Additional Ideas

This lesson could focus on a different source of data, such as the cost of a 30-second Super Bowl ad and the winning team player's share (data set available through Core Math Tools).

Exit Ticket

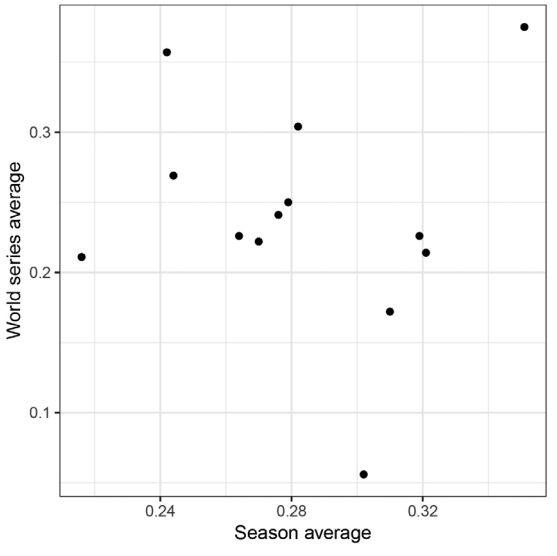
For the following four scatterplots:

» Summarize the relationship between the variables with a sentence.

» Determine if a linear approximation is appropriate. If so, estimate the correlation coefficient.

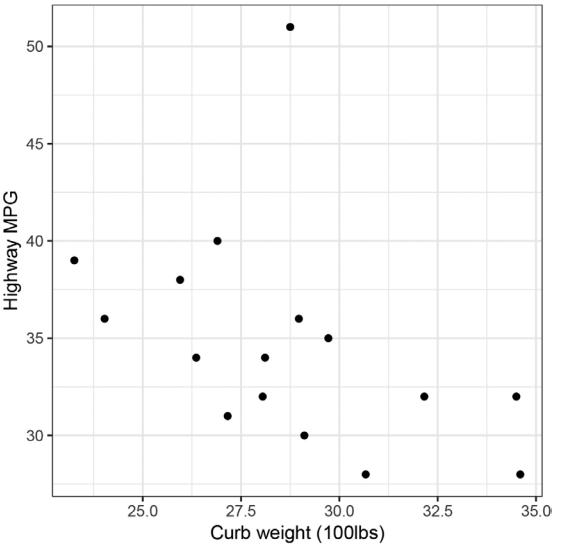
All data are sourced from Core Math Tools, available from *www.nctm.org/Classroom-Resources/Core-Math-Tools/Core-Math-Tools.*

1. World Series Average vs. Regular Season Average



Possible answer: *If a linear relationship exists, it would appear to be weak and maybe negative, indicating there is little to no linear relationship between a player's World Series batting average and regular season batting average. The actual value for r = -0.03.*

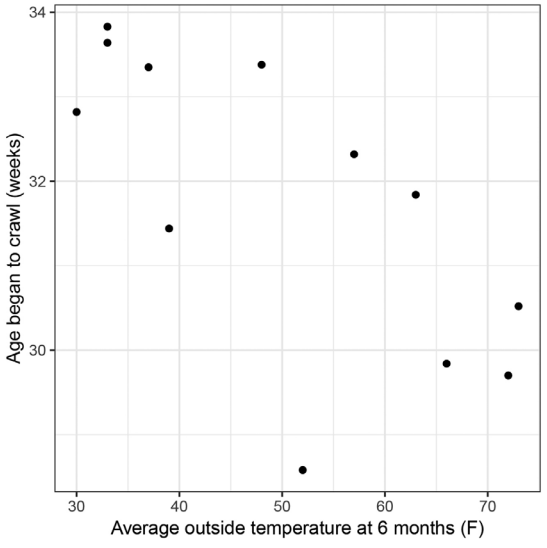
2. Highway mpg vs. Curb Weight for Selected Cars



Possible answer: *It appears there might be a moderate to weak negative correlation between highway mpg and curb weight in pounds for selected compact cars. The actual value for r = -0.43.*

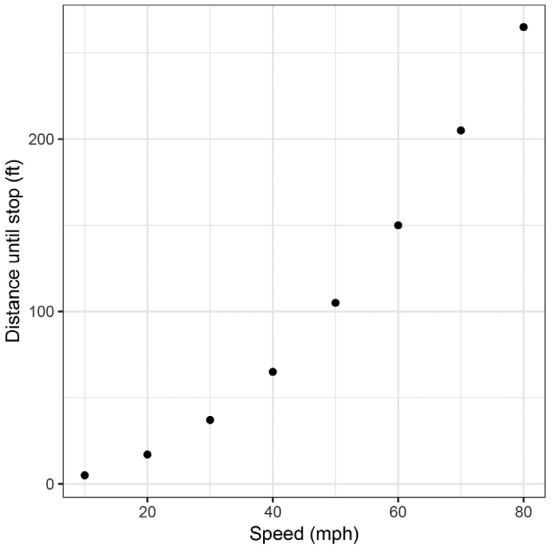
Exit Ticket Cont.

3. Age (Months) Baby Began to Crawl vs. Average Outside Temperature



Possible answer: *It appears there is a moderate negative linear relationship between the age babies began to crawl and the average outside temperature. The actual value for r = -0.7.*

4. Distance (feet) Until Car Stopped vs. Speed (mph)



Possible answer: *It appears there is a moderate to strong positive linear relationship between speed and the distance until stopped, although the scatterplot appears to have a slight curve. The actual value for r = 0.97.*

Further Explorations and Extensions

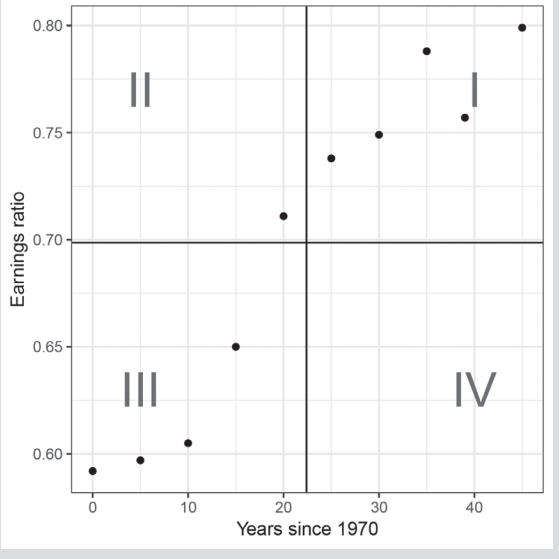
The formula for calculating Pearson's correlation coefficient is:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

In the formula, X_i and Y_i are the individual data points, $\bar{X}$ and $\bar{Y}$ are the means of the explanatory values and response values, and s_x and s_Y are the standard deviations of the explanatory values and response values. This formula is not necessary for students to memorize or calculate by hand.

The following explanation can be used to develop the conceptual understanding of this formula. Using technology, draw a horizontal line through the mean of the earnings ratio and draw a vertical line through the mean of the years since 1970.

Number each of the four quadrants, I (upper right), II (upper left), III (lower left), and IV (lower right).



Ask your students what they observe in terms of the number of points in each region.

Possible response: *Almost all the points are in regions I and III.*

Further Explorations and Extensions Cont.

Have students focus on the points in Region I. Ask what is true about the points in region I as compared with the point ($\bar{X}$ and $\bar{Y}$).

Answer: *These points have x values greater than the mean of the incomes for men and y values greater than the mean of the incomes for women.*

Point to the correlation coefficient formula. Ask how the ordered pairs in Region I result in a positive component in the sum part of the formula.

Answer: *The $(X_i - \bar{X})$ will be positive and $(Y_i - \bar{Y})$ will be positive so the product will be positive.*

Ask students what will happen in the formula for the points in Region III.

Answer: *The ordered pairs in Region III result in a positive component in the sum part of the formula. The $(X_i - \bar{X})$ will be negative and the $(Y_i - \bar{Y})$ will be negative, so the product will be positive.*

Ask students what will happen in the formula for the points in Region IV.

Answer: *The ordered pairs in Region IV result in a negative component in the sum part of the formula. The $(X_i - \bar{X})$ will be positive and $(Y_i - \bar{Y})$ will be negative, so the product will be negative.*

Ask students what will happen in the formula for the points in Region II.

Answer: *The $(X_i - \bar{X})$ will be negative and $(Y_i - \bar{Y})$ will be positive, so the product will be negative.*

Ask what the sign of the sum of all the products will be. How does this relate to the graph? Explain.

Answer: *The sign of the sum of all the products will be positive because there are only positive values since all points but one lie in quadrants I and III.*

Ask students when a correlation coefficient would be close to zero.

A correlation close to zero occurs when there are offsetting positive and negative products, resulting in a sum close to zero and no apparent trend in the data.

Investigation 7

Are Gender and Pay Related? Continued

Assessing Linear Fit

Overview

This investigation continues to explore relationships within bivariate data, using the same data set as in Investigation 6, which explored the relationship between median income and gender.

The previous investigation focused on using the correlation coefficient to determine the strength of the linear relationship between two quantitative variables (earnings ratio vs. years since 1970). This investigation builds on the idea of error developed in Investigation 5 (the amount of error between the actual and guessed number of calories in a small candy bar) and furthers this concept by defining residuals and exploring the use of residual plots as a tool to determine the appropriateness of using a line of fit for bivariate data.

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 activity.

The Analyze the Data section of the student worksheet is divided into three parts. The first part is optional, depending on the background knowledge of the students. It allows students to fit a line to data, determine the equation of the line, interpret the slope, and introduce the concept of error. The second part of this section uses statistical software to demonstrate the development of the concept of a residual. The third part continues to use statistical software to demonstrate how to construct residual plots.

Instructional Plan

Brief Overview

» Continue with the statistical question: "To what extent are the median income of males and the median income of females related?"

» Create linear models to fit a data set and explore the least squares regression line.

» Define residuals and create and use a residual plot to determine whether a linear model is appropriate for a data set.

» Answer the statistical question.

Part I: Fitting a Line to Data

Scenario

The scenario is the same as was explored in Investigation 6, which should be completed before this investigation.

Formulate a Statistical Question

(from Investigation 6)

Continue with the statistical question: "To what extent are the median income of males and the median income of females related?"

Collection of Data

(from Investigation 6)

Learning Goals

» Interpret the slope (rate of change) and intercept (constant term) of a linear model in the context of the data.

» Represent data on two quantitative variables on a scatterplot and describe how the variables are related.

» Fit a linear function for a scatterplot that suggests a linear association.

» Informally assess the fit of a function by plotting and analyzing residuals.

Mathematical Practices Through a Statistical Lens

MP7. Look for and make use of structure.

Students use structure to separate the 'signal' from the 'noise' in a set of data—the 'signal' being the structure, the 'noise' being the variability. They look for patterns in the variability around the structure and recognize these patterns can often be quantified.

Materials

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

Part I

» Straightedge (ruler or piece of spaghetti)

» Student Worksheet 7.1 Fitting a Line to Data (Used for Part I of Analyze the Data section)

Parts II and III

» Statistical technology that can do the following:

- Create a scatterplot
- Create horizontal and vertical lines at defined points
- Create a moveable line
- Show the residuals, squares of the residuals, and sum of the squares of the residuals
- Create the least squares regression line

(Core Math Tools is one piece of statistical software that can be used for demonstration: www.nctm.org/Classroom-Resources/Core-Math-Tools/Core-Math-Tools or http://nctm.org/resources/cmt/CoreMathTools.jar)

» Exit Ticket

Estimated Time

Two 50-minute class periods for parts I, II, and III. Part I is optional, depending on background knowledge of students.

One 50-minute class period for parts II and III. If only doing parts II and III, Part II may take a little longer (possibly 1.5 class periods total), depending on the comfort level of students with the technology being used.

Pre-Knowledge

Students should be able to:

» Construct scatterplots

» Given two ordered pairs, write the equation of a line in slope-intercept form

» Interpret the slope of a line in context of the given data

Table 7.1

Years since 1970	Median Income Men ($)	Median Income Women ($)	Earnings ratio
0	9,184	5,440	0.592
5	12,934	7,719	0.597
10	19,173	11,591	0.605
15	24,999	16,252	0.650
20	28,979	20,591	0.711
25	32,199	23,777	0.738
30	38,891	29,123	0.749
35	42,188	33,256	0.788
39	49,164	37,234	0.757
45	50,119	40,022	0.799

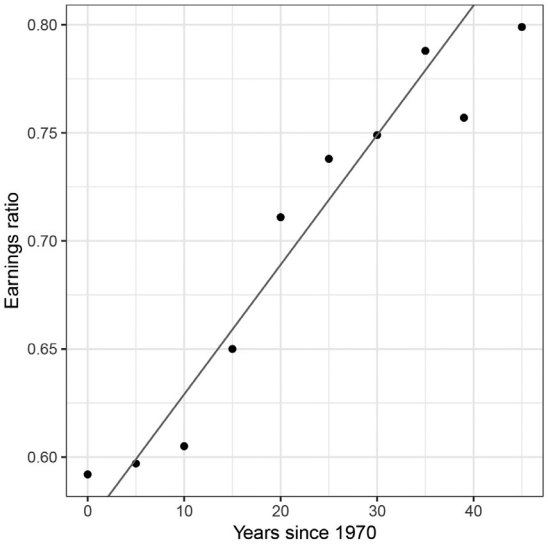


Figure 7.1: Scatterplot with example of line drawn

Analyze the Data

Part I: Informally Fitting a Line to Data

Note: If students have had experiences fitting a line to data, writing the equation of the line, and interpreting the slope of the line in context, this part can be skipped.

Distribute Student Worksheet 7.1 Fitting a Line to Data and a straightedge (piece of spaghetti or ruler).

Ask your students to complete numbers 1 to 5.

Example: *Figure 7.1*

1. Using a straightedge, draw a line on the scatterplot you think will summarize or fit the data. Explain how you decided to draw your line.

Possible answers:

» *Balanced same number of points above and*

below the line (misconception)

» *Connected a point in the bottom left corner to a point in the upper right corner*

» *Found the mean of the x's and y's to create a midpoint first, then fit a line through that point (unlikely but a great strategy)*

2. Locate two points on the line. Write each point as an ordered pair.
3. Using the two points, write the equation of your line in slope-intercept form.

Possible answer: *$\hat{y} = 0.006x + 0.569$*

4. Interpret the slope of your line in terms of the context.

Possible answer: *The slope means that for each additional year, the earnings ratio is predicted to increase about 0.006, on average. The earnings ratio is the median women's income to the median men's income, which means the women's income is increasing by $0.006 for every $1 of men's income each year.*

5. Interpret the y-intercept in terms of the context.

Possible answer: *The y-intercept of 0.569 would mean the predicted earnings ratio is 0.569 in 1970.*

Discuss the answers for questions 1 to 5.

Ask students to complete questions 6 to 10 on Student Worksheet 7.1.

6. Using 25 years from 1970 as your x value and the equation of your line, what is your prediction for the earnings ratio?

Possible answer: *0.006(25) + 0.569 = 0.719*

7. Look at the data table to find the actual earnings ratio of 25 years from 1970.

Answer: *0.738*

8. What is the difference between the actual earnings ratio for 25 years from 1970 and the earnings ratio your line predicted?

Possible answer: *0.738 – 0.719 = 0.019*

9. How well did your line predict the earnings ratio for 25 years from 1970? Did your line overestimate or underestimate the earnings ratio?

Possible answer: *close estimate but slightly underestimated*

10. Compare your prediction with others in class. Who was the best predictor of the earnings ratio for 25 years since 1970?

Possible answer would be to use the same strategy used in Investigation 5 to find the smallest error of the guesses.

Discuss answers to questions 6 to 10 with the whole class.

Note: If students need more experiences fitting a line to data, see Section 5 of *Bridging the Gap Between Common Core State Standards and Teaching Statistics* by Hopfensperger, Jacobbe, Lurie, and Moreno, published by the American Statistical Association in 2012 and available at *ww2.amstat.org/education/btg/index.cfm*.

Part II: Fitting a Line to Data Using Technology

Note: Part II is best done as a demonstration on a screen with the whole class while individual students or pairs of students have access to the same technology on a device. One suggestion is to have a student lead the class on the technology while the teacher gives supporting directions.

A free statistical application that can be used for the demonstration is Core Math Tools, available at *www.nctm.org/Classroom-Resources/Core-Math-Tools/Core-Math-Tools* or *http://nctm.org/*

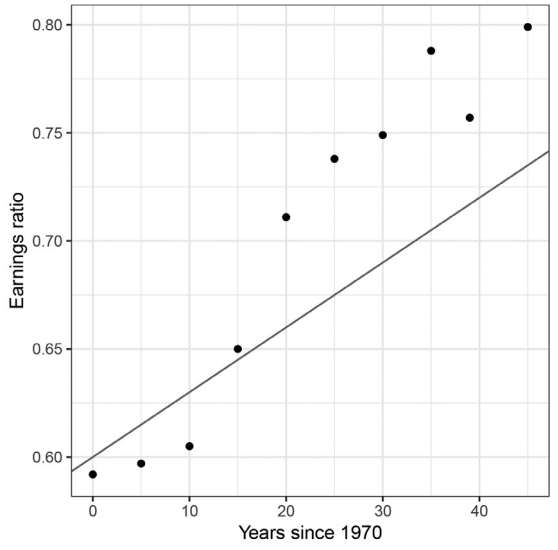


Figure 7.2: Scatterplot of earnings ratio vs. years since 1970 with a moveable line

resources/cmt/CoreMathTools.jar. However, any statistical technology can be used, as long as it meets the requirements in the materials list.

Other free tools:

» *www.rossmanchance.com/applets/RegShuffle.htm*

» *www.stapplet.com*

» *www.lock5stat.com/StatKey*

Before the demonstration, the data (years since 1970 and corresponding earnings ratio) need to be entered into the statistical software being used for the demonstration with years since 1970 as the independent and earnings ratio as the dependent variable.

Display the scatterplot of earnings ratio vs. years since 1970.

Remind your students that they determined there was a strong positive correlation (r=0.963) and the data appear to fit a line in Investigation 6.

Explain that the next step in the analysis of the data is to find the equation for a line that best fits the data.

Explain that we first want to draw an approximate line of fit using technology. To begin, we are going to "eyeball" a line.

Demonstrate how to add a moveable line to the scatterplot (See Figure 7.2).

After the line is on the graph, point out the equation of the line. (In this example, the equation is $\hat{y}$ = 0.006x + 0.569). Move the line around, changing both the position and the slope.

Note: Usually, the line can be moved by dragging the small square near the center of the line (changes the intercept) and the small squares toward the ends of the line (changes the slope). The equation of the line is in the upper right corner of the screen.

Ask students to direct which way(s) to move the line to fit the data.

Once students have decided the line is in the "best" place, have them interpret the slope and y-intercept of their line in context.

Possible answer: *For the example above, the slope means the earnings ratio is predicted to increase about 0.006, on average, for each additional year. The earnings ratio is the median women's income to the median men's income, which means the women's income is increasing by $0.006 for every $1 of men's income each year. The y-intercept of 0.569 would mean the predicted earnings ratio is 0.569 in 1970.*

Point out that this equation of the line of fit can be used to make predictions.

Write the equation of the line on the board using the symbol $\hat{y}$. For example, $\hat{y}$ = 0.006x+0.569. The symbol $\hat{y}$ (y-hat) stands

for the *predicted* value for y for a given value of x.

Ask your students to use the equation of the line and predict the value of the earnings ratio for 25 years since 1970.

Possible answer: *0.006(25) + 0.569 = 0.719*

Ask your students how well their line predicts the earnings ratio for 25 years from 1970. Did their line overestimate or underestimate the earnings ratio?

Possible answer: *Close estimate but slightly underestimated*

Ask your students if they think they have found the "best" line. Ask them how they could decide if they have the "best" line.

Ask your students how they determined who was the "best guesser" when guessing the number of calories in candy bars (Investigation 5)?

Possible answer: *Found the difference between actual calories and guessed number of calories, squared the differences, and found the sum of the differences. The lowest sum of the squared differences was determined to be the "best" guesser.*

Explain that this same strategy will be used to determine the line of "best" fit.

Introduce the term "residual" by sharing the definition and a drawing.

A *residual* is the vertical (signed) distance between a data point and the graph of the regression equation. The residual is positive if the data point is above the graph. The residual is negative if the data point is below the graph. The residual is 0 only when the graph passes through the data point.

Notation: (x_i,y_i) is used to represent any data point where *i* represents the i^{th} data point. For example, (x_2,y_2) is the second data point. The

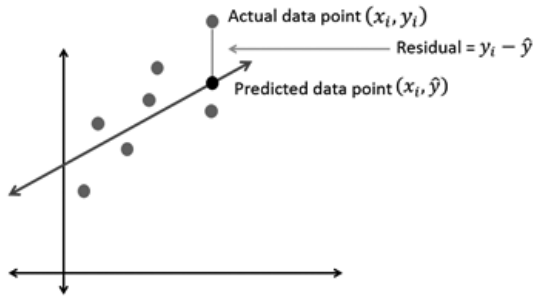


Figure 7.3: A drawing of a residual

notation of $\hat{y}$ (read as "y-hat") represents the predicted y-value from the graph of the regression line. Thus, the residual is $y_i - \hat{y}_i$, or *observed value – predicted value* (See Figure 7.3).

Explain that the line of best fit, referred to as the *least squares regression line,* is defined as the line that has the lowest sum of the squared residuals, $\sum (y_i - \hat{y}_i)^2$.

Use the software and display all the residuals on the moveable line. As the line moves, the residuals should change (See Figure 7.4).

Show the squares that are formed by the residuals (See Figure 7.5).

Remind your students they determined the best guesser of calories by finding the sum of

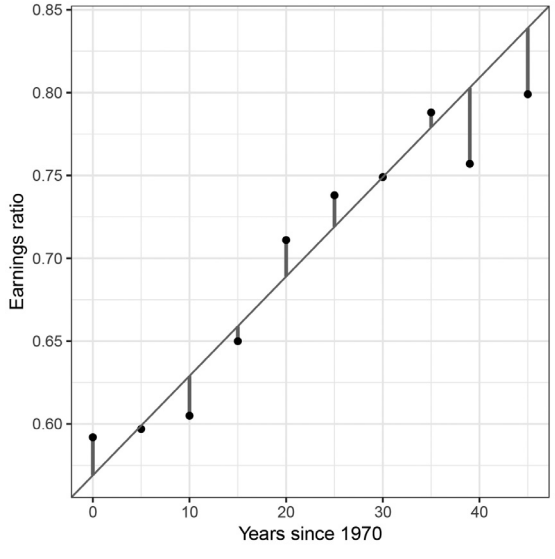


Figure 7.4: The residuals on the moveable line

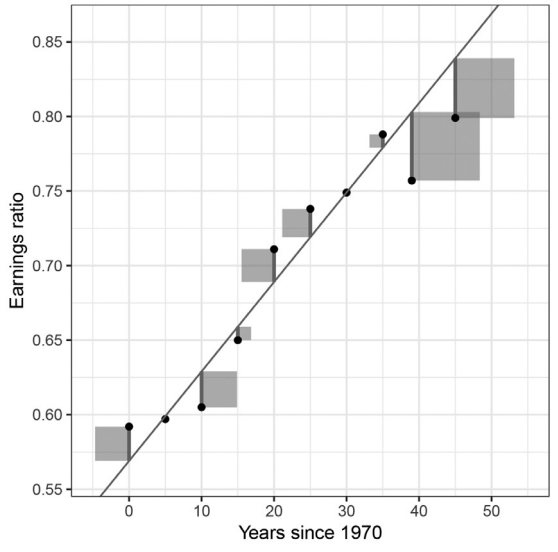


Figure 7.5: The squares formed by the residuals

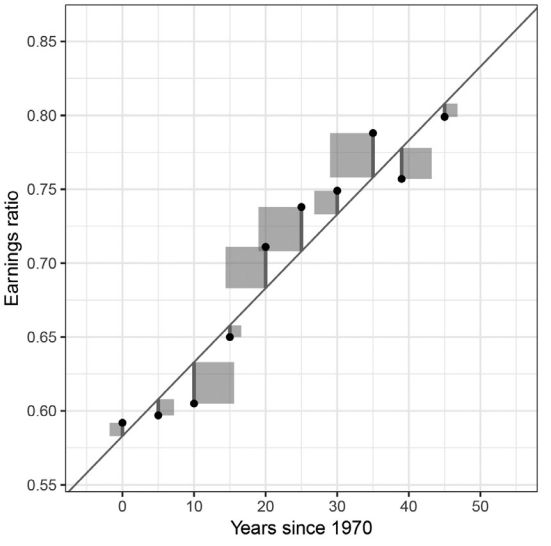


Figure 7.6: The least squares regression line

the squared differences between their guesses and the actual number of calories. Explain that, in a similar fashion, we want to find the sum of the squared residuals or the sum of the areas of the squares shown on the graph to find the "best" line.

Use the software to calculate and display the sum of the squared residuals.

Demonstrate that, as the line moves, the squares change and the sum of the residuals squared changes. Continue to move the line until students think the lowest sum of residuals squared has been found. If students are working on technology, allow them time to explore and find what they think is the lowest sum of squared residuals.

Explain that the software will actually find the line with the lowest sum of squared residuals.

Remove the moveable line. Display the least squares regression line. Show the squares and point out the sum of the squared residuals. Ask how well the class did in finding the least squares line using the moveable line feature of the software (See Figure 7.6).

Point out the equation $\hat{y} = 0.005x + 0.583$ or *predicted earnings ratio* = 0.583+0.005(*years since* 1970) is the *least squares regression line*. This is the line that minimizes the sum of the squared residuals. The symbol $\hat{y}$(y-hat) stands for the predicted value for y for a given value of x.

Part III: Residual Plots

When investigating the relationship between two quantitative variables, the first step was to construct a scatterplot and look for any relationship between the variables. If it appeared linear was an appropriate model, the correlation coefficient was used to determine the strength of a linear association. To get a closer look at the deviations from the line that may not be seen in the scatterplot, the next step is to construct a residual plot.

> A ***residual plot*** is a graph that shows the residuals on the vertical axis and the explanatory variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal line through the residual = 0 or $\hat{y} = 0$, a linear regression

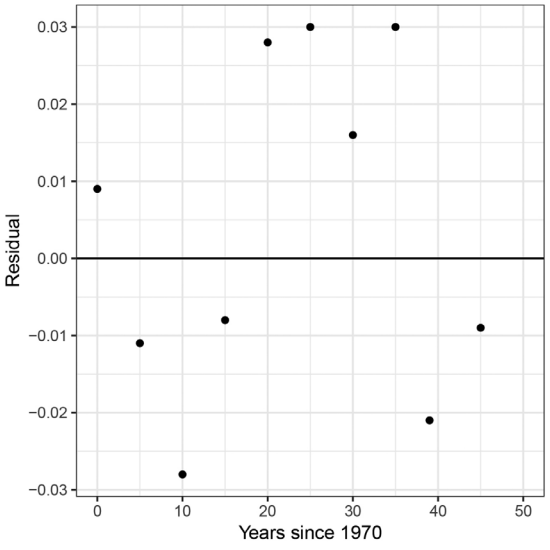


Figure 7.7: A residual plot

model is appropriate for the data; otherwise, a nonlinear model is more appropriate.

Create a residual plot using the statistical software (See Figure 7.7).

If a linear model is a good fit for the data, then the residual plot should show a random dispersion around the line, residual = 0, or $y_i - \hat{y}_i = 0$.

For this set of data, the residual plot shows no pattern in the residuals, further indicating a linear model is appropriate for the data.

Choose a point on the residual plot and ask students to explain what this point represents and how it is connected to the data set and least squares regression line. It might be useful to show the residual plot next to the graph that shows the residuals on the least squares regression line and point out where each residual can be seen on each graph (See Figure 7.8). Ask students to explain what the graph represents.

Interpret the Results in the Context of the Original Question

Ask your students to restate the statistical question. To what extent are the median income of males and median income of females related?

Ask your students to complete problems 11 and 12.

11. Talk with a partner and then write a paragraph to answer the statistical question based on the analysis that refers to the residuals.

Optional: What might be some social implications of the results?

Students might just add to the paragraphs they wrote in Investigation 6.

Possible answer: *A scatterplot of the data shows a positive linear pattern. The correlation coefficient is approximately 0.963, indicating a strong positive correlation between the years since 1970 and earnings ratio of median in-*

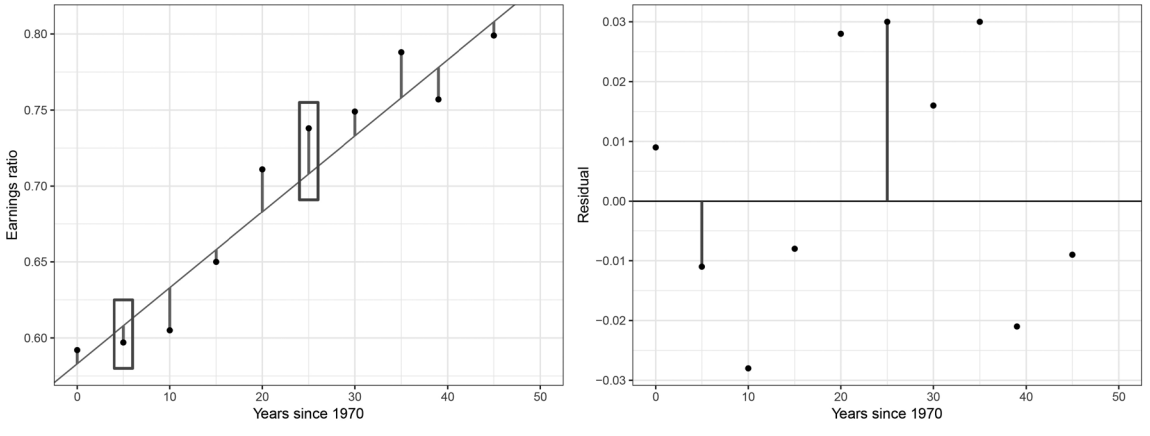


Figure 7.8: Residual plot next to the graph showing the residuals on the least squares regression line

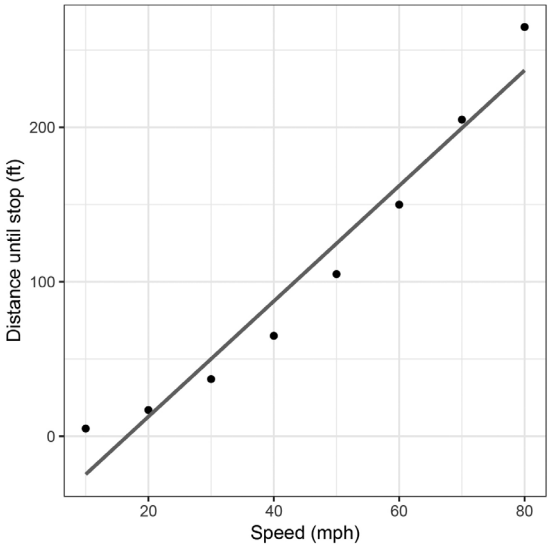


Figure 7.9: A scatterplot showing when linear would not be a good model

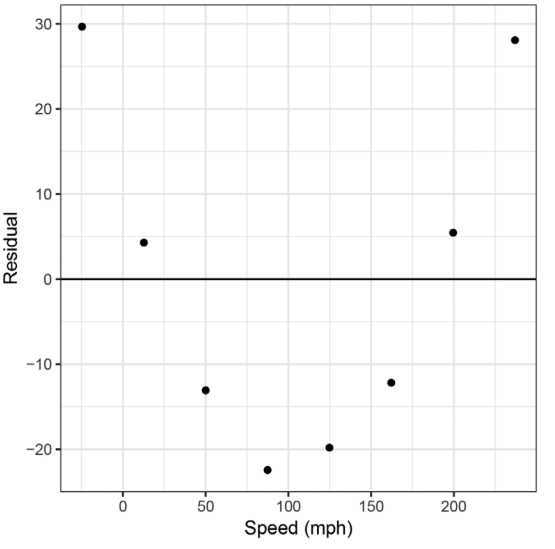


Figure 7.10: A residual plot showing when linear would not be a good model

comes of women to men. The equation for the line of best fit is (predicted earnings ratio) = 0.583+0.005(years since 1970), indicating the earnings ratio increases approximately .005 each year since 1970. The earnings ratio was 0.583 in 1970, according to the linear model. The residual plot shows no pattern, indicating a linear model is appropriate. If the pattern were allowed to continue, the earnings ratio will eventually be 1, indicating no gender gap in income.

12. When is the earnings ratio predicted to be 1? Do you think this will happen?

Ask students to find out when the earnings ratio would be 1 (when men and women have the same median income), as predicted by the line of best fit.

Possible answer: *The line of best fit predicts the earnings ratio will be 1 between 2053 and 2054. Answers will vary about whether this will happen. There are many variables that could affect the earnings ratio in the next 20+ years.*

Note: It might be useful to show a residual plot in which a pattern does occur. The scatterplot (Figure 7.9) and residual plot (Figure 7.10) show an example in which a pattern occurs in the residual plot and linear would not be a good model.

Example: Braking Road Test

Road tests under various conditions are likely to produce data like these that show speed (in mph) and distance until stopping (in feet).

While it appears a linear model might be appropriate from the graph of the data, the residual plot shows a pattern, indicating a linear model is not appropriate. The correlation coefficient can be calculated, and $r \approx 0.976$. This indicates linearity is plausible. Without looking at a residual plot, it might be difficult to tell a linear model is not appropriate.

Additional Ideas

Note that this lesson could focus on a different source of data, such as the cost of a 30-second Super Bowl ad and the winning team player's share (data set available through Core Math Tools).

Exit Ticket

For each scatterplot shown in problems 1 and 2:

a. Sketch a residual plot from the line of best fit
b. Determine if a linear model is appropriate
c. If a linear model is appropriate, estimate the correlation coefficient
d. Summarize the relationship between the variables
e. Interpret the values in the equation of the line of best fit in the context of the situation

Extra: What statistical question might the data answer?

1. Health and Nutrition

The scatterplot in Figure 7.11 shows how average daily food supply (in calories) is related to life expectancy (in years) in a sample of countries in the western hemisphere. The equation of the least squares regression line is $\hat{y} = 0.009x + 4681$.

Source: World Health Organization Global Health Observatory Data Repository, faostat3.fao.org

a. Sketch a residual plot from the line of best fit

Answer: *Figure 7.12*

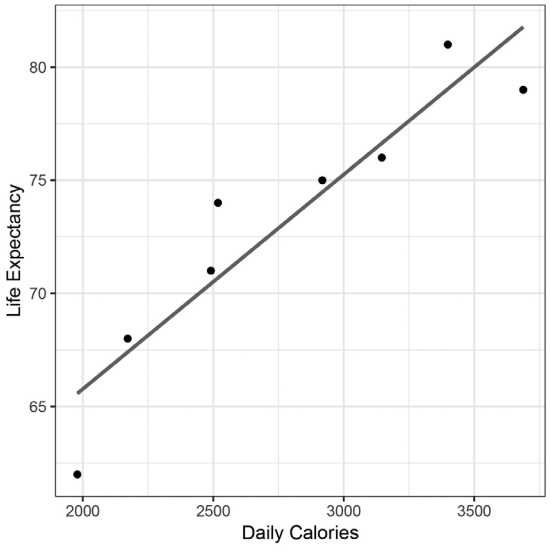


Figure 7.11: Average daily food supply (in calories) related to life expectancy (in years) in a sample of countries in the western hemisphere.

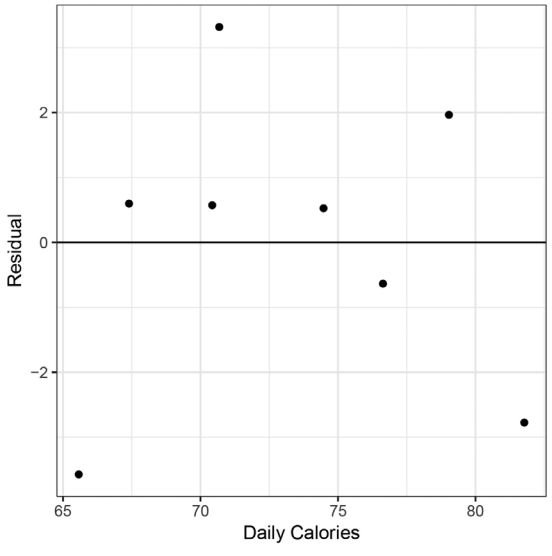


Figure 7.12: Sketch of a residual plot from the line of best fit

Exit Ticket Cont.

b. Is linear an appropriate model? Explain

Possible answer: *The residual plot is fairly scattered and shows no pattern, indicating a linear model is an appropriate model.*

c. Estimate for correlation coefficient

Possible answer: *The correlation coefficient could be estimated at approximately 0.8-0.9, ($r \approx 0.928$).*

d. Summary of the relationship

Possible answer: *The relationship indicated by the graph is a positive linear relationship, indicating the average life expectancy increases as daily caloric intake increases.*

e. Interpretation of slope and intercept

Possible answer: *The slope of 0.009 means the life expectancy increases 0.009 years for every 1 calorie, or 0.9 (almost 1) years of life expectancy is gained for every increase of 100 calories of daily intake. The y-intercept of about 47 means the life expectancy would be 47 years for someone with a daily caloric intake of 0 calories. In this case, the y-intercept does not make much sense in the context, as a person intaking 0 calories daily would probably not live for 47 years.*

Extra: A statistical question might be: "To what extent are daily caloric intake and life expectancy related?"

2. Crawling Age

The scatterplot in Figure 7.13 shows the average daily outside temperature when the babies were six months old, and the average age in weeks at which those babies began to crawl are reported. The equation of the least squares regression line is $\hat{y} = -0.078x + 35.68$.

(Source: Benson, Janette. "Infant Behavior and Development," 1993.)

a. Sketch a residual plot from the line of best fit

Answer: *Figure 7.14*

b. Is linear an appropriate model? Explain

Exit Ticket Cont.

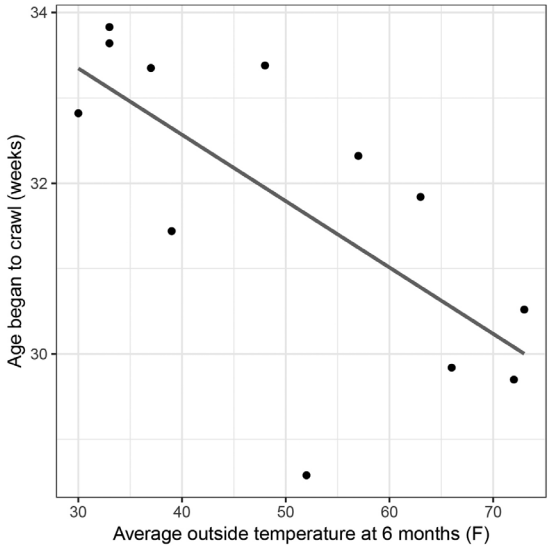


Figure 7.13: Average daily outside temperature when babies were six months old and average age in weeks at which those babies began to crawl

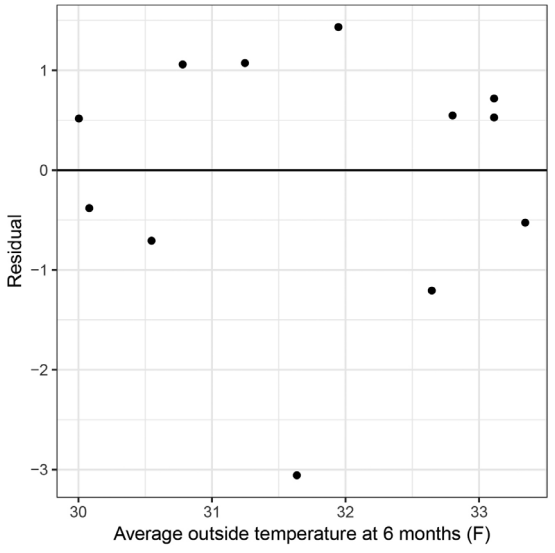


Figure 7.14: Sketch of a residual plot from the line of best fit

Possible answer: *The residual plot is fairly scattered and shows no pattern, indicating a linear model is an appropriate model.*

c. Estimate for correlation coefficient

Possible answer: *The correlation coefficient could be estimated between -0.6 and -0.8, ($r \approx -0.7$).*

d. Summary of the relationship

Possible answer: *The relationship indicated by the graph is a negative linear relationship, indicating the average age at which babies begin to crawl decreases as the average outside temperature increases.*

e. Interpretation of slope and intercept

Possible answer: *The slope of -0.078 means the average age at which a baby begins to crawl decreases by 0.078 weeks for every increase of 1 degree outside temperature. The y-intercept of about 36 means the average crawling age would be 36 weeks when the average outside temperature is 0 degrees. There are probably not many places where the average outside temperature is 0 degrees, but that might occur in locations closer to the north or south poles.*

Exit Ticket Cont.

Extra: A statistical question might be: "To what extent are average outside temperature and average crawling age of babies related?"

Note: Point out to students that even though there is a negative correlation between average outside temperature and crawling age, high outside temperatures do not necessarily cause the crawling age to decrease.

Investigation 8

How Long to Topple Dominoes?

Exploratory Lesson

Overview

This investigation offers two options for students. The first 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 a relationship between two quantitative variables. Several suggestions are included in this investigation, and students could be encouraged to come up with their own variables.

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, rubric, and competition information can be found at *www.amstat.org/asa/education/ASA-Statistics-Poster-Competition-for-Grades-K-12.aspx.*

The second option provides the set of directions for an investigation titled "How Long to Topple Dominoes?" This option is suggested for students who may need more scaffolding and direction when collecting and analyzing data. This option is based on Lesson 11, Exploring Linear Relations, available at *www.amstat.org/asa/files/pdfs/ddmseries/Exploring-LinearRelations.pdf*.

Both options for this investigation follow 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

» This lesson can be open-ended and allow students to choose a topic, as long as it involves two quantitative variables anticipated to have a linear relationship.

» Students follow the statistical problem-solving process, guided by the teacher. Provide support when needed.

» Students could be required to create a poster, a presentation, and/or a written report to communicate their process and results.

Instructions for Design 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 8.1 Design Your Own Investigation.

Formulate a Statistical Question

Students will brainstorm topics that might interest their group and include two variables. The two variables should be quantitative and ones in which students anticipate a linear relationship. Possible ideas include the following:

» Describe the relationship between two body measurements. Possible measurements

Learning Goals

» Design and collect data from an experiment

» Explore the relationship between two quantitative variables

» Apply techniques of finding a linear model

» Analyze the fit of the linear model

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*.

Option One

» Student Worksheet 8.1 Design Your Own Investigation

Option Two

» Student Worksheet 8.2 Topple Dominoes

» 200 dominoes (about 30 for each student group)

» Meter stick for each group

» Stopwatch for each group

» YouTube video of dominoes toppling: *www.youtube.com/watch?v=y4VJssQv_Qw*

Estimated Time

One to three 50-minute class periods, depending on the final product, amount of work required outside of class, and whether presentations occur.

Pre-Knowledge

Students should be able to:

» Find and interpret the equation of the least squares regression line

» Find and interpret the correlation coefficient

» Construct and interpret a residual plot

include width of head, width of shoulders, length of forearm (elbow to wrist), length of upper arm (elbow to shoulder), top of head to navel, top of head to tip of fingers (arms at sides), wrist circumference, neck circumference, height, arm span (arms out to sides, tip of fingers to tip of fingers), foot length, and stride length.

» Describe the relationship between ramp height and distance a matchbox car travels.

» Describe the relationship between number of people and length of time to pass a stack of books, pass a bucket of water, or bounce and pass a ball.

» Describe the relationship between length of time to say a tongue twister and the number of people.

» Describe the relationship between height of a catapult and how far a gummy bear is launched (Lesson 12, Exploring Linear Relations, *www.amstat.org/asa/files/pdfs/ddmseries/ExploringLinearRelations.pdf*).

Students should then develop the statistical question. Students could check in for approval before moving on to collect data.

Collect Appropriate Data

Have students describe the data-collection process, including possible complications and how these might be handled. Students should check in for approval before collecting data. Then, have students collect and organize the data.

Analyze the Data

Data analysis should include a scatterplot, description of the relationship between the two variables, interpretation of correlation coefficient, linear model, and residual plot.

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 (at least two graphs)

» Conclusions about the statistical question

The oral report should include the following:

» Reason the statistical question was chosen

» A description of the population sampled

Data Collection Table

Number of Dominoes	Time (Sec)	Time (sec)	Time (sec)	Mean Time (sec)
10				
15				
20				
25				
30				

» 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 How Long to Topple Dominoes?

Scenario

On November 13, 2009, World Domino Day 2009 saw the world record broken for the most dominoes toppled by a group when 4,491,863 dominoes were toppled. A total of 89 builders set up the dominoes in the WTC Expo Center in Leeuwarden, The Netherlands.

In April of 2017, a group of three students broke the unofficial world record for longest domino line with 15,524 dominoes!

View a YouTube video of the dominoes toppling at *www.youtube.com/watch?v=y4VJssQv_Qw.*

Formulate a Statistical Question

How long do you think it took for the line of 15,524 dominoes to fall over?

How long do you think it took for the 4,491,863 dominoes to fall over?

Statistical question: "What is the relationship between the number of dominoes in a line and the length of time for all the dominoes to topple over?"

Collect Appropriate Data

Divide students into groups of three or four. Distribute Student Worksheet 8.2 Topple Dominoes and 30 or more dominoes to each group. Each group should also have a meter stick and stopwatch. Students will need a flat and hard surface to set up dominoes. Carpeting does not work well.

Directions

On a flat and hard surface, stand up the dominoes on end in a straight line. Use the ruler or meter stick to make the spacing between the dominoes even. Space the dominoes about 2.5 cm apart. The only rule is that a domino can knock over only one other domino when it falls.

Set up 10 dominoes in a straight line and carefully time how long it takes for all 10 dominoes to topple. Repeat two more times. Record the three times in the data collection table.

Set up 15 dominoes in a straight line and carefully time how long it takes for all 15 dominoes to topple. Repeat two more times. Record the three times in the data collection table.

Set up 20 dominoes in a straight line and carefully time how long it takes for all 20 dominoes to topple. Repeat two more times. Record the three times in the data collection table.

Set up 25 dominoes in a straight line and carefully time how long it takes for all 25 dominoes to topple. Repeat two more times. Record the three times in the data collection table.

Set up 30 dominoes in a straight line and carefully time how long it takes for all 30 dominoes to topple. Repeat two more times. Record the three times in the data collection table.

Analyze the Data

1. Using technology construct a scatterplot of the mean time for the dominoes to fall versus the number of dominoes. Make a sketch of the scatterplot.
2. Is a linear model appropriate to describe the relationship between time for all the dominoes to fall and the number of dominoes? Use the correlation coefficient and a residual plot to explain your reasoning.
3. Find the equation of the least squares regression line.
4. Interpret the slope of the least squares regression line in context.

Interpret the Results in the Context of the Original Question

5. Using the linear model you developed, make a prediction for how long it would take 15,524 dominoes to fall. To make this prediction, what assumptions do you have to make about your model?
6. Watch the video at *www.youtube.com/watch?v=y4VJssQv_Qw* again and time how long it takes for all the dominoes to fall over.
7. How close was your prediction? What are some reasons why your prediction might have been off?

Extensions

Calculate the speed the dominoes travel as they topple.

Design and conduct an experiment to investigate the relationship between the distance between the dominoes and effect on time.

Investigation 9

Survey Says?

Analyzing Categorical Data in a Statistical Study

Overview

This investigation is the first of several lessons that focus on the analysis of two categorical variables. This investigation shares an initiative carried out by students at an urban high school that involved collecting data from the student body. This initiative involved creating a survey, obtaining a sample of completed surveys, and analyzing the sample to answer statistical questions posed by the students. The students addressed whether the sample was representative of the student body of their school.

The four components of statistical problem solving as put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report* are addressed in this investigation. 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 investigation is a GAISE Level B activity.

Instructional Plan

Brief Overview

» Define and give examples of categorical data.

» Read the scenario and study the survey questions.

» Develop a statistical question based on the survey questions.

» Discuss the four options to collect survey data.

» Discuss the collection plan used by high-school students.

» Summarize survey results.

Introduction to Categorical Data

Ask your students to give examples of *numerical* data they have analyzed. Have them share what types of graphs and calculations they did in their analysis of numerical data.

Possible answer: *Examples from previous investigations include height and arm span, length of baseball games in hours, time to complete a memory game, and homework times. Students would have constructed box plots, dot plots, and histograms and found means, medians, IQRs, and standard deviations.*

Discuss another type of data called *categorical data*. Categorical variables take on values that are names or labels. Share some examples of categorical data such as an answer to a true or false question, an answer to a multiple-choice question (A, B, C, or D), the size of a T-shirt (small, medium, or large), or the breed of a dog (shepherd, terrier, lab).

Ask your students to share other examples of categorical data.

Hand out Student Worksheet 9.1 Scenario

Explain that the case study presented was conducted by high-school students from an urban high school who analyzed categorical data.

Direct students to read the scenario from Student Worksheet 9.1.

Scenario

The administration at Rufus King High School, a United States urban high school of students in grades 9 to 12, was in the process of evaluating the school's academic and extracurricular programs. The high-school administration considered distributing and analyzing a survey addressing the school's programs that would be similar to the process businesses use to evaluate their products and services. They asked the students enrolled in an 11th grade mathematics class if they would help with the design, distribution, and analysis of a survey project.

Statistical studies about a school's services might result in decisions that alter a school's daily schedule, curriculum, course offerings, extracurricular opportunities, etc. Rufus King students wanted to be part of a study that might alter their school's academic and extracurricular programs. Students designed

Learning Goals

» Investigate methods for obtaining a representative sample of responses to a survey from a large population.

» Evaluate a random sample of students' responses to a survey.

» Summarize a population using the results of a random sample.

Mathematical Practices Through a Statistical Lens

MP2. Reason abstractly and quantitatively.

Statistically proficient students are able to summarize data to answer statistical questions. Students explain their summaries of data using proportions.

Materials

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

» Student Worksheet 9.1. Scenario

» Student Worksheet 9.2 Data Collection Methods

» Student Worksheet 9.3 Survey Results

» Exit Ticket

Estimated Time

One 50-minute class period. This lesson introduces a case study that is expanded in investigations 10 and 11.

Pre-Knowledge

Students should be able to convert a proportion to a percent.

Question 1: Indicate your gender:

___ Female (F) ___ Male (M) ___ Prefer not want to respond

Question 2: Indicate your grade level in high school:

___ 9th grade ___ 10th grade ___ 11th grade ___ 12th grade

Question 3: Do you consider yourself a dog person, a cat person, or neither?

___ A. I consider myself a dog person.

___ B. I consider myself a cat person.

___ C. I do not consider myself a dog or cat person.

Question 4: What is your main goal after completing high school?

___ A. To attend a college, university, or technical school.
___ B. To get a job.
___ C. To enlist in the military.
___ D. Other

Question 5: Do you participate in one or more of the athletic programs at your school (basketball, football, soccer, hockey, tennis, volleyball, etc.)

____ Yes (Y) ___ No (N)

Question 6: Do you exercise daily?

___ Yes (Y)___ No (N)

Question 7: Do you spend at least 1 hour a week involved in an outdoor activity (walking, running, playing a game etc.)?

___ Yes (Y)___ No (N)

Question 8: Are you involved in any community service activity?

____ Yes (Y) ___ No (N)

Figure 9.1: Survey questions developed by students at Rufus King High School

a survey they thought would address several important statistical questions related to the school's academic and extracurricular programs. A few of the survey questions are listed in Figure 9.1.

After students read the scenario, discuss the following questions:

1. Why do you think the students designing this survey wanted to know the grade level of a student completing the survey (Survey Question 2)?

Possible answer: *Several of the other questions might be answered differently based on a student's grade level. For example, is it possible 9th graders might be more or less involved in extracurricular activities than 12th graders? Does a student's plan after completing high school change over time? Might 12th graders answer Question 4 differently than 9th graders?*

2. Why do you think students included Survey Question 3? What was a possible reason to consider this question important?

Possible answer: *Survey Question 3 might be used to examine the growing interest in therapy pets to address stress or anxiety. If a program of this type were pursued, what type of pets would be selected? Do most students have a similar interest in the pets selected?*

3. Why might it be important to know if students exercise daily? Will most students understand what this question is asking? Will most students answer this question?

Possible answer: *Exercise may have different meanings to students. Some students may think of this as organized school activity. Other students may think of this as an individual activity involving walking, running, stretching, etc. The question was considered adequate by the Rufus King students, but whether they collected accurate responses was not clear. This question was unclear to some students answering the question and is a good example of how questions of this type have different interpretations. After the survey was distributed, students discussed that the following rewording of this question might have clarified some of the confusion: "Do you participate in at least 10 minutes a day of physical exercise either alone or as part of a group?"*

4. Why might it be important to know if students are involved in community service? Do students agree on what is meant by community service? Will most students answer this question?

Possible answer: *This question also needed more clarity. Community service was viewed by some students as service activities by the school; other students interpreted community service as an activity organized by other organizations or groups. Here again is an opportunity to discuss the importance of whether or not a question provided the intended information needed in the statistical study.*

Take the students through the process the Rufus King students followed.

Formulate a Statistical Question

The Rufus King students developed a series of statistical questions designed to provide a summary of the school's student population. Several possible statistical questions emerged from this project. For example, are students typically going to attend a college or university after high school or pursue other options? Are students likely to participate in the school's athletic programs? Do students typically spend at least one hour per week outdoors?

Discussion About Different Ways to Collect Appropriate Data

Explain that after the survey was designed and approved by the administration, a plan was needed to organize how students would complete the survey. It was possible, but not practical, to analyze completed surveys from more than 1,200 students enrolled in the high school. Students (under the direction of their teachers) discussed ways in which they might distribute surveys to obtain a sample that provided all students in the school the same opportunity to complete the survey.

Ask the students to read the four data collection options and answer the two questions for each option.

For each of the following four options, answer the two questions:

» Do you think this option will provide an accurate summary of the responses from students in the school?

» If this option is used, are there any groups of students who may not be represented? Explain your answer.

Option 1:

Consider placing computers at various locations around school (e.g., the cafeteria, library, computer lab) that are monitored by students from the mathematics class involved with this project. Students in the vicinity of the computers would be asked to complete the survey provided on the computer. After a student completed the survey, the students monitoring the computer would save the results and load a new survey for the next student to complete. At the end of the day, the

responses from the completed surveys would represent the representative sample for analyzing the questions.

Option 2:

There are 35 students in the mathematics class involved with this project. Each member of the class would be encouraged to anonymously complete the survey. The completed surveys would comprise the representative sample for analyzing the questions.

Option 3:

Students in the mathematics class involved with this project would post the survey online using a service provided by a private company. Each member of the class would encourage friends to complete the survey, both through word of mouth and also through their social media accounts. The online service would provide completed surveys that comprise the representative sample for analyzing the questions.

Option 4:

Students enrolled in the mathematics class would distribute surveys both before or after school at various locations in the school building. At the end of the day, the completed surveys would comprise the representative sample for analyzing the questions.

After the students have read the four options and answered the two questions for each, have the students share their answers for each option.

Option: Place students in groups. Have each group create a poster that lists the pros and cons of each method. Also, list their choice for a method.

Discussion Points

Discuss with students that each of the options would provide a sample, but there would be questions as to whether a representative sample of the student population would have completed the surveys. In general, these options result in a *convenience sample*, or a sample that did not provide an opportunity for a cross section of the school's students to complete the survey. It is important to indicate that a statistical study based on a convenience sample, or any sample not representative of the population, may result in *bias* that raises questions regarding any conclusions of the study.

Note: Consider discussing with your students how they would collect a representative sample at their school. Would any of the previous options provide a representative sample? What other options might be considered?

Design and Implementation of a Plan to Collect Data

Hand out Student Worksheet 9.2 Data Collection Methods

Have the students read the Plan to Collect Data section on Student Worksheet 9.2.

Plan to Collect Data

The following is a summary of the plan implemented at Rufus King High School.

All students attending Rufus King are required to take an English course. Students involved in the survey project arranged providing the option of completing a survey during an English class with the school's English teachers. They estimated it would take fewer than five minutes to complete the survey. A specific day was identified to complete the survey. Students were also told by their English teachers that they did not have to complete the survey. Students involved in organizing this project provided an explanation of the project to the students several days before the survey was distributed by way of an all-school announcement. In addition, a flier was sent home to inform parents and guardians about the project.

The number of students who completed the survey was 1103.

Each survey was collected and given a specific identification number. Identification numbers from 1 to 1103 were assigned to the completed surveys. It was decided that 50 randomly selected surveys would form the sample for this study. Students generated 50 random numbers from 1 to 1103 using a graphing calculator. The 50 numbers generated by the calculator represented the 50 identification numbers and the 50 surveys selected to form the sample.

Ask your students to answer questions 1 to 5.

1. Do you think the above plan resulted in a sample that provided all students an equal chance to be selected in the sample? Explain you answer.

Possible answer: *Essentially, every student who was in attendance had an opportunity to complete the survey. This plan would result in a sample in which each completed survey had an equal chance of selection.*

2. Why do you think it was important to inform students about the project before they received the survey?

Possible answer: *It is important to emphasize that data of this type must convince students their time to complete the survey and their responses are important. If this study had been a research study conducted by professionals, a careful review of the survey questions would need to be conducted by a team of advisers. This team would also be responsible for evaluating details about the research project and how it would be communicated to students and their parents or guardians. Authentic statistical studies are held to high standards of communication and review.*

3. Why do you think it was important to inform parents and guardians about the project?

Possible answer: *Emphasize again the importance of maintaining communication in a statistical study. Also, students in high school are considered minors. Involving parents and guardians was an important requirement of the administration.*

4. Using the plan described, which students would not have completed the survey?

Possible answer: *Students absent from school or students who opted out of completing the survey would not have been included.*

5. Do you think the sample of 50 completed surveys represents a representative sample of all students?

Possible answers: *Encourage students to express their opinions to this question. It is anticipated students may comment that a sample of only 50 students would likely not be representative of the school's population. Students may also indicate a sample of only 50 surveys would not be large enough to obtain adequate summaries of the survey questions.*

Analyze the Data

Ask your students what type of data was collected for each of the eight survey questions.

Answer: *Categorical data.*

Ask students how they would summarize the responses to Question 1. What measure would they use to communicate what they have collected?

Possible discussion points: *Students may initially focus on the counts of Male responses or Female responses. Although the count of each category is important, it is the proportion of the*

number of males or the number of females to the sample size that will be the more important summary of the question in this statistical study. A similar proportion would be calculated for each of the other questions in the survey.

Return to the discussion concerning whether the sample of 50 completed surveys represents a representative sample of all students. The main question is whether the selection of 50 surveys is large enough to estimate the proportions of the school population for each question. Will the proportion of females, proportion of students who exercise daily, or proportion of students who are involved in community service based on this sample of 50 students be the same as the school population?

The following example will help students understand that the random sample selected by students is likely to provide an adequate summary of the school population.

The sample of 50 students indicated 33 females and 17 males. A first step was to convert the number of females to a proportion—33/50 or 0.66 or 66% of the sample of 50 students was female and 17/50 or 0.34 or 34% of the students was male.

Share with your students the following summary of the school population posted on the school's website:

» Total enrollment (September 5): 1204 students

» Total number of females: 775

» Total number of males: 429

Based on this information summarizing the school population, the proportion of females at Rufus King High School at the time this project was conducted was 775/1204, or approximately 0.644 or 64%. The sample proportion of 66% was similar to the proportion of females of the school's population.

Interpret the Results in the Context of the Original Statistical Questions

Hand out Student Worksheet 9.3 Survey Results.

Direct students individually or in small groups to use the data presented on Student Worksheet 9.3 and answer Question 6.

Answers:

Q1 (Survey Question 1)

» Proportion of females: 33/50 or 0.66

» Proportion of males: 17/50 or 0.34

Q2 (Survey Question 2)

» Proportion of students in 9th grade: 15/50 or 0.30

» Proportion of students in 10th grade: 14/50 or 0.28

» Proportion of students in 11th grade: 16/50 or 0.32

» Proportion of students in 12th grade: 5/50 or 0.10

Q3 (Survey Question 3)

» Proportion of students who indicate they are a "dog person": 23/50 or 0.46

» Proportion of students who indicate they are a "cat person": 24/50 or 0.48

» Proportion of students who indicate they are neither: 3/50 or 0.06

Q4 (Survey Question 4)

» Proportion of students who plan to attend college after high school: 30/50 or 0.60

» Proportion of students who plan to get a job after high school: 9/50 or 0.18

» Proportion of students who plan to enlist in the military after high school: 6/50 or 0.12

» Proportion of students who selected other: 5/50 or 0.10

Q5 (Survey Question 5)

» Proportion of students who participate in the school's athletic program: 30/50 or 0.60

Q6 (Survey Question 6)

» Proportion of students who exercise daily: 30/50 or 0.60

Q7 (Survey Question 7)

» Proportion of students who spend at least one hour per week outdoors: 10/50 or 0.20

Q8 (Survey Question 8)

» Proportion of students involved in community service: 23/50 or 0.46

Direct students to individually answer questions 7 to 12. After they have completed the questions, discuss these questions with the whole class.

7. Based on the above summaries, provide a brief description of the students attending this high school.

Summary answer: Students can focus on one or two summaries they find interesting about the school. For example, slightly more students considered themselves a cat person or only 20% of the students spend at least one hour outdoors. Remind students they are using a representative sample to describe the students in the school population.

8. What is your estimate of the *number of* students who participate in an athletic program from the total enrollment of 1204 students? Do you think your estimate is the exact number of students who participate in an athletic program?

Answer: *Assume the proportion of the school population participating in an athletic program is the same as the proportion of the sample. Therefore, an estimate of 0.60, or 60%, of the 1204 students is 722 students. This estimate is likely not the exact number of students who participate in an athletic program.*

9. Why might it be important to know the number of students and the proportion of students who participate in a school athletic program?

Possible answer: *Results could be used to evaluate the interest in an athletic program and whether the school's facility can effectively address the interest. Estimating the number of students involved in an athletic program might be used to determine whether the facilities (e.g., gyms or volleyball courts or bathrooms) are sufficient.*

10. What is your estimate of the students who participate in community service? Do you think your estimate is the exact number of students who participate in community service?

Answer: *Assume the proportion of the school population participating in community service is the same as the proportion of the sample. Therefore, an estimate of 0.46, or 46%, of the 1204 students is approximately 553.84 or 554 students. This estimate is likely not the exact number of students who participate in community service.*

11. Why might it be important to know the number and proportion of students who participate in community service?

Possible answer: *If the school is considering improving students' participation in community service, it is important to determine the current involvement.*

12. Why might it be important to know the proportion of students who spend at least one hour involved in outdoor activities?

Possible answer: *Several research studies have linked outdoor activity to student achievement. This survey question might be connected to other items (e.g., exercise, gender, grade level) that examine whether there are noticeable differences in outdoor activity based on these other categories.*

Exit Ticket

1. Based on the estimate of the relative frequencies from the surveys and the summary of the high-school enrollment of 775 females and 361 ninth graders, determine an estimate for each of the following two questions. For each question, indicate how you determined your estimate.

a. How many females do you think are involved in an athletic program at King?

Summary answer: *Assume the proportion females participating in an athletic program is the same as the proportion of students participating in an athletic program based on the sample. Therefore, 60% of the 775 female students, or 465 female students, is an estimate of the number of females who participate in an athletic program.*

b. How many 9th graders do you think are involved in at least 1 hour of outdoor activities per week?

Answer: *Assume the proportion of 9th graders involved in at least one hour of outdoor activities is the same as the proportion of students involved in at least one hour of outdoor activities from the sample. Therefore, assume 20% of the 361 9th grade students, or approximately 72.2 or 73 students, is an estimate. This also assumes the proportion of 9th grade students involved in outdoor activities is the same as the proportion of other grade levels.*

Exit Ticket Cont.

2. Consider the following data collection option:

   Students in the mathematics class involved with the project would number each table in the cafeteria. They would select 10 random tables at each lunch period and ask everyone sitting at the selected table to answer the survey.

   Do you think this option will provide an accurate summary of the responses from students in the school? If this option is used, are there any groups of students who may not be represented? Explain your answer.

Possible answer: *Since students usually sit with their friends at the same lunch table, it is likely students at the table would answer the survey questions in a similar manner. Students who do not eat in the cafeteria or go out or home for lunch are not represented.*

Further Explorations and Extension

Interest in completing a survey project at your school may also be considered a viable extension of this investigation. If the entire project could not be completed (due to time constraints or other challenges), designing a plan to carry out a project at your school that is similar to the one in this investigation might also be a valuable discussion and an exploration to consider.

Investigation 10

Is There an Association?

Summarizing Bivariate Categorical Data

Overview

The process a group of high-school students used to collect data from the student body was explored in Investigation 9. The students collected data using a survey, and many of the survey questions resulted in collecting data on categorical variables. In this investigation, students examine two of the survey questions and are asked to evaluate whether they think there is a connection or association between the two variables based on the conditional relative frequencies.

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 investigation.

The activities and several questions are also based on *Lessons from Probability Through Data*, published by the American Statistical Association (original copyright by Dale Seymour Publications, 1999) and available at *www.amstat.org/ASA/Education*.

Instructional Plan

Brief Overview

» Present Scenario 1 (no association).

» Develop a statistical question based on survey questions 5 and 7.

» Construct a two-way table summarizing survey results.

» Construct a row conditional relative frequency table based on survey results.

» Interpret the statistical question based on the row conditional relative frequency table.

» Present Scenario 2 (an association).

» Interpret the statistical question based on the row conditional relative frequency table.

Hand out Student Worksheet 10.1 Scenario 1 and Student Worksheet 10.3 Questions and Results.

Note: Students are presented with two scenarios. Scenario 1 examines two categorical variables that have no differences in the conditional relative frequencies. If there are no differences, then one variable does not suggest a possible connection to the second variable, or the two variables appear not to be connected. Scenario 2 explores a connection between two variables that results in noticeable differences in the conditional relative frequencies. The variables discussed in Scenario 2 indicate a possible association.

Scenario 1

A recent internet posting indicated a key factor in improving academic success for high-school students is to spend time outdoors. Students in the Rufus King project thought

students involved in their school's athletic programs were more likely to spend time outdoors. To see if that was true, they incorporated a question into the survey (*www.health.harvard.edu/press_releases/spending-time-outdoors-is-good-for-you*).

After reading the scenario, discuss with students the following:

In your opinion, do students who participate in the organized sports program at their school and spend time outdoors have similar interests? Or, does participating in sports and spending time outdoors essentially have no connection?

Learning Goals

» Calculate and interpret conditional relative frequencies from two-way frequency tables involving two categorical variables.

» Evaluate whether the conditional relative frequencies are an indication of a possible association between the two categorical variables.

Mathematical Practices Through a Statistical Lens

MP2. Reason abstractly and quantitatively.

Statistically proficient students are able to summarize data to answer statistical questions. Students explain their summaries of data using relative frequencies and conditional relative frequencies.

Materials

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

» Student Worksheet 10.1 Scenario 1

» Student Worksheet 10.2 Scenario 2

» Student Worksheet 10.3 Questions and Results (same as Student Worksheet 9.3)

» Exit Ticket

Estimated Time

Two 50-minute class periods

Pre-Knowledge

» Examine and evaluate a random sample of students' responses to a survey.

» Summarize categorical data by relative frequencies and percent.

» Estimate summaries of a population using relative frequencies of a sample.

» Completion of Investigation 9.

Survey Response 1

Survey Number					Q5		Q7		
1					N		N		

Survey Response 2

Survey Number					Q5		Q7		
2					Y		Y		

Two of the questions on the Rufus King survey were the following:

» **Question 5:** Do you participate in one or more of the athletic programs at your school (basketball, football, soccer, hockey, tennis, volleyball, etc.)?

_____ Yes (Y) ___ No (N)

» **Question 7:** Do you spend at least one hour a week involved in an outdoor activity (walking, running, playing a game, etc.)?

___ Yes (Y) ___ No (N)

Point out that there was confusion expressed about the wording of "outdoor activity" in survey Question 7. Discuss this question with your students. Ask them what they think the question is asking. The intent by the students involved with this project was to determine whether a student spent active time outdoors (running, walking, hiking, gardening). If students agree this question was not clear and possibly did not collect the intended information, discuss possible revisions. It is important to review survey questions both before they are distributed and after the information is collected to evaluate the goals of a statistical study.

Formulate a Statistical Question

Ask the students to consider the following statistical question as an investigation of participation in the sports program and spending time outdoors.

Statistical question: "Is there a connection between participation in an athletic program and spending time outdoors?"

Analyze the Data

Direct students to examine specific survey numbers on worksheet 10.3 Questions and Results.

Ask your students to describe the student who completed this survey. Discuss with them the question, "If most students who answered no to Question 5 and also answered no to Question 7, as in the Survey Response 1 table, do you think there is a connection between participating in an athletic program and spending time outdoors?" At this point in the discussion, students' responses to this question are opinions of what they think the connection might indicate.

Continue a similar discussion with the next examples:

» Discuss with students the question, "If most students who answered yes to Question 5 also answered yes to Question 7, as in the Survey Response 2 table, do you think there is a connection between participating in an athletic program and spending time outdoors?"

Survey Response 3

Survey Number					Q5		Q7		
3					Y		N		

Survey Response 10

Survey Number					Q5		Q7		
10					N		Y		

» Discuss with students the question, "If most students who answered yes to Question 5 also answered no to Question 7, as in the Survey Response 3 table, do you think there is a connection between participating in an athletic program and spending time outdoors?"

» Discuss with students the question, "If most students who answered no to Question 5 also answered yes to Question 7, as in the Survey Response 10 table, do you think there is a connection between participating in an athletic program and spending time outdoors?"

Ask your students if there are any other ways students could have answered these two questions. As students look through the rest of the sample (Student Worksheet 10.3), point out that each of the remaining surveys is represented by one of the previous four examples.

Explain that we would like to summarize all the survey results for these two questions. Refer the students to the empty table on Student Worksheet 10.1 Scenario 1 following the four survey responses. Discuss with the students the labels needed to complete the *column labels* representing Question 7 (or "Spend time outdoors" and "Do not spend time outdoors"). Continue the discussion by summarizing Question 5 with appropriate *row labels* (or "Participate in an athletic program" and "Do not participate in an athletic program"). Have the students add the labels to the table.

Answer: *Table 10.1*

Ask the students to answer questions 1 to 4.

1. Based on your table, in what cell would you count Survey Response 1?

Answer: *Cell 5*

2. Based on your table, in what cell would you count Survey Response 2?

Answer: *Cell 1*

3. Based on your table, in what cell would you count Survey Response 3?

Answer: *Cell 2*

4. Based on your table, in what cell would you count Survey Response 10?

Answer: *Cell 4*

In groups, have the students go through worksheet 10.3 Results and complete the frequency table provided by determining in what cells each of the other surveys would be counted.

Note: Consider advising students to use tally marks in cells 1, 2, 4, and 5 for each of the 50 surveys. After the 50 surveys have been tallied, complete Frequency Table 10.2.

Table 10.1

	Spend time outdoors	Do not spend time outdoors	Total
Participate in an athletic program	Cell 1	Cell 2	Cell 3
Do not participate in an athletic program	Cell 4	Cell 5	Cell 6
Total	Cell 7	Cell 8	Cell 9

Frequency Table 10.2

	Spend time outdoors	Do not spend time outdoors	Total
Participate in an athletic program	Cell 1 6	Cell 2 24	Cell 3 30
Do not participate in an athletic program	Cell 4 4	Cell 5 16	Cell 6 20
Total	Cell 7 10	Cell 8 40	Cell 9 50

Analyze the Data by Measures and Graphs

Introduce the vocabulary pertaining to the two-way table.

Point out to the students there are two types of cells represented in this table that should be discussed. The shaded cells (1, 2, 4, and 5) are called *joint cells*, as they record the number of students responding to specific categories of two questions. The other cells (3, 6, 7, 8, and 9) are called *marginal cells*, as they represent the margins of the table and record the number of students responding to a specific category of one question, with the exception of cell 9, which records the total number of surveys completed in this sample. If necessary, identify a cell and ask students to explain what that cell indicates. For example, cell 2 indicates the number of students who participate in an athletic program but do not spend time outdoors.

Ask your students to use the frequency table and answer questions 5 and 6.

5. What proportion of the survey-takers answered that they spend time outdoors *and* participate in an athletic program?

Answer: *6/50 = 0.12, or 12%*

6. What proportion of the survey-takers answered that they do not spend time outdoors *and* do not participate in an athletic program?

Answer: *16/50 = 0.32, or 32%*

Explain that the 0.12 and 0.32 are called *relative frequencies*. Relative frequencies are the proportions of the counts in each cell to the total number of surveys completed in the sample (or 50 surveys for this example). The relative frequencies provide a description of the proportion or percent of students in each cell to the total number of students in the sample.

Explain that while the relative frequencies summarize the results of the sample, they do not help determine whether the responses to the two questions are connected. The relative frequencies do not specifically examine the

differences in the students who participate in an athletic program and do not participate in an athletic program.

Explain that to determine if there is a connection between participation in an athletic program and spending time outdoors, we need to find the proportion of those who participated in an athletic program who spent time outdoors and compare that to the proportion of those who did not participate in an athletic program who spent time outdoors.

Ask your students to highlight the first row of the frequency table by circling the entire row—those who participated in an athletic program.

Ask your students to answer questions 7 to 9.

7. Of the students in the survey, how many participated in an athletic program?

Answer: *30*

8. Of those who participated in an athletic program, how many spent time outdoors?

Answer: *6*

9. What proportion of students who participated in an athletic program spent time outdoors?

Answer: *6/30 = 0.20, or 20%*

Explain that this value of 0.20 is called a *row conditional relative frequency*. It is the proportion of those who participated in an athletic program who spent time outdoors.

Direct your students to enter the 0.20 into cell 1 of Conditional Relative Frequency Table 10.3.

Ask your students to calculate the proportion of the 30 students who participated in an athletic program who did not spend time outdoors and enter this value into cell 2.

Answer: *24/30=0.80, or 80%*

Ask your students to calculate the proportions in cells 4 and 5 of the students who spent time outdoors or did not spend time outdoors based on the condition they did not participate in the school's athletic program.

10. Ask your students to calculate the conditional relative frequencies for cells 3, 6, 7, 8, and 9.

The completed table is called a *row conditional relative frequency table* (See Table 10.3).

Optional: To help your students visualize the conditional relative frequencies of students who do and don't participate in athletics and spend time outdoors, demonstrate the construction of a segmented bar graph. Using the conditional relative frequency table, a segmented bar graph is shown in Figure 10.1. The conditional relative frequencies could also be visualized in side-by-side bar graphs.

Row Conditional Relative Frequency Table 10.3

	Spend time outdoors	Do not spend time outdoors	Total
Participate in an athletic program	Cell 1 6/30 = 0.20, or 20%	Cell 2 24/30 = 0.80, or 80%	Cell 3 30/30 = 1.00, or 100%
Do not participate in an athletic program	Cell 4 4/20 = 0.20, or 20%	Cell 5 16/20 = 0.80, or 80%	Cell 6 20/20 = 1.00, or 100%
Total	Cell 7 10/50 = 0.20, or 20%	Cell 8 40/50 = 0.80, or 80%	Cell 9 50/50 = 1.00, or 100%

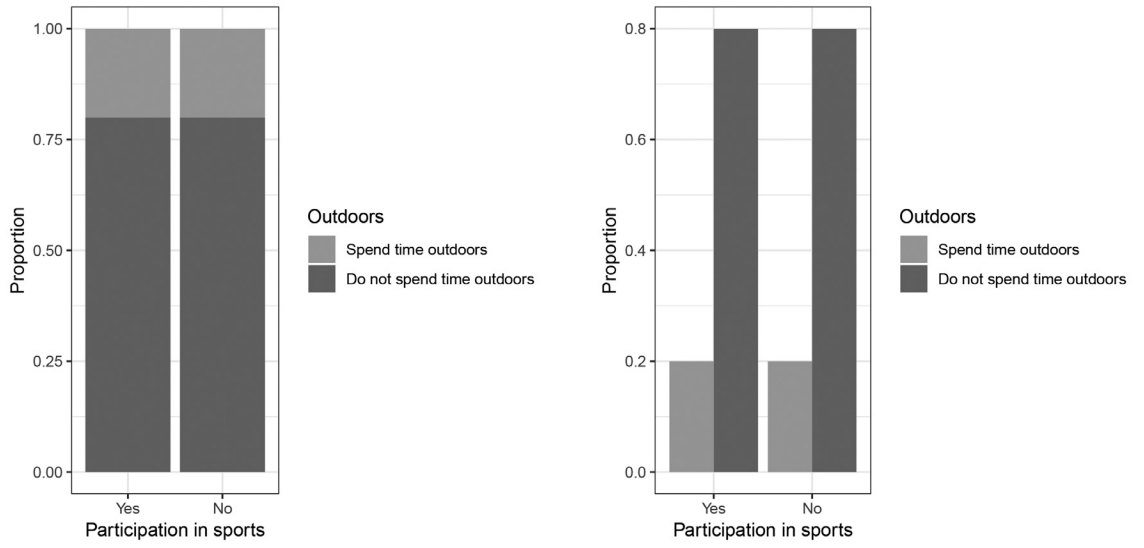


Figure 10.1: Segmented bar graph and side-by-side bar graph of athletic participation and time outdoors data

Interpret the Results in the Context of the Original Statistical Question

Ask your students to answer questions 11 to 17.

11. What is the proportion of students who spend time outdoors?

Answer: *0.20, or 20%*

12. What is the conditional relative frequency of the 30 students in an athletic program who spend time outdoors?

Answer: *0.20, or 20%*

13. What is the conditional relative frequency of the 20 students who do not participate in an athletic program who spend time outdoors?

Answer: *0.20, or 20%*

14. If a student completing the survey participated in an athletic program, what is your estimate she or he spends time outdoors?

Answer: *0.20, or 20%*

15. If a student completing the survey did not participate in an athletic program, what is your estimate she or he does not spend time outdoors?

Answer: *0.20, or 20%*

16. Does knowing whether a student participates or does not participate in an athletic program change your estimate of spending time outdoors?

Answer: *No*

17. Do you think the questions about participating in an athletic program and spending time outdoors are connected? Explain your answer.

Possible answer: *The data do not indicate that participation in the athletic program and spending time outdoors are connected.*

Hand out Student Worksheet 10.2 Scenario 2

Scenario 2

Many schools, nursing homes, and residential communities are making therapy pets available to address anxiety and depression. Dogs and cats are most commonly used in pet therapy. However, fish, guinea pigs, horses, and other animals that meet screening criteria can also be used.

The type of animal chosen depends on the therapeutic goals of a person's treatment plan. In selecting a therapy pet for a person, there is debate as to whether females and males have different preferences in the selection of their therapy pets.
Source: www.healthline.com/health/pet-therapy

Formulate a Statistical Question

Discuss with students an appropriate statistical question for this investigation.

Statistical Question

Is there a connection between gender and whether a person is a dog person or a cat person or neither a dog or cat person?

Analyze the Data

Frequency Table 10.4, involving the responses to item Survey Question 1 (Q1) and Survey Question 3 (Q3), is provided below.

Ask your students to answer Question 1.

1. Using the frequency table, complete the row conditional relative frequencies based on gender.

Answer: *Row Conditional Relative Frequency Table 10.5*

Optional: To help your students visualize the different percentages of females and males pet preferences, demonstrate the construction of a segmented bar graph. Using the conditional relative frequency table, a segmented bar graph is shown in figure 10.2. The conditional relative frequencies could also be visualized in side-by-side bar graphs.

Interpret the Results in the Context of the Original Statistical Question

Ask the students to answer questions 2 to 9.

2. What is the proportion of students who consider themselves a dog person?

Answer: *23/50 = 0.46, or 46%*

3. What is the proportion of males who consider themselves a dog person?

Answer: *15/17= 0.882, or 88.2%*

4. What is the proportion of females who consider themselves a dog person?

Answer: 8/33= *0.242, or 24.2%*

5. If a survey from the sample indicated the student completing the survey was male, what is your estimate he considers himself a dog person?

Answer: *0.882, or 88.2%*

6. If a survey from the sample indicated the student completing the survey was female, what is your estimate she considers herself a dog person?

Frequency Table 10.4

	A. Dog Person	B. Cat Person	C. Neither	Total
Female	8	22	3	33
Male	15	2	0	17
Total	23	24	3	50

Row Conditional Relative Frequency Table 10.5

	A. Dog Person	B. Cat Person	C. Neither	Total
Female	8/33 = 0.242, or 24.2%	22/33 = 0.667, or 66.7%	3/33 = 0.091, or 9.1%	33/33 = 1.00, or 100%
Male	15/17 = 0.882, or 88.2%	2/17 = 0.118, or 11.8%	0	17/17 = 1.00, or 100%
Total	23/50 = 0.46, or 46%	24/50 = 0.48, or 48%	3/50 = 0.06, or 6%	50/50 or 1.00, or 100%

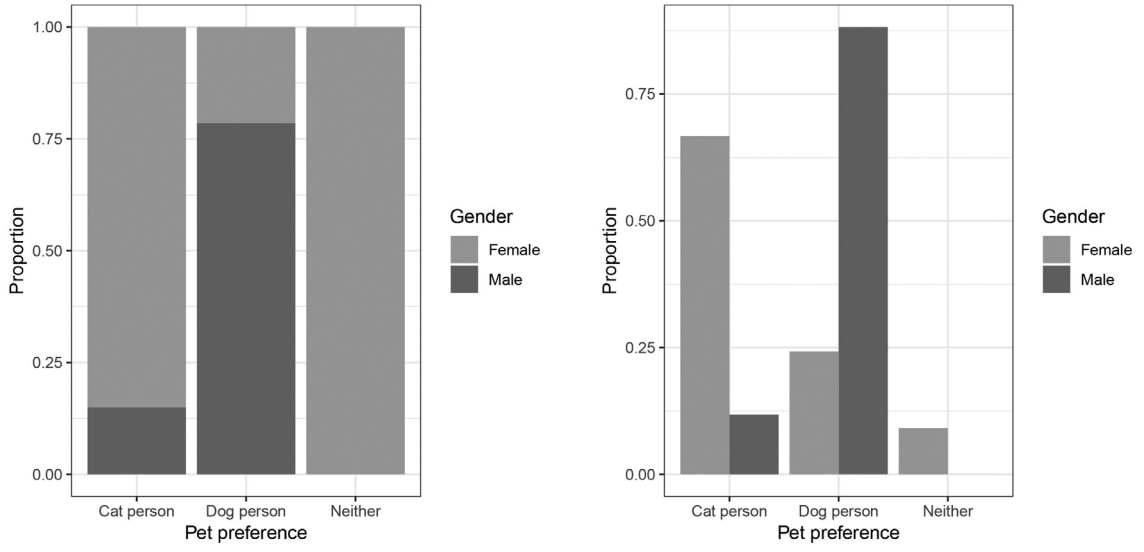


Figure 10.2: Segmented bar graph and side-by-side bar graph of females and males pet preferences data

Answer: *0.242, or 24.2%*

7. If a survey from the sample indicated the student completing the survey was male, what is your estimate he considers himself a cat person?

Answer: *2/17=0.118, or 11.8%*

8. If a survey from the sample indicated the student completing the survey was female, what is your estimate she considers herself a cat person?

Answer: *22/33=**0.667, or 66.7%***

9. Do you think the responses to the question about animal preference are connected to gender? Explain.

Possible answers: *The responses indicate there is a large difference in the proportion of males and the proportion of females who prefer a dog or a cat. There seems to be a connection in which males prefer dogs and females prefer cats.*

The differences in the conditional relative frequencies or proportions for each animal preference by gender indicates there is a possible connection. This connection is called an *association.*

Discuss with your students the general idea of association between two categorical variables.

Point out that the differences in conditional relative frequencies between two categorical variables indicate a possible association between the variables. The differences in the conditional relative frequencies indicate a connection is based on the conditions or responses to one of the survey questions.

It is important to remind students that whenever an association is noted in this type of statistical study, the connection suggested by the association is not *causal*. For this example, a person's gender does not cause the result of pet preference. Causation is important to analyze when examining experimental statistical studies.

Ask your students how might decisions that make therapy pets available for students be managed based on gender differences in pet preferences if there is a connection between gender and pet preference. As this is still a new challenge for schools, it is possible that students might suggest having both dogs and cats available is important (as opposed to just dogs or just cats).

Exit Ticket

Complete the following two-way frequency table for the variables of playing in the orchestra (Yes or No) and playing in the marching band (Yes or No) in which a sample of 50 students *indicates there is a possible association between the two variables*.

Why does your table indicate a possible association between the two questions involving participation in the orchestra and participation in the marching band?

	Play in the orchestra	Do not play in the orchestra	Total
Play in the marching band			20
Do not play in the marching band			30
Total	15	35	50

Possible Answer to the Exit Ticket

Answer will vary. As an example of how to evaluate answers, consider the following frequency table:

	Play in the orchestra	Do not play in the orchestra	Total
Play in the marching band	10	10	20
Do not play in the marching band	5	25	30
Total	15	35	50

Using the above frequency table, determine the conditional relative frequency table:

	Play in the orchestra	Do not play in the orchestra	Total
Play in the marching band	10/20 = 0.50, or 50%	10/20 = 0.50, or 50%	20/20 = 1.00, or 100%
Do not play in the marching band	5/30 = 0.167, or 16.7%	25/30 = 0.833, or 83.3%	30/30 = 1.00, or 100%
Total	15/50 = 0.30, or 30%	35/50 = 0.70, or 70%	50/50 = 1.00, or 100%

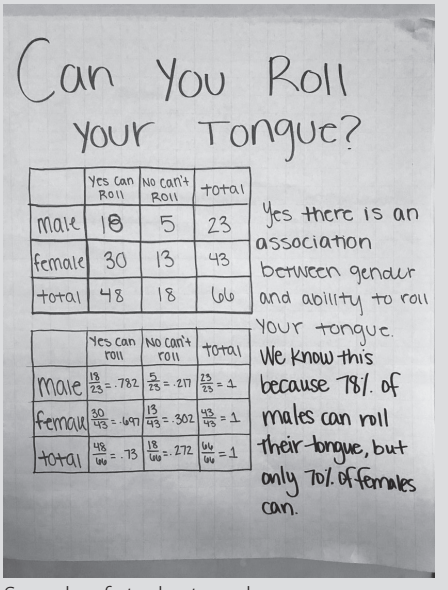
The conditional relative frequency table indicates there is a greater likelihood that if a survey is selected in which a student plays in the marching band, this student also plays in the orchestra than if a survey is selected in which the student does not play in the marching band and this student plays in the orchestra.

The rather large difference in the conditional relative frequencies indicates a possible connection or association of the variables involving orchestra and band. This particular example indicates there is a higher likelihood that if a student plays in the band, this student also plays in the orchestra.

Students should suggest values that add up to the marginal totals. Students should also select values that result in a noticeable or large difference in the conditional relative frequencies.

Further Explorations and Extensions

A blank template is provided for students to possibly explore two variables of their choice from the Rufus King High School data. (See Worksheet 10.3 Survey Questions Results.) Students should start by forming a statistical question based on selecting two questions from the survey. Students then organize the responses in a two-way frequency table. From the two-way frequency table, students create a relative frequency table and then a conditional relative frequency table. Based on the conditional relative frequency table, students would estimate whether the two variables they selected are possibly associated. They should summarize their statistical question based on the conditional relative frequencies.



	Yes can Roll	No can't Roll	total
male	18	5	23
female	30	13	43
total	48	18	66

	Yes can roll	No can't roll	total
male	$\frac{18}{23}=.782$	$\frac{5}{23}=.217$	$\frac{23}{23}=1$
female	$\frac{30}{43}=.697$	$\frac{13}{43}=.302$	$\frac{43}{43}=1$
total	$\frac{48}{66}=.73$	$\frac{18}{66}=.272$	$\frac{66}{66}=1$

Sample of student work

Investigation 11

Independent or Not Independent Events?

Comparing Conditional Relative Frequencies

Overview

Investigation 10 explored whether there was an association between two categorical variables by examining the differences in conditional relative frequencies. In Part I of this investigation, students investigate the connection between conditional relative frequencies and independent events (in probability). In Part II, students design and conduct a simulation to determine if two events are independent.

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 activity.

The activity is based on lessons from *Probability Through Data* published by the American Statistical Association, available at *www.amstat.org/ASA/Education* (original copyright by Dale Seymour Publications, 1999).

Instructional Plan

Brief Overview

Part I

» Construct a row conditional relative frequency table based on year in school and win/lose a computer game.

» Develop the definition of independent events

Part II

» Formulate a statistical question on the likelihood a random sample of 100 will produce 27.5% success rate.

» Design and conduct a simulation.

» Answer the statistical question based on results of the simulation.

Part I: What Are Independent Events?

In Investigation 10, students determined whether two categorical variables on a set of randomly selected data were associated by calculating and comparing conditional relative frequencies of categorical variables. In the framework of probability, events are categories and conditional relative frequencies of categories are estimates for probabilities of events.

Part I is designed to provide an understanding of the definition of independent events.

Hand out Student Worksheet 11.1 Independent or Not Independent Events.

Ask students to read the scenario and answer questions 1 to 3. Then, discuss student answers to questions 1 to 3.

Scenario

Games can involve chance, skill, strategy, or some mixture of them. This investigation is interested in games of chance such as Candy Land, Chutes and Ladders, and the card game War.

1. Identify a game that determines a win or loss by chance alone. Explain how chance

is involved and how skill or strategy are not involved.

2. Identify a game in which a win or loss is primarily determined by the skill of a player or players. Explain.
3. Identify a game in which a win or loss involves both chance and skill or strategies.

Have your students read Game: Over or Under.

Game: Over or Under

The computer science students at Rufus King High School designed a game to be played on a computer they call Over or Under. The directions were provided in the opening screen.

Learning Goal

Understand and interpret the connection between conditional relative frequencies and independent events in probability and the definition of independent events.

Mathematical Practices Through a Statistical Lens

MP2. Reason abstractly and quantitatively.

Statistically proficient students are able to summarize data to answer statistical questions. Students explain their summaries of data using relative frequencies and conditional relative frequencies as probabilities. Students interpret relative and conditional relative frequencies to reason about the population from which a sample was selected.

Materials

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

» Student Worksheet 11.1 Independent or Not Independent Events

» Student Worksheet 11.2 Simulation

» Student Worksheet 11.3 Template for Conducting the Simulations (on stock paper)

» For the simulation component of this investigation (Part II), students will need a small paper bag (one per small group) and the cut-out slips of paper from the template provided with this investigation.

» Exit Ticket

Estimated Time

Two 50-minute class periods. Part I develops an understanding of independent events. This section is expected to take one class period. Part II directs students in conducting a simulation of the scenario in the investigation.

Pre-Knowledge

Summarize data by relative frequencies, conditional relative frequencies, and percentages (completion of Investigation 10).

Each one of the numbers 0, 1, 2, 3, 4, and 5 is behind the following cards labeled A, B, C, D, E, and F. They are in random order. Each number is used with no repeats. Click on any three cards. If the sum of the numbers behind the cards is 6 or less, then you win the game. If the sum of the numbers is greater than 6, then you lose. Hit the Start icon to begin the game. Have fun!

Start

Justin played the game. The following opening screen starts the game.

Card A	Card B	Card C	Card D	Card E	Card F
?	?	?	?	?	?

Justin clicked on cards A, C, and D. The next screen indicated the following:

Card A		Card C		Card D		
3	+	0	+	2	=	5

You WIN!

If the sum had been greater than 6, Justin would have lost the game.

The computer science students decided to test out their game to determine if it would interest the students in their school. They were given permission to randomly select 100 students from their school and ask them several questions, including if they would play their game, what year in high school they were in (1st, 2nd, 3rd, or 4th), and whether the game was interesting. Each of 100 selected students agreed to play the game once and record whether they won or lost.

Table 11.1

	Number of students who won the game	Number of students who lost the game	Total number of students who played the game
1st- or 2nd-year Student	11	29	40
3rd- or 4th-year Student	19	41	60
Total	**30**	**70**	**100**

Exactly 100 students played the game once. Table 11.1 summarizes the results.

Students in the computer science class wanted to investigate if winning the game was connected to grade level. Are 3rd- or 4th-year students better at playing games of chance than 1st- or 2nd-year students?

Have your students complete Question 4.

Note: This question requires students to have completed Investigation 10.

4. Complete the conditional relative frequency table 11.2 of winning or losing the game categories based on year of student.

Answer: *Table 11.2*

Explain to your students that we are now going to use conditional relative frequencies of categories as estimates for probabilities of events. For example, what is the probability a randomly selected student wins the game?

Answer: *30/100=0.30, or 30%*

Have your students complete Question 5.

5. Use the conditional relative frequencies as estimates of conditional probabilities and complete Table 11.3, the conditional probability of events.

Table 11.2

	Conditional Relative Frequencies for Wins Based on Year	Conditional Relative Frequencies for Losses Based on Year	Totals
1st- or 2nd-Year Student	11/40 = 0.275, or 27.5%	29/40 = 0.725, or 72.5%	40/40 = 1.00, or 100%
3rd- or 4th-Year Student	19/60 = 0.317, or 31.7%	41/60 = 0.683, or 68.3%	60/60 = 1.00, or 100%
Totals	30/100 = 0.30, or 30%	70/100 = 0.70, or 70%	100/100 = 1.00

Table 11.3

	Conditional Probability of Winning Based on Year	Conditional Probability of Losing Based on Year	Totals
1st- or 2nd-Year Student	11/40 = 0.275, or 27.5%	29/40 = 0.725, or 72.5%	40/40 = 1.00, or 100%
3rd- or 4th-Year Student	19/60 = 0.317, or 31.7%	41/60 = 0.683, or 68.3%	60/60 = 1.00, or 100%
Totals	30/100 = 0.30, or 30%	70/100 = 0.70, or 70%	100/100 = 1.00

Answer: *Table 11.3*

Have your students answer questions 6 to 8.

Interpret the table of conditional probabilities.

6. If winning this game is totally based on chance and not connected to the year a student is in school, what is the probability that a randomly selected student wins the game?

Answer: *30%*

7. If winning this game is totally based on chance, what is the conditional probability that a 1st- or 2nd-year student would win the game?

Answer: *27.5%*

8. If winning the game is totally based on chance, what is the conditional probability that a 3rd- or 4th-year student would win the game?

Answer: *31.7%*

Note: You may want to explain to students that this probability is an empirical estimate. A theoretical probability could also be derived. However, for this investigation, the probability based on the above sample will be considered the probability of winning the game.

Discuss with students the following definition of *independent events*.

> Two events are *independent* when knowing that one event has occurred does not change the likelihood that the second event will occur.

Have the students answer Question 9.

9. If event A is "winning the game" and event B is "1st- or 2nd-year student," are A and B independent events?

Answer: *The probability that a randomly selected student wins the game is 30%. The probability that a 1st- or 2nd-year student wins the game is 27.5%, so winning the game is not independent of being a 1st- or 2nd-year student.*

Note: The conditional probability of winning the game for a 3rd- or 4th-year student is 31.7%. The conclusion is there is a higher probability that a student in their 3rd or 4th

Table 11.4

	Number of Students Who Won the Game	Number of Students Who Lost the Game	Total Number of Students Who Played the Game
1st- or 2nd-Year Student	12	28	40
3rd- or 4th-Year Student	18	42	60
Totals	30	70	100

Table 11.5

	Conditional Probability of Winning	Conditional Probability of Losing	Totals
1st- or 2nd-Year Student	12/40 = 0.30	28/40 = 0.70	40/40 = 1.00, or 100%
3rd- or 4th-Year Student	18/60 = 0.30	42/60 = 0.70	60/60 = 1.00, or 100%
Totals	30/100 = 0.30, or 30%	70/100 = 0.70, or 70%	100/100 = 1.00

year will win the game as compared to one in their 1st or 2nd year.

Direct your students to complete questions 10 and 11 based on the hypothetical results in Table 11.4.

10. Using the hypothetical results in Table 11.4, complete the row conditional probability of events based on year in school in Table 11.5.

Answer: *Table 11.5*

11. Using the hypothetical results, if event A is "winning the game" and event B is "1st- or 2nd-year student," are events A and B independent events?

Answer: *The probability that a randomly selected student wins the game is 30%. The probability that a 1st- or 2nd-year student wins the game is 30%, so winning the game is independent of being a 1st- or 2nd-year student.*

Part II: Using Simulation to Determine if Two Events Are Independent

Note: Part II simulates many samples of size 100 from a population of all Rufus King High School students in which the two events are assumed to be independent and uses that distribution to determine how likely a random sample of 100 students would produce 27.5% of 1st- or 2nd-year students winning the game. Recall that 27.5% was the observed percentage found in Investigation 10.

Refer your students to the sample the computer science students took. The row conditional relative frequency table based on year in school is shown in Table 11.6.

Ask your students to help complete a row conditional relative frequency (Example shown in Table 11.7) based on the year in school.

Table 11.6

	Number of students who won the game	Number of students who lost the game	Total number of students who played
1st- or 2nd-year Student	10	30	40
3rd- or 4th-year Student	20	40	60
Total	30	70	100

Draw a number line on the board and record the proportion of 1st- or 2nd-year students who would win the game.

Example:

Ask your students:

What does the 0.25 represent?

Answer: *Assuming the two events are independent, 25% of the 1st- or 2nd-year students won the game.*

Based on the sample result of 25%, are you convinced the two events are independent?

Explain that we need a large number of simulation results to draw a conclusion.

Place students in small groups. Each group will require a bag of the slips of paper cut out from the template. Have your students refer to Worksheet 11.2 Simulation.

Simulation Steps per Trial:

Step 1: Thoroughly mix the slips of paper in the paper bag.

Step 2: Pick 30 slips representing the students who won the game.

Step 3: Count the number of slips that have a 1 on them.

Step 4: Determine the estimated probability of a 1st- or 2nd-year student winning the game and record the estimated probability on a data recording sheet similar to the following:

Trial number	Number of slips representing 1st- or 2nd-year students winning the game	Probability estimate that a 1st- or 2nd-year student wins the game
Example	10	10/40 = 0.25
1		
2		
3		
4		
5		

Step 5: Repeat steps 1 to 4 at least four more times (for a total of five trials). Record each trial result on the recording sheet.

After each small group of students has collected these data for at least five trials, direct each group to add their results to the dot plot on the board.

Note: If more trials are needed, use the 25 simulations in Table 11.8 obtained by the above process.

Figure 11.1 is the dot plot of the 25 simulations.

Option: Direct students to obtain the results of the simulation from an applet or

Table 11.7

	Conditional Probability of Winning	Conditional Probability of Losing	Totals
1st- or 2nd-Year Student	10/40 = 0.25	30/40 = 0.75	40/40 = 1.00
3rd- or 4th-Year Student	20/60 = 0.33	40/60 = 0.66	60/60 = 1.00
Totals	30/100 = 0.30	70/100 = 0.70	100/100 = 1.00

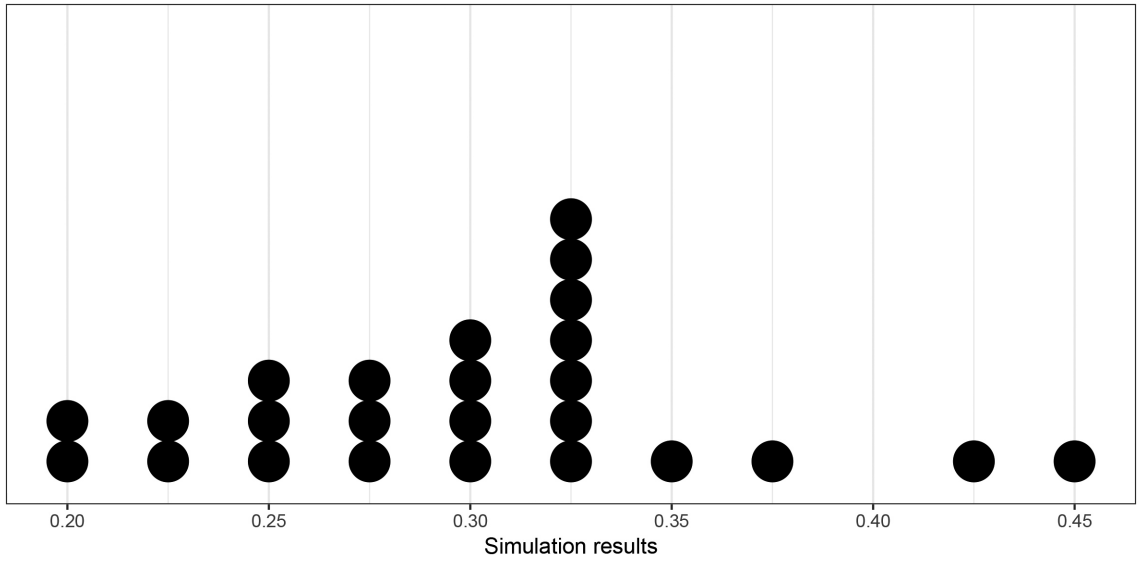


Figure 11.1: Dot plot of the 25 simulations

statistical software application like StatKey (*www.lock5stat.com/StatKey*).

Interpret the Results in the Context of the Original Question

Ask your students to answer questions 12 and 13 that estimate the likelihood of 1st- or 2nd-year students winning the game if the events are independent.

12. Based on the class dot plot of the simulated probabilities, what estimates of the proportion of a 1st- or 2nd-year student winning the game are most likely to occur under the assumption that the probability that 1st- or 2nd-year students win the game is 30%? Explain your answer.

Answer: *Students would identify the probabilities that occurred the most from the simulations. For the dot plot of 25 simulations, answers such as 0.25 to 0.32 would be expected. Note that most of the simulations were within that interval. As the number of simulations are added to the dot plot, the build-up of the values around 0.3 is more pronounced.*

13. Do you think the sample of 100 students collected by the computer science students could have come from a population in which the events of grade level and winning the game are independent? Explain your answer.

Answer: *A proportion 27.5% representing the probability of 1st- or 2nd-year students winning the game fits within the interval describing most of the expected proportions from the simulations. As a result, this sample is likely to have been drawn from a population in which the events are independent, or nearly independent.*

Table 11.8 Twenty-Five Simulations

Trial	1	2	3	4	5
Probability of a 1st- or 2nd-Year Student Winning	10/40 = 0.25	12/40 = 0.300	13/40 = 0.325	11/40 = 0.275	13/40 = 0.325
Trial	6	7	8	9	10
Probability of a 1st- or 2nd-Year Student Winning	13/40 = 0.325	15/40 = 0.375	13/40 = 0.325	8/40 = 0.200	13/40 = 0.325
Trial	11	12	13	14	15
Probability of a 1st- or 2nd-Year Student Winning	14/40 = 0.350	11/40 = 0.275	18/40 = 0.450	12/40 = 0.300	12/40 = 0.300
Trial	16	17	18	19	20
Probability of a 1st- or 2nd-Year Student Winning	17/40 = 0.425	9/40 = 0.225	13/40 = 0.325	12/40 = 0.300	10/40 = 0.250
Trial	21	22	23	24	25
Probability of a 1st- or 2nd-Year Student Winning	9/40 = 0.225	8/40 = 0.200	10/40 = 0.250	11/40 = 0.275	13/40 = 0.325

Exit Ticket

The students selected in the original sample collected by the computer science class also answered the question of whether a student participates in the school's extracurricular activities. Based on this sample, do you think the events of participation in a school's extracurricular activities and grade level are likely to be independent events in the population? Explain your answer.

	Participates in the School's Extracurricular Activities	Does Not Participate in the School's Extracurricular Activities	Total
1st- or 2nd-Year Student	34	6	40
3rd- or 4th-Year Student	6	54	60
Totals	40	60	100

Answer: *Students determine the conditional probabilities using the year in school of the students. (See table below.) The probability of 85% that a 1st- or 2nd-year participates in extracurricular activities is quite different than the 40% for all students participating in the school's extracurricular activities. This major difference indicates that the assumption the events are independent is not likely to be accurate. If a student is selected who is a 1st- or 2nd-year student, the estimate this student participates in extracurricular activities is higher than if the student selected was a 3rd- or 4th-year student.*

	Participates in the School's Extracurricular Activities	Does Not Participate in the School's Extracurricular Activities	Total
1st- or 2nd-Year Student	34/40 = 0.85, or 85%	6/40 = 0.15, or 15%	40/40 =1.00, or 100%
3rd- or 4th-Year Student	6/60 = 0.10, or 10%	54/60 = 0.90, or 90%	60/60 = 1.00, or 100%
Totals	40/100 = 0.40, or 40%	60/100 = 0.60, or 60%	100/100 = 1.00, or 100%

Extension

Close? Close Enough?

The conclusion in this investigation was the sample was likely to have been drawn from a population in which the events are independent, or nearly independent. We observed 27.5% for 1st- or 2nd-year students who won the game, and this is relatively close to 30%.

The probabilities in this investigation were estimated empirically based on a random sample of 100 students. Students were concerned, however, that the probabilities were not equal, and although within an interval that included most of the probabilities from a simulation, were the probabilities close enough to conclude they were independent events? What is "close enough"? It was noted by students that rarely will probabilities derived from the sample be equal.

Investigation 18, "How Stressed Are You?" investigates a similar statistical question in which the difference between two proportions based on two categories are close enough to conclude there is no significant difference in the categories. When is the difference close enough to indicate the categories are not significantly different? For now, students will use their best judgment to estimate whether they think the probabilities are close enough based on a comparison to simulated probabilities. Acknowledge, however, that estimating whether the probabilities are close is important and will be more precisely defined. Topics involving p-values or confidence levels will be addressed as they continue their study of statistics. The answers to these questions are especially important as students move to a more precise study of inferential statistics.