Section II: One-Variable Data Analysis

Investigation 1

Could You Be an Olympic Swimmer?

Graphical Displays

Overview

This investigation develops the concept of representing a distribution with a dot plot and a box plot and using these graphical representations to answer a statistical question. Students are asked to measure their height and arm span. They then compare the ratios of arm span to height of the class to the ratio of arm span to height of Michael Phelps. This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are: formulate a statistical question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level A activity.

Instructional Plan

Brief Overview

» Have students read and discuss the scenario concerning Michael Phelps.

» Develop a statistical question concerning the distribution of arm span to height ratios.

» Discuss with students how to measure their arm span and height.

» Students collect and record their measurements and find the ratios of arm span to height.

Wandering Albatross

» Students use class data to create a box plot and a dot plot.

» Students analyze the class ratios and compare their ratios to Michael Phelps's ratio.

Hand out Student Worksheet 1.1 Height and Arm Span

Scenario

Do you know what type of bird has the largest wingspan?

The "Wandering Albatross" has been declared the *bird with the largest wingspan* among all living birds. Its wingspan, on average, is from 8.2 to 11.5 feet. Its length is from 3.51 to 4.43 feet. This wandering albatross breeds in several islands north of the Antarctic Circle and feeds off the coast of New Zealand. One albatross that was banded and followed by scientists was reported to have traveled around 3,700 miles in just 12 days.

An ornithologist is a scientist who studies every aspect of birds. One aspect an ornithologist uses to compare bird species is the ratio of a bird's wingspan to the length of its body. What is the ratio of the Albatross's wingspan to length?

Answer: *Using the maximum values for the wingspan and length, the ratio is 11.5 to 4.43, or the Albatross's wingspan is about 2.5 times its length.*

Why do you think the Albatross has such a large wingspan when compared to its length?

Possible answer: *The Albatross travels great distances and needs a greater wingspan to conserve energy and be able to glide and not flap its wings as much.*

Learning Goals

- Describe the distribution of a quantitative variable from a dot plot and box plot.
- Draw appropriate conclusions based on box plots and dot plots.

Mathematical Practices through a statistical lens

MP2. Reason abstractly and quantitatively

Statistically proficient students reason in the presence of variability and anticipate, acknowledge, account for, and allow for variability in data as it relates to a context.

Materials

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

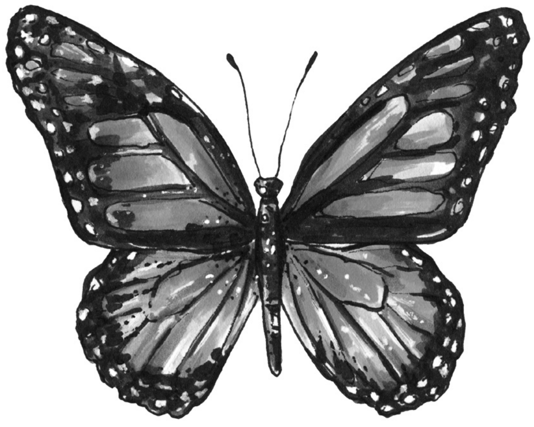
- Measuring tapes, meter sticks, rulers, masking tape, poster paper
- Student Worksheet 1.1 Height and Arm Span
- Student Worksheet 1.2 Measuring Directions
- Student Worksheet 1.3 Sample Data
- Student Exit Ticket
- Optional: Technology to find summary statistics and construct box plots and dot plots

Estimated Time

Two 50-minute class periods—approximately one period for data collecting and a second period for the analysis of the collected data

Pre-Knowledge

Students should already be able to construct a dot plot and a box plot of a set of data.



Monarch butterfly

While an ornithologist studies birds, a lepidopterist studies butterflies and moths. The wingspan of a Monarch butterfly is from 3 ½” to 4” and its length is from 1 ¼” to 1 3/8 ”.

What is the ratio of a Monarch butterfly's wingspan to length ratio?

Answer: *Using the maximum values for wingspan and length, the ratio is 4 to 1 3/8, or about 2.9. A Monarch's wingspan is almost 3 times its length.*

How does a ninth grader's arm span to height ratio compare to that of the wingspan to length ratio of the Albatross and Monarch butterfly? Do the students follow a similar pattern, in which the wingspan is many times the length?

Do you know who is the most-decorated Olympic swimmer?

Note: If time permits, show the video of Michael Phelps winning the 200m individual medley at the 2016 Rio Olympics: *www.youtube.com/watch?v=e-XGSYnhUjg*.

Michael Phelps is the most decorated Olympian of all time, with a total of 28 medals (as of the 2016 Olympics). Phelps also holds the all-time record for Olympic gold medals with 23. At the 2016 Summer Olympics in Rio de Janeiro, he won five gold medals and one silver, more than any other competitor for the fourth Olympics in a row.

Formulate a Statistical Question

Share a photo of Michael Phelps swimming that shows his arm span (one example is at *https://itraintherefore ieat.com/tag/michael-phelps*).

Discuss with your students how some people have suggested one reason Michael can swim so fast is because he has unusually long arms compared to his height. Michael stands 6’4” (193.04 cm) and has an arm span of 6’7” (200.66 cm). The ratio of his arm span to his height is 200.66/193.04 or 1.039.

Ask your students if they think his ratio is unusual.

Ask your students what might be some other reasons Michael has been so successful.

Possible answers: *Training, excellent physical shape, large hands and feet*

Ask your students if they think anyone in class has a greater ratio of arm span to height than Michael Phelps and, if so, mention that maybe they could be an Olympic champion. Encourage your students to consider the statistical question: “What is the typical arm span to height ratio of students in our class?” Also suggest they consider the question, “How does the ratio of arm span to height of Michael Phelps compare to the students in our class?”

Collect Appropriate Data

Following are three options for collecting the appropriate data.

Option 1: Students create procedures and help set up measuring stations.

Swimmer with a large arm span

Prior to placing students in groups and measuring their arm span and height, discuss some of the following points:

» Explain that they will be measuring their arm span and height and then investigating any patterns in the class *ratios* of arm span to height.

» Ask them how they could measure their arm span. Tip of finger to tip of finger? Tape a measuring tape or rulers to the wall? Make measuring strips?

» How do they plan to measure their height? Shoes off or on? Tape rulers to the wall? Or tape a measuring tape to the wall? Or their own measuring strip?

» What unit of measure should be used? How precise should the measurements be made? *Suggest the students use the metric system and measure to the nearest cm.*

After the class discussion, write the student-generated procedures on the poster paper and place near the measuring stations.

Provide each group with measuring tapes/rulers and masking tape.

Allow students time to collect their height and arm span. Draw a table on the board or set up a table on the classroom computer where students can record their arm span and height. Include Michael Phelps's measurements on the first row of the table.

Option 2: Students follow directions at teacher set-up stations.

Have an appropriate number of height measuring stations and arm span measurement stations set up before class or have students help set up the stations. You may wish to secure a tape measure to the wall, secure 2- or 3-meter sticks to the wall, or have students create their own measuring strip of paper.

Place your students into groups. Hand out Student Worksheet 1.2 Measuring Directions. This worksheet describes one method students can use to set up a measuring station for their height and arm span. If using a different method, Student Worksheet 1.2 Measuring Directions is not needed.

Discuss with your students how to set up each measuring station and how to measure their height and arm span.

Allow students time to collect their height and arm span. Draw a table on the board or set up a table/spreadsheet on a calculator/computer where students can record their arm span and height. Include Michael Phelps's measurements on the first row of the table. The students should record the ratio of their arm span to their height in the same row as their arm span and height measurements (see Table 1.1).

Option 3: If time or resources are a problem, provide students with sample data collected from the American Statistical Association Census at

Table 1.1 Sample Table

Student	Arm Span (cm)	Height (cm)
Michael Phelps	201	193
1		
2		

Table 1.2 Example

Student	Arm Span (cm)	Height (cm)	Arm Span / Height
Michael Phelps	201	193	1.04
1			
2			

School website: *ww2.amstat.org/censusatschool* (Student Worksheet 1.3 Sample Data).

Analyze the Data

After students have recorded their arm span and height, ask them to determine the ratio of the arm span to height for Michael Phelps. Then ask each student to determine their ratio of arm span to height. Record the ratios in the table (example in shown in Table 1.2) in the appropriate row.

Ask what Michael Phelps's ratio of 1.04 means.

***Answer:** This means Michael Phelps's arm span is just slightly larger than his height because the ratio is just a little more than 1.*

Explain that we want to create different graphical representations of the ratios of arm span to height and use these representations to compare the class ratios to Michael Phelps's ratio of 1.04.

Visualizing the Data with Dot Plots

Refer to Student Worksheet 1.1 Height and Arm Span. Ask the students to complete questions 1 to 4 while working in their groups.

Discuss the answers.

Answers using data from Student Worksheet 1.3 Sample Data

1. What does it mean if a person's arm span to height ratio is equal to 1? Less than 1? More than 1?

Answers:

» *A ratio equal to 1 means the arm span and height of an individual are equal.*

» *A ratio less than 1 means the arm span is less than the height.*

» *A ratio greater than 1 means the arm span is greater than the height.*

2. Construct a dot plot of the class ratios of arm span to height. Include Michael Phelps's ratio.

***Answer:** Figure1.1*

3. Describe the center, shape, and spread of the data.

***Answer:** The center is around 0.98 (median is 0.988 and mean is 0.974), the distribution is skewed left, and the data spread from 0.829 to 1.042 with a cluster of data from about 0.96 to 1.04.*

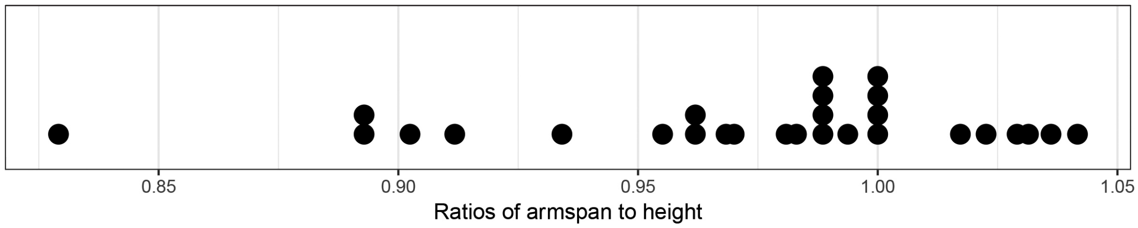


Figure 1.1: Dot plot of the class ratios of arm span to height

4. Using the dot plot, what can you conclude about the ratio of arm span to height for the students in class?

Possible answer: The shape of the distribution is skewed left with a possible outlier at 0.83. Most students have about the same height and arm span because the data center a bit less than 1.

Visualizing the Data with Box Plots

Working in their groups, ask the students to complete questions 5 and 6 on Student Worksheet 1.1.

Answers use data from Student Worksheet 1.3 Sample Data.

5. Construct a box plot of the class ratios of arm span to height. Include Michael Phelps's ratio in the class data. Use the same scale used for the dot plot.

Answers based on sample data:

» *Figure 1.2*

» *minimum = 0.829*

» *Q1 = 0.958*

» *median = 0.988*

» *Q3 = 1*

» *maximum =1.042*

6. What percent of the ratios are less than the lower quartile? What percent of the ratios are less than the upper quartile?

Answers:

» *Approximately 25% of the ratios are less than the lower quartile.*

» *Approximately 75% of the ratios are less than the upper quartile.*

Have your students answer questions 7 to 9.

Note: If students are not familiar with the concept of an outlier, demonstrate how to find an outlier using the following definition:

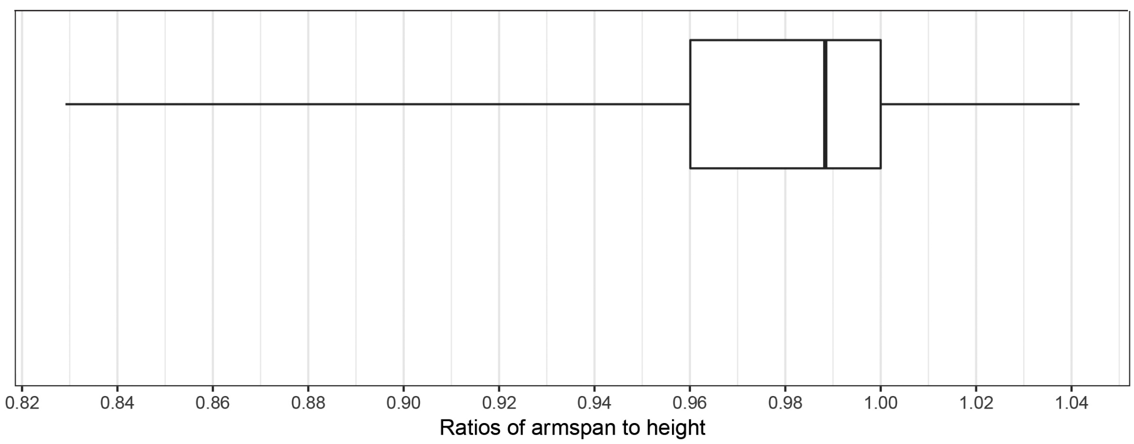


Figure 1.2: Box plot of the class ratios of arm span to height

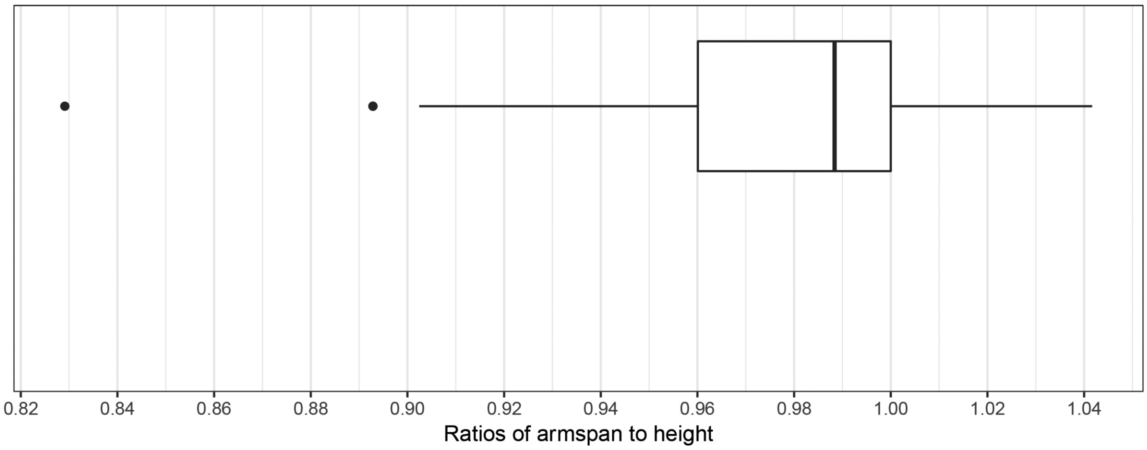


Figure 1.3: Box plot of the class ratios of arm span to height showing outliers

> A data point is an outlier if it falls more than 1.5*(IQR) above the upper quartile or more than 1.5* (IQR) below the lower quartile.

Note: The interquartile range (IQR) is the difference between the upper quartile (Q3) and lower quartile (Q1).

7. Use the given definition of an outlier and determine if there are any outliers. Is Michael Phelps's ratio an outlier?

Answers:

» *There is an outlier at 0.829 and 0.893 because they are less than 0.958-1.5(1-0.958) = 0.895.*

» *Michael Phelps is not an outlier because (using the sample data) 1.042 is not greater than 1+1.5(1-0.958) = 1.063.*

8. Using the box plot, what can you conclude about the ratio of arm span to height for the students in class?

Possible answer: *Students should comment on the variability of the data, including the middle 50%. They should recognize if there are outliers and comment about how Michael Phelps's ratio compares to the class ratios. For example: "The middle 50% of the students have ratios between 0.958 and 1, telling me that about half of the class has an arm span equal to or slightly less than their height. There are two outliers, indicating these ratios are not like the rest of the class."*

9. Explain how the box plot and dot plot each helped in comparing the class data with Michael Phelps's ratio of 1.04.

Possible answer: *The box plot helps us see whether there are any outliers and the overall spread of the data. The dot plot is helpful in describing the shape and estimating the mean of the distribution.*

If there are outliers, then demonstrate how to change the box plot to show outliers. Place an asterisk or x at each outlier and then draw the segment from the quartile to the most extreme data point that is not an outlier (see Figure 1.3).

Answer using data from Student Worksheet 1.3 Sample Data (shows outliers at 0.829 and 0.893; the segment or whisker ends at 0.902).

Interpret the Results in the Context of the Original Question

Allow time for your students to complete Question 10.

Discuss the answers to Question 10. Have the students share their summary with other groups.

10. Write a summary answering the statistical question. Your summary should include how your ratio compared to others in class and to Michael Phelps's ratio. Also, how did the class ratios compare to Michael Phelps's ratio?

Possible answer: *Students should comment on the position of Michael Phelps's ratio—is it an outlier, are there students who have a ratio greater than Michael Phelps? For example, "My ratio of ___ is (a lot more/less, a little more/less) compared to the rest of the class and (similar comparison statements) compared to Michael Phelps. This implies that Michael Phelps is not that unusual for his body ratio. There are probably other factors that contribute to his success. However, having arms much shorter than one's height might not make the best swimmer."*

Exit Ticket

How much of a tip should you leave at a restaurant? Currently, the standard tip is between 15% and 20% of the pre-tax amount of the bill. Sara, a high-school senior was working part- time at a local diner. Most items on the menu were from $6.50 to $12.00, and most of the tables she served had either one or two diners. One Saturday, Sara kept track of the tips she earned for 25 tables she served. Below (Figure 1.4) is a box plot and dot plot of the tip amounts.

Using the two graphical representations, describe the distribution of the amount of the tips Sara received on the Saturday. Include in your description the center and spread of the data.

Possible answer: *The median tip value was $1.70, and the mean was $2.08. The distribution is skewed right, with two outliers at $5.75 and $6.00. Most of Sara's tips were between $1.00 and $2.50.*

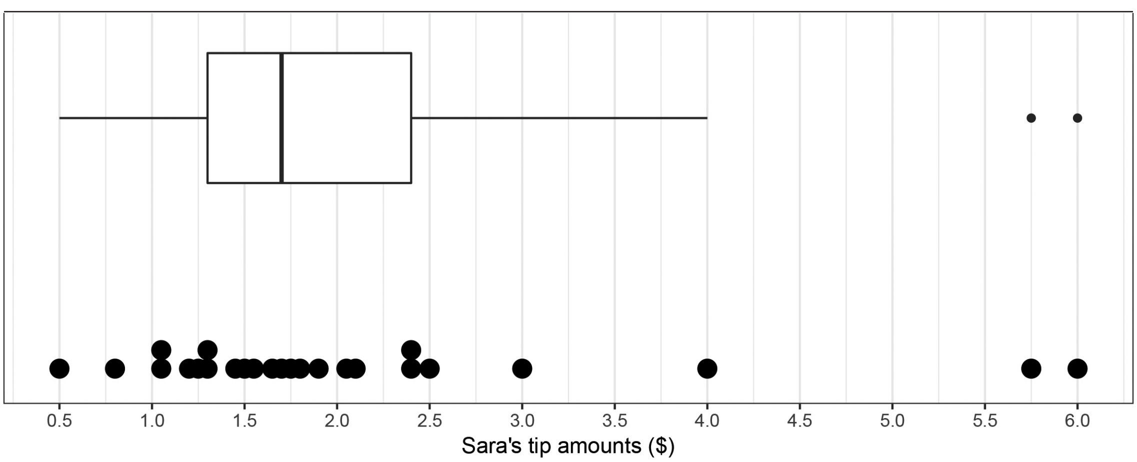


Figure 1.4: Box plot and dot plot of the tips Sara earned for 25 tables she served

Further Exploration

- Have students measure the distance from the top of their head to their chin. Have them investigate the relationship of the ratio of this distance to height for the students in their class. For an average adult, the total height is equivalent to 7 to 7.5 heads tall. How do the students' ratios compare to this standard?
- Have students measure the distance from the bottom of their nose to the outside corner of their right eye and the length of their right ear. For an average adult, the ratio of the distance to the eye and length of the ear is about 1. How do the students' ratios compare to this standard?
- Have students measure the width of their head and length of an eye. For an average adult, the ratio of head width to eye length is between 4 and 5. How do the students' ratios compare to this standard?

Investigation 2

Are Baseball Games Taking Longer?

Comparing Multiple Groups

Overview

This investigation asks students to use graphical displays—box plots and dot plots—to describe the distribution of a quantitative variable (length of major league baseball games). They compare the distributions of the length of games in baseball seasons from three decades using parallel box plots and summary statistics. The investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are formulate a statistical question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level A/B activity.

This investigation is based on an article in *USA Today,* "Average MLB Game Time Rises to Record High," published October 2, 2017. Read it at *www.usatoday.com/story/sports/mlb/2017/10/02/average-mlb-game-time-record-3-hours-5-minutes-this-season/106225166*.

Instructional Plan

Brief Overview

» Have students read and discuss the scenario about the length of baseball games from three years.

» Formulate a statistical question comparing the distribution of the length of baseball games in three years.

» Have students construct a box plot and dot plot for each of the three years.

» Construct parallel box plots and dot plots. Use the plots to compare the length of games in the three years.

Scenario

Have you been to a major league baseball (MLB) game recently? Or maybe watched one on television? What did you think about the length of the game? Was it too long, too short, or just about right?

The biggest complaint of many baseball fans—both young and old—is that the pace of the game is too slow. Throughout the decades, the length of major league games seems to have increased. Fans give all sorts of reasons for why the games might be getting longer. Some say it's due to more and longer TV commercial breaks; others say it's because of the multiple mid-inning pitching changes; still others suggest it's due to the use of replay by umpires to decide close calls.

What might be some other reasons the length of MLB games would increase?

Possible answer: *Constant stepping out of hitters from the batters' box and pitchers taking an inordinate amount of time between pitches with no one on base.*

What might be some suggestions for speeding up MLB games?

Possible suggestions: *Shorten the number of innings played and limit the number of pitching changes.*

Formulate a Statistical Question

Discuss the following with your students:

For a statistical study, a small group of high-school students wanted to investigate how the length of major league baseball games has changed over time. They decided to look

Learning Goals

- Describe the distribution of a quantitative variable from a dot plot and box plot.
- Compare the distributions of a quantitative variable using parallel box plots and summary statistics.

Mathematical Practices Through a Statistical Lens

MP6. Attend to Precision

Statistically proficient students are precise about choosing the appropriate analyses and representations that account for the variability in the data. They display carefully constructed graphs with clear labeling. As students interpret the analysis of the data, they are precise with their terminology and statistical language.

Materials

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

- Student Worksheet 2.1 Length of Baseball Games
- Student Worksheet 2.2 Analyzing the Times
- Student Exit Ticket
- Optional: Worksheet 2.3 to be used with Addition Ideas Section: Directions to Use New Zealand Census at School Website, *http://new.censusatschool.org.nz/tools/random-sampler*
- Optional: Technology to find summary statistics and construct dot plots and box plots

Estimated Time

Two 50-minute class periods

Pre-Knowledge

Students should be able to describe the shape, center, and variability of a distribution of a quantitative variable. They should also know how to construct a dot plot and box plot of a quantitative variable.

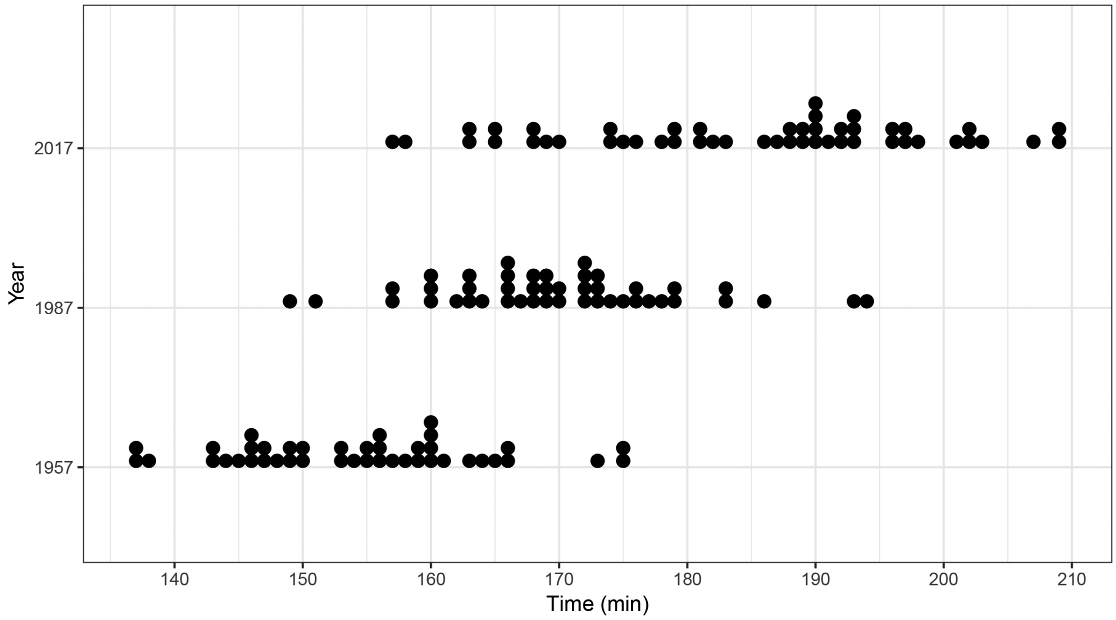


Figure 2.1: Dot plots for each of the three years included in the random sample

at one year toward the end of each of three decades—1950s, 1980s, and 2010s. Rather than try to manage the analysis of all the games played in 1957, 1987, and 2017, they decided to take a random sample of the length of the games from those three years.

The students used the data to investigate the statistical question, "Has there been an increase in the length of time to play regular season major league baseball games in 1957, 1987, and 2017?"

Collect Appropriate Data

Distribute Student Worksheet 2.1 Length of Baseball Games.

Point out that the data collected are from 50 randomly selected games from 1957, 46 games from 1987, and 43 games from 2017. The data were collected from samples of 9-inning games. Any games that lasted more than 9 innings, had rain delays, or were shortened due to weather were not included.

The data can be found at *www.baseball-reference.com/leagues/MLB/misc.shtml*.

Analyze the Data

Hand out Student Worksheet 2.2 Analyzing the Times.

Place students into groups of four and have the groups complete questions 1 to 7 on Student Worksheet 2.2 Analyzing the Times.

1. Work with members of your group to construct a dot plot of the sample lengths for baseball games in 1957. Use a scale from 130 minutes to 210 minutes.

Possible answer: *Figure 2.1 shows the dot plots for each of the three years.*

2. Using the dot plot for the year 1957, estimate the center of the distribution. Describe the spread of the data.

Possible answer: *The center of the distribution is approximately 155 min. The spread of the* *distribution goes from approximately 135 to 175 min. The distribution mounds up around 160 min.*

3. Work with members of your group to construct a dot plot of the sample lengths for baseball games in 1987. Place the dot plot above the dot plot for year 1957.

Possible answer: *See Figure 2.1.*

4. Using the dot plot for the year 1987, estimate the center of the distribution. Describe the spread of the data.

Possible answer: *The center of the distribution is approximately 170 min. The spread of the distribution goes from approximately 145 to 195 min. The distribution mounds up around 170 min.*

5. Work with members of your group to construct a dot plot of the sample lengths for baseball games in 2017. Place the dot plot above the dot plots for the years 1957 and 1987.

Possible answer: *See Figure 2.1.*

6. Using the dot plot for the year 2017, estimate the center of the distribution. Describe the spread of the data.

Possible answer: *The center of the distribution is approximately 185 min. The spread of the distribution goes from approximately 155 to 210 min. The distribution mounds up around 190 min.*

7. Using the three dot plots, what observations can you make concerning the length of the games in the three years? Comment on the center and spread of each dot plot.

Possible answer: *The centers have increased from 1957 to 2017. The minimums and maximums have also increased from 1957 to 2017.*

Have the students answer questions 8 to 10 and then discuss their answers.

8. To help compare the length of the games, work with members of your group to construct a box plot for the sample lengths of baseball games in each of the three years 1957, 1987, and 2017. Place the three box plots on the same number line with a scale from 130 minutes to 210 minutes to form parallel box plots.

Answer: *Figure 2.2 shows the box plots for each of the length of the games in the years 1957, 1987, and 2017.*

Using the parallel box plots of the samples of lengths of games, how do the length of major league games in 1957, 1987, and 2017 compare? Comment on the center and spread for each distribution.

Possible answer: *The length of the games appears to have been increasing from 1957 to 2017. Seventy-five percent of the lengths in 2017 are greater than all the games in the 1957 sample and greater than 75% of the 1987 games. The median of the 2017 games is greater than both the 1987 games and 1957 games. At least 75% of the 1987 games are greater than 75% of the 1957 games.*

9. What advantages and disadvantages do box plots have over dot plots for making comparisons between multiple groups of data?

Possible answer: *The parallel box plots make it easy to compare the centers (medians) and overall spread of the data. The dot plots give the overall shape of each distribution, which can be helpful when comparing distributions.*

Interpret the Results in the Context of the Original Question

Have the students answer Question 10.

10. Based on the three dot plots and parallel box plots you constructed, do

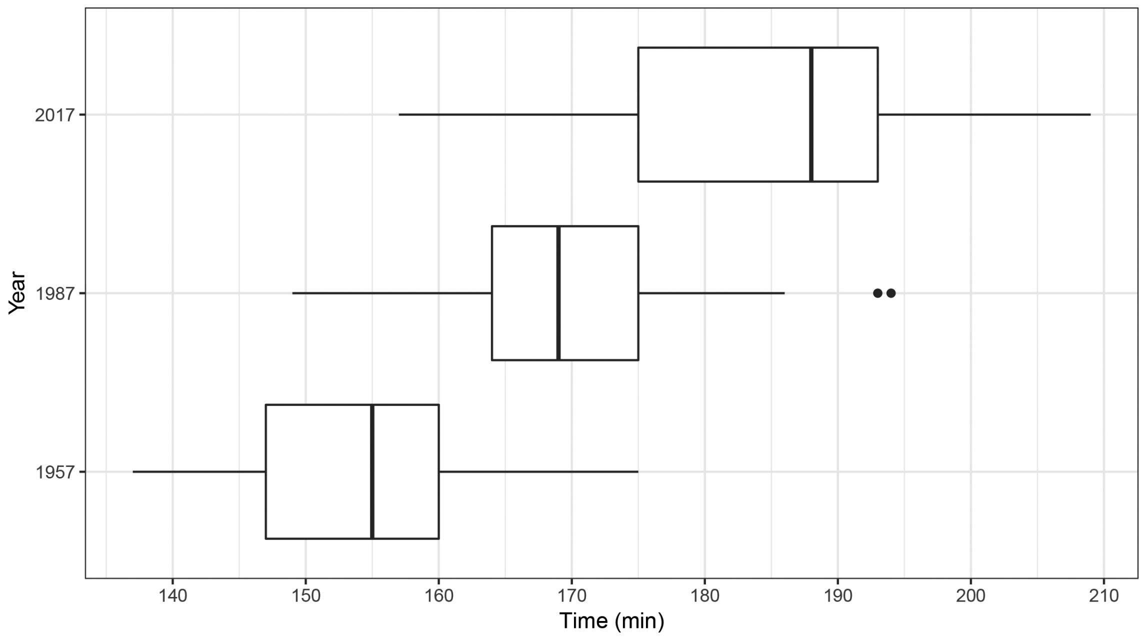


Figure 2.2: Box plots for each of the three years included in the random sample

you think the length of the games has changed by any meaningful amount? Explain your thinking.

Possible answer: *Yes, the median for 2017 is 17 minutes longer than the 1987 games and more than 30 minutes longer than the 1957 games. All the 2017 games are longer than the bottom half of the 1957 games. Seventy-five percent of the 2017 games are longer than the 1957 games.*

Additional Ideas

New Zealand Census at School hosts the random sampler (*http://new.censusatschool.org.nz/tools/random-sampler*) for international, New Zealand, and US data that have been "cleaned"—data entered incorrectly have been removed.

Using the random sampler at the New Zealand Census at School website, take a random sample of at least 75 ninth graders. Take the responses from Question 8—"What is the main method of transportation you typically use to get to school?"—and the answer to Question 9—"How long does it usually take you to travel to school?"—and create a table showing a summary of the responses. Then, using the table, create graphical representations that will help answer the statistical question, "How do the length of times to get to school compare for ninth graders who walk, ride a bus, or ride in a car?"

Directions for using New Zealand Census at School are on Student Worksheet 2.3.

Exit Ticket

On February 28, 1983, the final episode of M*A*S*H* aired on CBS. As of 2017, it remained the most-watched TV series finale. It was estimated that at least 105.9 million people watched this last show with a household rating of 60.2%. A 60.2% household rating means 60.2% of all households—homes with a TV set—were tuned to the final episode of M*A*S*H. Source: *https://en.wikipedia.org/wiki/List_of_most_watched_television_broadcasts_in_the_United_States*

The parallel box plots in Figure 2.3 show the household ratings in the past four decades for most of the top 66 TV series finale broadcasts. The data were sorted by the three major networks—ABC, CBS, and NBC—and a box plot was constructed of the household ratings for finale TV shows for each of the networks. Data are from 2017.

1. Describe the distribution of household ratings for each of the three networks. Include the center and spread of the data in your description.

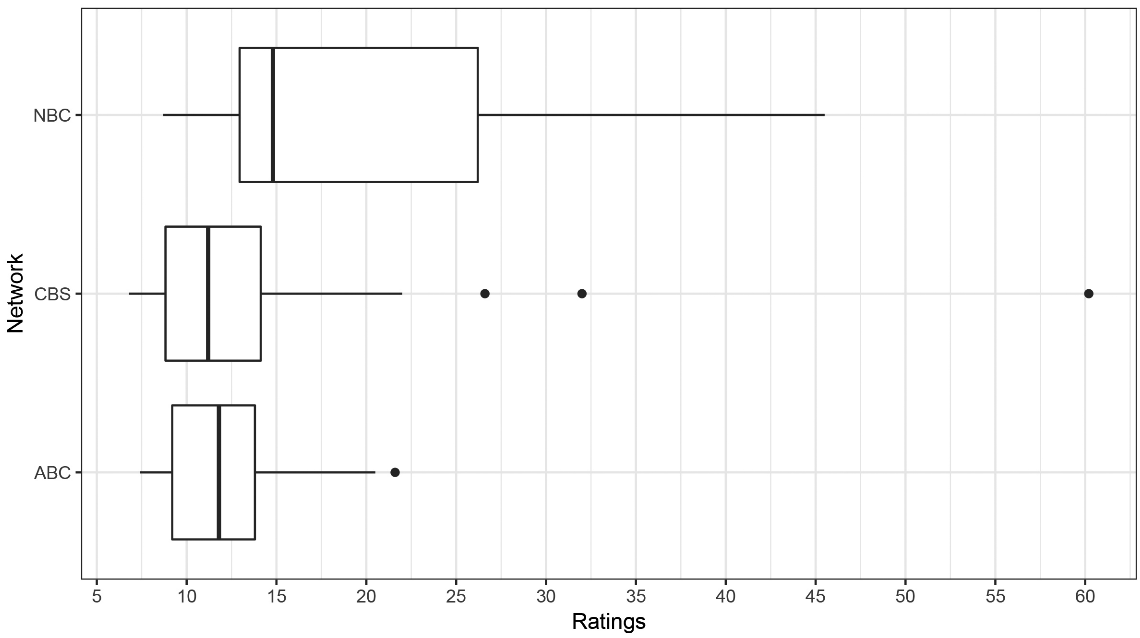


Figure 2.3: Box plots showing the household ratings in the past four decades for most of the top 66 TV series finale broadcasts

Possible answer:

- *ABC: The center has a rating of about 12 with an outlier at approximately 22. The spread of the data is small with the middle 50% of the data between 8 and 14, or an IQR equal to 6.*
- *CBS: The center has a rating of around 11 with three outliers at approximately 27, 33, and 61. The spread around the median is small with an IQR of approximately 6.*
- *NBC: The center has a rating around 14 with no outliers. NBC has a larger spread than the other two networks with an IQR of approximately 15.*

2. Which network would you rank as the top network when comparing the household ratings for the top 66 TV series finale broadcasts? Give reasons for your answer.

Possible answer: *NBC ranked highest because it has the highest median and the top 50% of their show ratings were greater than 75% of ABC's and CBS's shows.*

Further Explorations and Extensions

Common Core State Standard 7.SP.A.3 states: "Informally assess the degree of visual overlap of two numerical data distributions with similar variabilities, measuring the difference between the centers by expressing it as a multiple of a measure of variability."

Informally investigate this standard looking at the difference between the medians in parallel box plots and informally deciding if the medians represent a "significant" difference.

One informal method that can be used to determine if the difference between the medians represent a "significant" difference is to assess the degree of visual overlap of the box plots. This informal assessment can be done by determining how many IQRs separate the medians.

For example, the median of 1987 and the median of 1957 differ by 14 minutes. The IQR is 12 minutes for 1987 and 13 minutes for 1957. Divide the difference between the medians by the higher of the two IQRs, 14/13 or 1.08 IQR's. This doesn't appear to be an extremely large difference.

Have students compare the difference of the medians for 2017 with 1987 and 1957 to see if there is a meaningful difference between the medians.

Investigation 3

How Good Is Your Memory?

Standard Deviation

Overview

This investigation builds on the description of the spread of a distribution by developing a key measure of variability—the standard deviation. A procedure to calculate standard deviation is developed using data from the results of a memory test. Students use the mean and standard deviation to compare fourth grade memory test completion times (from the Census at School website) to their class memory test results.

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

This investigation is based on lessons from *Exploring Measurements* by Peter Barbella, James Kepner, and Richard Scheaffer, published in 1994 by Dale Seymour Publications as part of the Quantitative Literacy Series. Though out of print, the book listed is available through book resale sites and on Amazon.com.

The data used in this investigation were collected from the Census at School website using the random sampler at *ww2.amstat.org/CensusAtSchool*.

Instructional Plan

Brief Overview

» Formulate a statistical question concerning comparing times to complete a memory test of your class and a group of fourth graders from around the United States.

» Have students take the memory test on the Census at School website.

» Develop a procedure to calculate the standard deviation.

» Use the mean and standard deviation to compare the distributions of times to complete the memory test.

Scenario

Do you remember playing the Memory game or the game Concentration when you were in elementary school? This game usually consisted of a deck of pairs of matching cards. The cards were spread out on a table face down in rows. Players took turns turning over a pair of cards. If the cards matched, then the player kept the pair. If the cards did not match, they were returned face down to their positions, and it was the next player's turn. After all the cards had been turned over and the pairs found, the player with the most pairs won the game.

How do you think you would do playing this game now? If you were playing the game by yourself, do you think you could find all the matching pairs in a very short period of time?

Learning Goals

- Develop the measure of variability and standard deviation and interpret the standard deviation in context.
- Compare two or more distributions using the mean of the distributions as a measure of center and the standard deviation of the distributions as a measure of spread.

Mathematical Practices Through a Statistical Lens

MP7. Look for and make use of structure

Statistically proficient students look closely to discover a structure of pattern in a set of data as they attempt to answer a statistical question.

Materials

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

- Computers or tablets with internet access
- Software or statistical apps to find summary statistics and construct dot plots
- Student Worksheet 3.1 Directions for Taking the Memory Test at the Census at School Website
- Student Worksheet 3.2 Fourth Grade Memory Test Times
- Student Worksheet 3.3 Memory Test Investigation
- Optional: Student Worksheet 3.4 Sample Ninth Grade Memory Test Times (use if not having students take the memory test)
- Student Worksheet 3.5 Standard Deviation Match
- Student Exit Ticket

Estimated Time

Two 50-minute class periods

Pre-Knowledge

Students should be able to estimate the mean of a distribution visually and calculate and interpret the mean of a distribution.

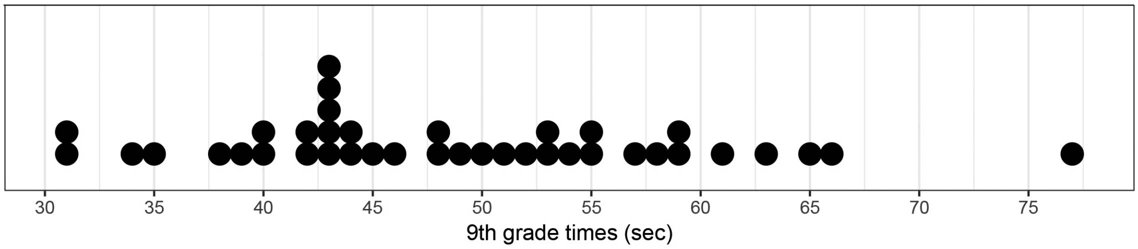


Figure 3.1: Ninth grade memory test completion times (sec.)

Formulate a Statistical Question

Explain to students that they will be taking a memory test on the Census at School website. The times to complete the test will be recorded and compared with fourth graders' times from around the United States. Do the students in class think they can beat the fourth grade completion times? Ask your students to consider the statistical question, "How do the memory test completion times of our class compare to the memory test completion times for fourth graders from around the United States?"

Collect Appropriate Data

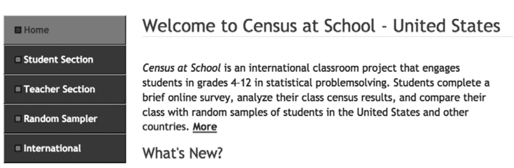
Note: If your students are not taking the memory test, then have them use the ninth grade data on Student Worksheet 3.4. The data were generated using the random sampler at the Census at School website. The data will be used to provide examples and answers to the questions.

Distribute Student Worksheet 3.1 Directions for Taking the Memory Test at the Census at School Website.

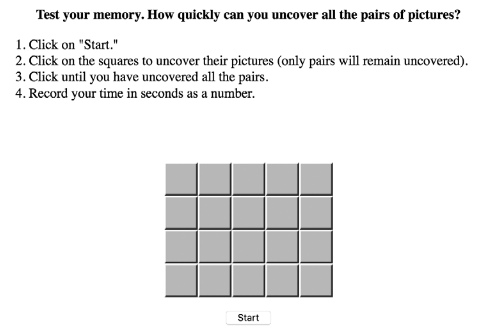
Demonstrate how to take the memory test on the Census at School website:

Go to the Census at School website at *ww2.amstat.org/CensusAtSchool*.

Choose the Student Section.



Choose the memory test on the right of the screen.



Press start and begin selecting boxes. Continue until all the matches have been found.

Direct your students to take the memory test. When they have completed the test, have them record their times as reported at the end of the test on the class dot plot and in the classroom calculator / spreadsheet / statistical software.

Note: While students are taking the memory test, draw a number line on the board for the class dot plot.

Ask your students to post their memory test results on the class dot plot.

Note: If your students did not take the memory test, then display the dot plot in Figure 3.1.

Answer using data from Worksheet 3.4 9th Grade Results.

Ask students to estimate the center of the distribution of their memory test times.

Possible answer: *Using the sample data on Worksheet 3.4, the center is approximately 50 seconds.*

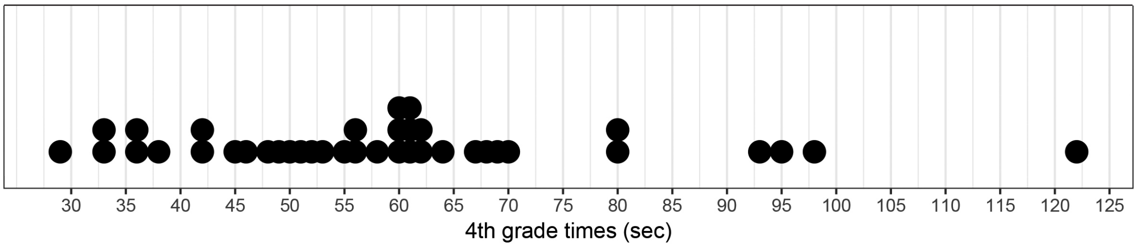


Figure 3.2: Fourth grade memory test completion times (sec.)

Analyze the Data

Distribute Student Worksheet 3.2 Fourth-Grade Memory Test Times.

Distribute Student Worksheet 3.3 Memory Test Investigation.

Ask your students to complete problems 1 to 3 on Worksheet 3.3 Memory Test Investigation.

1. To help answer the statistical question—"How do the memory test completion times of our class compare to the memory test completion times for fourth graders from around the United States?"—construct a dot plot of the fourth grade memory test times.

***Sample answer:** Figure 3.2.*

2. Describe the distribution of the fourth grade memory test times. Your description should include an estimate of the distribution's center and a description of the spread around the center.

***Possible answer:** The distribution centers around 60 sec. and has a spread from about 25 sec. to 120 sec., with much of the data from 35 sec. to 70 sec.*

3. Copy the dot plot of the class memory completion times above the dot plot of the fourth grade completion times.

***Sample answer:** Figure 3.3*

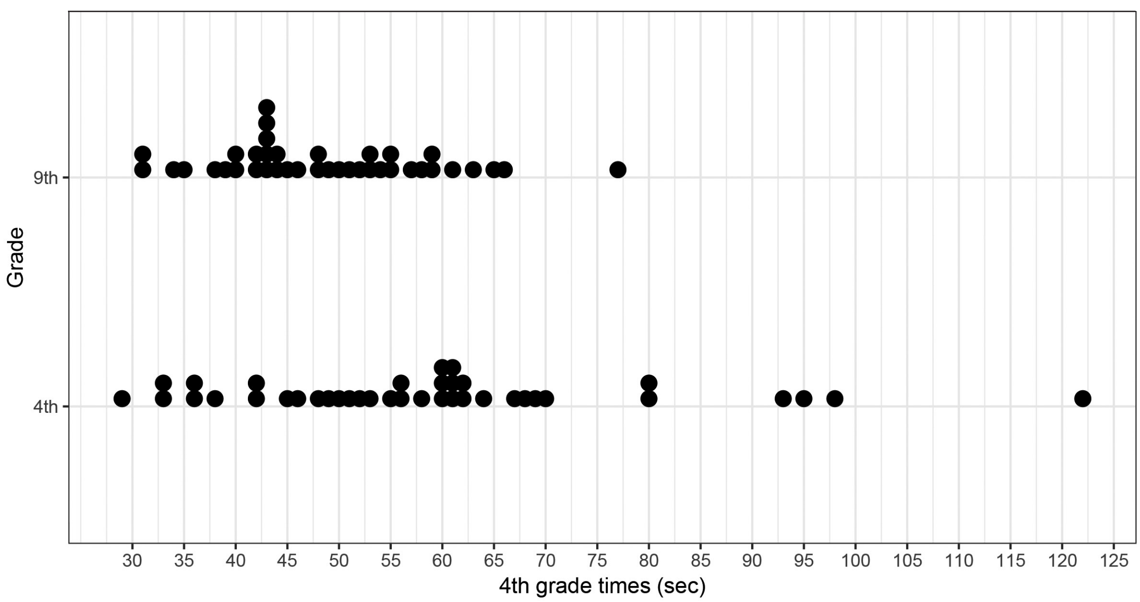


Figure 3.3: Fourth grade memory test completion times (sec.) with ninth grade memory test completion times (sec.) above.

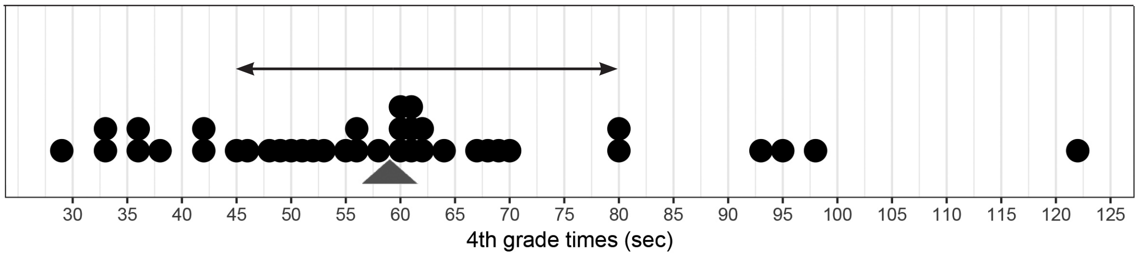


Figure 3.4: Fourth grade memory test times with the mean marked with a Δ

4. Using the dot plot of class memory completion times and the fourth grade dot plot of completion times, how do the two distributions compare?

Possible answer: *The center of the class times distribution is lower than the center of the fourth grade times. The distribution of class times is less spread out than the distribution of the fourth grade times. It appears the class times were typically better than the fourth grade times.*

When describing a distribution and comparing distributions, it is often useful to give a specific center such as the median or mean and a value that describes the variation around that center. The interquartile range (IQR) is one measure of variability that describes the variability around the median (see Investigation 2).

The next set of questions introduces a measure of variability that describes the variation about the mean. Have students work in their groups to complete questions 5 to 11.

5. Use technology (graphing calculator, spreadsheet, or app) and enter the fourth grade memory test times into a list. Use the technology to find the mean of the fourth grade memory test times. Mark the mean on the dot plot with a Δ, indicating the mean or the balance point of the distribution.

Answer: *Figure 3.4*

6. Draw an arrow from the mean of 59 sec. to the data point 45 sec. This arrow shows the distance the point 45 is below the mean. This distance is called the *deviation* from the mean.

Answer: *See Figure 3.4.*

7. Find the deviation from the mean for this data point (45) by subtracting the mean from the value of the data point.

Answer: *45-59 = -14*

8. This deviation is negative. What does this tell you about the data point in relation to the mean?

Answer: *45 sec. is 14 sec. below the mean of 59 sec.*

9. Draw an arrow from the mean of 59 sec. to the data point 80 sec. Find the deviation for the data point 80. This deviation is positive. What does this tell you about the data point in relation to the mean?

Answer: *80-59= 21 sec. 80 sec. is 21 sec. above the mean of 59 sec.*

10. Use technology (list on a graphing calculator or spreadsheet) to find the deviation from the mean for all the fourth grade times. If using a spreadsheet or lists in a graphing calculator, set a formula for a column that takes all the values of the fourth grade times minus the mean of 59.

Store these deviations in another list or column.

Table 3.1

4th-Grade Times (sec.)	Deviations from the Mean
68	9
80	21
49	-10
55	-4
69	10
62	3
98	39
42	-17
64	5
33	-26
58	-1
Etc.	Etc.

Sample of some of the answers (see Table 3.1):

11. Use technology to find the sum of all the deviations. Why does this value make sense?

Answer: 0, the mean is the balance point of the distribution, so the sum of the negative and positive deviations equals zero.

Explain to students that the objective is to find a number that summarizes all the deviations. Often, the mean is used as a summary of the data. Since the sum of the deviations is zero, the mean of the deviations would be zero. The mean of zero does not summarize the spread of all deviations from the mean. So, we need to deal with the negative deviations. We could take the absolute value of all the deviations or square of all the deviations—both turn all the negative deviations positive. If we find the absolute value of all the deviations and then the mean of these absolute deviations, we have found the value *MAD* (Mean Absolute Deviation). If we square the deviations instead of finding the absolute value, then the mean of these squared deviations is the *variance*. The *standard deviation* is the square root of the variance.

Table 3.2

4th-Grade Times (sec.)	Deviations	Squared Deviations
68	9	81
80	21	441
49	-10	100
55	-4	16
69	10	100
62	3	9
98	39	1521
42	-17	289
64	5	25
33	-26	676
58	-1	1
Etc.	Etc.	Etc.

Note: For more information about MAD, see the Further Explorations and Extensions at the end of this investigation. Also, see the ASA's *Bridging the Gap* Investigation 3.4.

The population *standard deviation* is represented by the Greek letter sigma σ and it is a measure of variability or spread of data around the mean μ of a population.

Ask your students to complete questions 12 to 15.

To find the population standard deviation—a measure that summarizes the spread of all the deviations or the typical deviation from the mean, complete the following steps using technology and the list or column of the fourth grade completion times.

12. Use technology and square each deviation found in Question 10. Store the squares in another column or list.

Sample answer: *Table 3.2*

13. Use technology and find the sum of the squared deviations.

Answer: *sum = 14428 sec.*^{2}

14. Use technology and find the mean of the squared deviations.

Answer: *369.95 sec.*^{2}

15. Take the square root of the mean of the squared deviations.

Answer: *19.23 seconds*

Explain that this value is called the *population standard deviation*. It is a measure used to describe the amount of variation or spread of a set of data values around the *population mean*. The fourth grade completion times had a mean of 59 seconds with a standard deviation of 19.2 seconds. This can be interpreted as the typical distance the data points are from the mean is approximately 19.2 seconds.

Point out to students that graphing calculators, spreadsheets, computer apps, and statistical software have a built-in standard deviation function. If the built-in function reports two standard deviation values, one is the symbol σ (sigma) for the population (division of the sum of the squared deviations by n) and the other is the letter s for the sample (division of the sum of the squared deviations by n-1). If we are using data collected from a sample, then the value of s should be used in describing a distribution. The symbol s represents the sample standard deviation.

Note: Careful study has shown using n-1 when calculating the standard deviation for a sample of data gives the best estimate for the standard deviation of the population.

See Further Explanations and Extensions at the end of this investigation for the standard deviation formulas.

Ask your students to complete questions 16 to 19.

16. Enter the class times into a calculator, spreadsheet, or statistical software. Use the built-in standard deviation function and find the mean and standard deviation of your class completion times. Since the class is taken to be a population in this investigation, report the population standard deviation.

Answer based on given ninth grade times: *mean = 47.9 sec. and standard deviation = 9.86 sec.*

17. Interpret the mean and standard deviation of your class memory test completion times.

Answer based on given ninth grade times: *The mean of 47.9 is the balance point of the distribution. If all the ninth graders had the same completion time, it would be 47.9. The standard deviation of approximately 9.9 seconds is the typical distance the data points are from the mean of 47.9 seconds.*

18. Compare the mean of your class completion times to the mean of the fourth grade completion times.

Answer: *The mean of the fourth grade times is 59 sec., and the mean of the ninth grade completion times is 47.9 sec. The ninth grade completion times are faster on average than the fourth grade completion times.*

19. Interpret and compare the standard deviation of your class times to the standard deviation of the fourth grade times. What does the value of the smaller standard deviation indicate?

Answers will vary, but based on given ninth grade sample times: *The standard deviation for the ninth grade times is much smaller than the standard deviation for the fourth grade times. The distribution of ninth grade times is not as spread out around its mean as the fourth grade times are around its mean.*

Interpret the Results in the Context of the Original Question

Ask your students to complete questions 20 and 21.

20. Using the results of your study of the fourth grade completion times and your class completion times, write a summary of your answer to the statistical question: "How do the memory test completion times of our class compare to the memory test completion times for fourth graders from around the United States?"

Answer based on given ninth grade times: *The mean of the ninth grade group is much lower than the fourth grade group—47.9 seconds compared to 59 seconds. The standard deviation of the ninth grade times is 9.9 seconds, while the fourth grade standard deviation is 19.2 seconds. This means the spread of the data in the ninth grade distribution is much less than the fourth grade, which means the times in the ninth grade are not as spread out as the fourth grade times. Based on this summary, the ninth grade class is generally able to complete the memory test in less time than the fourth grade group and, as a class, more consistently.*

21. The fourth grade times had an outlier at 122 seconds. Delete this point from the list/column of the fourth grade times and recalculate the mean and standard deviation. What effect did the outlier have on the mean and standard deviation?

Answer: *Both the mean and standard deviation decrease; the mean is now 57.3 sec. and the standard deviation is 16.3 sec.*

Summary

Ask your students what standard deviation measures.

Possible answer: *Standard deviation measures the spread of a data distribution. It measures the typical distance between each data point and the mean.*

Handout Student Worksheet 3.5 Matching Standard Deviation to Dot Plots.

Explain that the worksheet is designed for students to match each dot plot with the appropriate distribution number. Encourage them to estimate, rather than use the formula.

Additional Ideas

Have the class answer at least one of the questions 26, 27, 31, or 32 on the questionnaire from the Census at School website. (See the Teacher Resources section of this publication for more information about the Census at School website). Use the Random Sampler and select a random sample of at least 40 fourth graders (or another grade of your choice) who answered at least one of the questions 26, 27, 31, or 32. Compare the class results to the sample of 4th graders' results.

Note: Students should use *s*, the sample standard deviation in their analysis.

Census at School Questions:

26. How many hours of sleep per night do you usually get when you have school the next day?

27. How many hours of sleep per night do you usually get when you don't have school the next day?

31. About how many text messages did you send yesterday?

32. About how many text messages did you receive yesterday?

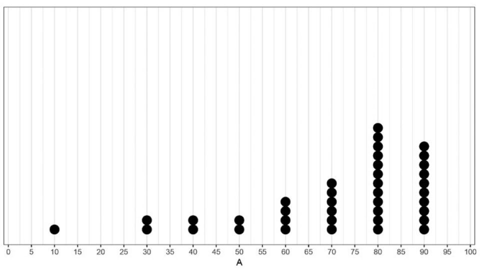
Student Worksheet 3.5 Matching Standard Deviation to Dot Plots

Consider the following group of dot plots and summary statistics. Each of the summary statistics for each data set 1 to 5 corresponds to one of the dot plots, lettered A to E.

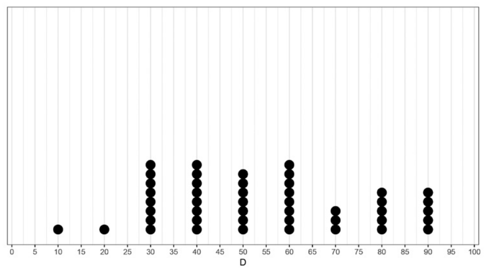
In the "Dot Plot" row of the table, write the letter of the matching dot plot next to the appropriate summary statistics and explain how you made your choice.

Summary	1	2	3	4	5
Mean	77.6	71	81.4	53.7	54.6
Median	80	80	80	50	50
Standard Deviation	9.3	19.8	8.5	21.2	9.7
Dot Plot					

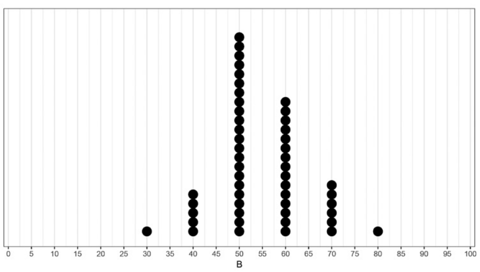
A



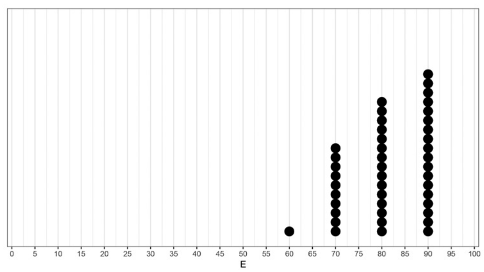
D



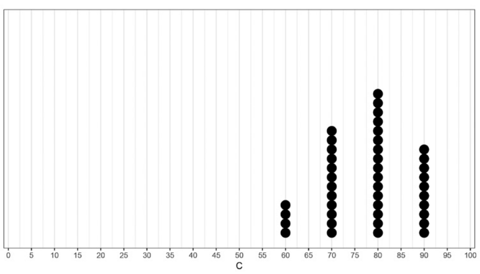
B



E



C



Answer: Table 3.3

Table 3.3

Summary	1	2	3	4	5
Mean	77.6	71	81.4	53.7	54.6
Median	80	80	80	50	50
Standard Deviation	9.3	9.7	8.5	21.2	9.7
Dot Plot	C	A	E	D	B

Exit Ticket

An algebra teacher wanted to determine whether ninth grade algebra students scored better when they took a math test in silence or when Mozart was being played. She randomly divided the students into two groups. One group took an algebra test in silence and the other group took the same test while a Mozart symphony was quietly playing in the room. The mean and standard deviation for the Mozart group were 55% and 17.4%, respectively, and the mean and standard deviation for the Silence group were 50.5% and 16.5%, respectively.
Data from Core Math Tools: www.nctm.org/coremathtools

The distribution of test scores from both groups is shown below.

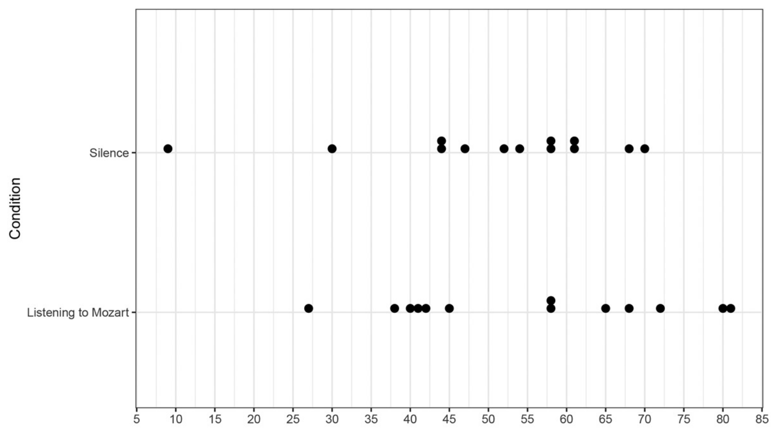


Figure 3.10: Distribution of test scores

1. Interpret the mean and standard deviation of the group that listened to Mozart.

Answer: *The mean of 55% is the balance point of the distribution. The sum of the deviations from 55% will be zero. The standard deviation of 17.4% is the typical distance a score on the math test is from the mean of 55%.*

2. Using the distribution and estimates for mean and standard deviation, did the ninth grade students in the Mozart group perform better on the math test than the group of ninth graders in the silence group?

Answer: *The scores for the ninth graders who listened to Mozart are slightly better than the scores for the group that took the test in silence. The mean of the Mozart group is around 55%, while the mean of the silence group is approximately 50%. The variation of both distributions is about the same. The Mozart group had the three highest scores, and the silence group had the lowest score.*

Further Explorations and Extensions

1. Present the formula for the population standard deviation: $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$
where

x is a value in the population

x-μ is a deviation of the value x, from the population mean, μ

$(x-\mu)^2$ is a squared deviation from the mean

$\sum(x-\mu)^2$ is the sum of the squared deviations

N is the population size

Compare each part of the formula with the steps used to calculate the standard deviation.

» Find the mean of all the data.

» Find the difference between each data point and the mean.

» Square each of the differences.

» Find the mean of the squared differences.

» Take the square root of the mean of the squared differences.

Point out when using the standard deviation application, the calculator or computer software often calculates two slightly different values for the standard deviation, sigma σ or the letter *s*. σ is the symbol for the population standard deviation and is found using the formula above. The *s* is the sample standard deviation and is found using the formula:

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

2. When asked what function to use to eliminate the negative signs for the negative deviations, students will often suggest using the absolute value function. A measure of variability that uses the absolute value function is called the mean absolute deviation or MAD. The MAD is interpreted in much the same way as standard deviation.

» Find the mean and the deviations from the mean.

» Take the absolute value of the deviations.

» Find the mean of the absolute deviations.

The formula for MAD is: $\frac{\sum|x-\bar{x}|}{n}$

Investigation 4

Do You Have Too Much Homework?

Exploratory Lesson

Overview

This investigation is an open-ended lesson designed to provide an opportunity for students to apply the four components of statistical problem solving. Students use the Random Sampler on the Census at School website and collect data on the number of hours fourth, eighth, and 12th grade students spend per week doing homework. (See Teacher Resource Section at the end of the book for more information about Census at School). After the data are collected, your students will have the opportunity to summarize the data using the techniques studied in investigations 1 to 3. The summaries could include written and oral presentations and/or construction of a poster to display the data and answer the statistical question.

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

Note: Census at School New Zealand hosts the Random Sampler (*http://new.censusatschool.org.nz/tools/random-sampler*) for international, New Zealand, and cleaned up US data. The American Statistical Association (ASA) hosts the "messy" US data (*ww2.amstat.org/CensusAtSchool*).

Use the New Zealand web address if you wish to have your students work with data that have been cleaned up (i.e., this database does not have missing data or data that have been entered incorrectly or collected using incorrect units).

Note: If you would like your students to work with "messy" data to provide them the opportunity to investigate "messy" data, then use the Random Sampler on the ASA website. "Messy" data means there are missing data, data collected incorrectly, or data collected using incorrect units.

This lesson is written using the "cleaned up" data from the New Zealand site.

Directions for using messy data on the ASA site can be found on Student Worksheet 4.1.

Brief Overview

» Develop a statistical question pertaining to the number of minutes fourth, eighth, and 12th grade students spend doing homework.

» Use the Census at School Random Sampler to collect data on the number of minutes fourth, eighth, and 12th grade students spent per week doing homework.

» Analyze the data and write a report, create a poster, and/or give an oral report that summarizes the data and answers the statistical question.

Hand out Student Worksheet 4.1 if students are downloading messy data.

Hand out Student Worksheet 4.2 if students are downloading cleaned up data.

Place students into groups of three. Have your students read the following scenario.

Scenario

The National Education Association (NEA) reported that survey data and anecdotal evidence show some students spend many hours nightly doing homework (*www.nea.org/tools/16938.htm*). According to research from the Brookings Institution and Rand Corporation, this homework overload is not the norm. Their researchers analyzed data from a variety of sources and concluded the majority of US

Learning Goal

Summarize numerical data sets by describing the center (mean and/or median), variability (mean absolute deviation and/or standard deviation), and shape (skewed or mound shaped).

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. Students must persevere through the process, adapting and adjusting each component as needed to arrive at a solution that adequately connects the interpretation of results to the statistical question posed.

Materials

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

» Student Worksheet 4.1 Directions for Using ASA Census at School Website (messy data)

» Student Worksheet 4.2 Directions for Using New Zealand Census at School Website (cleaned up data)

» Access to a computer with internet capability and a spreadsheet application such as Excel

» Statistical software or application capable of finding summary statistics and constructing graphs such as dot plots, box plots, and histograms

Estimated Time

One 50-minute class period to collect data and one class period to write and share a report.

Pre-Knowledge

Students should be able to construct a box plot, dot plot, and histogram using technology.

students spend less than an hour a day on homework, regardless of grade level, and this has held true for most of the past 50 years. In the last 20 years, the amount of homework has increased only in the lower grade levels.

Do you spend more time doing homework now than you did when you were in elementary or middle school? This investigation will look at the amount of time fourth, eighth, and 12th grade students spend on homework each week.

Formulate a Statistical Question

Discuss with your students that they will be investigating how many hours per week students in fourth, eighth, and 12th grade spend doing homework each week. Explain that they will take a random sample of students using the Census at School Random Sampler. These data will be used to investigate the statistical question: "How do the number of hours per week fourth graders, eighth graders, and 12th graders spend on homework compare with each other?"

Collect Appropriate Data

Have your students follow the steps outlined on Worksheet 4.1 or 4.2 to demonstrate how to use the Random Sampler on the Census at School site or the New Zealand site.

Note: The population of this study are the students who were involved in responding to this question (How many hours do you spend per week doing homework?) from the United States. The Random Sampler chooses a sample of US students who responded to the question.

Analyze the Data

After your students have taken a random sample from the three grades and used the software to create graphs and summary statistics for the number of hours of homework for each grade level data, ask them to begin their analysis of each grade level's reported number of hours doing homework. Suggest to them that their analysis should include graphs (dot plots, box plots, and/ or histograms) and calculations describing each grade level's distribution (shape, center, and spread). Using the graphs and calculations, students should then compare the hours of homework for each grade level and prepare a report.

Interpret the Results in the Context of the Original Question

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

» The statistical question that was investigated

» A description of the population sampled

» A summary of the sampling procedure

» Plots of the collected data

» Analysis and descriptions of the data, using calculations and the plots noting any unusual results

» A statement of conclusion about the statistical question

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

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

A data visualization poster is a display containing two or more related graphics that summarize a set of data, look at the data from different points of view, and answer specific statistical questions about the data.

The poster and presentation should include the following:

- The statistical question that was investigated as the title of the poster
- A description of the population sampled (in the oral report)
- A summary of the sampling procedure (in the oral report)
- The organized collected data—tables and plots (at least two graphs)
- Analysis and descriptions of the data, using calculations and the plots noting any unusual results (in the oral report)
- A statement of conclusion about the statistical question
- Recommendations for any follow-up studies or questions that may be investigated (in the oral report)

Note: A rubric for evaluating the posters can be found at *www.amstat.org*. The rubric is found under the Education tab and then K–12 Educators, Student Competitions. You can also download it from *www.amstat.org/asa/files/pdfs/EDU-PosterJudgingRubric.pdf*.