



Table of Contents¹

Statistics

Module Overview	3
Topic A: Understanding Distributions (6.SP.A.1, 6.SP.A.2, 6.SP.B.4, 6.SP.B.5b)	9
Lesson 1: Posing Statistical Questions	11
Lesson 2: Displaying a Data Distribution	19
Lesson 3: Creating a Dot Plot.....	27
Lesson 4: Creating a Histogram	35
Lesson 5: Describing a Distribution Displayed in a Histogram	49
Topic B: Summarizing a Distribution that is Approximately Symmetric Using the Mean and Mean Absolute Deviation (6.SP.A.2, 6.SP.A.3, 6.SP.B.4, 6.SP.B.5)	58
Lesson 6: Describing the Center of a Distribution Using the Mean	60
Lesson 7: The Mean as a Balance Point.....	70
Lesson 8: Variability in a Data Distribution.....	82
Lesson 9: The Mean Absolute Deviation (MAD).....	91
Lessons 10–11: Describing Distributions Using the Mean and MAD	102
Mid-Module Assessment and Rubric	119
<i>Topics A through B (assessment 1 day)</i>	
Topic C: Summarizing a Distribution that is Skewed Using the Median and the Interquartile Range (6.SP.A.2, 6.SP.A.3, 6.SP.B.4, 6.SP.B.5)	135
Lesson 12: Describing the Center of a Distribution Using the Median	137
Lesson 13: Describing Variability Using the Interquartile Range (IQR)	148
Lesson 14: Summarizing a Distribution Using a Box Plot	156
Lesson 15: More Practice with Box Plots.....	166
Lesson 16: Understanding Box Plots	177
Topic D: Summarizing and Describing Distributions (6.SP.B.4, 6.SP.B.5)	187
Lesson 17: Developing a Statistical Project	189
Lesson 18: Connecting Graphical Representations and Numerical Summaries.....	195

¹ Each lesson is ONE day and ONE day is considered a 45 minute period.

Lesson 19: Comparing Data Distributions 208

Lesson 20: Describing Center, Variability, and Shape of a Data Distribution from a Graphical Representation 219

Lesson 21: Summarizing a Data Distribution by Describing Center, Variability, and Shape 228

Lesson 22: Presenting a Summary of a Statistical Project 237

End-of-Module Assessment and Rubric 241

Topics A through D (assessment 1 day, remediation or further applications 1 day)

Grade 6 • Module 6

Statistics

OVERVIEW

In Grade 5, students used bar graphs and line plots to represent data and then solved problems using the information presented in the plots (**5.MD.B.2**). In this module, students move from simply representing data into analysis of data. In Topic A, students begin to think and reason statistically, first by recognizing a statistical question as one that can be answered by collecting data (**6.SP.A.1**). Students learn that the data collected to answer a statistical question has a distribution that is often summarized in terms of center, variability, and shape (**6.SP.A.2**). Beginning in Topic A, and throughout the module, students see and represent data distributions using dot plots and histograms (**6.SP.B.4**).

In Topics B and C, students study quantitative ways to summarize numerical data sets in relation to their context and to the shape of the distribution. The mean and mean absolute deviation (MAD) are used for data distributions that are approximately symmetric, and the median and interquartile range (IQR) are used for distributions that are skewed. Students apply their experience in writing, reading, and evaluating expressions in which letters stand for numbers (**6.EE.A.2**) as they learn to compute and interpret two pairs of statistical measures for center and spread (**6.SP.A.5**).

In Topic B, students study *mean* as a measure of center and *mean absolute deviation* as a measure of variability. Students learn that these measures are preferred when the shape of the distribution is roughly symmetric. Then, in Topic C, students study *median* as a measure of center and *interquartile range* as a measure of variability. Students learn that these measures are preferred when the shape of the distribution is skewed. Students develop in Topic B, and reinforce in Topic C, the idea that a measure of center provides a summary of all its values in a single number, while a measure of variation describes how values vary, also with a single number (**6.SP.A.3**).

Measures of center and variability for distributions that are approximately symmetric (mean and MAD) are covered before measures (median and IQR) for skewed data distributions. This choice was made because it is easier for students to understand measuring center and variability in the context of symmetric distributions.

For students, box plots are the most difficult of the graphical displays covered in this module. This is because they differ from dot plots and histograms in that they are not really a display of the data but rather a graph of five summary measures (minimum, lower quartile, median, upper quartile, and maximum). This graph conveys information on center and variability but is more difficult for students to interpret because, unlike histograms, where large area corresponds to many observations, in a box plot, large area indicates spread and small area indicates a large number of observations in a small interval. Box plots also require the calculation of quartiles and are best covered after quartiles have been introduced and used to calculate the IQR. For these reasons, box plots are introduced late in the module after the IQR and after students have already developed some fundamental understanding of data distributions, which is easier to do in the context of dot plots and histograms.

In Topic D, students synthesize what they have learned as they connect the graphical, verbal, and numerical summaries to each other within situational contexts, culminating with a major project (**6.SP.B.4**, **6.SP.B.5**). Students implement the four-step investigative process with their projects by stating their statistical questions, explaining the plan they used to collect data, analyzing data numerically and with graphs, and interpreting their results as related to their questions. The Mid-Module Assessment follows Topic B. The End-of-Module Assessment follows Topic D.

Focus Standards

Develop understanding of statistical variability.

- 6.SP.A.1** Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers. *For example, “How old am I?” is not a statistical question, but “How old are the students in my school?” is a statistical question because one anticipates variability in students’ ages.*
- 6.SP.A.2** Understand that a set of data collected to answer a statistical question has a distribution which can be described by its center, spread, and overall shape.
- 6.SP.A.3** Recognize that a measure of center for a numerical data set summarizes all of its values with a single number, while a measure of variation describes how its values vary with a single number.

Summarize and describe distributions.

- 6.SP.B.4** Display numerical data in plots on a number line, including dot plots, histograms, and box plots.
- 6.SP.B.5** Summarize numerical data sets in relation to their context, such as by:
 - a. Reporting the number of observations.
 - b. Describing the nature of the attribute under investigation, including how it was measured and its units of measurement.
 - c. Giving quantitative measures of center (median and/or mean) and variability (interquartile range and/or mean absolute deviation), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered.
 - d. Relating the choice of measures of center and variability to the shape of the data distribution and the context in which the data were gathered.

Foundational Standards

Perform operations with multi-digit whole numbers and with decimals to hundredths.

- 5.NBT.B.5** Fluently multiply multi-digit whole numbers using the standard algorithm.
- 5.NBT.B.6** Find whole-number quotients of whole numbers with up to four-digit dividends and two-digit divisors using strategies based on place value, the properties of operations and/or the relationship between multiplication and division. Illustrate and explain the calculation by using equations, rectangular arrays, and/or area models.
- 5.NBT.B.7** Add, subtract, multiply, and divide decimals to hundredths, using concrete models or drawings and strategies based on place value, properties of operations, and/or the relationship between addition and subtraction; relate the strategy to a written method and explain the reasoning used.

Represent and interpret data.

- 5.MD.B.2** Make a line plot to display a data set of measurements in fractions of a unit ($\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$). Use operations on fractions for this grade to solve problems involving information presented in line plots. *For example, given different measurements of liquid in identical beakers, find the amount of liquid each beaker would contain if the total amount in all the beakers were redistributed equally.*

Apply and extend previous understandings of arithmetic to algebraic expressions.

- 6.EE.A.2** Write, read, and evaluate expressions in which letters stand for numbers.
- Write expressions that record operations with numbers and with letters standing for numbers. *For example, express the calculation “Subtract y from 5” as $5 - y$.*
 - Identify parts of an expression using mathematical terms (sum, term, product, factor, quotient, coefficient); view one or more parts of an expression as a single entity. *For example, describe the expression $2(8 + 7)$ as a product of two factors; view $(8 + 7)$ as both a single entity and a sum of two terms.*
 - Evaluate expressions at specific values of their variables. Include expressions that arise from formulas used in real-world problems. Perform arithmetic operations, including those involving whole-number exponents, in the conventional order when there are no parentheses to specify a particular order (Order of Operations). *For example, use the formulas $V = s^3$ and $A = 6s^2$ to find the volume and surface area of a cube with sides of length $s = \frac{1}{2}$.*

Focus Standards for Mathematical Practice

- MP.1** **Make sense of problems and persevere in solving them.** Students make sense of problems by defining them in terms of a statistical question and then determining what data might be collected in order to provide an answer to the question and therefore a solution to the problem.

- MP.2 Reason abstractly and quantitatively.** Students pose statistical questions and reason about how to collect and interpret data in order to answer these questions. Students use graphs to summarize the data and to answer statistical questions.
- MP.3 Construct viable arguments and critique the reasoning of others.** Students examine the shape, center, and variability of a data distribution. They communicate the answer to a statistical question in the form of a poster presentation. Students also have an opportunity to critique poster presentations made by other students.
- MP.4 Model with mathematics.** Students create graphs of data distributions. They select an appropriate measure of center to describe a typical data value for a given data distribution. They also calculate and interpret an appropriate measure of variability based on the shape of the data distribution.
- MP.6 Attend to precision.** Students interpret and communicate conclusions in context based on graphical and numerical data summaries. Students use statistical terminology appropriately.

Terminology

New or Recently Introduced Terms

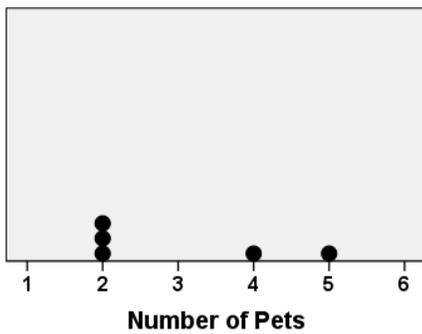
- **Statistical Question** (A question that anticipates variability in the data that would be collected in order to answer the question.)
- **Median** (A measure of center appropriate for skewed data distributions. It is the middle value when the data are ordered from smallest to largest if there are an odd number of observations and half way between the middle two observations if the number of observations is even.)
- **Mean** (A measure of center appropriate for data distributions that are approximately symmetric. It is the average of the values in the data set. Two common interpretations of the mean are as a “fair share” and as the balance point of the data distribution.)
- **Dot Plot** (A plot of numerical data along a number line.)
- **Histogram** (A graphical representation of a numerical data set that has been grouped into intervals. Each interval is represented by a bar drawn above that interval that has a height corresponding to the number of observations in that interval.)
- **Box Plot** (A graph of five numerical summary measures: the minimum, lower quartile, median, upper quartile, and the maximum. It conveys information about center and variability in a data set.)
- **Variability** (Variability in a data set occurs when the observations in the data set are not all the same.)
- **Deviations from the Mean** (The differences calculated by subtracting the mean from the observations in a data set.)
- **Mean Absolute Deviation (MAD)** (A measure of variability appropriate for data distributions that are approximately symmetric. It is the average of the absolute value of the deviations from the mean.)
- **Interquartile Range (IQR)** (A measure of variability appropriate for data distributions that are skewed. It is the difference between the upper quartile and the lower quartile of a data set and describes how spread out the middle 50% of the data are.)

Familiar Terms and Symbols²

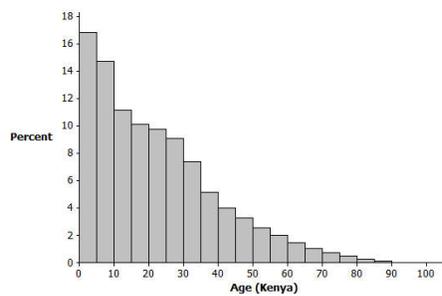
- Line Plot or Dot Plot

Suggested Tools and Representations

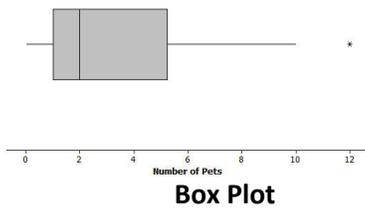
- Dot Plots
- Histograms
- Box Plots



Dot Plot



Histogram



Box Plot

² These are terms and symbols students have seen previously.

Assessment Summary

Assessment Type	Administered	Format	Standards Addressed
Mid-Module Assessment Task	After Topic B	Constructed response with rubric	6.SP.A.1, 6.SP.A.2, 6.SP.A.3, 6.SP.B.4, 6.SP.B.5
End-of-Module Assessment Task	After Topic D	Constructed response with rubric	6.SP.A.1, 6.SP.A.2, 6.SP.A.3, 6.SP.B.4, 6.SP.B.5
Project	Topic D: Lessons 17 and 22	Poster or other form of presentation	6.SP.A.1, 6.SP.A.2, 6.SP.A.3, 6.SP.B.4, 6.SP.B.5



Topic A:

Understanding Distributions

6.SP.A.1, 6.S.AP.2, 6.SP.B.4, 6.SP.B.5b

Focus Standard:	6.SP.A.1	Recognize a statistical question as one that anticipates variability in the data related to the question and accounts for it in the answers. For example, “How old am I?” is not a statistical question, but “How old are the students in my school?” is a statistical question because one anticipates variability in students’ ages.
	6.SP.A.2	Understand that a set of data collected to answer a statistical question has a distribution, which can be described by its center, spread, and overall shape.
	6.SP.B.4	Display numerical data in plots on a number line, including dot plots, histograms, and box plots.
	6.SP.B.5b	Summarize numerical data sets in relation to their context, such as by: <ol style="list-style-type: none"> b. Describing the nature of the attribute under investigation, including how it was measured and its units of measurement.
Instructional Days:	5	
	Lesson 1:	Posing Statistical Questions (P) ¹
	Lesson 2:	Displaying a Data Distribution (P)
	Lesson 3:	Creating a Dot Plot (P)
	Lesson 4:	Creating a Histogram (P)
	Lesson 5:	Describing a Distribution Displayed in a Histogram (P)

In Topic A, students begin a study of statistics by learning to recognize a statistical question. They develop an understanding of what data could be collected to answer a statistical question and anticipate variability in the data collected to answer the question. In Lesson 1, statistical questions are introduced in the context of a four-step process for posing and answering questions based on data. As students begin to explore data, they see the need to organize and summarize data. In Lesson 2, students are introduced to the idea that a data distribution can be represented graphically and that there are several different types of graphs, including dot plots and histograms, commonly used to represent a distribution of numerical data. This lesson then builds

¹ Lesson Structure Key: P-Problem Set Lesson, M-Modeling Cycle Lesson, E-Exploration Lesson, S-Socratic Lesson

on students' previous work with line plots, introducing them to dot plots (line plots, but in a data context where students are to think about the distribution of data rather than to think of individual points plotted on a number line). In Lesson 3, students construct dot plots and begin to describe data distributions. In Lesson 4, students are introduced to histograms as another way of representing a data distribution graphically and the advantages and disadvantages of histograms relative to dot plots are discussed. Additionally, students begin to consider the shape of a data distribution (symmetric versus skewed) in this lesson and are introduced to the idea that different numerical summary measures of center and variability are used to describe data distributions that are approximately symmetric than the measures used to describe data distributions that are skewed. This is an important distinction and is the basis for the content introduced in Topics B and C. Lesson 5 gives students additional practice in constructing and describing histograms and introduces relative frequency histograms (histograms where relative frequency rather than frequency is used for the vertical scale).



Lesson 1: Posing Statistical Questions

Student Outcomes

- Students distinguish between statistical questions and those that are not statistical.
- Students formulate a statistical question and explain what data could be collected to answer the question.
- Students distinguish between categorical data and numerical data.

Classwork

Statistics is about using data to answer questions. In this module, the following four steps will summarize your work with data:

Step 1: Pose a question that can be answered by data.

Step 2: Determine a plan to collect the data.

Step 3: Summarize the data with graphs and numerical summaries.

Step 4: Answer the question posed in Step 1 using the data and summaries.

You will be guided through this process as you study these lessons. This first lesson is about the first step – what is a statistical question, and what does it mean that a question can be answered by data?

Example 1 (15 minutes): What is a Statistical Question?

Example 1: What is a Statistical Question?

Jerome, a 6th grader at Roosevelt Middle School, is a huge baseball fan. He loves to collect baseball cards. He has cards of current players and of players from past baseball seasons. With his teacher's permission, Jerome brought his baseball card collection to school. Each card has a picture of a current or past major league baseball player, along with information about the player. When he placed his cards out for the other students to see, they asked Jerome all sorts of questions about his cards. Some asked:

- How many cards does Jerome have altogether?
- What is the typical cost of a card in Jerome's collection?
- Where did Jerome get the cards?

Introduce the situation described in Example 1 to the class. You may want to show the students an example of a baseball card so that they can see the varied amount of information on the card.

Then, consider the questions that follow the description and ask the students:

- Which of these questions do you think might be statistical questions?
- What do you think I mean when I say “a statistical question”?

Students do not have a definition or understanding of what a statistical question is at this point. Allow them to discuss and make conjectures about what that might mean before guiding them to the following:

A statistical question is one that can be answered with data and for which it is anticipated that the data (information) collected to answer the question will vary.

The second and third questions are statistical questions because the answer for each card in the collection could vary. The 1st question, “How many cards do you have in your collection?” is not a statistical question because we do not anticipate any variability in the data collected to answer this question. There is only one data value and no variability.

Convey the main idea that a question is statistical if it can be answered with data that varies. Point out to the students the concept of variability in the data means that not all data values have the same value.

The question, “How old am I?” is not a statistical question because it is not answered by collecting data that vary. The question, “How old are the students in my school?” is a statistical question because when you collect data on the ages of students at the school, the ages will vary – not all students are the same age.

Ask students if the following questions would be answered by collecting data that vary:

- How tall is your 6th grade math teacher?
- What is your hand span (measured from tip of the thumb to the tip of the small finger)?

Ask students which of these data sets would have the most variability.

- Number of minutes students in your class spend getting ready for school.
- Number of pockets on the clothes of students in your class.

After arriving at this understanding as a class, post the informal definition of *statistical question* on the board for students to refer to for the remainder of the class.

Exercises 1–5 (10 minutes)

These question sets are designed to reinforce the definition of a statistical question. The main focus is on whether there is variability in the data that would be used to answer the question. You may want to have students share their answers to Exercise 3 with a partner and have the partner decide whether or not the question is a statistical question.

Exercises 1–5

1. For each of the following, determine whether or not the question is a statistical question. Give a reason for your answer.
 - a. Who is my favorite movie star?
No, not answered by collecting data that vary.
 - b. What are the favorite colors of 6th graders in my school?
Yes, colors will vary.
 - c. How many years have students in my school's band or orchestra played an instrument?
Yes, number of years will vary.
 - d. What is the favorite subject of 6th graders at my school?
Yes, subjects will vary.
 - e. How many brothers and sisters does my best friend have?
No, not answered by collecting data that vary.
2. Explain why each of the following questions is not a statistical question.
 - a. How old am I?
Not answered by data that vary.
 - b. What's my favorite color?
Not answered by data that vary – I just have one favorite color.
 - c. How old is the principal at our school?
The principal has just one age at the time I ask the principal's age. Answered by data that does not vary.
3. Ronnie, a 6th grader, wanted to find out if he lived the farthest from school. Write a statistical question that would help Ronnie find the answer.
What is a typical distance from home to school (in miles) for students at my school?
4. Write a statistical question that can be answered by collecting data from students in your class.
What is the typical number of pets owned by students in my class?
How many hours each day does a typical student in my class play video games?
5. Change the following question to make it a statistical question: "How old is my math teacher?"
What is the typical age of teachers in my school?

Example 2 (10 minutes): Types of Data

To answer statistical questions, we collect data. In the context of baseball cards, we might record the cost of a card for each of 25 baseball cards. This would result in a data set with 25 values. We might also record the age of a card or the team of the player featured on the card.

Example 2: Types of Data

We use two types of data to answer statistical questions: numerical data and categorical data. If we recorded the age of 25 baseball cards, we would have numerical data. Each value in a numerical data set is a number. If we recorded the team of the featured player for 25 baseball cards, you would have categorical data. Although you still have 25 data values, the data values are not numbers. They would be team names, which you can think of as categories.

- What are other examples of categorical data? Eye color, the month in which you were born, and the number that may be used to identify your classroom are examples of categorical data.
- What are other examples of numerical data? Height, number of pets, and minutes to get to school are all examples of numerical data.

To help students distinguish between the two data types, encourage them to think of possible data values. If the possible data values include words or categories, then the variable is categorical.

Suppose that you collected data on the following. What are some of the possible values that you might get?

- Eye color
- Favorite TV show
- Amount of rain that fell during storms
- High temperatures for each of 12 days

Exercises 6–7 (5 minutes)

Have students complete the exercise to reinforce students' understanding of the two types of data.

Exercises 6–7

6. Identify each of the following data sets as categorical (C) or numerical (N).
- Heights of 20 6th graders N
 - Favorite flavor of ice cream for each of 10 6th graders C
 - Hours of sleep on a school night for 30 6th graders N
 - Type of beverage drank at lunch for each of 15 6th graders C
 - Eye color for each of 30 6th graders C
 - Number of pencils in each desk of 15 6th graders N

7. For each of the following statistical questions, students asked Jerome to identify whether the data are numerical or categorical. Explain your answer, and list four possible data values.

a. How old are the cards in the collection?

Numerical data, as I anticipate data will be a number.

Possible data values: 2 years, $2\frac{1}{2}$ years, 4 years, 20 years

b. How much did the cards in the collection cost?

Numerical data, as I anticipate data will be a number.

\$ 0.20, \$ 1.50, \$ 10.00, \$ 35.00

c. Where did you get the cards?

Categorical, as I anticipate the data represents the name of a place.

(e.g., a store, a garage sale, from my brother, from a friend)

Lesson Summary

A **statistical question** is one that can be answered by collecting data that vary (i.e., not all of the data values are the same).

There are two types of data: numerical and categorical. In a **numerical data set**, every value in the set is a number. **Categorical data sets** can take on non-numerical values, such as names of colors, labels, etc. (e.g., “large,” “medium,” or “small”).

Statistics is about using data to answer questions. In this module, the following 4 steps will summarize your work with data:

Step 1: Pose a question that can be answered by data.

Step 2: Determine a plan to collect the data.

Step 3: Summarize the data with graphs and numerical summaries.

Step 4: Answer the question posed in Step 1 using the data and the summaries.

Exit Ticket (10 minutes)



Name _____

Date _____

Lesson 1: Posing Statistical Questions

Exit Ticket

1. Indicate whether each of the following two questions is a statistical question. Explain why or why not.
 - a. How much does Susan's dog weigh?

 - b. How much do the dogs belonging to students at our school weigh?

2. If you collected data on the weights of dogs, would the data be numerical or categorical? Explain how you know it is numerical or categorical.

Exit Ticket Sample Solutions

1. Indicate whether each of the following two questions is a statistical question. Explain why or why not.

a. How much does Susan's dog weigh?

This is not a statistical question. This question is not answered by collecting data that vary.

b. How much do the dogs belonging to students at our school weigh?

This is a statistical question. This question would be answered by collecting data on weights of dogs. There is variability in these weights.

2. If you collected data on the weights of dogs, would the data be numerical or categorical?

Numerical

Problem Set Sample Solutions

1. For each of the following, determine whether the question is a statistical question. Give a reason for your answer.

a. How many letters are in my last name?

No, this question is not answered by collecting data that vary.

b. How many letters are in the last names of the students in my 6th grade class?

Yes, there is variability in the lengths of the last names.

c. What are the colors of the shoes worn by the students in my school?

Yes, we expect variability in the colors.

d. What is the maximum number of feet that roller coasters drop during a ride?

Yes, we expect variability in the feet to drop for different roller coasters; they are not all the same.

e. What are the heart rates of the students in a 6th grade class?

Yes, we expect variability – not all 6th graders have exactly the same heart rate.

f. How many hours of sleep per night do 6th graders usually get when they have school the next day?

Yes, we do not expect that all 6th graders sleep the same number of hours.

g. How many miles per gallon do compact cars get?

Yes, we expect variability in the miles per gallon – not all compact cars get the same miles per gallon.

2. Identify each of the following data sets as categorical (C) or numerical (N). Explain your answer.
- Arm spans of 12 6th graders
N; the arm span can be measured as number of inches for example, so the data is numerical.
 - Number of languages spoken by each of 20 adults
N; number of languages is clearly numerical.
 - Favorite sport of each person in a group of 20 adults
C; a sport falls into a category, such as "soccer" or "hockey" and cannot be measured numerically.
 - Number of pets for each of 40 3rd graders
N; number of pets is clearly numerical.
 - Number of hours a week spent reading a book for a group of middle school students
N; number of hours is clearly numerical.
3. Rewrite each of the following questions as a statistical question.
Answers will vary
- How many pets does your teacher have?
How many pets do students in our school have?
 - How many points did the high school soccer team score in its last game?
How many points did the high school soccer team score in soccer games this season?
 - How many pages are in our math book?
How many pages are in the books in the school library?
 - Can I do a handstand?
Can most 6th graders do a handstand?
4. Write a statistical question that would be answered by collecting data from the 6th graders in your classroom.
Answers will vary. Check if the question would be answered by collecting data that vary.
5. Are the data you would collect to answer that question categorical or numerical? Explain your answer.
Answers will vary.



Lesson 2: Displaying a Data Distribution

Student Outcomes

- Given a dot plot, students begin describing the distribution of the points on the dot plot in terms of center and variability.

Classwork

Example 1 (10 minutes): Heart Rate

Example 1: Heart Rate

Mia, a 6th grader at Roosevelt Middle School, was thinking about joining the middle school track team. She read that Olympic athletes have lower resting heart rates than most people. She wondered about her own heart rate and how it would compare to other students. Mia was interested in investigating the statistical question: “What are the heart rates of the students in my 6th grade class?”

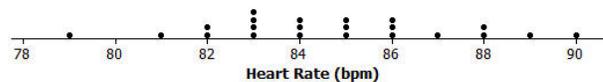
Heart rates are expressed as bpm (or beats per minute). Mia knew her resting heart rate was 80 beats per minute. She asked her teacher if she could collect the heart rates of the other students in her class. With the teacher’s help, the other 6th graders in her class found their heart rates and reported them to Mia. Following are the heart rates (in beats per minute) for the 22 other students in Mia’s class:

89 87 85 84 90 79 83 85 86 88 84 81 88 85 83 83 86 82 83 86 82 84

To learn about the heart rates, a good place to start is to make a graph of the data. There are several different graphs that could be used, including the three types of graphs that you will learn in this module: dot plots, histograms, and box plots. In this lesson, you will learn about dot plots.

Mia noticed that there were many different heart rates. She decided to make a dot plot to show the different heart rates. She drew a number line and started numbering from 78 to 92. She then placed a dot above the number on the number line for each heart rate. If there was already a dot above a number she added another dot above the one already there. She continued until she had added one dot for each heart rate.

Dot Plot of Heart Rate



This example uses the scenario of students’ resting heart rate. As you discuss the scenario with students, you may want to demonstrate how a pulse is taken, and (if time permits) have your students find their resting heart rate and use that data to make a dot plot. Note: This lesson is not intended to teach students how to construct a dot plot, but rather to show them a dot plot as a graph of the distribution of the data collected to answer a statistical question. Emphasize thinking about the center of the data and the spread of the data.

MP.1

At this point, students may differ on what “center” of the data means. Write down students’ suggestions of what “center” means so that you can refer back to their initial ideas throughout the module. Exercise 6 asks students to describe the center. Some students may choose a number that occurs most often while others may pick a number that is in the “middle” (i.e., halfway between the highest and lowest data values). Some students may say it’s the “average.” The intent is not to calculate any specific value, but to gauge your students’ thinking about center. Formal measures of center will be developed starting in Lesson 6.

Exercise 8 asks students to think about what a “typical” heart rate is for 6th graders. Similar to the idea of center, students will vary to what they think is typical. The idea is to have students begin to discuss where there are clusters of data and where the data centers.

Ask students the following questions as you develop this example:

- What can you tell me about the heart rates of the 6th grade students?
- Where do the heart rates tend to center? Why did you choose that number?
- How much spread do you see in the heart rates?

Exercises 1–10 (15 minutes)

These ten questions are designed to have students recognize the details that can be observed in a dot plot. They should be able to find the lowest heart rate, the highest heart rate, and the most common heart rate, and describe the approximate location of the center.

Allow the students about 5–8 minutes to work independently or in small groups. Then, bring the groups together to summarize their answers.

Exercises 1–10

1. What was the heart rate for the student with the lowest heart rate?

79

2. What was the heart rate for the student with the highest heart rate?

90

3. How many students had a heart rate greater than 86?

5

4. What fraction of the students had a heart rate less than 82?

$\frac{2}{22}$ or $\frac{1}{11}$

5. What is the most common heart rate?

83

6. What heart rate describes the center of the data?

85 (Answers may vary, but students’ responses should be around the center.)

7. What heart rates are the most unusual heart rates?

79 and 90

8. If Mia’s teacher asked what the typical heart rate is for 6th graders in the class, what would you tell Mia’s teacher?

Most students had a heart rate between 82 and 86. The most common was 83.

9. On the dot plot add a dot for Mia’s heart rate.

Add a dot above 80.

10. How does Mia’s heart rate compare with the heart rates of the other students in the class?

Her heart rate is lower than all but one of the students.

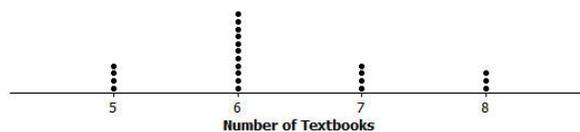
Example 2 (10 minutes): Seeing the Spread in Dot Plots

Example 2: Seeing the Spread in Dot Plots

Mia’s class collected data to answer several other questions about her class. After they collected data, they drew dot plots of their findings.

Here is a dot plot showing the data collected to answer the question: “How many textbooks are in the desks of 6th graders?”

Dot Plot of Number of Textbooks

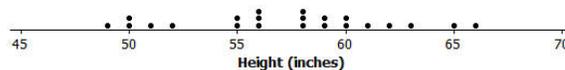


When the students thought about this question, many said that they all had about the same number of books in their desk since they all take the same subjects in school.

The class noticed that the graph was not very spread out since there were only four different answers that students gave, with most of the students answering that they had 6 books in their desk.

Another student wanted to ask the question: “How tall are the 6th graders in our class?” When students thought about this question, they thought that the heights would be spread out since there were some shorter students and some very tall students in class. Here is a dot plot of the students’ heights:

Dot Plot of Height



In this example, the focus is on the spread of the data. Display the two dot plots (one of the number of textbooks in their desks and the other for the heights of 6th graders). Discuss with students the values shown on the number line. For the number of textbooks, the data span from 5 to 8, while the heights go from 49 to 66.

Ask students the following question as you develop this example:

- Why do you think the data for the number of textbooks go from 4 to 8, while the heights span from 49” to 66”?

Exercises 11–14 (10 minutes)

This exercise is a matching problem where students are given statistical questions to which they should match a dot plot. Stress to students that they need to explain why they matched the questions and the dot plots as they did.

Allow the students about 3–5 minutes to work independently or in small groups. Bring the groups together to summarize their answers.

Exercises 11–14

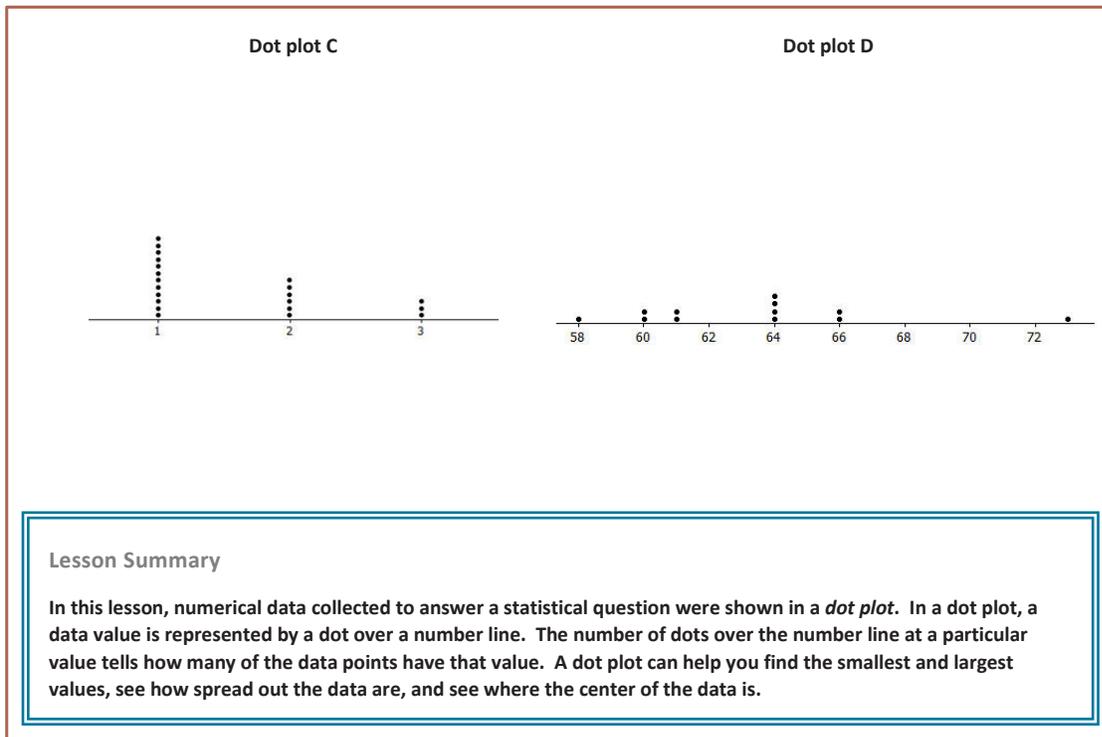
Listed are four statistical questions and four different dot plots of data collected to answer these questions. Match each statistical question with the appropriate dot plot. Explain each of your choices.

Statistical Question:

- What are the ages of 4th graders in our school?
A - Many 4th graders are around 9 or 10 years old.
- What are the heights of the players on the 8th grade boys’ basketball team?
D - The guys on an 8th grade basketball team can vary in height. Generally, there is a tall player (73 inches), while most others are around 5 feet, or 60 inches.
- How many hours do 6th graders in our class watch TV on a school night?
B - Answers vary. I think a few of the students may watch a lot of TV. Most students watch two hours or less.
- How many different languages do students in our class speak?
C - Most students know one language – English. Many of the students in our class also study another language, or live in an environment where their family speaks another language.

Dot plot A

Dot plot B



Exit Ticket (5 minutes)

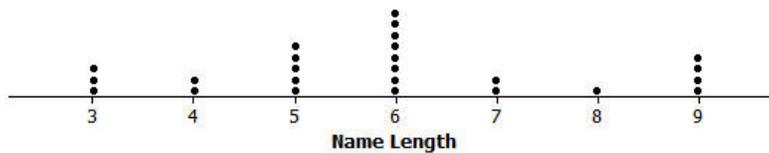
Name _____

Date _____

Lesson 2: Displaying a Data Distribution

Exit Ticket

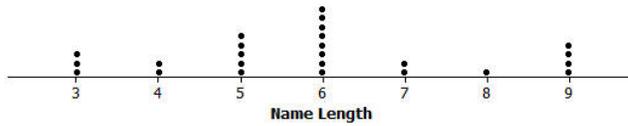
A 6th grade class collected data on the number of letters in the first names of all the students in class. Here is the dot plot of the data they collected:



1. How many students are in the class?
2. What is the shortest name length?
3. What is the longest name length?
4. What is the most common name length?
5. What name length describes the center of the data?

Exit Ticket Sample Solutions

A 6th grade class collected data on the number of letters in the first names of all the students in class. Here is the dot plot of the data they collected:

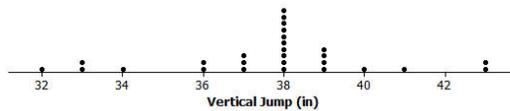


- How many students are in the class?
25
- What is the shortest name length?
3 letters
- What is the longest name length?
9 letters
- What is the most common name length?
6 letters
- What name length describes the center of the data?
6 letters

Problem Set Sample Solutions

- The dot plot below shows the vertical jump of some NBA players. A vertical jump is how high a player can jump from a standstill.

Dot Plot of Vertical Jump



- What statistical question do you think could be answered using these data?
What is the vertical jump of NBA players?
- What was the highest vertical jump by a player?
43 inches

c. What was the lowest vertical jump by a player?

32 inches

d. What is the most common vertical jump?

38 inches

e. How many players jumped that high?

10

f. How many players jumped higher than 40 inches?

3

g. Another NBA player jumped 33 inches. Add a dot for this player on the dot plot. How does this player compare with the other players?

This player jumped the same as two other players and jumped higher than only one player.

2. Listed are two statistical questions and two different dot plots of data collected to answer these questions. Match each statistical question with its dot plot. Explain each of your choices.

Statistical questions:

a. What is the number of fish (if any) that students in class have in an aquarium at their home?

A; some students may not have any fish (0 from the dot plot) while another student has 10 fish.

b. How many pockets do the 6th graders have in the pants that they are wearing at school on a particular day?

B; the dot plot displays the values 2, 3, 4, 5, 6, which are all reasonable within the context of the question. Pants generally have at least 2 pockets.

Dot Plot A



Dot Plot B



3. Read each of the following statistical questions. Write a description of what the dot plot of the data collected to answer the question might look like. Your description should include a description of the spread of the data and the center of the data.

a. What is the number of hours 6th grade students are in school during a typical school day?

Most students are in school for the same number of hours. Differences may exist for those students who travel or participate in a club or afterschool activity. Students' responses vary based on their estimate of the number of hours students spend in school.

b. What is the number of video games owned by the 6th graders in our class?

These data would have a very big spread. Some students might have no video games, while others could have a large number of games. A typical value of 5 (or something similar) would identify a center. In this case, the center is based on the number most commonly reported by students.



Lesson 3: Creating a Dot Plot

Student Outcomes

- Students create a dot plot of a given data set.
- Students summarize a given data set using equal length intervals and construct a frequency table.
- Based on a frequency table, students describe the distribution.

Classwork

Example 1 (5 minutes): Hours of Sleep

Example 1: Hours of Sleep

Robert, a 6th grader at Roosevelt Middle School, usually goes to bed around 10:00 p.m. and gets up around 6:00 a.m. to get ready for school. That means that he gets about 8 hours of sleep on a school night. He decided to investigate the statistical question: How many hours per night do 6th graders usually sleep when they have school the next day?

Robert took a survey of 29 6th graders and collected the following data to answer the question:

7 8 5 9 9 9 7 7 10 10 11 9 8 8 8 12 6 11 10 8 8 9 9 9 8 10 9 9 8

Robert decided to make a dot plot of the data to help him answer his statistical question. Robert first drew a number line and labeled it from 5 to 12 to match the lowest and highest number of hours slept.



He then placed a dot above 7 for the first piece of data he collected. He continued to place dots above the numbers until each number was represented by a dot.



MP.1

This example begins with the statistical question: How many hours per night do 6th graders usually sleep when they have school the next day? The data shown come from a random sample of 6th graders collected from the Census At School website (<http://www.amstat.org/censusatschool/>). The beginning steps to make a dot plot are presented and students are asked to complete the plot. It is important to point out to students that as they determine labels for the number line, they must list the numbers sequentially using the same interval. A common mistake is that students may not list a number if there is no data for that value. If there is a large gap between the values, student may also skip numbers. Emphasize that it is important to keep the vertical spacing the same to better understand the distribution that the dot plot summarizes. Lined paper can be useful to students who need help keeping vertical spacing consistent when constructing a dot plot.

As you develop this example, pose the following questions to students:

- Why is the number line labeled from 5 to 12? Could we have labeled the number line from 8 to 16? Could we have labeled the number line from 0 to 15?
- If there is no data for a particular value, do you have to show that value on the number line? For example, if your data are 1, 2, 3, 4, 8, 9, 10. Can you skip 5, 6, and 7?

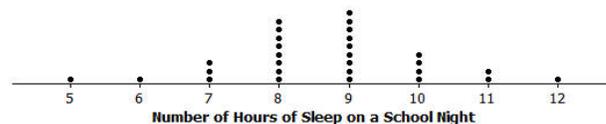
Exercises 1–9 (20 minutes)

The first five Exercises are designed to have students complete the dot plot that is started and to answer questions about the most common value and the center of the distribution. Exercise 6 is designed to have students make a dot plot without any prompts. Some students may need help with making the number line. For those students, have them find the lowest and highest value and suggest that they use these values to start and end their number line. Exercise 9 is designed to have students begin to compare two distributions. This comparison of distribution will be a focus of lessons later in this unit.

Allow students about 15 minutes to work independently or in small groups. Bring the groups together to summarize their answers.

Exercises 1–9

1. Complete Robert's dot plot by placing a dot above the number on the number line for each number of hours slept. If there is already a dot above a number, then add another dot above the dot already there.



2. What are the least and the most hours of sleep reported in the survey of 6th graders?
The least is 5, and the most is 12.
3. What is the most common number of hours slept?
9 is the most common.
4. How many hours of sleep describes the center of the data?
The center is around 8 or 9.

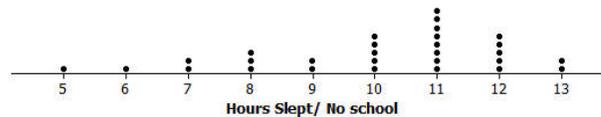
5. Think about how many hours of sleep you usually get on a school night. How does your number compare with the number of hours of sleep from the survey of 6th graders?

Answers will vary. Analyze answers based on students' responses.

Here are the data for the number of hours 6th graders sleep when they don't have school the next day:

7 8 10 11 5 6 12 13 13 7 9 8 10 12 11 12 8 9 10 11 10 12 11 11 11 12 11 11 10

6. Make a dot plot of the number of hours slept when there is no school the next day.



7. How many hours of sleep with no school the next day describe the center of the data?

Around 11 hours.

8. What are the least and most hours slept with no school the next day reported in the survey?

The least is 5, and the most is 13.

9. Do students sleep longer when they don't have school the next day than they do when they do have school the next day? Explain your answer using the data in both dot plots.

Yes, because more of the data points are in the 10, 11, 12, 13 categories in the "no school" dot plot than in the "have school" dot plot.

Example 2 (10 minutes): Building and Interpreting a Frequency Table

Example 2: Building and Interpreting a Frequency Table

A group of 6th graders investigated the statistical question: "How many hours per week do 6th graders spend playing a sport or outdoor game?"

Here are the data the students collected from a sample of 26 6th graders showing the number of hours per week spent playing a sport or a game outdoors:

3 2 0 6 3 3 3 1 1 2 2 8 12 4 4 4 3 3 1 1 0 0 6 2 3 2

To help organize the data, the students placed the number of hours into a frequency table. A frequency table lists items and how often each item occurs.

To build a frequency table, first draw three columns. Label one column "Number of Hours Playing a Sport/Game," label the second column "Tally," and the third column "Frequency." Since the least number of hours was 0, and the most was 12, list the numbers from 0 to 12 under the "Number of Hours" column.

Number of Hours Playing a Sport/Game	Tally	Frequency
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		

As you read each number of hours from the survey, place a tally mark opposite that number. The table shows a tally mark for the first number 3.

MP.4

The data shown come from a random sample of 6th graders collected from the Census at School website (<http://www.amstat.org/censusatschool/>). The format for the frequency table is presented, and students are directed on how to complete the table. It is important to point out to students when listing values under the number column that the numbers must be listed sequentially with no missing numbers or gaps in the numbers. Students should be able to draw a dot plot from the frequency table and build a frequency table from the dot plot. After students have completed the frequency table and the dot plot of the data, discuss with them what each representation tells about the data.

As you develop this example, pose the following question to students:

- What information is available in the frequency table that is not readily available in the dot plot?

Exercises 10–15 (10 minutes)

In Exercises 10 and 11, students complete the frequency table. In Exercise 12, students are directed to make a dot plot of the data. Encourage students to use the frequency table to help build the dot plot.

Exercise 15 is designed to have students begin to analyze the data as it is presented in two different representations. They should focus on the center and spread of the data as they answer this question.

Exercises 10–15

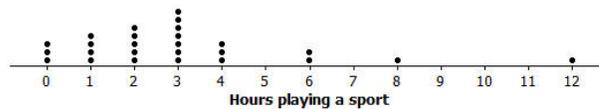
10. Complete the tally mark column.

Number of hours	Tally	Frequency
0		3
1		4
2		5
3		7
4		3
5		0
6		2
7		0
8		1
9		0
10		0
11		0
12		1

11. For each number of hours, find the total number of tally marks and place this in the frequency column.

See table above.

12. Make a dot plot of the number of hours playing a sport or playing outdoors.



13. What number of hours describes the center of the data?

Around 3.

14. How many 6th graders reported that they spend eight or more hours a week playing a sport or playing outdoors?

Only 2 students.

15. The 6th graders wanted to answer the question, “How many hours do 6th graders spend per week playing a sport or playing an outdoor game?” Using the frequency table and the dot plot, how would you answer the 6th graders’ question?

Most 6th graders spend about 2 to 4 hours per week playing a sport or playing outdoors.

Lesson Summary

This lesson described how to make a *dot plot*. This plot starts with a number line labeled from the smallest to the largest value. Then, a dot is placed above the number on the number line for each value in your data.

This lesson also described how to make a *frequency table*. A frequency table consists of three columns. The first column contains all the values of the data listed in order from smallest to largest. The second column is the tally column, and the third column is the number of tallies for each data value.

Exit Ticket (5 minutes)



Name _____

Date _____

Lesson 3: Creating a Dot Plot

Exit Ticket

A biologist collected data to answer the question: “How many eggs do robins lay?”

The following is a frequency table of the collected data:

Number of Eggs	Tally	Frequency
1		
2		
3		
4		
5		

- Complete the frequency column.
- Draw a dot plot of the number of eggs a robin lays.
- What number of eggs describes the center of the data?

Exit Ticket Sample Solutions

This Exit Ticket is designed to assess if a student can complete a frequency table and draw a dot plot from a given frequency table.

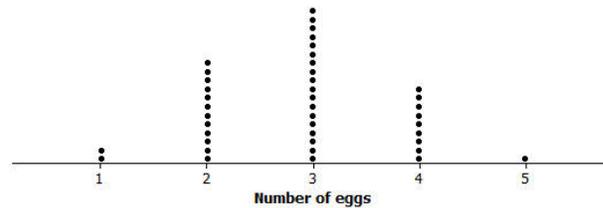
A biologist collected data to answer the question: “How many eggs do robins lay?”

The following is a frequency table of the collected data:

- Complete the frequency column.

Number of Eggs	Tally	Frequency
1		2
2		12
3		18
4		9
5		1

- Draw a dot plot of the number of eggs a robin lays.



- What number of eggs describes the center of the data?

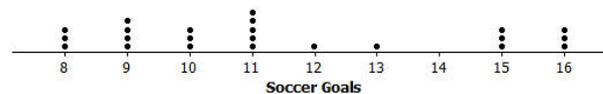
3

Problem Set Sample Solutions

- The data below is the number of goals scored by a professional indoor soccer team over their last 23 games.

8 16 10 9 11 11 10 15 16 11 15 13 8 9 11 9 8 11 16 15 10 9 12

- Make a dot plot of the number of goals scored.



- What number of goals describes the center of the data?

Around 11 or 12

- What is the least and most number of goals scored by the team?

8 is the least, and 16 is the most.

- Over the 23 games played, the team lost 10 games. Circle the dots on the plot that you think represent the games that the team lost. Explain your answer.

Students should circle the lowest 10 scores.

2. A 6th grader rolled two number cubes 21 times. The student found the sum of the two numbers that he rolled each time. The following are the sums of the 21 rolls of the two number cubes:

9 2 4 6 5 7 8 11 9 4 6 5 7 7 8 8 7 5 7 6 6

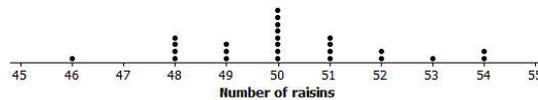
- a. Complete the frequency table.

Sum rolled	Tally	Frequency
2		1
3		0
4		2
5		3
6		4
7		5
8		3
9		2
10		0
11		1
12		0

- b. What sum describes the center of the data?
7
- c. What was the most common sum of the number cubes?
7

3. The dot plot below shows the number of raisins in 25 selected small boxes of raisins.

Dot Plot of Number of Raisins



- a. Complete the frequency table.

Number of Raisins	Tally	Frequency
46		1
47		0
48		4
49		3
50		8
51		4
52		2
53		1
54		2

- b. Another student opened up a box of raisins and reported that it had 63 raisins. Did this student have the same size box of raisins? Why or why not?

No, the boxes opened had at most 54 raisins, and 63 is too high.



Lesson 4: Creating a Histogram

Student Outcomes

- Students construct a frequency histogram.
- Students recognize that the number of intervals may affect the shape of a histogram.

Classwork

This lesson organizes the development of the student outcomes in three examples. Example 1 introduces frequency tables with intervals. Example 2 discusses how to create a histogram from the data that is organized in the interval frequency table from Example 1. Example 3 discusses another feature of a histogram – its shape. Students are introduced to a mound/symmetric shape and a skewed shape. Following each example is an exercise set designed for independent or small group work to reinforce the main objectives of constructing and interpreting a histogram. Teacher selection of problems is encouraged. If all problems are completed, this lesson may take longer than one class period.

Example 1 (10 minutes): Frequency Table with Intervals

Example 1: Frequency Table with Intervals

The boys and girls basketball teams at Roosevelt Middle School wanted to raise money to help buy new uniforms. They decided to sell hats with the school logo on the front to family members and other interested fans. To obtain the correct hat size, the students had to measure the head circumference (distance around the head) of the adults who wanted to order a hat. The following data represents the head circumferences, in millimeters (mm), of the adults:

513, 525, 531, 533, 535, 535, 542, 543, 546, 549, 551, 552, 552, 553, 554, 555, 560, 561, 563, 563, 565, 565, 568, 568, 571, 571, 574, 577, 580, 583, 583, 584, 585, 591, 595, 598, 603, 612, 618

The hats come in six sizes: XS, S, M, L, XL, and XXL. Each hat size covers a span of head circumferences. The hat manufacturer gave the students the table below that shows the interval of head circumferences for each hat size. The interval $510 < 530$ represents head circumferences from 510 to 530, not including 530.

Hat Sizes	Interval of Head Circumferences (mm)	Tally	Frequency
XS	$510 < 530$		
S	$530 < 550$		
M	$550 < 570$		
L	$570 < 590$		
XL	$590 < 610$		
XXL	$610 < 630$		

This example begins with data from the GAISE (Guidelines for Assessment and Instruction in Statistics Education) Report published by the American Statistical Association (<http://www.amstat.org/education/gaise/index.cfm>). The example presents the data in a frequency table, but the head circumferences are grouped into intervals. You may want to display a frequency table from Lesson 3 and discuss with the students the similarities and differences. It is also important that students understand that each interval should be the same width and that they should not skip intervals even if there is no data for an interval.

MP.1 As students complete the following exercises, pose the following questions:

- How is the frequency table with intervals similar to the frequency tables from Lesson 3? How is it different?
- What is the span or width of each interval? Are all the intervals the same width?
- What patterns do you see in the interval column?

Exercises 1–4 (10 minutes)

The four exercises that follow are designed to help students understand the idea of grouping data in intervals.

Exercises 1–4

1. If someone has a head circumference of 570, what size hat would they need?
Large
2. Complete the tally and frequency columns in the table to determine the number of each size hat the students need to order for the adults who wanted to order a hat.

Hat Sizes	Interval of Head Circumferences (mm)	Tally	Frequency
XS	510–< 530		2
S	530–< 550		8
M	550–< 570		15
L	570–< 590		9
XL	590–< 610		4
XXL	610–< 630		2
3. What hat size does the data center around?
Medium
4. Describe any patterns that you observe in the frequency column?
The numbers start small but increase to 15 and then go back down.

Example 2 (15 minutes): Histogram

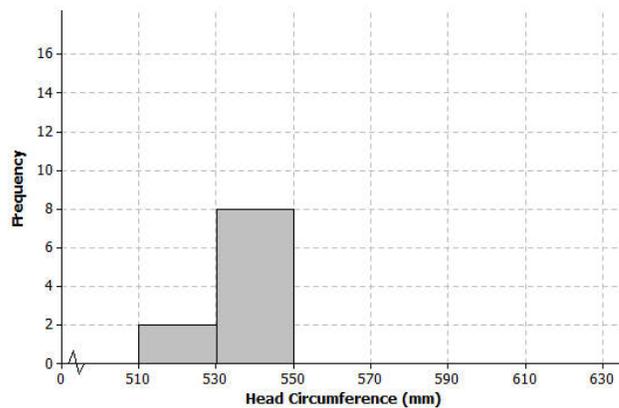
Example 2: Histogram

One student looked at the tally column and said that it looked somewhat like a bar graph turned on its side. A histogram is a graph that is like a bar graph, except that the horizontal axis is a number line that is marked off in equal intervals.

To make a histogram:

- Draw a horizontal line and mark the intervals.
- Draw a vertical line and label it “frequency.”
- Mark the frequency axis with a scale that starts at 0 and goes up to something that is greater than the largest frequency in the frequency table.
- For each interval, draw a bar over that interval that has a height equal to the frequency for that interval.

The first two bars of the histogram have been drawn below.



The students are introduced to a histogram in this example. They use the data that was organized in a frequency table with intervals in Example 1. You may want to begin this lesson by showing the students an example of a bar graph. For example, show a bar graph showing favorite pizza toppings. Point out the horizontal axis is *not* a number line, but contains categories. The vertical axis is the frequency (or count) of how many people chose the particular pizza topping. As you present the histogram to the students, point out the main difference is the horizontal axis is a number line, and the intervals are listed in order from smallest to largest. Some students may struggle with the notation for the intervals. Point out to the students that the interval labeled $510 < 530$, represents any head circumference from 510 mm to 530, not including 530. A head circumference of 530 is counted in the bar from $530 - 550$ and not in the bar from $510 - 530$.

Pose the following questions as students develop the following exercise:

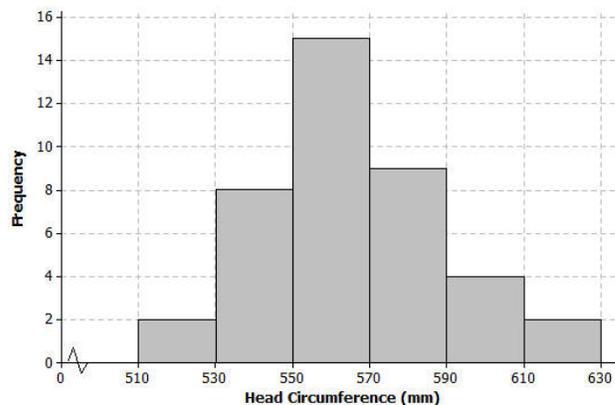
- Why should the bars touch each other in the histogram?
- How are histograms and bar graphs similar? How are they different?

Exercises 5–9 (10 minutes)

In the first problem, students are asked to complete the histogram. Emphasize that the bars should touch each other and be the same width. Also point out the jagged line (or “scissor cut”), and explain that it is used to indicate a cutting of the horizontal axis. (A “scissor cut” could also be used on a vertical axis.) The cut is used to show the graph by “pulling in” unused space.

Exercises 5–9

5. Complete the histogram by drawing bars whose heights are the frequencies for those intervals.



6. Based on the histogram, describe the center of the head circumferences.

Around 560 mm.

7. How would the histogram change if you added head circumferences of 551 and 569?

The bar in the 550 to 570 interval would go up to 17.

8. Because the 40 head circumference values were given, you could have constructed a dot plot to display the head circumference data. What information is lost when a histogram is used to represent a data distribution instead of a dot plot?

In a dot plot, you can see individual values. In a histogram, you only see the total number of values in an interval.

9. Suppose that there had been 200 head circumference measurements in the data set. Explain why you might prefer to summarize this data set using a histogram rather than a dot plot.

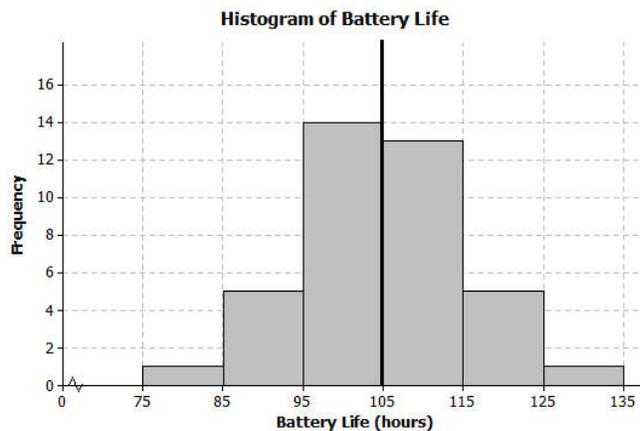
There would be too many dots on a dot plot, and it would be hard to read. A histogram would work for a large data set because the scale can be adjusted.

Example 3 (10 minutes): Shape of the Histogram

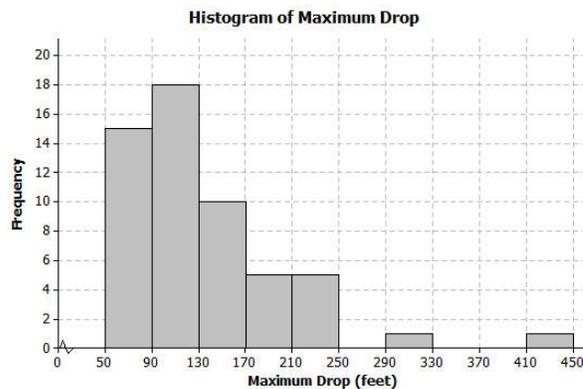
Example 3: Shape of the Histogram

A histogram is useful to describe the shape of the data distribution. It is important to think about the shape of a data distribution because depending on the shape, there are different ways to describe important features of the distribution, such as center and variability.

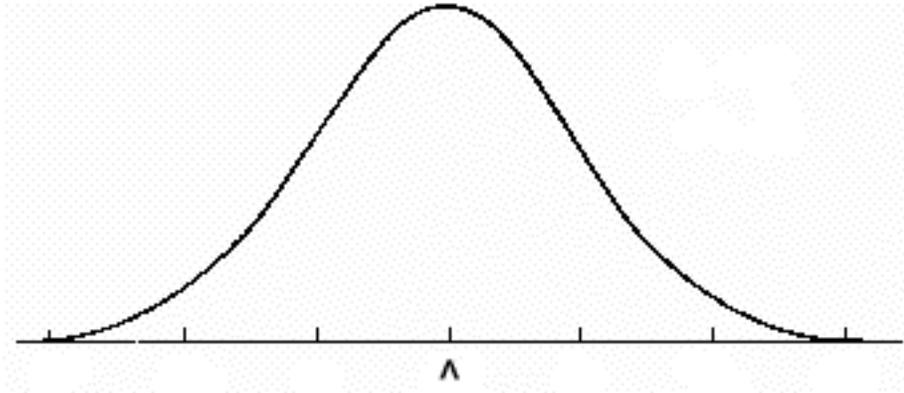
A group of students wanted to find out how long a certain brand of AA batteries lasted. The histogram below shows the data distribution for how long (in hours) that some AA batteries lasted. Looking at the shape of the histogram, notice how the data “mounds” up around a center of approximately 105. We would describe this shape as mound shaped or symmetric. If we were to draw a line down the center, notice how each side of the histogram is approximately the same or mirror images of each other. This means the graph is approximately symmetrical.



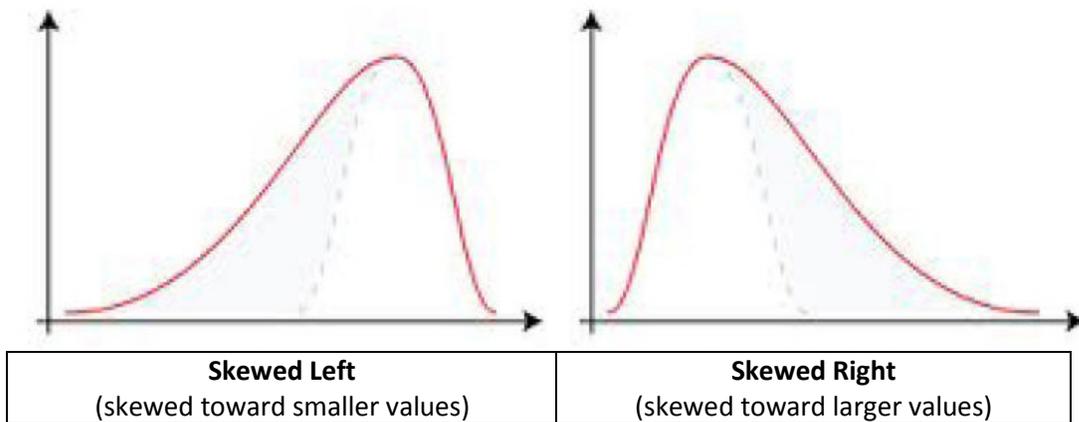
Another group of students wanted to investigate the maximum drop length for roller coasters. The histogram below shows the maximum drop (in feet) of a selected group of roller coasters. This histogram has a skewed shape. Most of the data are in the intervals from 50 to 170. But there are two values that are unusual (or not typical) when compared to the rest of the data. These values are much higher than most of the data.



MP.4 This example discusses the concept of the shape of a distribution and how it relates to center and variability. Two shapes are introduced, mound-shaped or symmetric and skewed. Below is an example of a symmetric (i.e., mound-shaped) distribution. The emphasis is on the approximate symmetry in the histogram.



Below are two examples of skewed distributions:



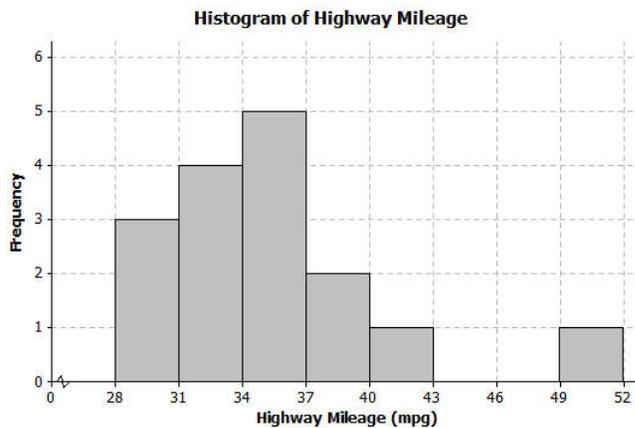
Point out to the students that a skewed distribution has values that are not typical of the rest of the data. They either could be data much greater than the rest of the data or much lower than the rest of the data. The graph will have a tail that is longer on one side than the other.

Exercises 10–12 (10 minutes)

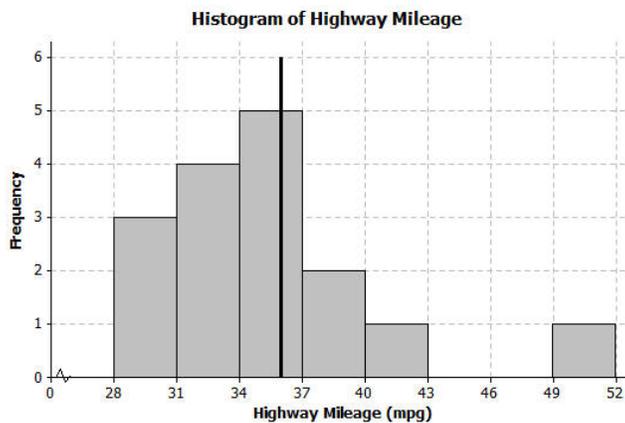
The next three questions are designed to help students classify a distribution as approximately symmetric (i.e., mound-shaped) or skewed.

Exercises 10–12

10. The histogram below shows the highway miles per gallon of different compact cars.



- a. Describe the shape of the histogram as approximately symmetric, skewed left, or skewed right.
Skewed right toward the larger values.
- b. Draw a vertical line on the histogram to show where the “typical” number of miles per gallon for a compact car would be.
Around 36.

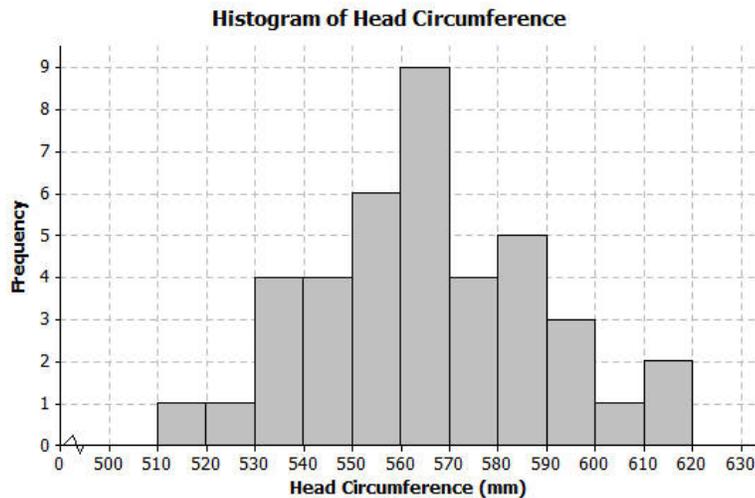


- c. What does the shape of the histogram tell you about miles per gallon for compact cars?
Most cars get around 31 to 40 mpg. But there was one car that got between 49 and 52 mpg.

11. Describe the shape of the head circumference histogram that you completed in Exercise 5 as approximately symmetric, skewed left, or skewed right.

Approximately symmetric.

12. Another student decided to organize the head circumference data by changing the width of each interval to be 10 instead of 20. Below is the histogram that the student made.



- a. How does this histogram compare with the histogram of the head circumferences that you completed in Exercise 5?

Answers will vary; same shape and center, but not as symmetric.

- b. Describe the shape of this new histogram as approximately symmetric, skewed left, or skewed right.

Approximately symmetric.

- c. How many head circumferences are in the interval from 570 to 590?

9

- d. In what interval would a head circumference of 571 be included? In what interval would a head circumference of 610 be included?

571 is in the interval from 570 to 580; 610 is in the interval from 610 to 620.

Lesson Summary

A histogram is a graph that represents the number of data values falling in an interval with a bar. The horizontal axis shows the intervals and the vertical axis shows the frequencies (how many data values are in the interval). Each interval should be the same width and the bars should touch each other.

Exit Ticket (7–10 minutes)



Name _____

Date _____

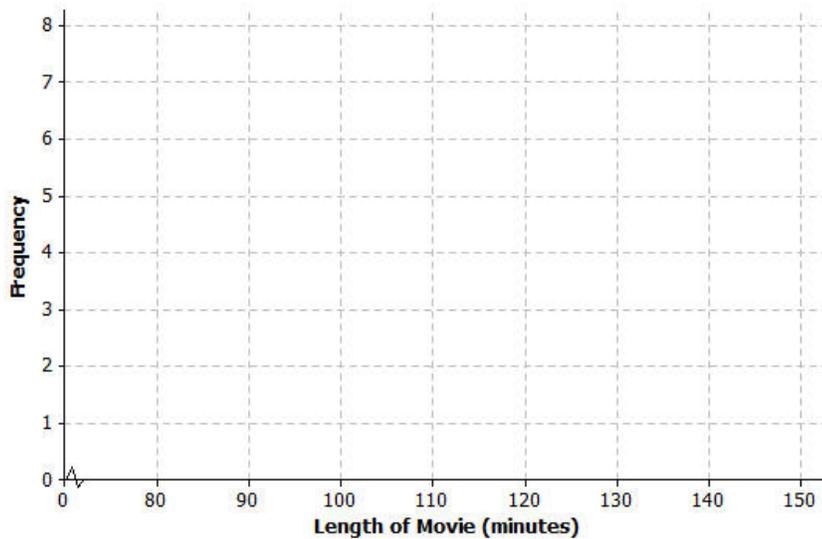
Lesson 4: Creating a Histogram

Exit Ticket

The frequency table below shows the length of selected movies shown in a local theater over the past six months.

Length of Movie (min)	Tally	Frequency
80–< 90		1
90–< 100		4
100–< 110	++++	7
110–< 120	++++	5
120–< 130	++++	7
130–< 140		3
140–< 150		1

- Construct a histogram for the length of movies data.



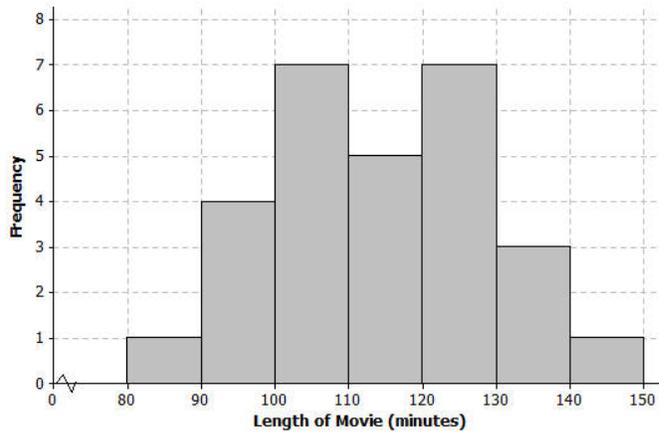
- Describe the shape of the histogram.
- What does the shape tell you about the length of movies?

Exit Ticket Sample Solutions

The frequency table below shows the length of selected movies shown in a local theater over the past six months.

Length of Movie (min)	Tally	Frequency
80–< 90		1
90–< 100		4
100–< 110		7
110–< 120		5
120–< 130		7
130–< 140		3
140–< 150		1

- Construct a histogram for the length of movies data.

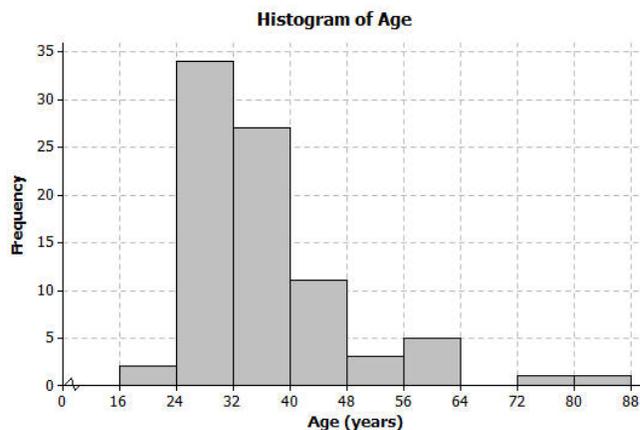


- Describe the shape of the histogram.
Mound shaped or approximately symmetric.
- What does the shape tell you about the length of movies?
Most movies lengths were between 100 and 130 minutes.

Problem Set Sample Solutions

Note that teacher discretion is encouraged for assigning problems from this problem set. Problems are provided to address the varying interests of students.

1. The following histogram shows ages of the actresses whose performances have won in the Best Leading Actress category at the annual Academy Awards (Oscars).

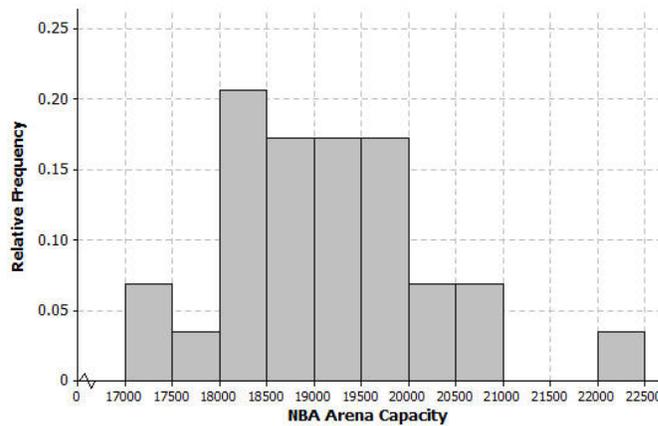


- Which age interval contains the most actresses? How many actresses are represented in that interval?
The interval 24 to 32 contains the most actresses. There are 34 actresses whose age falls into that category.
- Describe the shape of the histogram.
Skewed to the right.
- What does the shape tell you about the ages of actresses who win the Oscar for best actress award?
Most of the ages are between 24 and 40, with two ages much larger than the rest.
- Which interval describes the center of the ages of the actresses?
32 to 40
- An age of 72 would be included in which interval?
It is in the interval from 72 to 80.

2. The frequency table below shows the seating capacity of arenas for NBA basketball teams.

Number of seats	Tally	Frequency
17000–< 17500		2
17500–< 18000		1
18000–< 18500		6
18500–< 19000		5
19000–< 19500		5
19500–< 20000		5
20000–< 20500		2
20500–< 21000		2
21000–< 21500		0
21500–< 22000		0
22000–< 22500		1

a. Draw a histogram of the number of seats in NBA arenas. Use the histograms you have seen throughout this lesson to help you in the construction of your histogram.



b. What is the width of each interval? How do you know?

The width of each interval is 500.

Subtract the values identifying an interval.

c. Describe the shape of the histogram.

Skewed to the right.

d. Which interval describes the center of the number of seats?

19,000 to 19,500

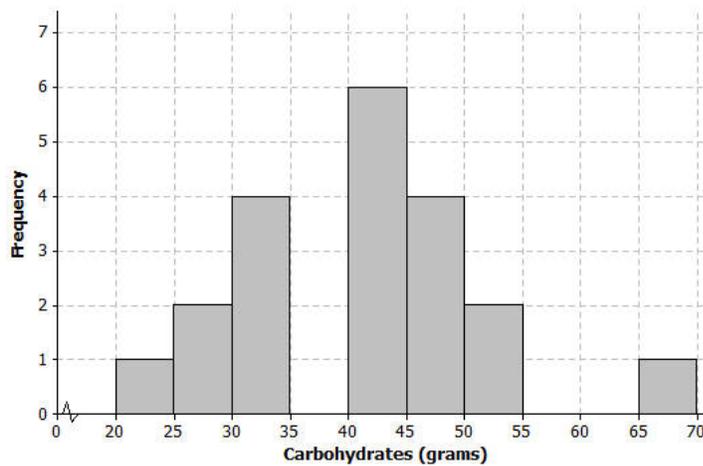
3. Listed are the grams of carbohydrates in hamburgers at selected fast food restaurants.

33 40 66 45 28 30 52 40 26 42
42 44 33 44 45 32 45 45 52 24

a. Complete the frequency table with intervals of width 5.

Number of carbohydrates (grams)	Tally	Frequency
20–< 25		1
25–< 30		2
30–< 35		4
35–< 40		0
40–< 45		6
45–< 50		4
50–< 55		2
55–< 60		0
60–< 65		0
65–< 70		1

b. Draw a histogram of the carbohydrate data.



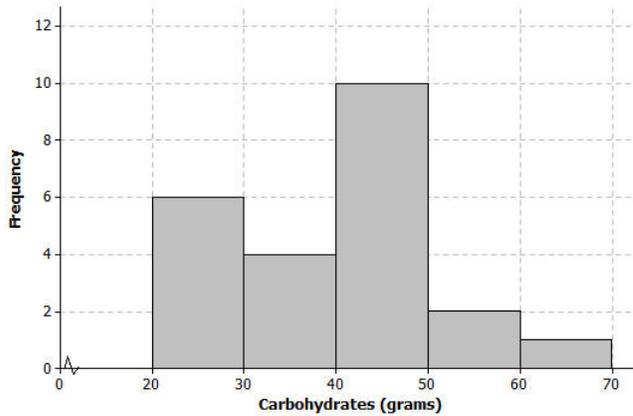
c. Describe the center and shape of the histogram.

Center is around 40; it is mound shaped.

d. In the frequency table below, the intervals are changed. Using the carbohydrate data above, complete the frequency table with intervals of width 10.

Number of carbohydrates (grams)	Tally	Frequency
20–< 30		3
30–< 40		4
40–< 50	+	10
50–< 60		2
60–< 70		1

e. Draw a histogram.



4. Use the histograms that you constructed in question 3 parts (b) and (e) to answer the following questions.

a. Why are there fewer bars in the histogram in question 3 part (e) than the histogram in part (b)?

There are fewer bars because the width of the interval changed from 5 grams to 10 grams, so there are fewer intervals.

b. Did the shape of the histogram in question 3 part (e) change from the shape of the histogram in part (b)?

Generally, both are mound shaped, but the histogram in question 3 part (b) has gaps.

c. Did your estimate of the center change from the histogram in question 3 part (b) to the histogram in part (e)?

The centers of the two histograms are about the same.



Lesson 5: Describing a Distribution Displayed in a Histogram

Student Outcomes

- Students construct a relative frequency histogram.
- Students recognize that the shape of a histogram does not change when relative frequency is used compared to when frequency is used to construct the histogram.

Lesson Notes

This lesson may take longer than one class period.

Classwork

Example 1 (10 minutes): Relative Frequency Table

Example 1: Relative Frequency Table

In Lesson 4, we investigated the head circumferences that the boys and girls basketball teams collected. Below is the frequency table of the head circumferences that they measured.

Hat Sizes	Interval of Head Circumferences (mm)	Tally	Frequency
XS	510–< 530		2
S	530–< 550		8
M	550–< 570		15
L	570–< 590		9
XL	590–< 610		4
XXL	610–< 630		2
		Total	40

Isabel, one of the basketball players, indicated that most of the hats were small, medium, or large. To decide if Isabel was correct, the players added a relative frequency column to the table. Relative frequency is the value of the frequency in an interval divided by the total number of data values.

This example begins with the frequency table of head circumferences that students used in Lesson 4. At the start of the lesson, display the frequency table and ask:

- What does the 15 in the frequency column represent, and how many hats need to be ordered?
- What percent of the total order are medium-size hats?
 - This question leads into the vocabulary of relative frequency as the ratio of the frequency for an interval divided by the total number of data values.*

Explain the concept of *relative frequency* by working through the calculation of the first two rows in the table:

- There are 2 people in the XS hat size interval (head circumferences from 510 – 529). The relative frequency for this interval is 2 divided by the total number 40, or $\frac{2}{40}$, which is 0.05 or 5%.
- In the interval from 530– 549 the frequency is 8. The relative frequency for this interval is $\frac{8}{40}$, which is 0.2 or 20%.

Ask the students:

- How do you find the total number of data values?
- What will the sum of the relative frequency column equal?
- What is the difference between a frequency table and a relative frequency table?
- How are the two types of table similar?

The students should write the relative frequency as a decimal. Converting the decimal to a percent helps to interpret the value. When writing the relative frequency, have the students write their answers to three decimal places. For some exercises they may have to round to the nearest thousandth.

Exercises 1–4 (15 minutes)

In this exercise students are asked to complete the relative frequency table and begin to interpret the values in the table. Let students work in pairs and confirm answers as a class.

Exercises 1–4

1. Complete the relative frequency column in the table below.

Hat Sizes	Interval of Head Circumferences (mm)	Tally	Frequency	Relative Frequency
XS	510–< 530		2	$\frac{2}{40} = 0.05$
S	530–< 550		8	$\frac{8}{40} = 0.20$
M	550–< 570		15	0.375
L	570–< 590		9	0.225
XL	590–< 610		4	0.10
XXL	610–< 630		2	0.05
		Total	40	

2. What is the total of the relative frequency column?
100%

3. Which interval has the greatest relative frequency? What is the value?
Medium size hats 550– 569; relative frequency = 0.375.

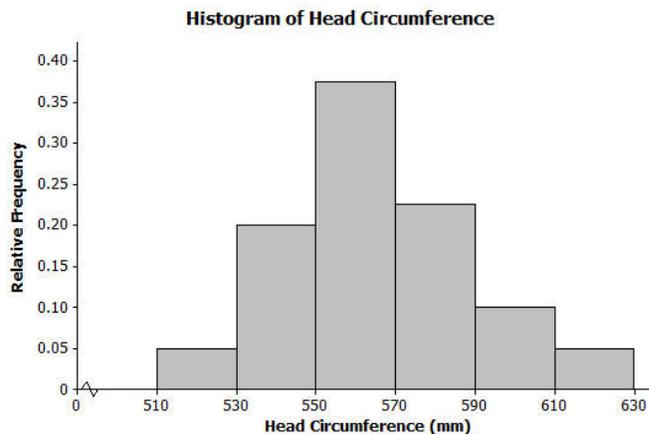
4. What percent of the head circumferences is between 530 and 589? Show how you determined the answer.
 $0.20 + 0.375 + 0.225 = 0.80$ or 80%

Example 2 (15 minutes): Relative Frequency Histogram**Example 2: Relative Frequency Histogram**

The players decided to construct a histogram using the relative frequencies instead of the frequencies.

They noticed that the relative frequencies in the table ranged from close to 0 to about 0.40. They drew a number line and marked off the intervals on that line. Then, they drew the vertical line and labeled it relative frequency. They added a scale to this line by starting at 0 and counting by 0.05 until they reached 0.40.

They completed the histogram by drawing the bars so the height of each bar matched the relative frequency for that interval. Here is the completed relative frequency histogram:

**MP.1**

In this example students consider the connection between the table and relative frequency histogram. Display the frequency histogram of head circumferences from Lesson 4. Remind students of the importance of the intervals being the same width.

Alongside this frequency histogram, demonstrate how to construct a relative frequency histogram. The labeling of the horizontal axis is the same as for the frequency histogram. The vertical axis scale changes to represent the relative frequency. Students may struggle with the scaling along this vertical axis since you are counting by a decimal rather than by a whole number.

After drawing the relative frequency histogram, ask the students to compare the two histograms. They should notice that the center and shape are the same.

- What do you notice about the shape and center?
 - *They are the same.*
- What is the greatest number that could be on the vertical axis in a relative frequency histogram?
 - 1.00
- What is the relative frequency for the large hat sizes? What does this number mean?
 - *About 0.24; approximately 24% of the people measured would wear a large hat.*

Exercises 5–6 (15 minutes)

The first exercise asks students to compare the two types of histograms: a frequency histogram and a relative frequency histogram. In the second exercise, students are asked to calculate relative frequencies and to construct a relative frequency histogram. Let students work in small groups. Allow for calculator usage.

Exercises 5–6

5. Answer the following questions.

a. Describe the shape of the relative frequency histogram of head circumferences from Example 2.

Slightly skewed to the right.

b. How does the shape of this histogram compare with the frequency histogram you drew in Exercise 5 of Lesson 4?

The shape looks the same.

c. Isabel said that most of the hats that needed to be ordered were small, medium, and large. Was she right? What percent of the hats to be ordered is small, medium, or large?

She was right. The total percentage was 80%. (Small 20%, Medium 37.5%, Large 22.5% for a total of 80%.)

6. Here is the frequency table of the seating capacity of arenas for the NBA basketball teams.

Number of seats	Tally	Frequency	Relative Frequency
17,000–< 17,500		2	0.069
17,500–< 18,000		1	0.034
18,000–< 18,500	+++	6	0.207
18,500–< 19,000	+++	5	0.172
19,000–< 19,500	+++	5	0.172
19,500–< 20,000	+++	5	0.172
20,000–< 20,500		2	0.069
20,500–< 21,000		2	0.069
21,000–< 21,500		0	0
21,500–< 22,000		0	0
22,000–< 22,500		1	0.034

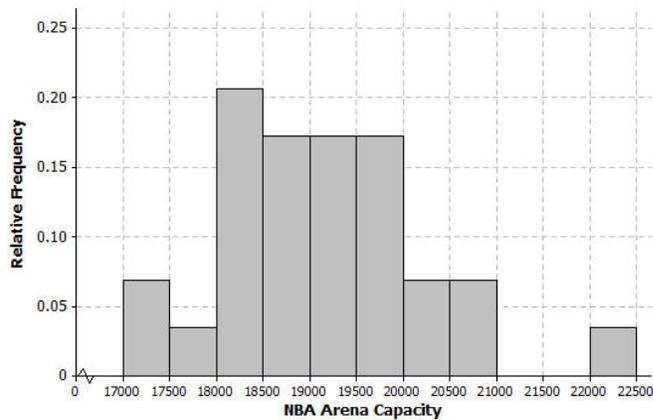
a. What is the total number of NBA arenas?

29

b. Complete the relative frequency column. Round to the nearest thousandth.

See table above.

c. Construct a relative frequency histogram. Round to the nearest thousandth.



d. Describe the shape of the relative frequency histogram.

Slightly skewed to the right.

e. What percent of the arenas has a seating capacity between 18,500 and 19,999 seats?

0.516 or 51.6%

f. How does this relative frequency histogram compare to the frequency histogram that you drew in problem 2 of the Problem Set in Lesson 4?

It has the same shape.

Lesson Summary

A **relative frequency histogram** uses the same data as a frequency histogram but compares the frequencies for each interval frequency to the total number of items. For example, if the first interval contains 8 out of the total of 32 items, the relative frequency of the first interval $\frac{8}{32}$ or $\frac{1}{4} = 0.25$.

The only difference between a frequency histogram and a relative frequency histogram is that the vertical axis uses relative frequency instead of frequency. The shapes of the histograms are the same as long as the intervals are the same.

Exit Ticket (5 minutes)



Name _____

Date _____

Lesson 5: Describing a Distribution Displayed in a Histogram

Exit Ticket

Calculators are allowed for completing your problems.

Hector's mom had a rummage sale, and after she sold an item, she tallied for how much money she sold the item. Following is the frequency table Hector's mom created:

Amount of Money the Item sold for	Tally	Frequency	Relative Frequency
\$0–< 5		2	
\$5–< \$10		1	
\$10–< \$15		4	
\$15–< \$20	+++ +++	10	
\$20–< \$25	+++	5	
\$25–< \$30		3	
\$30–< \$35		2	

- What was the total number of items sold at the rummage sale?
- Complete the relative frequency column. Round to the nearest thousandth.
- What percent of the items Hector's mom sold was sold for \$15 or more, but less than \$20?

Exit Ticket Sample Solutions

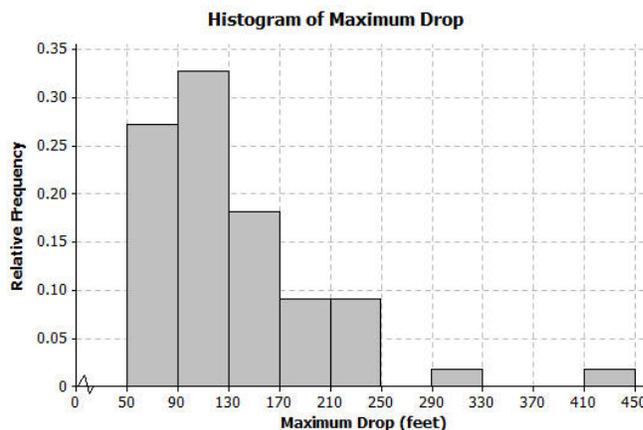
Hector’s mom had a rummage sale, and after she sold an item, she tallied for how much money she sold the item. Following is the frequency table Hector’s mom created:

Amount of Money the Item sold for	Tally	Frequency	Relative Frequency
\$0–< \$5		2	0.074
\$5–< \$10		1	0.037
\$10–< \$15		4	0.148
\$15–< \$20		10	0.370
\$20–< \$25		5	0.185
\$25–< \$30		3	0.111
\$30–< \$35		2	0.074

- What was the total number of items sold at the rummage sale?
27 items
- Complete the relative frequency column. Round to the nearest thousandth.
See table above.
- What percent of the items Hector’s mom sold was sold for \$15 or more, but less than \$20?
0.37 or 37%

Problem Set Sample Solutions

1. Below is a relative frequency histogram of the maximum drop (in feet) of a selected group of roller coasters.



- Describe the shape of the relative frequency histogram.
Skewed to the right.

b. What does the shape tell you about the maximum drop (in feet) of roller coasters?

Most of the roller coasters have a maximum drop that is between 50 and 170 feet.

c. Jerome said that more than half of the data is in the interval from 50 – 130 feet. Do you agree with Jerome? Why or why not?

Yes, that span has 60% of the data.

2. The frequency table below shows the length of selected movies shown in a local theater over the past 6 months.

Length of Movie (min)	Tally	Frequency	Relative Frequency
80–< 90		1	0.036
90–< 100		4	0.143
100–< 110		7	0.25
110–< 120		5	0.179
120–< 130		7	0.25
130–< 140		3	0.107
140–< 150		1	0.036

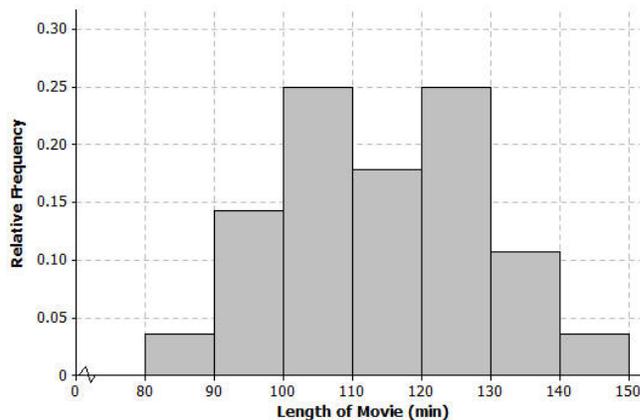
a. Complete the relative frequency column. Round to the nearest thousandth.

See table above.

b. What percent of the movie lengths is greater than or equal to 130 minutes?

0.143 = 14.3%

c. Draw a relative frequency histogram.



d. Describe the shape of the relative frequency histogram.

Mound shaped/approximately symmetric.

e. What does the shape tell you about the length of movie times?

The length of most movies is between 100 and 130 minutes.

3. The table below shows the highway mile per gallon of different compact cars.

Mileage	Tally	Frequency	Relative Frequency
128-< 31		3	0.188
31-< 34		4	0.250
34-< 37		5	0.313
37-< 40		2	0.125
40-< 43		1	0.063
43-< 46		0	0
46-< 49		0	0
49-< 52		1	0.063

a. What is the total number of compact cars?

16

b. Complete the relative frequency column. Round to the nearest thousandth.

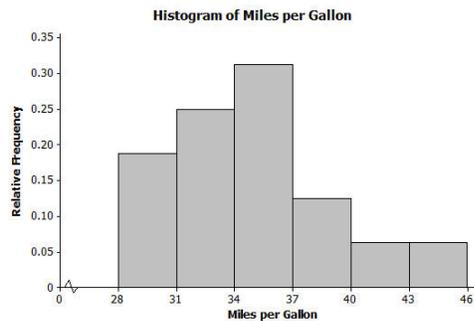
See table above.

c. What percent of the cars gets between 31 and up to but not including 37 miles per gallon on the highway?

$0.563 = 56.3\%$

d. Juan drew the relative frequency histogram of the miles per gallon of the compact cars, shown on the right. Do you agree with the way Juan drew the histogram? Explain your answer.

No, Juan skipped the intervals 43-< 46 and 46-< 49.





Topic B:

Summarizing a Distribution that is Approximately Symmetric Using the Mean and Mean Absolute Deviation

6.SP.A.2, 6.SP.A.3, 6.SP.B.4, 6.SP.B.5

Focus Standard:	6.SP.A.2	Understand that a set of data collected to answer a statistical question has a distribution, which can be described by its center, spread, and overall shape.
	6.SP.A.3	Recognize that a measure of center for a numerical data set summarizes all of its values with a single number, while a measure of variation describes how its values vary with a single number.
	6.SP.B.4	Display numerical data in plots on a number line, including dot plots, histograms, and box plots.
	6.SP.B.5	Summarize numerical data sets in relation to their context, such as by: <ol style="list-style-type: none"> Reporting the number of observations. Describing the nature of the attribute under investigation, including how it was measured and its units of measurement. Giving quantitative measures of center (median and/or mean) and variability (interquartile range and/or mean absolute deviation), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered. Relating the choice of measures of center and variability to the shape of the data distribution and the context in which the data were gathered.

Instructional Days: 6**Lesson 6:** Describing the Center of a Distribution Using the Mean (P)¹**Lesson 7:** The Mean as a Balance Point (P)**Lesson 8:** Variability in a Data Distribution (P)**Lesson 9:** The Mean Absolute Deviation (MAD) (P)**Lessons 10–11:** Describing Distributions Using the Mean and MAD (P,P)

In Topic B, students begin to summarize data distributions numerically. In Topic A, students have represented data distributions graphically and have described distributions informally in terms of shape, center, and variability. In this topic, students are introduced to a measure of center (the mean) and a measure of variability (the mean absolute deviation (MAD)) that are appropriate for describing data distributions that are approximately symmetric. In Lesson 6, students learn to calculate the mean and to understand the “fair share” interpretation of the mean. In Lesson 7, students develop an understanding of the mean as a balance point of a data distribution—the point where the sum of distances of points to the right of the mean and the sum of distances of points to the left of the mean are equal. This understanding provides a foundation for considering distances from the mean, which are used in calculating the MAD, a measure of variability around the mean. Lessons 8 and 9 introduce the MAD as a measure of variability, and students calculate and interpret the value of the MAD. Lessons 10 and 11 give students the opportunity to use both graphical and numerical summaries to describe data distributions, to compare distributions, and to answer questions in context using information provided by a data distribution.

¹ Lesson Structure Key: **P**-Problem Set Lesson, **M**-Modeling Cycle Lesson, **E**-Exploration Lesson, **S**-Socratic Lesson



Lesson 6: Describing the Center of a Distribution Using the Mean

Student Outcomes

- Students define the center of a data distribution by a “fair share” value called the mean.
- Students connect the “fair share” concept with a mathematical formula for finding the mean.

Lesson Notes

In earlier grades, students may have heard the term *average* (or *mean*) to describe a measure of center, although it is not part of Common Core Grades K–5. If they have heard the term, typically their understanding of it is the “add up and divide” formula. The goal of Lesson 6 is to bring students an understanding of what *mean* is, not just how to find it.

For example, if students hear that their class had a mean score of 74 on a test, we want them to immediately understand that 74 is the score each student in the class would receive! That’s the “fair share” interpretation of mean. So, when the term *mean* is mentioned, we want students to think initially of its “fair share” meaning and not of its mathematical formula, although they do go hand-in-hand.

Some students have difficulty understanding what “characterizing a data distribution” means. The idea expressed in this lesson is that single numbers are sought to characterize some feature (e.g., the “center”) of the distribution and that there may be several different ways to characterize a given specific feature. If students suggest the *mode* and *median* as measures of center, that’s great, although we are not going to pursue them in this lesson. The term *mode* is not discussed much at all in the Common Core, and *median* will be covered in a later lesson.

Teachers should be prepared to distribute some type of manipulative, like Unifix cubes, for group work in Exercise 3. Students use cubes to represent data and manipulate them to develop a measure of center, namely the “fair share” interpretation of the mean. Each group will need to have 90 units of the manipulative, so teachers should plan accordingly.

MP.4

Classwork

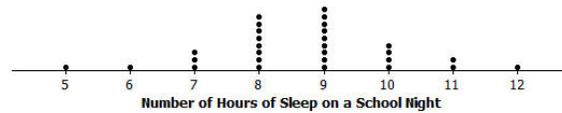
Example 1 (5–7 minutes)

Example 1

Recall that in Lesson 3, Robert, a 6th grader at Roosevelt Middle School, investigated the number of hours of sleep sixth grade students get on school nights. Today, he is to make a short report to the class on his investigation. Here is his report.

“I took a survey of 29 6th graders asking them ‘How many hours of sleep per night do you usually get when you have school the next day?’ The first thing I had to do was to organize the data. I did this by drawing a dot plot.

Dot Plot of Number of Hours of Sleep



Part of our lessons last week was to identify what we thought was a centering point of the data, the spread of the data, and the shape of the data. So, for my data, looking at the dot plot, I would say that the typical number of hours sixth-grade students sleep get when they have school the next day is around 8 or 9 because that is what most students said and the values are kind of in the middle. I also noticed that the data were spread out from the center by about three or four hours in both directions. The shape of the distribution is kind of like a mound.”

Michelle is Robert’s classmate. She liked his report but has a really different thought about determining the center of the number of hours of sleep. Her idea is to even out the data in order to determine a typical or center value.

Read the introductory paragraph to the class. Choose a student to read “Robert’s thought process” out loud. Then ask students:

- How is Robert thinking about the center?
 - *When asked to characterize the “center” of the hours of sleep data as represented in a dot plot, many students (including Robert) are drawn to the data point that occurs most often (the mode) or to the middle of the data set (the median).*

Read through the last two sentences of the example. Then ask students:

- What do you think Michelle means by evening out the data to determine a typical or center value?
 - *Michelle wants to get students thinking a bit more deeply about determining a center. The bottom line is that her view of “center” is an equal sharing of the data (i.e., a “fair share” in which the “fair share” process terminates when all subjects have the same amount of data).*

Note: To help explain what Michelle means, it’s easier to use a smaller number of data points. Robert’s data set is too big to work with.

Exercises 1–6 (15 minutes)

Work through Exercises 1–2 as a class. Briefly summarize Michelle’s “fair share method” from the text. Then, split students up into groups to work on Exercises 3–6, with each group getting 90 cubes.

Exercises 1–6

Suppose that Michelle asks ten of her classmates for the number of hours they usually sleep when there is school the next day.

Suppose they responded (in hours): 8 10 8 8 11 11 9 8 10 7

1. How do you think Robert would organize his data? What do you think Robert would say is the center of these ten data points? Why?

Dot plot; the center is around 8 hours because it is the most common value.

2. Do you think his value is a good measure to use for the “center” of Michelle’s data set? Why or why not?

Answers will vary; it is a good measure as most of the values are described by 8 hours, or it is not a good measure because half of the values are greater than 8 hours.

Michelle’s “center” is called the mean. She finds the total number of hours of sleep for each of the ten students. That is 90 hours. She has 90 Unifix cubes (Snap cubes). She gives each of the ten students the number of cubes that equals the number of hours of sleep each had reported. She then asks each of the ten students to connect their cubes in a stack and put their stacks on a table to compare them. She then has them share their cubes with each other until they all have the same number of cubes in their stacks when they are done sharing.

3. Work in a group. Each group of students gets 90 cubes. Make ten stacks of cubes representing the number of hours of sleep for each of the ten students. Using Michelle’s Method, how many cubes are in each of the ten stacks when they are done sharing?

There will be 9 cubes in each of the 10 stacks.

4. Noting that one cube represents one hour of sleep, interpret your answer to Exercise 3 in terms of “number of hours of sleep.” What does this number of cubes in each stack represent? What is this value called?

If all ten students slept the same number of hours, it would be 9 hours. The 9 cubes for each stack represent the 9 hours of sleep for each student if this was a fair share. This value is called the mean.

5. Suppose that the student who told Michelle he slept 7 hours changes his data entry to 8 hours. You will need to get one more cube from your teacher. What does Michelle’s procedure now produce for her center of the new set of data? What did you have to do with that extra cube to make Michelle’s procedure work?

The extra cube must be split into 10 equal parts. The mean is now $9\frac{1}{10}$.

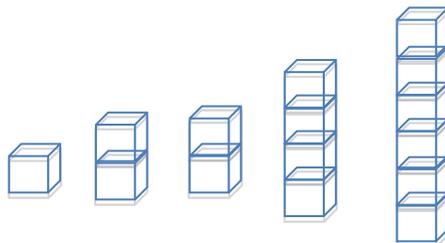
6. Interpret Michelle’s “fair share” procedure by developing a mathematical formula that results in finding the fair share value without actually using cubes. Be sure that you can explain clearly how the fair share procedure and the mathematical formula relate to each other.

Answers may vary. The “fair share” procedure is the same as adding all of the values and dividing by the number of data points.

Example 2 (5 minutes)

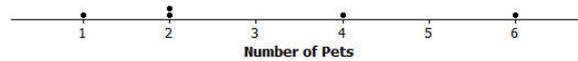
Example 2

Suppose that Robert asked five sixth graders how many pets each had. Their responses were 2, 6, 2, 4, 1. Robert showed the data with cubes as follows:



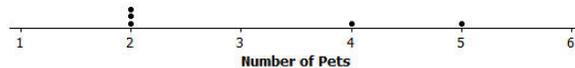
Note that one student has one pet, two students have two pets each, one student has four pets, and one student has six pets. Robert also represented the data set in the following dot plot.

Dot Plot of Number of Pets

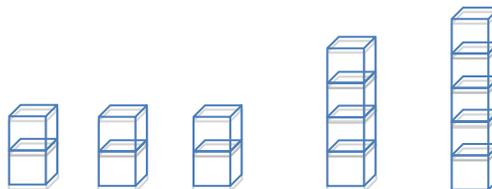


Robert wants to illustrate Michelle’s fair share method by using dot plots. He drew the following dot plot and said that it represents the result of the student with six pets sharing one of her pets with the student who has one pet.

Dot Plot of Number of Pets



Robert also represented the data with cubes as shown below.

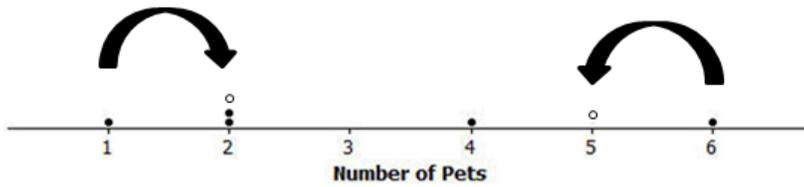


This example gets students to distinguish between the representations of a data set using cubes versus a dot plot. It also reinforces the concept of *sharing* – one student gives a pet to another that needs one.

Read through the first scenario with students, and then ask:

- Does the original stack of cubes match the dot plot representation? Explain.
 - Yes, the number of cubes in each stack corresponds with a dot on the plot.

Read through the next part of the example with students, where one student *shares* a pet with another. Demonstrate this step visually on the board or by using an overhead projector. Display the numerical representation next to the dot plot.



1	→	2
2		2
2		2
4		4
+ 6	→	5
15		15

Then ask:

- Is Robert’s new dot plot correct?
 - Yes
- How does the dot plot change?
 - *The student who had six pets now has five (new dot), and the student who had one pet now has two (new dot) – the dots are moving towards each other.*
- Are the stacks of cubes correct?
 - Yes
- Do the dot plot and stacks represent a “fair share” or mean?
 - *No, there is no typical or center value yet.*

Exercises 7–10 (12–15 minutes)

Let students work in pairs to complete Exercises 7–10.

Exercises 7–10

Now continue distributing the pets based on the following steps.

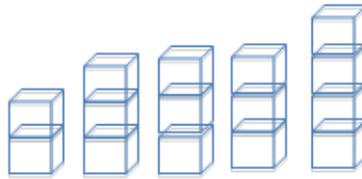
7. Robert does a fair share step by having the student with five pets share one of her pets with one of the students with two pets.

a. Draw the cubes representation that shows Robert’s fair share step.

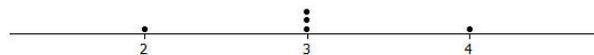
b. Draw the dot plot that shows Robert’s fair share step.

8. Robert does another fair share step by having one of the students who has four pets share one pet with one of the students who has two pets.

a. Draw the cubes representation that shows Robert’s fair share step.

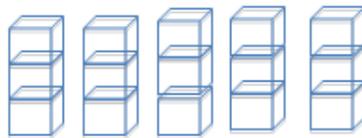


b. Draw the dot plot that shows Robert’s fair share step.



9. Robert does a final fair share step by having the student who has four pets share one pet with the student who has two pets.

a. Draw the cubes representation that shows Robert’s final fair share step.



b. Draw the dot plot representation that shows Robert’s final fair share step.



10. Explain in your own words why the final representations using cubes and a dot plot show that the mean number of pets owned by the five students is 3 pets.

The result of the sharing produces three pets each for the five students. The cube representation shows that after sharing, each student has a “fair share” of three pets. The dot plot representation should have all of the data points at the same point on the scale, the mean. In this problem, the mean number of pets is 3 for the five students, so there should be five dots above 3 on the horizontal scale.

Lesson Summary

In this lesson, you developed a method to define the center of a data distribution. The method was called the “fair share” method, and the center of a data distribution that it produced is called the mean of the data set. The reason it is called the fair share value is that if all the subjects were to have the same data value, it would be the mean value.

Mathematically the “fair share” term comes from finding the total of all of the data values and dividing the total by the number of data points. The arithmetic operation of division divides a total into equal parts.

Exit Ticket (5 minutes)

Exit Ticket Sample Solutions

1. If a class of 27 students had a mean of 72 on a test, interpret the mean of 72 in the sense of a “fair share” measure of the center of the test scores.

72 would be the test score that all 27 students would have, were all 27 students to have the same score.

2. Suppose that your school’s soccer team has scored a mean of 2 goals in each of 5 games.
 - a. Draw a representation using cubes that displays that your school’s soccer team has scored a mean of 2 goals in each of 5 games. Let one cube stand for one goal.

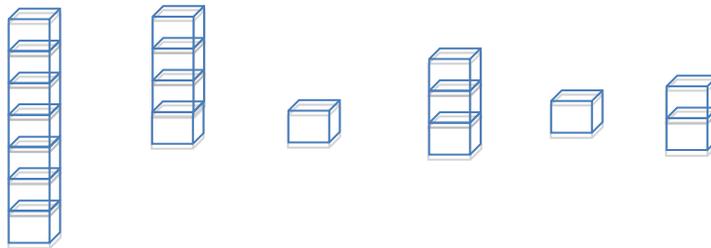


- b. Draw a dot plot that displays that your school’s soccer team has scored a mean of 2 goals in each of 5 games.

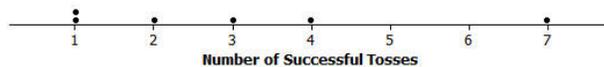


Problem Set Sample Solutions

1. A game was played where ten tennis balls are tossed into a basket from a certain distance. The number of successful tosses for six students were: 4, 1, 3, 2, 1, 7.
 - a. Draw a representation of the data using cubes where one cube represents one successful toss of a tennis ball into the basket.

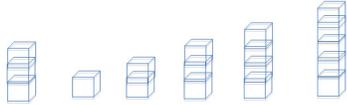
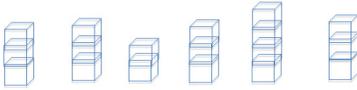


- b. Draw the original data set using a dot plot.

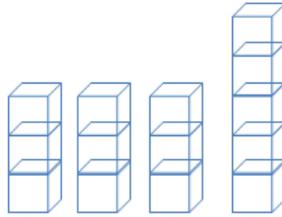


2. Find the mean number of successful tosses for this data set by Michelle’s fair share method. For each step, show the cubes representation and the corresponding dot plot. Explain each step in words in the context of the problem. You may move more than one successful toss in a step, but be sure that your explanation is clear. You must show two or more steps.

Clearly, there are several ways of getting to the final cube representation that each of the six stacks will contain three cubes. Ideally, students will move one cube at a time since for many students the leveling is seen more easily in that way. If a student shortcuts the process by moving several cubes at once, that’s okay, as long as the graphic representations are correctly done and the explanation is clear. The table provides one possible representation:

Step described in words	“Fair Share” cube representation	Dot plot
<p><i>Share two of the cubes in the 7-cube stack with one of the 1-cube stacks. The result would be: 5, 4, 3, 3, 1, 2. The 7-stack went from 7 successful tosses to 5 successful tosses, and one of the 1-stacks went from 1 successful toss to 3 successful tosses.</i></p>		
<p><i>Suppose that the student who has 5 successful tosses shares two tosses with the student who had one successful toss. The student with 5 successful tosses went down two tosses to 3 successful tosses, and the student with one successful toss went up two tosses to 3 successful tosses.</i></p>		
<p><i>Finally, the student with 4 successful tosses shares one of them with the student who has 2 successful tosses. The final step of Michelle’s fair share method shows an even number of tosses for each of the six students. So, the mean number of successful tosses for these six students is 3 tosses.</i></p>		

3. The number of pockets in the clothes worn by four students to school today is 4, 1, 3, 6. Paige produces the following cube representation as she does the fair share process. Help her decide how to finish the process of 3, 3, 3, and 5 cubes.



It should be clear to the student that there are two “extra” cubes in the stack of five cubes. Those two “extras” need to be distributed among the four students. That requires that each of the extra cubes needs to be split in half to produce four halves. Each of the four students gets half of a pocket to have a fair share mean of three and one-half pockets.

4. Suppose that the mean number of chocolate chips in 30 cookies is 14 chocolate chips.
- Interpret the mean number of chocolate chips in terms of fair share.
If each of the 30 cookies were to have the same number of chocolate chips, each would have 14 chocolate chips.
 - Describe the dot plot representation of the fair share mean of 14 chocolate chips in 30 cookies.
The dot plot consists of 30 dots stacked over the number 14 on the number line.
5. Suppose that the following are lengths (in millimeters) of radish seedlings grown in identical conditions for three days: 12 11 12 14 13 9 13 11 13 10 10 14 16 13 11.
- Find the mean length for these 15 radish seedlings.
The mean length is $12\frac{2}{15}$ mm.
 - Interpret the value from part (a) in terms of the “fair share” center length.

If each of the 15 radish seedlings were to have the same length, each would have a length of $12\frac{2}{15}$ mm.

Note: Students should realize what the cube representation for these data would look like but also realize that it may be a little cumbersome to move cubes around in the fair share process. Ideally, they would set up the initial cube representation and then use the mathematical approach of summing the lengths to be 182 mm which, when distributed evenly to 15 plants, by division would yield $12\frac{2}{15}$ mm as the fair share mean length.



Lesson 7: The Mean as a Balance Point

Student Outcomes

- Students characterize the center of a distribution by its mean in the sense of a balance point.
- Students understand that the mean is a balance point by calculating the distances of the data points from the mean and call the distances, *deviations*.
- Students understand that the mean is the value such that the sum of the deviations is equal to zero.

Lesson Notes

You may want to introduce this lesson by recalling Lessons 3 and 6. In Lesson 3, Robert gathered data from sixth grade students regarding the amount of sleep they get on school nights. He drew a dot plot of the data and decided informally on a value for the *center* of the distribution. In Lesson 6, Michelle formalized a *center* value to be the number of hours that all subjects in the sample would sleep if they all had the same number, called the mean.

MP.4 In this lesson, students will interpret the mean as a “balance point” by using a ruler and pennies to represent data. The objective of this lesson is for students to discover that if they were to draw a dot plot of the original data set that it would balance at the mean. Also, if in the process of moving data, total movement of points to the left equals the total movement of points to the right, then the balance point does not change. Therefore, it remains at the mean of the original data set.

A word of caution before beginning this lesson: Many rulers have holes in them. When data are not symmetric around 6 inch mark of a 12 inch ruler, the holes will affect the balancing at the correct value for the mean. Balancing pennies can be problematic. Students will probably have to tape on pennies (or whatever object is used). Also, if balancing on the eraser end of a pencil proves too difficult, try using a paper towel tube cut in half lengthwise (as suggested in Connected Mathematics, *Data Distributions*, Pearson).

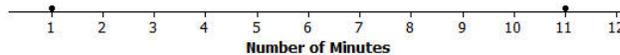
In physics, the underlying principle that pertains to the balance interpretation of mean is called *Archimedes’ Law of the Lever*. Recall that the Law states that the sum of the products of weights and their distances to the left of the balance point equals the sum of the products of weights and their distances to the right of the balance point. Our use of the Law is a special case since all of our weights (data points) are considered to be equal. Therefore, for us, the sum of the distances from the balance point to points left of the balance point equals the sum of the distances from the balance point to points right of the balance point. Moreover, the mean of the data is the value where the balance point must be to balance the lever. In statistics, deviations are calculated. A deviation is $x - \bar{x}$, where x is a data point and \bar{x} is the mean of the data. Data values to the left of the mean will have negative deviations; data values to the right of the mean will have positive deviations. The sum of all the deviations will be 0. Further, note that Archimedes’ lever has to be weightless. Clearly, a ruler is not weightless, so when students try to balance various data distributions on a ruler, the balance point may not be the exact value it should be.

Classwork

In Lesson 3, Robert gave us an informal interpretation of the center of a data set. In Lesson 6, Michelle developed a more formal interpretation of the center as a “fair share” mean, a value that every person in the data set would have if they all had the same value. In this lesson, Sabina will show us how to interpret the mean as a “balance point.”

Example 1 (7 minutes): The Mean as a Balance Point**Example 1: The Mean as a Balance Point**

Sabina wants to know how long it takes students to get to school. She asks two students how long it takes them to get to school. It takes one student 1 minute and the other student 11 minutes. Sabina represents these data on a ruler putting a penny at 1 and another at 11 and shows that the ruler balances on the eraser end of a pencil at 6. Note that the mean of 1 and 11 is also 6. Sabina thinks that there might be a connection between the mean of two data points and where they balance on a ruler. She thinks the mean may be the balancing point. What do you think? Will Sabina's ruler balance at 6? Is the mean of 1 and 11 equal to 6? Sabina shows the result on a dot plot.

Dot Plot of Number of Minutes

Sabina decides to move the penny at 1 to 4 and the other penny from 11 to 8 on the ruler, noting that the movement for the two pennies is the same distance but in opposite directions. She notices that the ruler still balances at 6. Sabina decides that if data points move the same distance but in opposite directions, the balancing point on the ruler does not change. Does this make sense? Notice that this implies that the mean of the time to get to school for two students who take 4 minutes and 8 minutes to get to school is also 6 minutes.

Sabina continues by moving the penny at 4 to 6. To keep the ruler balanced at 6, how far should Sabina move the penny at 8 and in what direction? Since the penny at 4 moved two to the right, to maintain the balance the penny at 8 needs to move two to the left. Both pennies are now at 6, and the ruler clearly balances there. Note that the mean of these two values (6 minutes and 6 minutes) is still 6 minutes.

Recall the scenarios from Lessons 3 and 6 (i.e., Robert's informal interpretation of the center value, Michelle's "fair share" mean). Now Sabina will describe the mean as a "balance point."

Read through the first paragraph as a class. Ask students the following questions and then display the ruler example.

- Will Sabina's ruler balance at 6?
 - Yes. (Now show the ruler balancing on a pencil.)
- Is 6 the mean of 1 and 11?
 - Yes

Read through the second paragraph as a class. As the scenario is being read, tell students to mark the new positions of the pennies on the dot plot provided in the example. Ask students:

- Do you think the balance point will remain at 6?
 - Yes. (Now show the ruler balancing on a pencil.)

Read through the last paragraph as a class. Note that Sabina is moving the first penny from 4 to 6. Ask students the following:

- How far is she moving the penny?
 - *2 inches.*
- How far should she move the other penny to keep the ruler in balance?
 - *2 inches. (If needed, remind students it needs to be moved in the opposite direction.)*
- If she is moving the penny from 8, where should it be placed?
 - *At 6 inches. (If students say 10 inches, again remind them they need to move the pennies in opposite directions.)*

Exercises 1–2 (7 minutes)

Let students work in pairs on Exercises 1–2.

Exercises 1–2

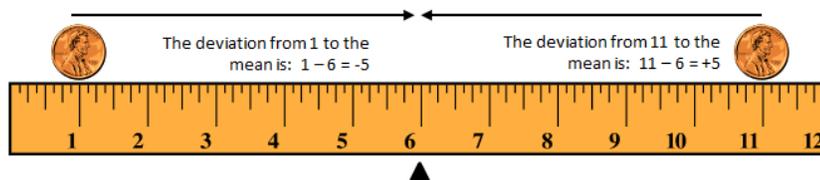
Now it is your turn to try balancing two pennies on a ruler.

1. Tape one penny at 2.5 on your ruler.
 - a. Where should a second penny be taped so that the ruler will balance at 6?
At 9.5 inches.
 - b. How far is the penny at 2.5 from 6? How far is the other penny from 6?
Each is 3.5 inches away.
 - c. Is the mean of the two locations of the pennies equal to 6?
Yes.
2. Move the penny that is at 2.5 two inches to the right.
 - a. Where will the point be placed?
At 4.5 inches.
 - b. What do you have to do with the other data point to keep the balance point at 6?
Move it 2 inches to the left.
 - c. What is the mean of the two new data points? Is it the same value as the balancing point of the ruler?
The mean is 6; it is the same.

Example 2 (5 minutes): Understanding Deviations**Example 2: Understanding Deviations**

In the above example using two pennies, it appears that the balance point of the ruler occurs at the mean location of the two pennies. We computed the distance from the balance point to each penny location and treated the distances as positive numbers. In statistics, we calculate a difference by subtracting the mean from the data point and call it the deviation of a data point from the mean. So, points to the left of the mean are less than the mean and have a negative deviation. Points to the right of the mean are greater than the mean and have a positive deviation.

Let's look at Sabina's initial placement of pennies at 1 and 11 with a mean at 6 on the graph below. Notice that the deviations are +5 and -5 . What is the sum of the deviations?



Similarly, when Sabina moved the pennies to 4 and 8, the deviation of 4 from 6 is $4 - 6 = -2$, and the deviation of 8 from 6 is $8 - 6 = +2$. Here again, the sum of the two deviations is 0, since $-2 + 2 = 0$. It appears that for two data points the mean is the point when the sum of its deviations is equal to 0.

This example introduces the very important concept of deviation of a data point x from its mean \bar{x} , namely the difference $x - \bar{x}$. Explain that a deviation is calculated by subtracting the mean *from* the data point, i.e., *deviation = data point - mean*. Students need to understand the correct order when subtracting. Be sure that students realize that data to the left of the mean will have negative deviations and those to the right will have positive deviations.

Examine the graphic with students, talk about the deviations and ask:

- What is the sum of the deviations?
 - 0

Discuss how the deviations change when Sabina moves the pennies again and ask:

- In a data distribution, what is the sum of all of the deviations?
 - 0
- What does this say about the mean of a data set regarding balance?
 - *The mean balances the sum of the positive deviations with the sum of the negative deviations, i.e., the sum of all deviations equals 0.*

Exercises 3–4 (5 minutes)

Let students work in pairs.

Exercises 3–4

Refer back to Exercise 2, where one penny was located at 2.5 and the mean was at 6.

3. Where was the second penny located?

At 9.5 inches.

4. Calculate the deviations of the two pennies and show that the sum of the deviations is 0.

The deviation of 2.5 to 6: $2.5 - 6 = -3.5$.

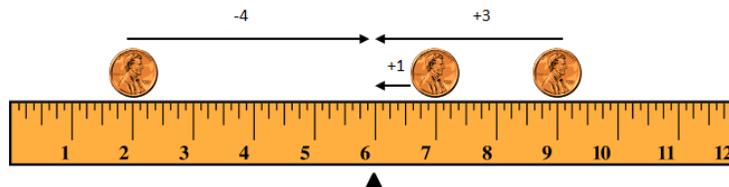
The deviation of 9.5 to 6: $9.5 - 6 = 3.5$.

The sum of $-3.5 + 3.5$ is 0.

Example 3 (5 minutes): Balancing the Mean

Example 3: Balancing the Mean

Sabina wants to know what happens if there are more than two data points. Suppose there are three students. One student lives 2 minutes from school, and another student lives 9 minutes from school. If the mean time for all three students is 6 minutes, she wonders how long it takes the third student to get to school. She tapes pennies at 2 and 9 and by experimenting finds the ruler balances with a third penny placed at 7. To check what she found, she calculates deviations.



The data point at 2 has a deviation of -4 from the mean. The data point at 7 has a deviation of $+1$ from the mean. The data point at 9 has a deviation of $+3$ from the mean. The sum of the three deviations is 0, since $-4 + 1 + 3 = 0$. So, the mean is indeed 6 minutes.

Robert says that he found out that the third penny needs to be at 7 without using his ruler. He put 2 and 9 on a dot plot. He says that the sum of the two deviations for the points at 2 and 9 is -1 , since $-4 + 3 = -1$. So, he claims that the third data point would require a deviation of $+1$ to make the sum of all three deviations equal to 0. That makes the third data point 1 minute above the mean of 6 minutes, which is 7 minutes.

This example extends the data set from containing two data points to three. The main idea is that it does not matter how many data points there are. Whether the data points are represented as pennies on a ruler or as dots on a dot plot, the mean balances the sum of the negative deviations with the sum of the positive deviations.

Read through the example as a class and study the diagram. Note that the sum of the deviations is 0. Then ask students:

- Can the concept of the mean as the balance point be extended to more than two pennies on a ruler?
 - Yes. (Try it if time permits.)
- Is the concept of the mean as the balance point true if you put multiple pennies on a single location on the ruler?
 - Yes. The balancing process is applicable to stacking pennies or having multiplicity of data points on a dot plot.

Exercises 5–7 (7 minutes)

Students should continue working in pairs. If time is running short, choose just one problem for students to attempt.

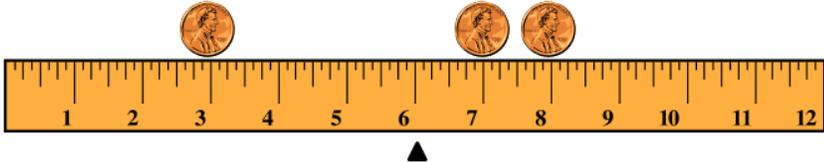
Exercises 5–7

Imagine you are balancing pennies on a ruler.

5. Suppose you place one penny each at 3, 7, and 8 on your ruler.

a. Sketch a picture of the ruler. At what value do you think the ruler will balance? Mark the balancing point with the symbol Δ .

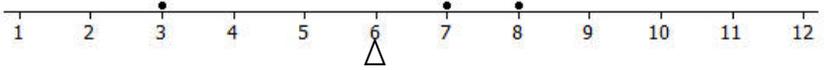
Students should represent the pennies at 3, 7, and 8 on the ruler with a balancing point at 6.



b. What is the mean of 3, 7, and 8? Does your ruler balance at the mean?

The mean is 6. Yes, it balances at the mean.

c. Show part (a) on a dot plot. Mark the balancing point with the symbol Δ .



d. What are the deviations from each of the data points to the balancing point? What is the sum of the deviations? What is the value of the mean?

The deviation of 3 to 6: $3 - 6 = -3$

The deviation of 7 to 6: $7 - 6 = +1$

The deviation of 8 to 6: $8 - 6 = +2$

The sum of the deviations ($-3 + 1 + 2$) is 0.

The mean is 6.

6. Now suppose you place a penny each at 7 and 9 on your ruler.

a. Draw a dot plot representing these two pennies.

See below.

b. Estimate where to place a third penny on your ruler so that the ruler balances at 6 and mark the point on the dot plot above. Mark the balancing point with the symbol Δ .

The third penny should be placed at 2 inches.



c. Explain why your answer in part (b) is true by calculating the deviations of the points from 6. Is the sum of the deviations equal 0?

The deviation of 2 to 6: $2 - 6 = -4$

The deviation of 7 to 6: $7 - 6 = +1$

The deviation of 9 to 6: $9 - 6 = +3$

The sum of the deviations ($-4 + 1 + 3$) is 0.

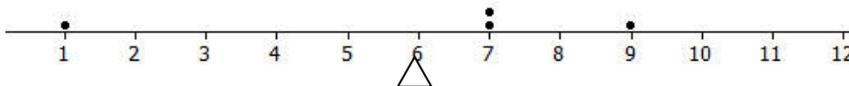
7. Suppose you place two pennies at 7 and one penny at 9 on your ruler.

a. Draw a dot plot representing these three pennies.

See below.

b. Estimate where to place a fourth penny on your ruler so that the ruler balances at 6 and mark the point on the dot plot above. Mark the balancing point with the symbol Δ .

The fourth penny should be placed at 1 inch.



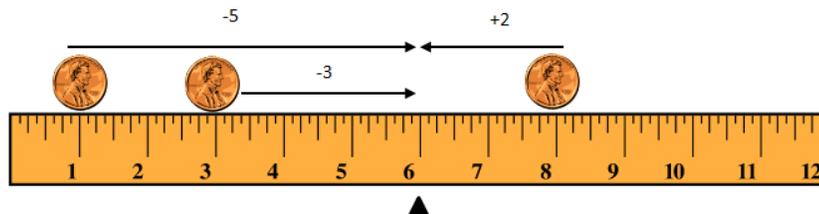
c. Explain why your answer in part (b) is true by calculating the deviations of the points from 6. Does the sum of the deviations equal 0?

The negative deviation is -5 . The positive deviations are $+1$, $+1$, and $+3$. The sum of the deviations is 0, so the mean is still 6.

Example 4 (5 minutes): Finding the Mean

Example 4: Finding the Mean

Not all data distributions on a ruler are going to have a “fair share” mean, or “balance point” of 6. What if the data were 1, 3, and 8? Will your ruler balance at 6? Why not?



Notice that the deviation of 1 from 6 is -5 . The deviation of 3 from 6 is -3 . The deviation of 8 from 6 is $+2$. The sum of the deviations is -6 , since $-5 + (-3) + 2 = -6$. The sum should be 0. Therefore, the mean is not at 6. Is the mean greater than 6 or less than 6? The sum of the deviations is negative. To decrease the negative deviations and increase the positive deviations, the balance point would have to be less than 6.

Let's see if the balance point is at 5. The deviation of 1 from 5 is -4 . The deviation of 3 from 5 is -2 . The deviation of 8 from 5 is $+3$. The sum of the three deviations is -3 , since $-4 + (-2) + 3 = -3$. That's closer to 0 than before.

Let's keep going and try 4 as the balance point. The deviation of 1 from 4 is -3 . The deviation of 3 from 4 is -1 . The deviation of 8 from 4 is $+4$. The sum of the deviations is 0, since $-3 + (-1) + 4 = 0$. The balancing point of the data distribution of 1, 3, and 8 shown on your ruler or on a dot plot is at 4. The mean of 1, 3, and 8 is 4.

This example looks at a data set whose mean is not 6. Read through the example with students. Then ask:

- Is 6 the balancing point?
 - No
- Why not?
 - *The sum of the deviations is not 0.*

If time permits, read through the remainder of the example and explain how to find the mean by using *intelligent guessing* and checking the guess by calculating the sum of deviations. Then ask:

- If you guess a value for the mean and the sum of the deviations is positive, should your next guess be lower or higher? Since the sum of the deviations is positive, the guess was too low. To decrease the positive sum, the next guess needs to be higher.

If the sum is negative, then the next guess should be lower in order to decrease the negative sum. If the sum is positive, then the next guess should be higher in order to decrease the positive sum.

Exercise 8 (7–10 minutes)

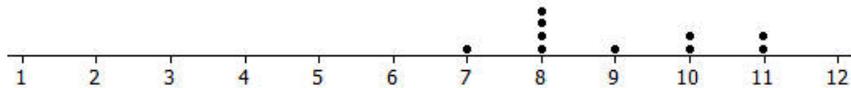
If time permits, let students work in pairs on Exercise 8.

Exercise 8

Use what you have learned about the mean to answer the following questions.

8. Recall in Lesson 6 that Michelle asked ten of her classmates for the number of hours they usually sleep when there is school the next day. Their responses (in hours) were 8, 10, 8, 8, 11, 11, 9, 8, 10, 7.

- a. It's hard to balance ten pennies. Instead of actually using pennies and a ruler, draw a dot plot that represents the data set.



- b. Use your dot plot to find the balance point. What is the sum of the deviations of the data points from the fair share mean of 9 hours?

A balance point of 9 would mean the deviations are $-2, -1, -1, -1, -1, 0, +1, +1, +2, +2$. The sum of these deviations is 0.

Note to teacher: Demonstrate how crossing out zero pairs (for example, -1 and $+1$) is a good strategy when trying to find the sum of a large number of deviations.

Lesson Summary

In this lesson, the “balance” process was developed to provide another way in which the mean characterizes the “center” of a distribution.

- The mean is the balance point of the data set when the data are shown as dots on a dot plot (or pennies on a ruler).
- The difference formed by subtracting the mean from a data point is called its deviation.
- The mean can be defined as the value that makes the sum of all deviations in a distribution equal to zero.
- The mean is the point that balances the sum of the positive deviations with the sum of the negative deviations.

Exit Ticket (5 minutes)

Name _____

Date _____

Lesson 7: The Mean as a Balance Point

Exit Ticket

1. If a class of 27 students has a mean score of 72 on a test, what is the sum of the 27 deviations of the scores from 72?

2. The dot plot below shows the number of goals scored by a school's soccer team in 7 games so far this season.



Use the “balancing” process to explain why the mean number of goals scored is 3. List all of the deviations and calculate the sum of the deviations. Explain your answer.

Exit Ticket Sample Solutions

1. If a class of 27 students has a mean score of 72 on a test, what is the sum of the 27 deviations of the scores from 72?

The sum is 0.

2. The dot plot below shows the number of goals that a school's soccer team has scored in 7 games so far this season.



Use the “balancing” process to explain why the mean number of goals scored is 3. List all of the deviations and calculate the sum of the deviations. Explain your answer.

The deviation from 0 to 3 is -3 ; from 2 to 3 is -1 ; from 5 to 3 is $+2$, for each of the two data points. The sum of the deviations is 0, since $-3 + (-1) + 2(+2) = 0$. The mean is 3.

Problem Set Sample Solutions

1. The number of pockets in the clothes worn by four students to school today is 4, 1, 3, 4.

- a. Perform the “fair share” process to find the mean number of pockets for these four students. Sketch the cube representations for each step of the process.

Each of the 4's gives up a pocket to the person with one pocket, yielding three common pockets. The mean is 3 pockets.

- b. Find the sum of the deviations to prove the mean found in part (a) is correct.

The 1-pocket data point has a deviation of -2 . Each of the two 4-pocket data points has a deviation of $+1$. So, the sum of deviations is 0.

2. The times (rounded to the nearest minute) it took each of six classmates to run a mile are 7, 9, 10, 11, 11, and 12 minutes.

- a. Draw a dot plot representation for the times. Suppose that Sabina thinks the mean is 11 minutes. Use the sum of the deviations to show Sabina that the balance point of 11 is too high.

7 has a deviation of -4 from 11; 9 has a deviation of -2 ; 10 has a deviation of -1 ; each of the 11's has a deviation of 0; 12 has a deviation of $+1$. The sum of the deviations is -6 . That indicates that 11 is too high.

- b. Sabina now thinks the mean is 9 minutes. Use the sum of the deviations to verify that 9 is too small to be the mean number of minutes.

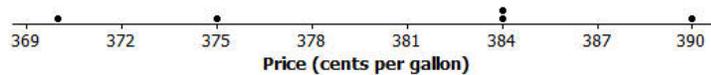
7 has a deviation of -2 from 9; 9 has a deviation of 0; 10 has a deviation of $+1$; each of the 11's has a deviation of $+2$; 12 has a deviation of $+4$. The sum of the deviations is $+7$; therefore, 9 is too low for the mean.

- c. Sabina asks you to find the mean by using the balancing process. Demonstrate that the mean is 10 minutes.

As 9 is too low, and 11 too high, try 10. The sum of the deviations is 0. So, the mean is 10 minutes.

3. The prices per gallon of gasoline (in cents) at five stations across town on one day are shown in the following dot plot. The price for a sixth station is missing, but the mean price for all six stations was reported to be 380 cents per gallon. Use the “balancing” process to determine the price of a gallon of gasoline at the sixth station?

Dot Plot of Price (cents per gallon)



The sum of the negative deviations from 380 is $(370 - 380) + (375 - 380) = -15$ cents. The sum of the positive deviations from 380 is $2(384 - 380) + (390 - 380) = +18$. So, the sixth station has to have a deviation that will cause the sum of the negative deviations plus the sum of the positive deviations to be 0. The deviation from 380 for the sixth station has to be -3 . Therefore, the price of gasoline at the sixth station must be 377 cents.

Note: Try to keep your students from using the mathematical formula for the mean to solve this problem. They could, however, use it to check the answer they get from the balancing process.

4. The number of phones (landline and cell) owned by the members of each of nine families is 3, 5, 5, 5, 6, 6, 6, 6, 8.
- a. Use the mathematical formula for the mean (sum the data points and divide by the number of data points) to find the mean number of phones owned for these nine families.

The mean is $\frac{50}{9} = 5\frac{5}{9}$ phones.

- b. Draw a dot plot of the data and verify your answer in part (a) by using the “balancing” process and finding the sum of the deviations.

The sum of the negative deviations from $5\frac{5}{9}$ is: $(3 - 5\frac{5}{9}) + 3(5 - 5\frac{5}{9}) = -4\frac{2}{9}$.

The sum of the positive deviations from $5\frac{5}{9}$ is: $4(6 - 5\frac{5}{9}) + (8 - 5\frac{5}{9}) = 4\frac{2}{9}$.

The sum of the deviations is 0, so the mean is $5\frac{5}{9}$ phones.



Lesson 8: Variability in a Data Distribution

Student Outcomes

- Students interpret the mean of a data set as a “typical” value.
- Students compare and contrast two small data sets that have the same mean but different amounts of variability.
- Students see that a data distribution is not characterized only by its center. Its spread or variability must be considered as well.
- Students informally evaluate how precise the mean is as an indicator of the typical value of a distribution, based on the variability exhibited in the data.
- Students use dot plots to order distributions according to the variability around the mean for each of the data distributions.

Classwork

Example 1 (5 minutes): Comparing Two Distributions

Example 1: Comparing Two Distributions

Robert’s family is planning to move to either New York City or San Francisco. Robert has a cousin in San Francisco and asked her how she likes living in a climate as warm as San Francisco. She replied that it doesn’t get very warm in San Francisco. He was surprised, and since temperature was one of the criteria he was going to use to form his opinion about where to move, he decided to investigate the temperature distributions for New York City and San Francisco. The table below gives average temperatures (in degrees Fahrenheit) for each month for the two cities.

City	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
New York City	39	42	50	61	71	81	85	84	76	65	55	47
San Francisco	57	60	62	63	64	67	67	68	70	69	63	58

Read through the introductory paragraph as a class. Give students a moment to examine the table and then ask:

- How would you describe the temperatures in New York City?
 - *The temperatures change a lot through the year.*
- How would you describe the temperatures in San Francisco?
 - *The temperatures do not change much.*

Exercises 1–2 (5 minutes)

Let students work independently and confirm their answer with a neighbor. Encourage calculator use when working with larger data sets.

Exercises 1–2

Use the table above to answer the following:

1. Calculate the annual mean monthly temperature for each city.

New York City: 63 degrees

San Francisco: 64 degrees

2. Recall that Robert is trying to decide to which city he wants to move. What is your advice to him based on comparing the overall annual mean monthly temperatures of the two cities?

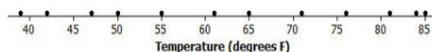
Since the means are almost the same, it looks like Robert could move to either city.

Example 2 (5 minutes): Understanding Variability**Example 2: Understanding Variability**

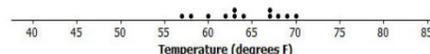
In Exercise 2, you found the overall mean monthly temperatures in both the New York City distribution and the San Francisco distribution to be about the same. That didn't help Robert very much in making a decision between the two cities. Since the mean monthly temperatures are about the same, should Robert just toss a coin to make his decision? Is there anything else Robert could look at in comparing the two distributions?

Variability was introduced in an earlier lesson. Variability is used in statistics to describe how spread out the data in a distribution are from some focal point in the distribution (such as the mean). Maybe Robert should look at how spread out the New York City monthly temperature data are from its mean and how spread out the San Francisco monthly temperature data are from its mean. To compare the variability of monthly temperatures between the two cities, it may be helpful to look at dot plots. The dot plots for the monthly temperature distributions for New York City and San Francisco follow.

Dot Plot of Temperature for New York City



Dot Plot of Temperature for San Francisco



Read through the first paragraph as a class. Since the means are about the same, it would be helpful if Robert had more information as basis for a decision. He needs to go beyond comparing centers to incorporating variability into his decision-making process. Ask students:

MP.1

- Should he just toss a coin to make a decision?
 - *Answers will vary.*
- What else do you think Robert could use to make a decision?
 - *He could consider the range or variety of temperatures in each city.*

Read though the second paragraph (above) and define variability. In this example, we want students to become familiar with the concept of variability by viewing how spread out the data are from their mean in dot plots. Give students a moment to examine the dot plots and ask:

- How are the two dot plots different?
 - *The temperatures for New York City are spread out, while the temperatures for San Francisco are clustered together.*

Exercises 3–7 (5–7 minutes)

Let students work independently and compare answers with a neighbor.

Exercises 3–7

Use the dot plots above to answer the following:

3. Mark the location of the mean on each distribution with the balancing Δ symbol. How do the two distributions compare based on their means?

Place Δ at 63 for New York City and at 64 for San Francisco. The means are about the same.

4. Describe the variability of the New York City monthly temperatures from the mean of the New York City temperatures.

The temperatures are widespread around the mean. From a low of around 40, to a high of 85.

5. Describe the variability of the San Francisco monthly temperatures from the mean of the San Francisco monthly temperatures.

The temperatures are compact around the mean. From a low of 57, to a high of 70.

6. Compare the amount of variability in the two distributions. Is the variability about the same, or is it different? If different, which monthly temperature distribution has more variability? Explain.

The variability is different. The variability in New York City is much greater compared to San Francisco.

7. If Robert prefers to choose the city where the temperatures vary the least from month to month, which city should he choose? Explain.

He should choose San Francisco because the temperatures vary the least, from a low of 57 to a high of 70. New York City has temperatures with more variability, from a low of 40, to a high of 85.

Example 3 (7–9 minutes): Using Mean and Variability in a Data Distribution**Example 3: Using Mean and Variability in a Data Distribution**

The mean is used to describe the “typical” value for the entire distribution. Sabina asks Robert which city he thinks has the better climate? He responds that they both have about the same mean, but that the mean is a better measure or a more precise measure of a typical monthly temperature for San Francisco than it is for New York City. She’s confused and asks him to explain what he means by this statement.

Robert says that the mean of 63 degrees in New York City (64 in San Francisco) can be interpreted as the typical temperature for any month in the distributions. So, 63 or 64 degrees should represent all of the months’ temperatures fairly closely. However, the temperatures in New York City in the winter months are in the 40s and in the summer months are in the 80s. The mean of 63 isn’t too close to those temperatures. Therefore, the mean is not a good indicator of typical monthly temperature. The mean is a much better indicator of the typical monthly temperature in San Francisco because the variability of the temperatures there is much smaller.

MP.3

The concept in this example may be challenging for some students. When Robert talks about the precision of the mean, Sabina asks him to explain what he means by a mean being precise.

Although the means are about the same for the two distributions, Robert is suggesting that the mean of 64 degrees for San Francisco is a better indicator of the city’s typical monthly temperature, than the mean of 63 degrees is as an indicator of a typical monthly temperature in New York City. He bases this on the variability of the monthly temperatures in each city. He says that a mean is a only precise indicator of monthly temperatures if the variability in the data is very low. The higher the variability gets, the less precise the mean is as an indicator of typical monthly temperatures.

If there is still confusion, draw two dot plots similar to Example 3 on the board and ask the following:

- Which dot plot has greater variability?
- If data points have a lot of variability, is the mean a good indicator of a “typical” value in the data set?
 - No.
- If the data points are clustered around the mean, is the mean a good indicator of a “typical” value in the data set?
 - Yes.

Exercises 8–11 (5 minutes)

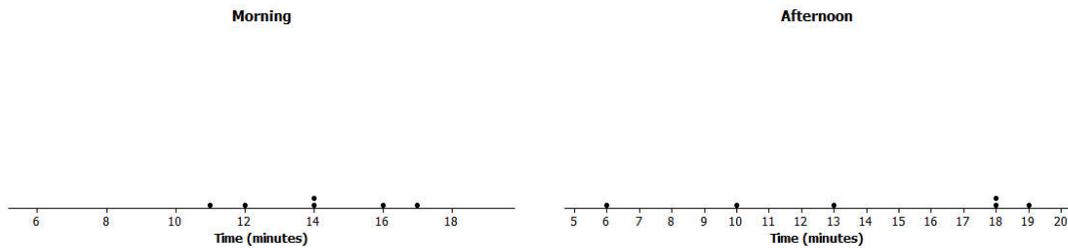
Let students work independently and confirm their answer with a neighbor.

Exercises 8–14

Consider the following two distributions of times it takes six students to get to school in the morning and to go home from school in the afternoon.

	Time (minutes)					
Morning	11	12	14	14	16	17
Afternoon	6	10	13	18	18	19

8. To visualize the means and variability, draw dot plots for each of the two distributions.



9. What is the mean time to get from home to school in the morning for these six students?

The mean is 14 minutes. (Note: It is visible from the graphs.)

10. What is the mean time to get from school to home in the afternoon for these six students?

The mean is 14 minutes. (Note: The sum of the negative deviations is -13 , and the sum of the positive deviations is $+13$.)

11. For which distribution does the mean give a more precise indicator of a typical value? Explain your answer.

The morning mean is a more precise indicator. The spread of the afternoon data is far greater around the mean.

Exercises 12–14 (7 minutes)

Let students work in pairs or small groups. If time allows, discuss Exercise 13 as a class.

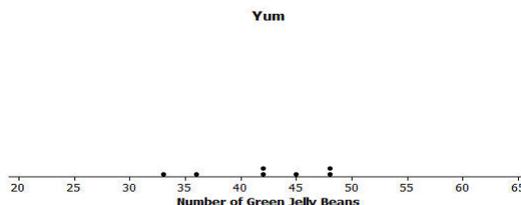
Distributions can be ordered according to how much the data values vary around their means.

Consider the following data on the number of green jellybeans in seven bags of jellybeans from each of five different candy manufacturers (AllGood, Best, Delight, Sweet, Yum). The mean in each distribution is 42 green jellybeans.

	1	2	3	4	5	6	7
AllGood	40	40	41	42	42	43	46
Best	22	31	36	42	48	53	62
Delight	26	36	40	43	47	50	52
Sweet	36	39	42	42	42	44	49
Yum	33	36	42	42	45	48	48

12. Draw a dot plot of the distribution of number of green jellybeans for each of the five candy makers. Mark the location of the mean on each distribution with the balancing Δ symbol.

The dot plots should each have a balancing Δ symbol located at 42.





Name _____

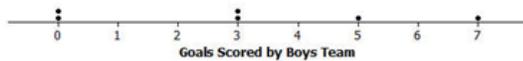
Date _____

Lesson 8: Variability in a Data Distribution

Exit Ticket

- Consider the following statement: Two sets of data with the same mean will also have the same variability. Do you agree or disagree with this statement? Explain.

- Suppose the dot plot on the left shows the number of goals a boys' soccer team has scored in 6 games so far this season, and the dot plot on the right shows the number of goals a girls' soccer team has scored in 6 games so far this season.



- Compute the mean number of goals for each distribution.

- For which distribution, if either, would the mean be considered a better indicator of a typical value? Explain your answer.

Exit Ticket Sample Solutions

1. Consider the following statement: Two sets of data with the same mean will also have the same variability. Do you agree or disagree with this statement? Explain.

Students should disagree with this statement. There were many examples in this lesson that could be used as the basis for an explanation.

2. Suppose the dot plot on the left shows the number of goals a boys' soccer team has scored in 6 games so far this season, and the dot plot on the right shows the number of goals a girls' soccer team has scored in 6 games so far this season.



- a. Compute the mean number of goals for each distribution.

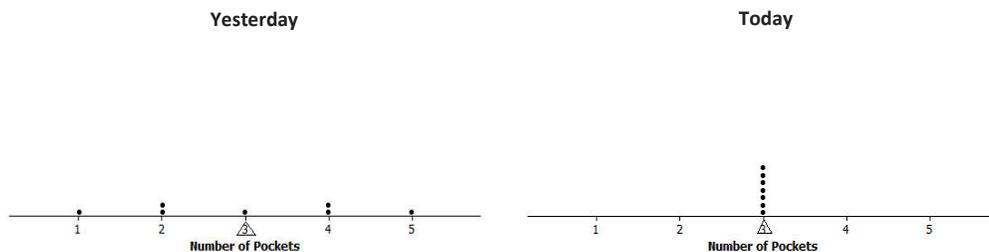
The mean for each is 3 goals. If your students found the mean by the formula, have them verify the answer by summing the negative and positive deviations.

- b. For which distribution, if either, would the mean be considered a better indicator of a typical value? Explain your answer.

Variability in the girls' distribution is less than in the boys', so the mean of 3 goals for the girls' is more precise.

Problem Set Sample Solutions

1. The number of pockets in the clothes worn by seven students to school yesterday were 4, 1, 3, 4, 2, 2, 5. Today those seven students each had three pockets in their clothes.
 - a. Draw one dot plot for what the students wore yesterday, and another dot plot for what the students wore today. Be sure to use the same scales. Show the means by using the balancing Δ symbol.



- b. For each distribution, find the mean number of pockets worn by the seven students.

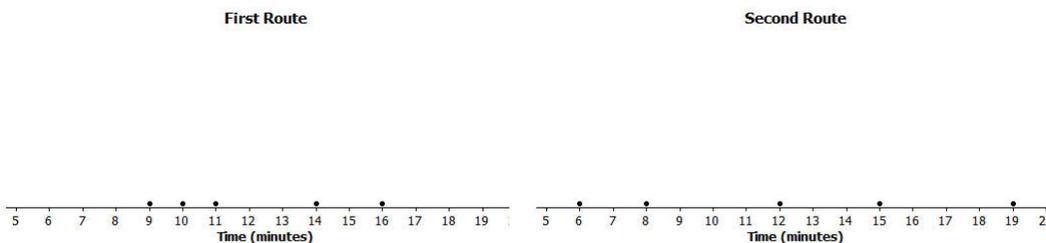
Students should not need to use the formula to calculate the means for either of these distributions. The first is clearly balanced around 3 pockets, and the second only has 3 as its data value.

- c. For which distribution is the mean number of pockets a better indicator of what is “typical?” Explain.

There is certainly variability in the data for the yesterday’s distribution, whereas today’s distribution has none. The mean of 3 pockets is a better indicator (more precise) for today’s distribution.

2. The number of minutes (rounded) it took to run a certain short cross-country route was recorded for each of five students. The resulting data were 9, 10, 11, 14, and 16 minutes. The number of minutes (rounded to the nearest minute) it took the five students to run a different cross-country route was also recorded, resulting in the following data: 6, 8, 12, 15, and 19 minutes.

- a. Draw dot plots for the two distributions of the time it takes to run a cross-country route. Be sure to use the same scale on both dot plots.



- b. Do the distributions have the same mean?

Yes, both distributions have the same mean, 12 minutes.

- c. In which distribution is the mean a better indicator of the typical amount of time taken to run its cross-country route? Explain.

Looking at the dot plots, the times completing the second route are more varied than those in the first route. So, the mean in the first route is a better indicator (more precise) of a typical value.

3. The following table shows the prices per gallon of gasoline (in cents) at five stations across town as recorded on Monday, Wednesday, and Friday of a certain week.

Day	R&C	Al’s	PB	Sam’s	Ann’s
Monday	359	358	362	359	362
Wednesday	357	365	364	354	360
Friday	350	350	360	370	370

- a. The mean price per day over the five stations is the same for the three days. Without doing any calculation and simply looking at Friday’s prices, what must the mean price be?

Friday’s prices are symmetric around 360 cents. So, the mean is 360 cents.

- b. In which daily distribution is its mean a better indicator of the typical price per gallon for the five stations? Explain.

Note that the data are not in numerical order across the stations for Monday and Wednesday prices. So, encourage students to draw dot plots to help them answer this question. From the dot plots, the mean for Monday is the most precise (least variability), and the mean for Friday is the least precise (most variability).



Lesson 9: The Mean Absolute Deviation (MAD)

Student Outcomes

- Students calculate the mean absolute deviation (MAD) for a given data set.
- Students interpret the MAD as the average distance of data values from the mean.

Lesson Notes

Variability was discussed informally in Lesson 8. This lesson focuses on developing a more formal measure of variability in a data distribution called the mean absolute deviation, denoted by MAD. The concept of deviation from the mean should be clear to students by now, since previous lessons used deviations to develop the idea of the mean as a balance point. This lesson challenges students to answer why absolute values of deviations are used in calculating the MAD.

Mean absolute deviation is the measure of variability used in the middle school curriculum. At the high school level, deviations are squared instead of using the absolute value. This leads to other important measures of variability called the variance and the standard deviation.

Classwork

Example 1 (5 minutes): Variability

Example 1: Variability

In Lesson 8, Robert tried to decide to which of two cities he would rather move, based on comparing their mean annual temperatures. Since the mean yearly temperature for New York City and San Francisco turned out to be about the same, he decided instead to compare the cities based on the variability in their monthly temperatures from the overall mean. He looked at the two distributions and decided that the New York City temperatures were more spread out from their mean than were the San Francisco temperatures from their mean.

Read through Example 1 as a class, and recall the main idea of Lesson 8. Then ask students:

- What is variability?
 - *The spread of data in a distribution from some focal point in the distribution (such as the mean).*
- What does a distribution that has no variability look like?
 - *All of the data points are the same.*
- What does a distribution that has a lot of variability look like?
 - *The data points are spread far apart.*

Suggest a visual way to order several data sets from the one with least variability to the one with most variability.

Exercises 1–3 (7–10 minutes)

Let students work in small groups on this exercise. Then confirm answers as a class. The discussion of Exercise 3 leads into Example 2.

Students are asked to order the seven data sets from least variability to most variability. Students will no doubt suggest different orderings. Several orderings are reasonable – focus on the students’ explanations for ordering the distributions. What is important is not their suggested orderings but rather their arguments to support their orderings. Also, the goal for this example is for students to realize that they need to have a more formal way of deciding the best ordering. Sabina suggests that a formula is needed, and she proceeds in this lesson to develop one.

MP.3

Exercises 1–3

The following temperature distributions for seven other cities all have a mean temperature of approximately 63 degrees. They do not have the same variability. Consider the following dot plots of the mean yearly temperatures of the seven cities in degrees Fahrenheit.

<p>City A</p>	<p>City B</p>	<p>City C</p>
<p>City D</p>	<p>City E</p>	<p>City F</p>
<p>City G</p>		

- Which distribution has the smallest variability of the temperatures from its mean of 63 degrees? Explain your answer.
City A, because all points are the same.
- Which distribution(s) seems to have the most variability of the temperatures from the mean of 63 degrees? Explain your answer.
One or more of the following is acceptable: Cities D, E, and F. They appear to have data points spread furthest from the mean.

3. Order the seven distributions from least variability to most variability. Explain why you listed the distributions in the order that you chose.

Several orderings are reasonable. Focus on students' explanations for choosing the order.

Example 2 (5 minutes): Measuring Variability

Example 2: Measuring Variability

Based on just looking at the distributions, there are different orderings of variability that seem to make some sense. Sabina is interested in developing a formula that will give a number that measures the variability in a data distribution. She would then use the formula for each data set and order the distributions from lowest to highest. She remembers from a previous lesson that a deviation is found by subtracting the mean from a data point. The formula was summarized as: $\text{deviation} = \text{data point} - \text{mean}$. Using deviations to develop a formula measuring variability is a good idea to consider.

No doubt students had different orderings of variability for the seven cities in Exercise 2. Sabina suggests that, in this example, a formula is needed to give a formal ordering. Since variability is being viewed from the mean, it seems reasonable that a formula should be based on how far data points are from the mean. Recall that a deviation results from subtracting the mean from a data point, *or deviation = data point – mean*. She concludes that it seems to be a good idea to use deviations in developing a formula for a measure of variability.

Ask students:

- Do you think using deviations is a good basis for a formula to measure variability?
 - *Yes. A deviation measures how far a data point is from the mean of its distribution. That certainly addresses the concept of variability.*
- When are deviations negative?
 - *When they are located to the left of the mean.*
- When are deviations positive?
 - *When they are located to the right of the mean.*

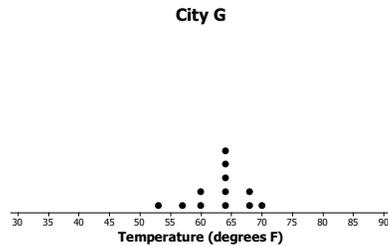
Exercises 4–6 (5 minutes)

Let students work in pairs.

In this exercise, City *G* is used to focus on calculating deviations and verifying that the sum of deviations is equal to zero. This means summing deviations is not a good measure of variability because it always turns out to be zero (by the development of the mean as a balance). A graph is drawn of City *G* to illustrate the values of the deviations.

Exercises 4–6

The dot plot for the temperatures in City G is shown below. Use the dot plot and the mean temperature of 63 degrees to answer the following questions.



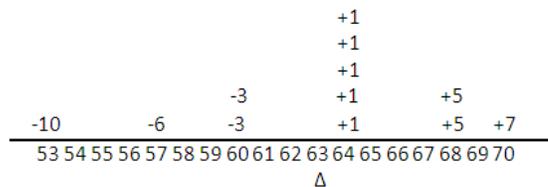
4. Fill in the following table for City G temperature deviations.

Temp	Deviation	Result
53	$53 - 63$	-10
57	$57 - 63$	-6
60	$60 - 63$	-3
60	$60 - 63$	-3
64	$64 - 63$	+1
64	$64 - 63$	+1
64	$64 - 63$	+1
64	$64 - 63$	+1
68	$68 - 63$	+5
68	$68 - 63$	+5
70	$70 - 63$	+7
Sum		0

5. Why should the sum of your deviations column be equal to zero? (Hint: Recall the balance interpretation of the mean of a data set.)

The mean is the value that makes the sum of the positive and negative deviations 0. It is the balance point.

6. Another way to graph the deviations is to write them on a number line as follows. What is the sum of the positive deviations (the deviations to the right of the mean)? What is the sum of the negative deviations (the deviations to the left of the mean)? What is the total sum of the deviations?



Sum of the positive deviations = +22

Sum of the negative deviations = -22

Sum of all of the deviations = 0

Example 3 (5–7 minutes): Finding the Mean Absolute Deviation (MAD)

Example 3: Finding the Mean Absolute Deviation (MAD)

By the balance interpretation of the mean, the sum of the deviations for any data set will always be zero. Sabina is disappointed that her idea of developing a measure of variability using deviations isn't working. She still likes the concept of using deviations to measure variability, but the problem is that the sum of the positive deviations is cancelling out the sum of the negative deviations. What would you suggest she do to keep the deviations as the basis for a formula but to avoid the deviations cancelling out each other?

MP.1

This example asks students how they could still use deviations in developing a measure of variability but correct the difficulty of having the negative deviations offset the positive deviations when the deviations are summed. The operation of absolute value should come to mind because it is part of what students have previously studied in mathematics. The example leads them through the calculation of the deviations and then to taking the mean of the absolute deviations.

Ask students:

- If we were to treat the negative deviations as distances, what mathematical operation would do that?
 - *Finding the absolute value of a negative deviation makes it a positive distance. Emphasize that the concept of deviation has been maintained.*
- If we use the absolute value of all the deviations, what will happen to the sum?
 - *It will not be zero.*

Define the mean absolute deviation as the sum of the absolute values of the deviations divided by the number of deviations. (See Exercise 7, part (c).) Teachers may wish to work through Exercise 7, parts (a)–(c) as a class to develop this concept.

Exercises 7–8 (10–12 minutes)

Let students continue to work in pairs or small groups. As previously indicated, teachers may decide to work through Exercise 7, parts (a)–(c) as a class.

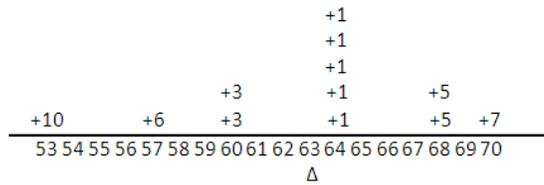
Exercises 7–8

7. One suggestion to possibly help Sabina is to take the absolute value of the deviations.

a. Fill in the following table.

Temp	Deviation	Result	Abs
53	53 – 63	–10	+10
57	57 – 63	–6	+6
60	60 – 63	–3	+3
60	60 – 63	–3	+3
64	64 – 63	+1	+1
64	64 – 63	+1	+1
64	64 – 63	+1	+1
64	64 – 63	+1	+1
64	64 – 63	+1	+1
68	68 – 63	+5	+5
68	68 – 63	+5	+5
70	70 – 63	+7	+7

b. From the following graph, what is the sum of the absolute deviations?



The sum of the absolute deviations is +44.

c. Sabina suggests that the mean of the absolute deviations could be a measure of the variability in a data set. Its value is the average distance that all the data values are from the mean temperature. It is called the Mean Absolute Deviation and is denoted by the letters, MAD. Find the MAD for this data set of City G temperatures. Round to the nearest tenth.

The mean absolute deviation is $\frac{44}{12}$ or 3.7 degrees to the nearest tenth of a degree.

d. Find the MAD for each of the temperature distributions in all seven cities, and use the values to order the distributions from least variability to most variability. Recall that the mean for each data set is 63 degrees. Does the list that you made in Exercise 2 by just looking at the distributions match this list made by ordering MAD values?

**If time is a factor in completing this lesson, assign a city to individual students. After each student has calculated the mean deviation, organize results for the whole class. Direct students to calculate the MAD to the nearest tenth of a degree.*

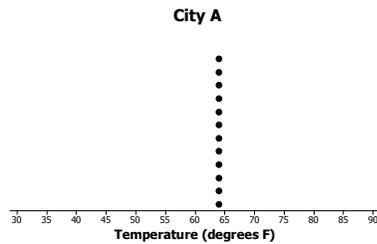
MAD values:		
City A = 0	E = 10.5	The order from least to greatest is: A, G, and B, C (tied), and D, E, and F (all tied).
City B = 5.3	F = 10.5	
City C = 5.3	G = 3.7	
City D = 10.5		

e. Which of the following is a correct interpretation of the MAD?

- i. The monthly temperatures in City G are spread 3.7 degrees from the approximate mean of 63 degrees.
- ii. The monthly temperatures in City G are, on average, 3.7 degrees from the approximate mean temperature of 63 degrees.
- iii. The monthly temperatures in City G differ from the approximate mean temperature of 63 degrees by 3.7 degrees.

Answer is (ii).

8. The dot plot for City A temperatures follows.



- a. How much variability is there in City A's temperatures? Why?

No variability. The deviations are all 0.

- b. Does the MAD agree with your answer in part (a)?

Yes. The mean absolute deviation is 0.

Lesson Summary

In this lesson, a formula was developed that measures the amount of variability in a data distribution.

- The absolute deviation of a data point is how far away that data point is from the mean.
- The Mean Absolute Deviation (MAD) is computed by finding the mean of the absolute deviations in the distribution.
- The value of MAD is the average distance that all the data values are from the mean.
- A small MAD indicates that the distribution has very little variability.
- A large MAD indicates that the data points are spread far away from the mean.

Exit Ticket (5 minutes)



Name _____

Date _____

Lesson 9: The Mean Absolute Deviation (MAD)

Exit Ticket

- The Mean Absolute Deviation (MAD) is a measure of variability for a data set. What does a data distribution look like if its MAD equals zero? Explain.
- Is it possible to have a negative value for the MAD of a data set?
- Suppose that seven students have the following number of pets: 1, 1, 1, 2, 4, 4, 8.
 - The mean number of pets for these seven students is three pets. Use the following table to find the MAD number of pets for this distribution of number of pets.

Student	# of Pets	Deviations	Absolute Deviations
1	1		
2	1		
3	1		
4	2		
5	4		
6	4		
7	8		
Sum			

- Explain in words what the MAD means for this data set.

Exit Ticket Sample Solutions

- The Mean Absolute Deviation (MAD) is a measure of variability for a data set. What does a data distribution look like if its MAD equals zero? Explain.

If the MAD is zero, then all of the deviations are zero. For example, City A had a dot plot with all the same temperatures. They were all the same, so there was no variability, since the MAD measures average temperature from the mean. And it is zero, because all the deviations are zero.

- Is it possible to have a negative value for the MAD of a data set?

Because a MAD is the average of the absolute values of the deviations, it is always a positive value.

- Suppose that seven students have the following number of pets: 1, 1, 1, 2, 4, 4, 8.

- The mean number of pets for these seven students is three pets. Use the following table to find the MAD number of pets for this distribution of number of pets.

The MAD number of pets is $+\frac{14}{7} = 2$ pets.

Student	# of Pets	Deviations	Absolute Deviations
1	1	$1 - 3 = -2$	+2
2	1	$1 - 3 = -2$	+2
3	1	$1 - 3 = -2$	+2
4	2	$2 - 3 = -1$	+1
5	4	$4 - 3 = +1$	+1
6	4	$4 - 3 = +1$	+1
7	8	$8 - 3 = +5$	+5
Sum		0	+14

- Explain in words what the MAD means for this data set.

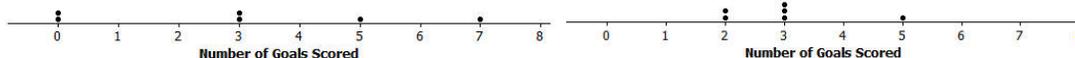
On average, these students' number of pets differ by 2 pets from their mean of 3 pets.

Problem Set Sample Solutions

- Suppose the dot plot on the left shows the number of goals a boys' soccer team has scored in six games so far this season, and the dot plot on the right shows the number of goals a girls' soccer team has scored in six games so far this season. The mean for both of these teams is 3.

Dot Plot of Number of Goals Scored by Boys' Team

Dot Plot of Number of Goals Scored by Girls' Team FINAL -



- Before doing any calculations, which dot plot has the larger MAD? Explain how you know.

The graph of the Boys' team is more spread out and has the larger deviations from the mean.

- b. Use the following tables to find the MAD number of goals for each distribution. Round your calculations to the nearest hundredth.

Boys' Team		
#Goals	Deviations	Absolute Deviations
0	-3	+3
0	-3	+3
3	3 - 3	0
3	3 - 3 = 0	0
5	5 - 3 = 2	+2
7	7 - 3 = 4	+4
Sum		+12

Girls' Team		
#Goals	Deviations	Absolute Deviations
2	2 - 3 = -1	+1
2	2 - 3 = -1	+1
3	3 - 3 = 0	0
3	3 - 3 = 0	0
3	3 - 3 = 0	0
5	5 - 3 = 2	+2
Sum		+4

- c. Based on the computed MAD values, for which distribution is the mean a better indication of a typical value? Explain your answer.

The Girls' team, because the measure of variability given by the MAD is lower (0.67 goals) than the Boys' MAD (2 goals). Visually, the data in the Girls' dot plot are more compact around the mean than they are in the Boys' dot plot.

2. Recall Robert's problem of deciding whether to move to New York City or to San Francisco. The table of temperatures (in degrees Fahrenheit) and deviations for the New York City distribution is as follows:

NYC	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Temp	39	42	50	61	71	81	85	84	76	65	55	47
Deviation	-24	-21	-13	-2	8	18	22	21	13	2	-8	-16

- a. The dot plot below is written with the deviations above each of the monthly temperatures. What is the sum of all of the deviations? Are you surprised? Explain.

The sum of the negative deviations is: $-24 + -21 + -16 + -13 + -8 + -2 = -84$. The sum of the positive deviations is: $2 + 8 + 13 + 18 + 21 + 22 = +84$. The sum of all of the deviations is $-84 + 84 = 0$. Students should not be surprised, since the sum of the deviations of any data set around its mean is 0.

-24	-21	-16	-13	-8	-2	2	8	13	18	21	22
39	42	47	50	55	61	65	71	76	81	84	85

- b. The absolute deviations for the monthly temperatures are shown below. Use this information to calculate the MAD. Explain the MAD in words for this problem.

The sum of the absolute deviations is $2(84) = 168$ degrees. The average of the absolute deviations, MAD, is $\frac{168}{12} = 14$ degrees. On average, the monthly temperatures in New York City differ from the mean of 63 degrees by 14 degrees.

+24	+21	+16	+13	+8	+2	2	8	13	18	21	22
39	42	47	50	55	61	65	71	76	81	84	85

- c. Complete the following table and then use the values to calculate the MAD for the San Francisco data distribution.

First of all, note that the sum of the negative deviations is -21 , and the sum of the positive deviations is $+21$, as it should be. The sum of the absolute deviations is $2(21) = 42$. The MAD is the mean of the absolute deviations, which equals $\frac{42}{12} = 3.5$ degrees.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Temp	57	60	62	63	64	67	67	68	70	69	63	58
Deviations	-7	-4	-2	-1	0	+3	+3	+4	+6	+5	-1	-6
Absolute Deviations	+7	+4	+2	+1	0	+3	+3	+4	+6	+5	+1	+6

- d. Comparing the MAD values for New York City and San Francisco, which city would Robert choose to move to if he is interested in having a lot of variability in monthly temperatures? Explain using the MAD.

New York City has a MAD of 14 degrees, as compared to 3.5 degrees in San Francisco. Robert should choose New York City if he wants to have more variability in monthly temperatures.

3. Consider the following data of the number of green jellybeans in seven bags sampled from five different candy manufacturers (Awesome, Delight, Finest, Sweeties, YumYum). Note that the mean in each distribution is 42 green jellybeans.

	Bag 1	Bag 2	Bag 3	Bag 4	Bag 5	Bag 6	Bag 7
Awesome	40	40	41	42	42	43	46
Delight	22	31	36	42	48	53	62
Finest	26	36	40	43	47	50	52
Sweeties	36	39	42	42	42	44	49
YumYum	33	36	42	42	45	48	48

- a. Complete the following table of the deviations of the number of green jellybeans from the mean number of green jellybeans in the seven bags.

	Bag 1	Bag 2	Bag 3	Bag 4	Bag 5	Bag 6	Bag 7
Awesome	-2	-2	-1	0	0	+1	+4
Delight	-20	-11	-6	0	+6	+11	+20
Finest	-16	-6	-2	+1	+5	+8	+10
Sweeties	-6	-3	0	0	0	+2	+7
YumYum	-9	-6	0	0	+3	+6	+6

- b. Based on what you learned about MAD, which manufacturer do you think will have the lowest MAD? Calculate the MAD for the manufacturer you selected.

Use the MAD for each manufacturer to evaluate students' responses.

	Bag 1	Bag 2	Bag 3	Bag 4	Bag 5	Bag 6	Bag 7	SUM	MAD
Awesome	+2	+2	+1	0	0	+1	+4	+10	1.4
Delight	+20	+11	+6	0	+6	+11	+20	+74	10.6
Finest	+16	+6	+2	+1	+5	+8	+10	+48	6.9
Sweeties	+6	+3	0	0	0	+2	+7	+18	2.6
YumYum	+9	+6	0	0	+3	+6	+6	+30	4.3



Lesson 10: Describing Distributions Using the Mean and MAD

Student Outcomes

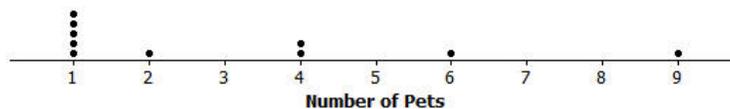
- Students calculate the mean and MAD for a data distribution.
- Students use the mean and MAD to describe a data distribution in terms of center and variability.

Classwork

Example 1 (8 minutes): Describing Distributions

Example 1: Describing Distributions

In Lesson 9, Sabina developed the mean absolute deviation (MAD) as a number that measures variability in a data distribution. Using the mean and MAD with a dot plot allows you to describe the center, spread, and shape of a data distribution. For example, suppose that data on the number of pets for ten students is shown in the dot plot below.



There are several ways to describe the data distribution. The mean number of pets each student has is three, which is a measure of center. There is variability in the number of pets the students have, which is an average of 2.2 pets from the mean (the MAD). The shape of the distribution is heavy on the left and it thins out to the right.

Introduce the data set and explain that distributions can be described by their center, spread, and shape. Note that the mean is 3 pets and the MAD is 2.2 pets. The shape is described as well.

MP.4

In your discussion, you want your students to begin to conceptualize the measures. Have them draw a triangle over the 3 on the number line and see that the distribution is balanced there with the sum of negative deviations (-11) balancing the sum of positive deviations ($+11$). Then ask:

- Without extensive calculating, how is the MAD 2.2 pets?
 - *The total distance the data are from the mean is $2(11) = 22$, and the mean of the absolute deviations is $\frac{22}{10} = 2.2$ pets.*

Exercises 1–4 (9–12 minutes)

Let students work with a partner. Then discuss and confirm answers to Exercises 1–3.

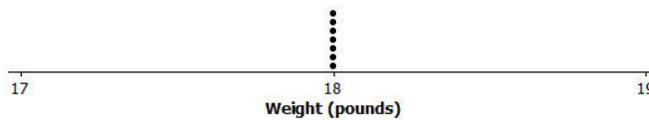
Exercises 1–4

1. Suppose that the weights of seven middle-school students' backpacks are given below.

a. Fill in the following table.

Student	Alan	Beth	Char	Damon	Elisha	Fred	Georgia
Weight (lbs.)	18	18	18	18	18	18	18
Deviations	0	0	0	0	0	0	0
Absolute Deviations	0	0	0	0	0	0	0

b. Draw a dot plot for these data and calculate the mean and MAD.



The mean is 18 pounds.

The MAD is 0.

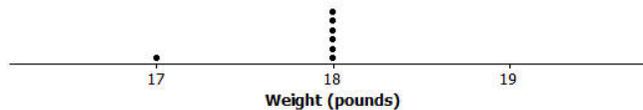
c. Describe this distribution of weights of backpacks by discussing the center, spread, and shape.

The mean is 18 pounds. There is no variability.

All of the data is centered.

2. Suppose that the weight of Elisha's backpack is 17 pounds, rather than 18.

a. Draw a dot plot for the new distribution.



b. Without doing any calculation, how is the mean affected by the lighter weight? Would the new mean be the same, smaller, or larger?

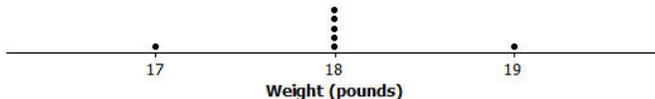
The mean will be smaller because the new point is smaller.

c. Without doing any calculation, how is the MAD affected by the lighter weight? Would the new MAD be the same, smaller, or larger?

Now there is variability, so the MAD is greater than zero.

3. Suppose that in addition to Elisha’s backpack weight having changed from 18 to 17 lb., Fred’s backpack weight is changed from 18 to 19 lb.

a. Draw a dot plot for the new distribution.



b. Without doing any calculation, what would be the value of the new mean compared to the original mean?

The mean is 18 lbs.

c. Without doing any calculation, would the MAD for the new distribution be the same, smaller, or larger than the original MAD?

Since there is more variability, the MAD is larger.

d. Without doing any calculation, how would the MAD for the new distribution compare to the one in Exercise 2?

There is more variability, so the MAD is greater than in Exercise 2.

4. Suppose that seven second-graders’ backpack weights were:

Student	Alice	Bob	Carol	Damon	Ed	Felipe	Gale
Weight (lbs.)	5	5	5	5	5	5	5

a. How is the distribution of backpack weights for the second-graders similar to the original distribution for sixth-graders given in Exercise 1?

Both have no variability, so the MAD is zero. The dot plots look the same.

b. How are the distributions different?

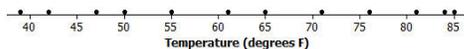
The means are different. One mean is 18 and the other is 5.

Example 2 (5 minutes): Using the Mean Versus the MAD

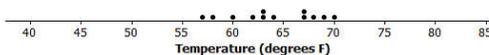
Example 2: Using the Mean Versus the MAD

Decision-making by comparing distributions is an important function of statistics. Recall that Robert is trying to decide whether to move to New York City or to San Francisco based on temperature. Comparing the center, spread, and shape for the two temperature distributions could help him decide.

Dot Plot of Temperature for New York City



Dot Plot of Temperature for San Francisco



From the dot plots, Robert saw that monthly temperatures in New York City were spread fairly evenly from around 40 degrees to the 80s, but in San Francisco the monthly temperatures did not vary as much. He was surprised that the mean temperature was about the same for both cities. The MAD of 14 degrees for New York City told him that, on average, a month's temperature was 14 degrees above or below 63 degrees. That is a lot of variability, which was consistent with the dot plot. On the other hand, the MAD for San Francisco told him that San Francisco's monthly temperatures differ, on average, only 3.5 degrees from the mean of 64 degrees. So, the mean doesn't help Robert very much in making a decision, but the MAD and dot plot are helpful.

Which city should he choose if he loves hot weather and really dislikes cold weather?

MP.2

Read through the example as a class. Note that, although the mean provides useful information, it does not give an accurate picture of the *spread* of monthly temperatures for New York City. It is important to consider the center, spread, and shape of distributions when making decisions.

Let students answer the questions:

- Which city should he choose if he loves hot weather and really dislikes cold weather?
 - *San Francisco, because there is little variability and it does not get as cold as New York city.*
- What measure of the data would justify your decision? Why did you choose that measure?
 - *The mean absolute deviation (MAD) as it provides a measure of the variability. On average, the monthly temperatures in San Francisco do not vary as much from the mean monthly temperature.*

Exercises 5–7 (15–20 minutes)

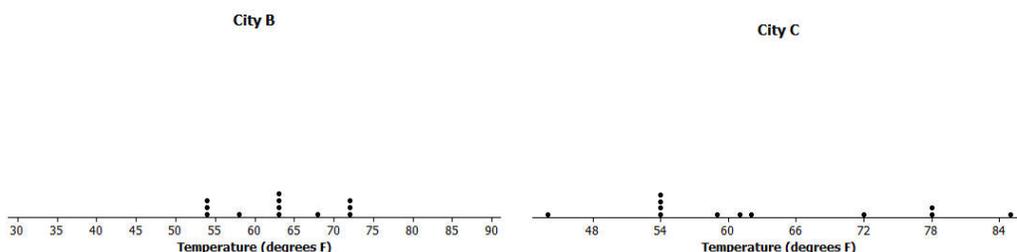
Give students an opportunity to work independently. Let them confirm answers with a neighbor as needed. If time allows, discuss answers as a class. Allow students to use calculators for this exercise. Prioritize your discussion of questions, as Exercise 7 is the first time students see a dot plot with the same MAD and mean.

Exercises 5–7

5. Robert wants to compare temperatures for Cities B and C.

	J	F	M	A	M	J	J	A	S	O	N	D
City B	54	54	58	63	63	68	72	72	72	63	63	54
City C	54	44	54	61	63	72	78	85	78	59	54	54

a. Draw a dot plot of the monthly temperatures for each of the cities.



b. Verify that the mean monthly temperature for each distribution is 63 degrees.

The data is nearly symmetrical around 63 degrees for City B.

The sum of positive deviations is +61, and the sum of the negative deviations is -61 around the mean of 63 for City C.

c. Find the MAD for each of the cities. Interpret the two MADs in words and compare their values.

The MAD is 5.3 degrees for City B, which means, on average, a month's temperature differs from 63 degrees by 5.3 degrees.

The MAD is 10.2 for City C, which means, on average, a month's temperature differs from 63 degrees by 10.2 degrees.

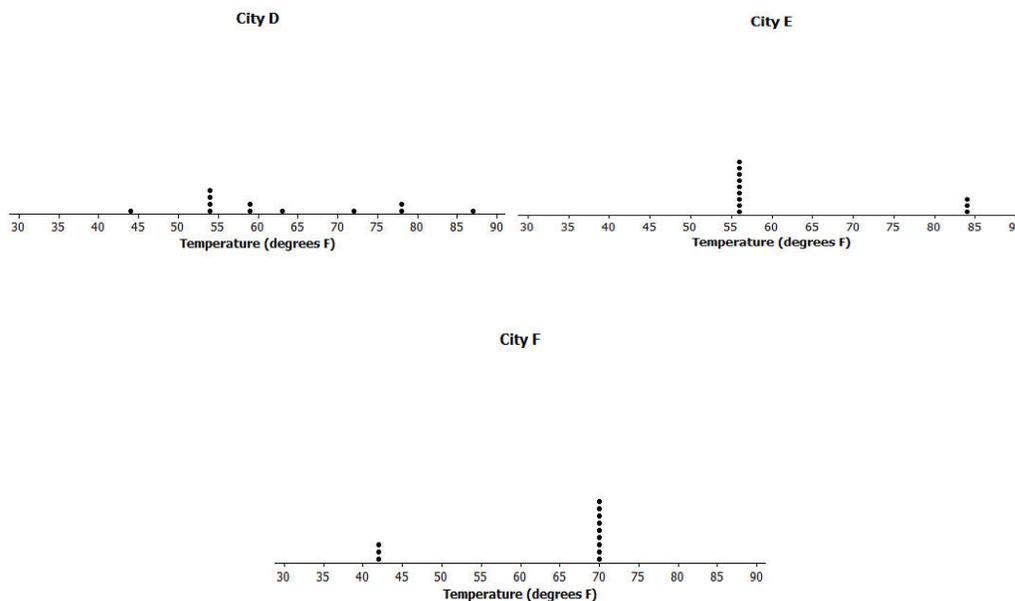
6. How would you describe the differences in the shapes of the monthly temperature distributions of the two cities?

The temperatures are nearly symmetric around the mean in City B. The temperatures are compact to the left of the mean for City C and then spread out to the right (skewed right).

7. Suppose that Robert had to decide between Cities D, E, and F.

	J	F	M	A	M	J	J	A	S	O	N	D	Mean	MAD
City D	54	44	54	59	63	72	78	87	78	59	54	54	63	10.5
City E	56	56	56	56	56	84	84	84	56	56	56	56	63	10.5
City F	42	42	70	70	70	70	70	70	70	70	70	42	63	10.5

a. Draw dot plots for each distribution.



b. Interpret the MAD for the distributions. What does this mean about variability?

The MADs are all the same, so Robert needs to look more at the shapes of the distributions to help him make a decision.

- c. How will Robert decide to which city he should move? List possible reasons Robert might have for choosing each city.

City D – Appears to have “four seasons” with widespread temperatures.

City E – Has mainly cold weather and is only hot for 3 months.

City F – Has mainly moderate weather and only a few cold months.

Lesson Summary

A data distribution can be described in terms of its center, spread, and shape.

- The center can be measured by the mean.
- The spread can be measured by the mean absolute deviation (MAD).
- A dot plot shows the shape of the distribution.

Exit Ticket (5 minutes)

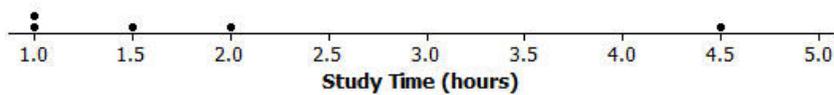
Name _____

Date _____

Lesson 10: Describing Distributions Using the Mean and MAD

Exit Ticket

1. A dot plot of times that five students studied for a test is displayed below.



- a. Use the table to determine the mean number of hours that these five students studied. Then, complete the table.

Student	Aria	Ben	Chloe	Dellan	Emma
Number of study hours	1	1	1.5	2	4.5
Deviations				0	
Absolute deviations					

- b. Find and interpret the MAD for this data set.
2. The same five students are preparing to take a second test. Suppose that the data were the same except that Ben studied 2.5 hours for the second test (1.5 hours more) and Emma studied only 3 hours for the second test (1.5 hours less.)
- a. Without doing any calculations, is the mean for the second test the same, higher, or lower than the mean for the first test? Explain your reasoning.
- b. Without doing any calculations, is the MAD for the second test the same, higher, or lower than the MAD for the first test? Explain your reasoning.

Exit Ticket Sample Solutions

1. A dot plot of times that five students studied for a test is displayed below.



a. Use the table to determine the mean number of hours that these five students studied. Then, complete the table.

The mean is 2 hours since the deviation around 2 hours is 0.

Student	Aria	Ben	Chloe	Dellan	Emma
Number of study hours	1	1	1.5	2	4.5
Deviations	-1	-1	-1.5	0	2.5
Absolute deviations	1	1	-0.5	0	2.5

b. Find and interpret the MAD for this data set.

The MAD is $\frac{1 + 1 + 0.5 + 2.5}{5} = 1$ hour.

On average the students studied 1 hour away from the group mean of 2 hours.

2. The same five students are preparing to take a second test. Suppose that the data were the same, except that Ben studied 2.5 hours for the second test (1.5 hours more), and that Emma studied only 3 hours for the second test (1.5 hours less).

a. Without doing any calculations, is the mean for the second test the same, higher, or lower than the mean for the first test? Explain your reasoning.

The mean would be the same since the distance that one data point moved to the right was matched by the distance another data point moved to the left. The distribution is balanced at the same place.

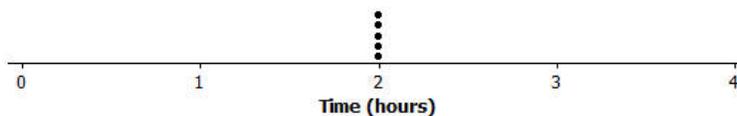
b. Without doing any calculations, is the MAD for the second test the same, higher, or lower than the MAD for the first test? Explain your reasoning.

The MAD would be smaller since the data points are clustered closer to the mean.

Problem Set Sample Solutions

1. Draw a dot plot of the times that five students studied for a test if the mean time they studied was two hours and the MAD was zero hours.

Since the MAD is 0, all data points are the same and that would be the mean value.



2. Suppose the times that five students studied for a test is as follows:

Student	Aria	Ben	Chloe	Dellan	Emma
Time (hrs.)	1.5	2	2	2.5	2

Michelle said that the MAD for this data set is 0 because the dot plot is balanced around 2. Without doing any calculation, do you agree with Michelle? Why or why not?

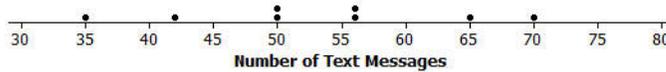
No, Michelle is wrong. There is variability within the data set, so the MAD is greater than zero.

Note: If students agree with Michelle, then they have not yet mastered that the MAD is measuring variability. They need to grasp that if data points differ in a distribution, whether the distribution is symmetric or not, then there is variability. Therefore, the MAD cannot be zero.

3. Suppose that the number of text messages eight students receive on a typical day is as follows:

Student	1	2	3	4	5	6	7	8
Number	42	56	35	70	56	50	65	50

a. Draw a dot plot for the number of text messages received on a typical day by these eight students.



b. Find the mean number of text messages these eight students receive on a typical day.

Since the distribution appears to be somewhat symmetrical around a value in the 50s, one could guess a value for the mean, such as 52 or 53, and then check sums of positive and negative deviations. Using the formula, the mean is $\frac{424}{8} = 53$ text messages.

c. Find the MAD number of text messages and explain its meaning using the words of this problem.

The sum of the positive deviations from 53 is: $2(56 - 53) + (65 - 53) + (70 - 53) = 6 + 12 + 17 = 35$. So, $\frac{2(35)}{8}$ yields a MAD of 8.75 text messages.

This means that, on average, the number of text messages these eight students receive on a typical day is 8.75 messages away from the group mean of 53 messages.

d. Describe the shape of this data distribution.

The shape of this distribution is fairly symmetrical (balanced) around the mean of 53 messages.

e. Suppose that in the original data set, Student 3 receives an additional five more text messages per day, and Student 4 receives five fewer messages per day.

i. Without doing any calculation, does the mean for the new data set stay the same, increase, or decrease as compared to the original mean? Explain your reasoning.

The mean would remain at 53 messages because one data point moved the same number of units to the right as another data point moved to the left. So, the balance point of the distribution does not change.

ii. Without doing any calculation, does the MAD for the new data set stay the same, increase, or decrease as compared to the original MAD? Explain your reasoning.

Since the lowest data point moved closer to the mean and the highest data point moved closer to the mean, the variability in the resulting distribution would be more compact than the original distribution. So, the MAD would decrease.



Lesson 11: Describing Distributions Using the Mean and MAD

Student Outcomes

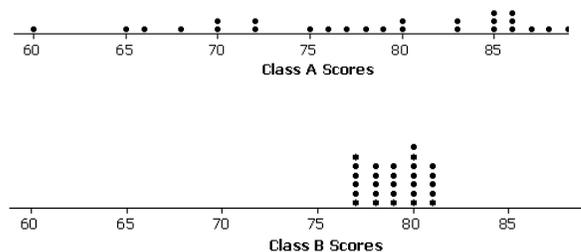
- Students use the mean and MAD to describe a data distribution in terms of center and variability.
- Students use the mean and MAD to describe similarities and differences between two distributions.

Classwork

Example 1 (5 minutes): Comparing Distributions with the Same Mean

Example 1: Comparing Distributions with the Same Mean

In Lesson 10, a data distribution was characterized mainly by its center (mean) and variability (MAD). How these measures help us make a decision often depends on the context of the situation. For example, suppose that two classes of students took the same test and their grades (based on 100 points) are shown in the following dot plots. The mean score for each distribution is 79 points. Would you rather be in Class A or Class B if you had a score of 79?



In Lesson 10, a data distribution was characterized mainly by its center (mean) and variability (MAD). How these measures help us make a decision often depends on the context of the situation. This example shows two distributions of test scores with the same mean, 79, but clearly very different variability. Students will need to be able to explain their reasoning for making decisions based not only on the mean but also the variability of the distributions.

MP.6

Introduce the data sets and ask students:

- In which class would you rather be if you scored a 79?
- How would you describe the shape of the data?

Exercises 1–3 (5–7 minutes)

Let students work independently. Then discuss and confirm answers as a class.

Exercises 1–6

- Looking at the dot plots, which class has the greater MAD? Explain without actually calculating the MAD.

Class A. The data for Class A has a much wider spread. Thus, it has greater variability and a larger MAD.

- If Liz had one of the highest scores in her class, in which class would she rather be? Explain your reasoning.

She would rather be in Class A. This class had higher scores in the 90s, whereas Class B had a high score of only 81.

- If Logan scored below average, in which class would he rather be? Explain your reasoning.

Logan would rather be in Class B. The low scores in B were in the 70s, whereas Class A had low scores in the 60s.

Exercises 4–6 (10–12 minutes)

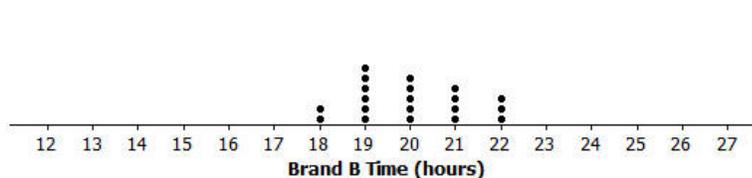
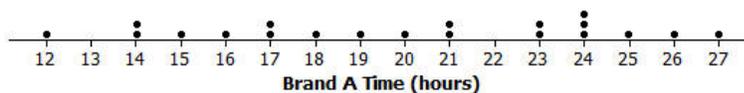
Let students work in pairs. Discuss answers to Exercises 5–6 as a class.

Exercises 4–6

Your little brother asks you to replace the battery in his favorite remote control car. The car is constructed so that it is difficult to replace its battery. Your research of the lifetimes (in hours) of two different battery brands (A and B) shows the following data for 20 batteries from each brand:

A	12	14	14	15	16	17	17	18	19	20	21	21	23	23	24	24	24	25	26	27
B	18	18	19	19	19	19	19	19	20	20	20	20	20	21	21	21	21	22	22	22

- To help you decide which battery to purchase, start by drawing a dot plot for each brand.



- Find the mean battery life for each brand and compare them.

The mean of Brand A is 20 hours.

The mean of Brand B is 20 hours.

6. Looking at the variability of each data set shown in its dot plot, give one reason you would choose Brand A. What is one reason you would choose Brand B? Explain your reasoning.

Answers will vary.

Brand A: take a risk and hope the battery lasts longer. You may get a good, long-lasting battery.

Brand B: the battery range is around 20 hours. Most of the batteries in this brand last that long.

Example 2 (5–7 minutes): Comparing Distributions with Different Means

Example 2: Comparing Distributions with Different Means

You have been comparing distributions that have the same mean, but different variability. As you have seen, deciding whether large variability or small variability is best depends on the context and on what is being asked. For example, in Exercise 2, Liz preferred to be in the distribution with more variability because she had one of the highest scores in the class. Thus, her score would have been higher had she been in Class A than had she been in Class B. Logan, on the other hand, preferred the class with lesser variability (i.e., Class B), since his score was below average.

If two data distributions have different means, how does a measure of variability play a part in making decisions?

Recount the pairs of distributions from Example 1 and Exercises 1–6 that had the same means and how variability played a role in making decisions.

MP.2

Then answer the question posed in the text: If two data distributions have different means, how does a measure of variability play a part in making decisions?

- Comparing the variability in distributions with different means is not really any different, as variability is not concerned with location.
- Note that when the magnitude of the data sets is substantially different, it may not be possible to graph both sets using the same scaling. In such cases, students should be careful in drawing a conclusion concerning variability because what may appear to be differing amounts of spread may not be that different at all. Encourage your students to draw dot plots using a scale that covers the span of both distributions whenever possible.
- If the magnitude of the two distributions is substantially different so that using the same scale is not practical, then advise them to be very careful comparing variability visually. That case requires the use of a measure, such as the MAD.

Exercises 7–9 (10–12 minutes)

Let students continue to work in pairs to complete Exercises 7–9. Then confirm answers to Exercises 8 and 9 as a class. Calculators are needed for this exercise.

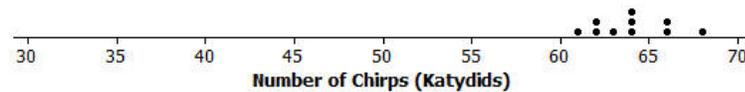
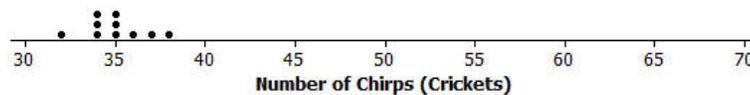
Exercises 7–9

Suppose that you wanted to answer the following question: Are field crickets better predictors of atmospheric temperature than katydids are? Both species of insect make chirping sounds by rubbing their front wings together.

The following data are the number of chirps (per minute) for 10 insects each. All the data were taken on the same evening at the same time.

Insect	1	2	3	4	5	6	7	8	9	10
Crickets	35	32	35	37	34	34	38	35	36	34
Katydids	66	62	61	64	63	62	68	64	66	64

7. Draw dot plots for these two data distributions using the same scale, going from 30 to 70. Visually, what conclusions can you draw from the dot plots?



Visually you can see a higher value for the mean of the katydids. The variability looks to be similar.

8. Calculate the mean and MAD for each distribution.

Crickets: The mean is 35 chirps per minute.

The MAD is 1.2 chirps per minute.

Katydids: The mean is 64 chirps per minute.

The MAD is 1.6 chirps per minute.

9. The outside temperature T can be predicted by counting the number of chirps made by these insects.

- a. For crickets, T is found by adding 40 to its mean number of chirps per minute. What value of T is being predicted by the crickets?

The predicted temperature is $35 + 40$ or 75 degrees.

- b. For katydids, T is found by adding 161 to its mean number of chirps per minute and then dividing the sum by 3. What value of T is being predicted by the katydids?

The predicted temperature is $\frac{(64+161)}{3}$ or 75 degrees.

- c. The temperature was 75 degrees when these data were recorded, so using the mean from each data set gave an accurate prediction of temperature. If you were going to use the number of chirps from a single cricket or a single katydid to predict the temperature, would you use a cricket or a katydid? Explain how variability in the distributions of number of chirps played a role in your decision.

The crickets had a smaller MAD. This indicates that an individual cricket is more likely to have a number of chirps that is close to the mean.

Lesson Summary

This lesson focused on comparing two data distributions based on center and variability. It is important to consider the context when comparing distributions. In decision-making, drawing dot plots and calculating means and MADs can help you make informed decisions.

Exit Ticket (5 minutes)



Name _____

Date _____

Lesson 11: Describing Distributions Using the Mean and MAD

Exit Ticket

You need to decide which of two brands of chocolate chip cookies to buy. You really love chocolate chip cookies. The numbers of chocolate chips in each of five cookies from each brand are as follows:

Cookie	1	2	3	4	5
ChocFull	17	19	18	18	18
AllChoc	22	15	14	21	18

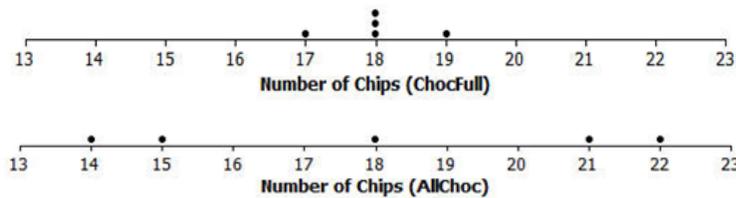
- Draw a dot plot for each set of data that shows the distribution of number of chips for each brand. Use a scale for your dot plots that covers the same span for both distributions.
- Find the mean number of chocolate chips for each of the two brands. Compare the means.
- Looking at your dot plots and considering variability, which brand do you prefer? Explain your reasoning.

Exit Ticket Sample Solutions

You need to decide which of two brands of chocolate chip cookies to buy. You really love chocolate chip cookies. The numbers of chocolate chips in each of five cookies from each brand are as follows:

Cookie	1	2	3	4	5
ChocFull	17	19	18	18	18
AllChoc	22	15	14	21	18

- a. Draw a dot plot for each set of data that shows the distribution of number of chips for each brand. Use a scale for your dot plots that cover the same span for both distributions.



- b. Find the mean number of chocolate chips for each of the two brands. Compare the means.

Students should look at both graphs and immediately determine that the means are both 18 chips, since the distributions are symmetric around 18.

- c. Looking at your dot plots and considering variability, which brand do you prefer? Explain your reasoning.

Students could argue either way:

- *Students who prefer ChocFull may argue that they are assured of getting 18 chips most of the time, with no fewer than 17 chips, and a bonus once in a while of 19 chips. With AllChoc, they may sometimes get more than 20 chips, but would sometimes get only 14 or 15 chips.*
- *Students who prefer AllChoc are the risk-takers who are willing to tolerate getting only 14 or 15 chips for the chance of getting 21 or 22 chips.*

Problem Set Sample Solutions

1. Two classes took the same mathematics test. Summary measures for the two classes are as follows:

	Mean	MAD
Class A	78	2
Class B	78	10

- a. Suppose that you received the highest score in your class. Would your score have been higher if you were in Class A or Class B? Explain your reasoning.

Class B, because the means are the same. And the variability, as measured by the MAD, is higher in that class than it is in Class A.

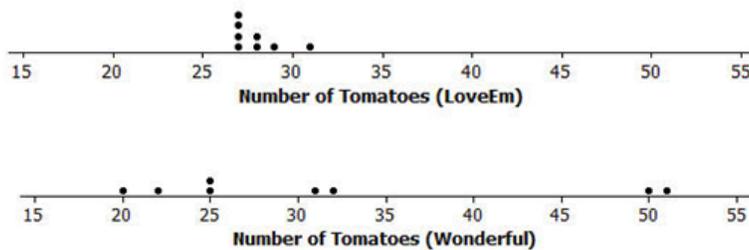
- b. Suppose that your score was below the mean score. In which class would you prefer to have been? Explain your reasoning.

Class A because the variability, as measured by the MAD, indicates a more compact distribution around the mean. Whereas, a score below the mean in Class B could be far lower than in Class A.

2. Eight tomato plants each of two varieties, LoveEm and Wonderful, are grown under the same conditions. The numbers of tomatoes produced from each plant of each variety are shown:

Plant	1	2	3	4	5	6	7	8
LoveEm	27	29	27	28	31	27	28	27
Wonderful	31	20	25	50	32	25	22	51

- a. Draw dot plots to help you decide which variety is more productive.



- b. Calculate the mean number of tomatoes produced for each variety. Which one produces more tomatoes on average?

Guessing a mean, and checking by summing deviations, is not as obvious for these distributions. So, using the formula is probably more efficient. The mean number of LoveEm tomatoes is 28, and the mean number of Wonderful tomatoes is 32.

- c. If you want to be able to accurately predict the number of tomatoes a plant is going to produce, which variety should you choose – the one with the smaller MAD, or the one with the larger MAD? Explain your reasoning.

LoveEm produces fewer tomatoes on average but is far more consistent. Looking at the dot plots, its variability is far less than that of Wonderful tomatoes. Based on these data sets, choosing LoveEm should yield numbers in the high 20s consistently, but the number from Wonderful could vary wildly from lower yields in the low 20s, to huge yields around 50.

- d. Calculate the MAD of each plant variety.

The MAD for LoveEm is 1 tomato.

The MAD for Wonderful is 9.25 tomatoes.

2. There are nine judges currently serving on the Supreme Court of the United States. The following table lists how long (number of years) each judge has been serving on the court as of 2013.

Judge	Length of service
Antonin Scalia	27
Anthony Kennedy	25
Clarence Thomas	22
Ruth Bader Ginsburg	20
Stephen Breyer	19
John Roberts	8
Samuel Alito	7
Sonia Sotomayor	4
Elena Kagan	3

- a. Calculate the mean length of service for these nine judges. Show your work.
- b. Calculate the mean absolute deviation (MAD) of the lengths of service for these nine judges. Show your work.
- c. Explain why the mean may not be the best way to summarize a typical length of service for these nine judges.

3. The following table displays data on calories for several Chinese foods (from *Center for Science in the Public Interest*, tabulated by the *Philadelphia Inquirer*).

Dish	Dish size	Calories	Dish	Dish size	Calories
Egg roll	1 roll	190	House lo mein	5 cups	1059
Moo shu pork	4 pancakes	1228	House fried rice	4 cups	1484
Kung Pao chicken	5 cups	1620	Chicken chow mein	5 cups	1005
Sweet and sour pork	4 cups	1613	Hunan tofu	4 cups	907
Beef with broccoli	4 cups	1175	Shrimp in garlic sauce	3 cups	945
General Tso's chicken	5 cups	1597	Stir-fried vegetables	4 cups	746
Orange (crispy) beef	4 cups	1766	Szechuan shrimp	4 cups	927
Hot and sour soup	1 cup	112			

- a. Round the *Calories* values to the nearest 100 calories, and use these rounded values to produce a dot plot of the distribution of the calories in these dishes.
- b. Describe the distribution of the calories in these dishes.
- c. Suppose you wanted to report data on calories per cup for different Chinese foods. What would the calories per cup be for Kung Pao chicken?

- d. Could you calculate calories per cup for all of the foods listed in the table? Explain why or why not.
- e. If you wanted to compare the healthiness of these foods in terms of calories, would you compare the calorie amounts or the calories per cup? Explain your choice.
4. A father wanted some pieces of wood that were 10 inches long for a building project with his son. He asked the hardware store to cut some longer pieces of wood into 10 inch pieces. However, he noticed that not all of the pieces given to him were the same length. He then took the cut pieces of wood home and measured the length (in inches) of each piece. The table below summarizes the lengths that he found.

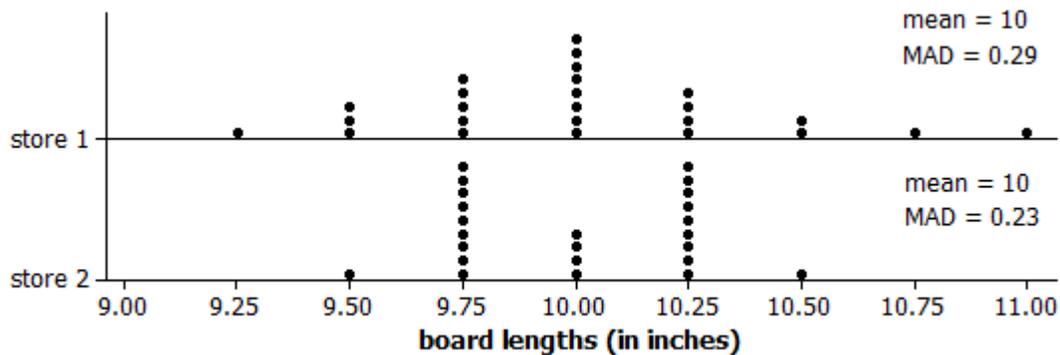
Length (inches)	8.50– < 8.75	8.75– < 9.00	9.00– < 9.25	9.25– < 9.50	9.50– < 9.75	9.75– < 10.00	10.00– < 10.25	10.25– < 10.50	10.50– < 10.75	12.00– < 12.25
Frequency	1	2	2	4	3	2	5	6	1	1

- a. Create a histogram for these data.
- b. Describe the shape of the histogram you created.

- c. The father wanted to know whether the mean length was equal to 10 inches or if the wood cutter cut pieces that tended to be too long or tended to be too short. Without calculating the mean length, explain based on the histogram whether the mean board length should be equal to 10 inches, greater than 10 inches, or less than 10 inches. Explain what strategy you used to determine this.

- d. Based on the histogram, should the mean absolute deviation (MAD) be larger than 0.25 inches or smaller than 0.25 inches? Explain how you made this decision.

- e. Suppose this project was repeated at two different stores, and the following two dot plots of board lengths were found. Would you have a preference for one store over the other store? If so, which store would you prefer and why? Justify your answer based on the displayed distributions.



A Progression Toward Mastery

Assessment Task Item		STEP 1 Missing or incorrect answer and little evidence of reasoning or application of mathematics to solve the problem.	STEP 2 Missing or incorrect answer but evidence of some reasoning or application of mathematics to solve the problem.	STEP 3 A correct answer with some evidence of reasoning or application of mathematics to solve the problem, <u>or</u> an incorrect answer with substantial evidence of solid reasoning or application of mathematics to solve the problem.	STEP 4 A correct answer supported by substantial evidence of solid reasoning or application of mathematics to solve the problem.
1	a 6.SP.A.1	Student simply provides a numerical response to the question.	Student incorrectly identifies this as an invalid question <u>OR</u> identifies this as a valid question but with an incorrect justification.	Student identifies this as a valid question but does not give a complete justification that distinguishes it from other questions, e.g., “We can record this.”	Student identifies this as a valid question and justifies the choice based on the variability in answers (amount watched last night) among students at the school.
	b 6.SP.A.1	Student only provides a guess for the answer to the question.	Student incorrectly identifies this as a valid question <u>OR</u> identifies this as an invalid question with an incorrect justification.	Student identifies this as an invalid question but fails to give a clear explanation, e.g., “There is just one answer.”	Student identifies this as an invalid question and justifies the choice by the lack of variation in the responses for students at the school.
	c 6.SP.A.1	Student simply provides a yes/no response to the question.	Student identifies this as an invalid question with an incorrect justification.	Student identifies this as a valid question but fails to give a clear explanation <u>OR</u> identifies this as an invalid question assuming that every student at the school would have the same answer.	Student identifies this as a valid question and justifies the choice based on the variability in answers (whether or not students watch at least one hour on weekend) among students at the school.
	d 6.SP.A.1	Student simply provides a numerical response to the question.	Student identifies this as a valid question with an incorrect justification.	Student identifies this as an invalid question but fails to give a clear explanation <u>OR</u> identifies this as a valid question assuming that every student at the school would respond with	Student identifies this as an invalid question and justifies the choice based on the lack of ability to gather data from the students to address the question.

				different guesses.	
2	a 6.SP.B.5c	Student does not provide a sensible answer (e.g., outside $3 - 27$).	Student makes a major calculation error such as reporting the median.	Student correctly calculates mean but does not show work or has a minor, but traceable, calculation error.	Student calculates $\frac{27+\dots+3}{9} = 15$ years.
	b 6.SP.B.5c	Student does not provide a sensible answer (e.g., larger than 22).	Student demonstrates understanding of measuring spread but not of “deviation.”	Student calculates deviations from mean but does not combine them correctly (e.g., only the sum or does not use absolute values and gets zero).	Student calculates $[(27 - 15) + \dots + (15 - 3)]/9 = 76/9 = 8.44$ years.
	c 6.SP.B.5d	Student response confirms mean as a measure of center of a distribution.	Student discusses disadvantages of mean in general but does not relate to context (e.g., not best with skewed data).	Student discusses the possibility of the mean being thrown off by outliers but does not address the bimodal shape of this distribution.	Student comments that there seem to be two clusters of data, one below 15 and one above 15, but that there are no judges with length of service right around 15.
3	a 6.SP.B.4	Student fails to create a graph displaying the distribution of calories.	Student produces a graph other than a dot plot for the calories data.	Graph is poorly labeled or poorly scaled, or student makes major errors in rounding.	Student correctly rounds the values to the nearest 100 (or makes a minor error) and constructs, scales, and labels a dot plot.
	b 6.SP.B.5	Student does not provide a reasonable description of the graph constructed.	Student only addresses one aspect (e.g., center) of describing the distribution.	Student comments on numerous features of the distribution but does not describe the distribution as a whole in terms of tendency and variability.	Student comments on the shape, center, and variability of the distribution. Distribution is not very symmetric, center is around 1000 calories, and dishes range from about 200 to 1766 calories.
	c 6.SP.B.5b	Student fails to correctly perform calculation.	Student reports cup/calorie value.	Student gives value but has minor calculator error or does not show work.	Student reports $1620/5 = 324$ calories/cup.
	d 6.SP.B.5b	Student does not attempt question.	Student says yes without considering all the dishes.	Student recognizes the need to have a common scale among the dishes but does not notice the two dishes without cup sizes.	Student recognizes that we do not have “per cup” results for every food item. (Egg roll and Moo shu pork are not clearly single servings as the other food items are.)

	e 6.SP.B.5b	Student does not understand the goal of comparing calorie amounts across these different dishes.	Student relates comments to context but relies on external information rather than the information presented in the table.	Student recognizes that the comparisons should be made on an equivalent scale but does not specifically answer question.	Student selects the calories per cup as a more reasonable way to compare across the different size dishes. Student could select calories with an assumption that original values corresponded to equivalent serving sizes.
4	a 6.SP.B.4	Student fails to use provided information to construct a histogram.	Student produces a type of dot plot or box plot from the data.	Student produces a histogram but does not scale x -axis appropriately (e.g., does not leave gap between 10.75 and 12).	Student produces a complete and well labeled (“lengths”) histogram using all 10 frequencies.
	b 6.SP.A.2	Student does not address shape of distribution.	Student’s description of shape is not consistent with graph.	Student describes the distribution in detail but does not use accepted language to efficiently describe shape.	Student’s description of shape is consistent with constructed graph. This may or may not include separate comments on outliers.
	c 6.SP.B.5c	Student’s response does not relate to the center of the distribution.	Student only attempts to calculate mean and ignores tallies.	Student only attempts to calculate mean using tallied information but does not arrive at a reasonable answer, <u>OR</u> student’s explanation only applies to determining the location of the median.	Student uses the tallied information and/or histogram and the idea of balancing to conclude that the mean is less than 10 inches.
	d 6.SP.B.5c	Student’s response does not relate to the spread of the distribution.	Student only attempts to calculate MAD and ignores tallies, <u>OR</u> student confuses MAD with the bin widths of the histogram.	Student only attempts to calculate MAD using tallied information but does not arrive at a reasonable answer, <u>OR</u> student’s explanation only applies to determining the value of MAD.	Student uses the tallied information and/or histogram and the idea of balancing deviations to draw a consistent conclusion about the value of MAD (for the mean identified in (c)). Student may notice that 0.25 is too small as it would only encompass about 5 of the 27 values but some values are much further out.

	e 6.SP.A.3	Student does not use information from graph to address question.	Student only justifies the choice based on store one having more values at 10.00.	Student’s explanation is not consistent with the choice but attempts to make use of dot plot, mean, and MAD information.	Student justification relates to the dot plots and the mean and MAD values. Student may prefer store with smaller MAD or store with all but two values between 9.75 and 10.25.
5	a 6.SP.B.5c	Student’s answer does not use the information provided.	Student does not recognize the relationship between mean and total time.	Student makes a minor calculation error.	Student provides the correct answer: Total time = $5(2.75) = 13.75$ seconds.
	b 6.SP.B.5c	Student is unable to begin problem.	Student makes a calculation error and arrives at a nonsensical answer.	Student makes a minor calculation error but answer is still reasonable (between 2.5 and 3.5).	Student provides the correct answer: New mean = $\frac{13.75-3.6}{4} = 2.54$ seconds.

Name _____

Date _____

1. For each of the following, identify whether or not it would be a valid *statistical question* you could ask about people at your school. Explain for each why it is, or is not, a statistical question.

- a. What was the mean number of hours of television watched by students at your school last night?

This is a statistical question
because the number of TV hours
will vary from student to student.

- b. What is the school principal's favorite television program?

This is not a statistical question
because there is just one principal
at this school and the answer each student
would get is the same.

- c. Do most students at your school tend to watch at least one hour of television on the weekend?

This is a statistical question
because some students will have
watched at least one hour and
some will not → answers vary.

- d. What is the recommended amount of television specified by the American Pediatric Association?

This is not a statistical
question because we would not
ask different students at the
school this question. There is just
one answer.

2. There are nine judges currently serving on the Supreme Court of the United States. The following table lists how long (number of years) each judge has been serving on the court as of 2013.

Judge	Length of service
Antonin Scalia	27
Anthony Kennedy	25
Clarence Thomas	22
Ruth Bader Ginsburg	20
Stephen Breyer	19
John Roberts	8
Samuel Alito	7
Sonia Sotomayor	4
Elena Kagan	3

- a. Calculate the mean length of service for these nine judges. Show your work.

$$\frac{27 + 25 + 22 + 20 + 19 + 8 + 7 + 4 + 3}{9} = 15 \text{ years}$$

- b. Calculate the mean absolute deviation (MAD) of the lengths of service for these nine judges. Show your work.

$$\begin{array}{l} 27 - 15 = 12 \quad 20 - 15 = 5 \quad |8 - 15| = 7 \quad |4 - 15| = 11 \\ 25 - 15 = 10 \quad 19 - 15 = 4 \quad |7 - 15| = 8 \quad |3 - 15| = 12 \\ 22 - 15 = 7 \end{array}$$

$$\frac{12 + 10 + 7 + 5 + 4 + 7 + 8 + 11 + 12}{9} = \frac{76}{9} \approx 8.44 \text{ years}$$

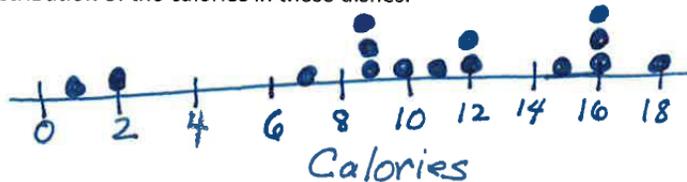
- c. Explain why the mean may not be the best way to summarize a typical length of service for these nine judges.

There are no judges around 15 years. In fact we have 2 clumps of judges - one clump is 3-8 years and the other is 19-27 years. Giving one center is not very useful here.

3. The following table displays data on calories for several Chinese foods (from *Center for Science in the Public Interest*, tabulated by the *Philadelphia Inquirer*).

Dish	Dish size	Calories	Dish	Dish size	Calories
Egg roll	1 roll	190	House lo mein	5 cups	1059
Moo shu pork	4 pancakes	1228	House fried rice	4 cups	1484
Kung Pao chicken	5 cups	1620	Chicken chow mein	5 cups	1005
Sweet and sour pork	4 cups	1613	Hunan tofu	4 cups	907
Beef with broccoli	4 cups	1175	Shrimp in garlic sauce	3 cups	945
General Tso's chicken	5 cups	1597	Stir-fried vegetables	4 cups	746
Orange (crispy) beef	4 cups	1766	Szechuan shrimp	4 cups	927
Hot and sour soup	1 cup	112			

- a. Round the *Calories* values to the nearest 100 calories, and use these rounded values to produce a dot plot of the distribution of the calories in these dishes.



- b. Describe the distribution of the calories in these dishes.

A typical number of calories is around 900 calories with several dishes around 1600 calories as well. There is a lot of variability in the number of calories, from 112 to 1766. Two dishes have a lot fewer calories than the rest.

- c. Suppose you wanted to report data on calories per cup for different Chinese foods. What would the calories per cup be for Kung Pao chicken?

$$\frac{1620 \text{ calories}}{5} = 324 \text{ calories per cup}$$

d. Could you calculate calories per cup for all of the foods listed in the table? Explain why or why not.

No. We don't have "cups" for egg rolls and moo shu pork.

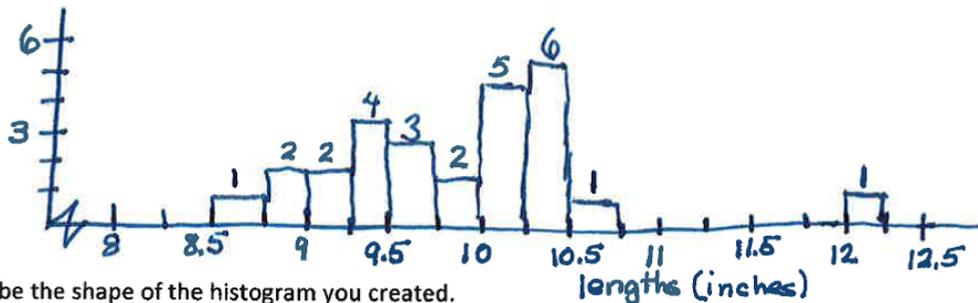
e. If you wanted to compare the healthiness of these foods in terms of calories, would you compare the calorie amounts or the calories per cup? Explain your choice.

Calories per cup seems more fair because the number of cups vary across the dishes.

4. A father wanted some pieces of wood that were 10 inches long for a building project with his son. He asked the hardware store to cut some longer pieces of wood into 10 inch pieces. However, he noticed that not all of the pieces given to him were the same length. He then took the cut pieces of wood home and measured the length (in inches) of each piece. The table below summarizes the lengths that he found.

Length (inches)	8.50-<8.75	8.75-<9.00	9.00-<9.25	9.25-<9.50	9.50-<9.75	9.75-<10.00	10.00-<10.25	10.25-<10.50	10.50-<10.75	12.00-<12.25
Frequency	1	2	2	4	3	2	5	6	1	1

a. Create a histogram for these data.



b. Describe the shape of the histogram you created.

The histogram appears more skewed than symmetric.

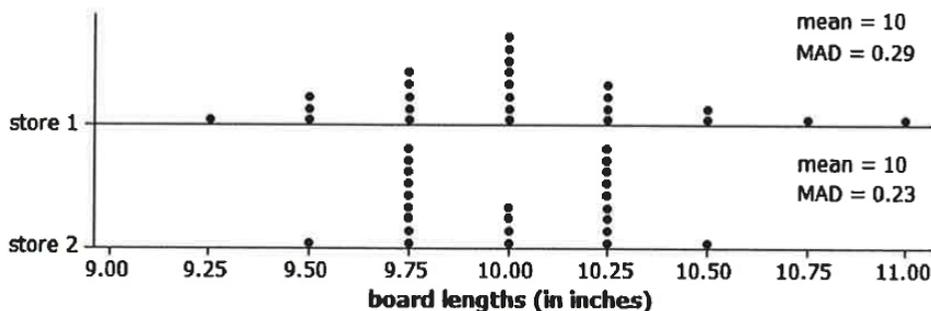
- c. The father wanted to know whether the mean length was equal to 10 inches or if the wood cutter cut pieces that tended to be too long or tended to be too short. Without calculating the mean length, explain based on the histogram whether the mean board length should be equal to 10 inches, greater than 10 inches, or less than 10 inches. Explain what strategy you used to determine this.

There are more observations below 10 inches (14) than above 10 inches. Most of the above 10 inches are within .5 inches, but the ones below are more spread out and less than 10 inches.

- d. Based on the histogram, should the mean absolute deviation (MAD) be larger than 0.25 inches or smaller than 0.25 inches? Explain how you made this decision.

Too many data values are more than .25 inches from mean, so $MAD > .25$.

- e. Suppose this project was repeated at two different stores, and the following two dot plots of board lengths were found. Would you have a preference for one store over the other store? If so, which store would you prefer and why? Justify your answer based on the displayed distributions.



I like the second store. All but 2 boards are within .25 inches of 10 inches. Some of the boards in the top graph are very far from 10 inches.

5. Suppose you are timing how long it takes a car to race down a wood track placed at a forty-five degree angle. The times for five races are recorded. The mean time for the five races is 2.75 seconds.

a. What was the total time for the five races (the times of the five races summed together)?

$$\begin{aligned} \text{mean} &= \frac{\text{total}}{5}, \text{ so } \text{total} = 5 \times \text{mean} \\ &= 5 \times 2.75 \text{ seconds} \\ &= 13.75 \text{ seconds} \end{aligned}$$

b. Suppose you learn that the timer malfunctioned on one of the five races. The result of the race had been reported to be 3.6 seconds. If you remove that time from the list and recomputed the mean for the remaining four times, what do you get for the mean? Show your work.

$$\begin{array}{r} 13.75 \text{ seconds} \\ - 3.6 \text{ seconds} \\ \hline 10.15 \text{ seconds} \end{array} \Rightarrow \frac{10.15 \text{ seconds}}{4} = 2.5375 \text{ seconds}$$



Topic C:

Summarizing a Distribution that is Skewed Using the Median and the Interquartile Range

6.SP.A.2, 6.SP.A.3, 6.SP.B.4, 6.SP.B.5

Focus Standard:	6.SP.A.2	Understand that a set of data collected to answer a statistical question has a distribution, which can be described by its center, spread, and overall shape.
	6.SP.A.3	Recognize that a measure of center for a numerical data set summarizes all of its values with a single number, while a measure of variation describes how its values vary with a single number.
	6.SP.B.4	Display numerical data in plots on a number line, including dot plots, histograms, and box plots.
	6.SP.B.5	Summarize numerical data sets in relation to their context, such as by: <ol style="list-style-type: none"> Reporting the number of observations. Describing the nature of the attribute under investigation, including how it was measured and its units of measurement. Giving quantitative measures of center (median and/or mean) and variability (interquartile range and/or mean absolute deviation), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered. Relating the choice of measures of center and variability to the shape of the data distribution and the context in which the data were gathered.

Instructional Days: 5**Lesson 12:** Describing the Center of a Distribution Using the Median (P)¹**Lesson 13:** Describing Variability Using the Interquartile Range (IQR) (P)**Lesson 14:** Summarizing a Distribution Using a Box Plot (P)**Lesson 15:** More Practice with Box Plots (P)**Lesson 16:** Understanding Box Plots (P)

In Topic C, students are introduced to a measure of center (the median) and a measure of variability (the interquartile range (IQR)) that are appropriate for describing data distributions that are skewed. Box plots are also introduced in this topic. In Lesson 12, students learn to calculate and interpret the median. Quartiles are introduced in Lesson 13, and the quartiles are then used to calculate the IQR. Students also learn to interpret the IQR as a measure of variability in a data distribution. Lessons 14–16 introduce box plots. Boxplots are often difficult for students to interpret, as they are not a graph of a data distribution (as are dot plots and histograms), but rather are a graph of five key summary statistics of a data set (the minimum, lower quartile, median, upper quartile, and the maximum). Lesson 16 has students use box plots to compare groups, setting the stage for future work on comparing groups in Grade 7.

¹ Lesson Structure Key: **P**-Problem Set Lesson, **M**-Modeling Cycle Lesson, **E**-Exploration Lesson, **S**-Socratic Lesson



Lesson 12: Describing the Center of a Distribution Using the Median

Student Outcomes

- Given a data set, students calculate the median of the data.
- Students estimate the percent of values above and below the median value.

Lesson Overview

The focus of this lesson is the median as a summary statistic to describe a data set. Students report the number of observations for both odd and even numbered sets of data. Informally, they consider the variability among three different data sets to assess a claim about typical behavior. In preparation for a later lesson on finding quartiles, students calculate the median of the values below the median and the median of the values above the median and estimate the approximate count/percent of values above and below the median. This lesson provides the background for the development of a box plot; however, this lesson is not about creating a box plot.

In this lesson students construct arguments and critique the reasoning of others. They respond to the reasoning of others in some of the tasks, distinguish correct reasoning from flawed reasoning, and explain why it is flawed. They also model with mathematics, apply mathematics to problems from everyday life, and interpret results in the context of the situation.

It should be noted that students should have access to calculators throughout this module.

Classwork

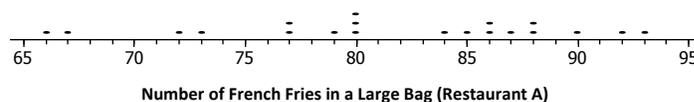
How do we summarize a data distribution? What provides us with a good description of the data? The following exercises help us to understand how a numerical summary answers these questions.

Example 1 (2 minutes): The Median—A Typical Number

The activity begins with a set of data displayed in a dot plot. Introduce the data presented in the example.

Example 1: The Median – A Typical Number

Suppose a chain restaurant (Restaurant A) advertises that a typical number of french fries in a large bag is 82. The graph shows the number of french fries in selected samples of large bags from Restaurant A.



Using the data shown in the plot, students are asked to think about when it might be useful to separate a set of data into two parts that have the same number of elements. In other words, when would it be useful to know the point that separates the top half from the bottom half? The notion of median is developed by a set of questions and then defined.

Let students work independently on the exercises and confirm answers with a neighbor.

Exercises 1–3 (5 minutes)

Exercises 1–3

1. You just bought a large bag of fries from the restaurant. Do you think you have 82 french fries? Why or why not?

The number seems to vary greatly from bag to bag. No bag even had 82 fries, so mine probably will not. The restaurant sells french fries in bags that have from 66 to 93 per bag, so the claim that they typically have 82 fries doesn't seem right.

2. How many bags were in the sample?

20

3. Which of the following statements would seem to be true given the data? Explain your reasoning.

- Half of the bags had more than 82 fries in them.
- Half of the bags had fewer than 82 fries in them.
- More than half of the bags had more than 82 fries in them.
- More than half of the bags had fewer than 82 fries in them.
- If you got a random bag of fries, you could get as many as 93 fries.

(a) and (b) are true because there are 10 bags above 82 fries and 10 bags below 82 fries. (e) is true because that happened once and so probably could happen again.

Example 2 (3 minutes): The Median

Read through the text with students.

Example 2: The Median

Sometimes it is useful to know what point separates a data distribution into two equal parts, where one part represents the larger “half” of the data values and the other part represents the smaller “half” of the data values. This point is called the **median**. When the data are arranged in order from smallest to largest, the same number of values will be above the median as are below the median.

As a class, work through the exercises one at a time. As students work through the problems, ask the following questions:

- How do you find the median if there are an even number of data points?
- About what fraction of the data values should be above the median? What fraction should be below the median?

Exercises 4–5 (5 Minutes)

Exercises 4–5

4. Suppose you were trying to convince your family that you needed a new pair of tennis shoes. After checking with your friends, you argued that half of them had more than four pairs of tennis shoes, and you only had two pairs. Give another example of when you might want to know that a data value is a half-way point? Explain your thinking.

Possible responses: When the data are about how much people earn, it would be interesting to know the amount that is less than what half of the people earn; if you are looking at the number of points earned in a competition, it would be good to know what number separates the top half of the competitors from the bottom.

5. Use the information from the dot plot in Example 1. The median number of fries was 82.

- a. What percent of the bags have more fries than the median? Less than the median?

50 percent or $\frac{1}{2}$ of the bags have more fries than the median, and 50% or $\frac{1}{2}$ have fewer fries than the median.

- b. Suppose the bag with 93 fries was miscounted and there were only 85 fries. Would the median change? Why or why not?

The median would not change because there would still be 10 bags with fewer than 82 fries and 10 bags with more than 82 fries.

- c. Suppose the bag with 93 fries really only had 80 fries. Would the median change? Why or why not?

The median would change because now there would be 11 bags that would have fewer than 82 fries and only 9 that have more than 82 instead of the same number in both directions.

Exercises 6–7 (15 minutes): A Skewed Distribution

In this activity, students have to order the data before they find the median. There are 19 values, so the median is the 10th value with 9 counts above and 9 counts below. Another way to determine the median after ordering the data is to cross out the maximum and minimum values continuously until students reach one number in the middle if there are an odd number of data values, or two numbers for an even number of values. Students would then find the mean of the two values. The questions are designed to help students confront some common misconceptions: not ordering the data before counting to the middle; confusing median and mode (most frequent value); confusing median and midrange (half way between the maximum and the minimum). They also compute the mean and compare the median to the mean, noting that several bags with a low number of french fries pulled the mean down and so the median might be more reflective of the typical number of fries.

MP.3

Consider the following questions as students are completing this exercise:

- Why is it necessary to order the data before you find the median?
- Is the median connected to the range (maximum-minimum) of the data? Why or why not?
- What is the difference in the effect of very extreme values on the mean and on the median?

Exercises 6–7: A Skewed Distribution

6. The owner of the chain decided to check the number of french fries at another restaurant in the chain. Here is the data for Restaurant B: 82, 83, 83, 79, 85, 82, 78, 76, 76, 75, 78, 74, 70, 60, 82, 82, 83, 83, 83.

- a. How many bags of fries were counted?

19

- b. Sallee claims the median is 75 as she sees that 75 is the middle number in the data set listed above. She thinks half of the bags had fewer than 75 fries. Do you think she would change her mind if the data were plotted in a dot plot? Why or why not?

You cannot find the median unless the data are ordered by size. Plotting the number of fries in each bag on a number line in a dot plot would order the data so you would probably get a different halfway point because the data above is not ordered from smallest to largest.

- c. Jake said the median was 83. What would you say to Jake?

83 is the most common number of fries in the bags (5 bags had 83 fries), but it is not in the “middle” of the data, marking where the number of bags with fries more than and less than are the same.

- d. Betse argued that the median was halfway between 60 and 85 or 72.5. Do you think she is right? Why or why not?

She is wrong because the median is not connected to the distance between points on the number line but are connected to finding a point that separates the data into two parts with the same number of values in each part.

- e. Chris thought the median was 82. Do you agree? Why or why not?

Chris is correct because if you order the numbers, the middle number will be the 10th number, with 9 bags that have more than 82 fries and 9 bags with fewer than 82 fries.

7. Calculate the mean and compare it to the median. What do you observe about the two values? If the mean and median are both measures of center, why do you think one of them is lower than the other?

The mean is 78.6 and the median is 82. The bag with the 60 fries lowered the value of the mean.

Exercises 8–10 (15 minutes): Finding Medians from Frequency Tables

MP.4

In this example, students find the median using a frequency table halfway between the 13th and 14th counts. They also find the medians of the top and bottom halves, the 7th value from the top and from the bottom, as a precursor to finding an interquartile range in a later lesson. They will encounter repeated values in finding the quartiles. You may want students to write out the individual counts in a long ordered list. For example, the first 13 counts would be as follows:

Median of the lower half

75 75 76 77 77 78 **78** 78 79 79 79 79 ...

Then have students find the medians of each half by counting from the top and bottom of the list, noting that a value for bags with the same count can be in both halves. It might help to think about the individual bags – one of the bags with 78 fries is in the first half, one of the bags with 78 fries is in the second half, and one of the bags divides the two halves and marks the median of the data set. At this point, the important idea is that students get a sense of how to find a median: order the values and find a midpoint for the ordered values.

Exercises 8–10: Finding Medians from Frequency Tables

8. A third restaurant (Restaurant C) tallied a sample of bags of french fries and found the results below.

Number of fries	Frequency
75	
76	
77	
78	
79	
80	
81	
82	
83	
84	
85	
86	

- a. How many bags of fries did they count?

26

- b. What is the median number of fries for the sample of bags from this restaurant? Describe how you found your answer.

79.5; I took half of 26, which was 13 and then counted 13 tallies from 86 down to get to 80. I also counted 13 up from 75 to get to 79. The point halfway between 79 and 80 is the median.

9. Robere decided to divide the data into four parts. He found the median of the whole set.

- a. List the 13 values of the bottom half. Find the median of these 13 values.

75 75 76 77 77 78 78 78 79 79 79 79 79

Median of this half is 78.

- b. List the 13 values of the top half. Find the median of these 13 values.

80 80 80 80 81 82 84 84 84 85 85 85 86

Median of this is 84.

10. Which of the three restaurants seems most likely to really have 82 fries in a typical bag? Explain your thinking.

Answers will vary: Restaurant B seems to have most bags closest to 82 because the middle half of the number of fries in bags covers a span of 7 fries (from 76 to 83) with a median of 82; Restaurant A goes from 87.5 to 75 fries for a span of 12.5 with a median of 80, and Restaurant C goes from 78 to 84 for a span of 6 but the median is 79. Some students might say Restaurant B because the median is 82.

Closing (3 minutes)

Lesson Summary

In this lesson, you learned about a summary measure for a set of data called the *median*. To find a median you first have to order the data. The median is the midpoint of a set of ordered data; it separates the data into two parts with the same number of values below as above that point. For an even number of data values, you find the average of the two middle numbers; for an odd number of data values, you use the middle value. It is important to note that the median might not be a data value and that the median has nothing to do with a measure of distance. Medians are sometimes called a measure of the center of a frequency distribution but do not have to be the middle of the spread or range (maximum-minimum) of the data.

Exit Ticket (5 minutes)



Name _____

Date _____

Lesson 12: Describing the Center of a Distribution Using the Median

Exit Ticket

1. What is the median age for the following data set representing the age of students requesting tickets for a summer band concert?

13 14 15 15 16 16 17 18 18

2. What is the median number of diseased trees from a data set of diseased trees on 10 city blocks?

11 3 3 4 6 12 9 3 8 8 8 1

3. Describe how you would find the median for a set of data that has 35 values. How would this be different if there were 36 values?

Exit Ticket Sample Solutions

1. What is the median age for the following data set representing the age of students requesting tickets for a summer band concert?

13 14 15 15 16 16 17 18 18

The median is the 5th value, or 16 years old, as there are 4 values less than 16 and 4 values greater than or equal to 16.

2. What is the median number of diseased trees from a data set of diseased trees on 10 city blocks?

11 3 3 4 6 12 9 3 8 8 1

To find the median, the values first need to be ordered: 1 3 3 3 4 6 8 8 8 9 11 12

As there is an even number of data values, the median would be the mean of the 6th and 7th values: $\frac{6+8}{2}$, or 7 diseased trees.

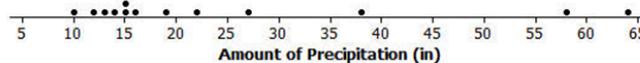
3. Describe how you would find the median for a set of data that has 35 values. How would this be different if there were 36 values?

Answers will vary; first you would order the data from smallest to largest. Because there are 35 values, you would look for the 18th value from the top or bottom. This would be the median with 17 values above and 17 values below. If the set had 36 values, you would go halfway between the 18th and 19th values.

Problem Set Sample Solutions

1. The amount of precipitation in the western states in the U.S. is given in the table as well as the graph.

State	Amount of Precipitation (in)
WA	38.4
OR	27.4
CA	22.2
MT	15.3
ID	18.9
WY	12.9
NV	9.5
UT	12.2
CO	15.9
AZ	13.6
NM	14.6
AK	58.3
HI	63.7



Data Source: <http://www.currentresults.com/Weather/US/average-annual-state-precipitation.php>

- a. How do the amounts vary across the states?

Answers will vary: The spread is pretty large: 54.2 inches. Nevada has the lowest at 9.5 inches per year. Hawaii, Alaska, and Washington have more rain than most of the states; Hawaii has the most with 63.7 inches followed by Alaska at 58.3 inches.

- b. Find the median. What does the median tell you about the amount of precipitation?

The median is 15.9 inches; half of the states have more than 15.9 inches of precipitation per year and half have less.

- c. Use the median and the range to describe the average monthly precipitation in western states in the U.S.

The amount of precipitation varies from 63.7 to 9.5 inches per year. Half of the states have from 9.5 to 15.9 inches per year, but only two have more than 40 inches.

- d. Do you think the mean or median would be a better description of the typical amount of precipitation? Explain your thinking.

The mean at 24.8 inches reflects the extreme values, while the median seems more typical at 15.9 inches.

2. Identify the following as true or false. If a statement is false, give an example showing why.

- a. The median is always equal to one of the values in the data set.

False. If the numbers are 1 and 5, the median is 3 and it is not in the set.

- b. The median is the midpoint between the smallest and largest values in the data set.

False. Look at the number of french fries per bag for Restaurant A above where the median is 82, which is not halfway between 66 and 93 (79.5).

- c. At most, half of the values in a data set have values less than the median.

True.

- d. In a data set with 25 different values, if you change the two smallest values of a data set to smaller values, the median will not be changed.

True.

- e. If you add 10 to every element of a data set, the median will not change.

False. The median will increase by 10 as well. If the set is 1, 2, 3, 4, 5, the median is 3; for the set 11, 12, 13, 14, 15, the median will be 13.

3. Make up a data set such that the following is true:

- a. The set has 11 different values and the median is 5.

Answers will vary depending on whether the numbers are whole numbers or fractions. If the numbers are whole numbers, the set would be 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

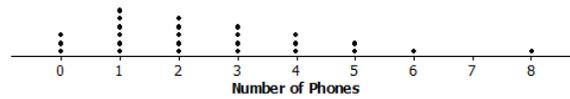
- b. The set has 10 values and the median is 25.

Answers will vary. One answer is to have all 25's.

- c. The set has 7 values and the median is the same as the smallest value.

Answers will vary. One answer is to have 1, 1, 1, 1, 2, 3, 4.

4. The dot plot shows the number of landline phones that a sample of people have in their homes.



a. How many people were in the sample?

25

b. Why do you think three people have no landline phones in their homes?

Possible answers: Some people might only have cell phones, or some people may not be able to afford a phone.

c. Find the median number of phones for the people in the sample.

The median number of phones per home is 2.

d. Use the median and the range (maximum-minimum) to describe the distribution of the number of phones.

Possible answer: The median number of phones was 2 per home, and over half of the people have fewer than 3 phones in their homes. Three had none, and one house had 8 phones.

5. The salaries of the Los Angeles Lakers for the 2012–2013 basketball season are given below.

Player	Salary (\$)
Kobe Bryant	\$27,849,149
Dwight Howard	\$19,536,360
Pau Gasol	\$19,000,000
Steve Nash	\$8,700,000
Metta World Peace	\$7,258,960
Steve Blake	\$4,000,000
Jordan Hill	\$3,563,600
Chris Duhon	\$3,500,000
Jodie Meeks	\$1,500,000
Earl Clark	\$1,240,000
Devin Ebanks	\$1,054,389
Darius Morris	\$962,195
Antawn Jamison	\$854,389
Robert Sacre	\$473,604
Darius Johnson-Odom	\$203,371

Data Source: www.basketball-reference.com/contracts/LAL.html

a. Just looking at the data, what do you notice about the salaries?

Possible answer: A few of the salaries for the big stars like Kobe are really big, while others are very small in comparison.

b. Find the median salary, and explain what it tells you about the salaries.

\$3,500,000 for Chris Duhon. Half of the players make more than \$3,500,000 and half make less than that.

c. Find the median of the lower half of the salaries and the median of the upper half of the salaries.

\$962,195 for the bottom half of the salaries; \$8,700,000 for the top half of the salaries.

- d. Find the width of each of the following intervals. What do you notice about the size of the interval widths, and what does that tell you about the salaries?
- minimum salary to median of the lower half: **\$758,824**
 - median of the lower half to the median of the whole set: **\$2,537,805**
 - median of the whole set to the median of the upper half: **\$5,200,000**
 - median of the upper half to the highest salary: **\$19,149,149**

The largest width is from the median of the upper half to the highest salary. The smaller salaries are closer together than the larger ones.

6. Use the salary table from above to answer the following.

- a. If you were to find the mean salary, how do you think it would compare to the median? Explain your reasoning.

Possible answer: The mean will be a lot larger than the median because when you add in the really big salaries, the size of the mean will increase a lot.

- b. Which measure do you think would give a better picture of a typical salary for the Lakers, the mean or the median? Explain your thinking.

Possible answer: The median seems better as it is more typical of most of the salaries.



Lesson 13: Describing Variability Using the Interquartile

Range (IQR)

Student Outcomes

- Given a set of data, students describe how the data might have been collected.
- Students describe the unit of measurement for observations in a data set.
- Students calculate the median of the data.
- Students describe the variability in the data by calculating the interquartile range.

Lesson Notes

Students should develop understanding of statistical variability, in particular recognizing that a measure of center for a numerical data set summarizes all of its values with a single number, while a measure of variation describes how the values in the data set vary.

Students are also expected to summarize numerical data sets in relation to their context. This is done by reporting the number of observations and describing the nature of the attribute under investigation. Students indicate how data were measured and the units of measurement. They provide quantitative measures of center (median) and variability (interquartile range), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered.

In this lesson students are engaged in making sense of problems and solving them. As the contexts change, they are asked to pay attention to the units for each. They are also engaged in working in groups, sharing their reasoning, and critiquing the reasoning of others as they create contexts that satisfy given constraints.

Classwork

The median was used to describe the typical value of our data in Lesson 12. Clearly, not all of the data is described by the value. How do we find a description of how the data vary? What is a good way to indicate how the data vary when we use a median as our typical value? These questions are developed in the following exercises.

Exercises 1–4 (15 minutes): More French Fries

MP.1

This exercise returns to the data from Lesson 12, raising questions about how the data might have been collected and whether any bias (the formal word is not used) might be inherent in the process. Students examine work from Lesson 12 finding the medians of the lower half and upper half of the data and use them to calculate the interquartile range for the three data sets.

Be sure that students understand that a quartile is the cut off point for the median of either the top half or bottom half of the data and not the length of the segment from the minimum (maximum) to the quartile. They should be able to

approximate the number of elements in each section in terms of $\frac{1}{4}$ or 25% of the data or by giving an estimate of the actual number data values as well as knowing that $\frac{1}{2}$ or 50% of the data values are between the two quartiles. They should also recognize that the IQR is a measure of spread around the median (the length of the interval that captures the middle 50% of the data).

Consider the following questions as you discuss the example questions with students:

- Approximately how many data values would be between the quartiles? Explain your reasoning.
 - *Answers will vary depending on the data.*
- What other measures of center and spread have you studied and how do you think they would compare to the median and IQR? (If you want to pursue this question, you could have students compute the mean and MAD for one of the data sets.)
 - *This question provides students an opportunity to discuss what they might recall about the mean as a center and the MAD as a measure of variability.*

Data from Lesson 12: Number of french fries

Restaurant A: 80, 72, 77, 80, 90, 85, 93, 79, 84, 73, 87, 67, 80, 86, 92, 88, 86, 88, 66, 77

Restaurant B: 83, 83, 83, 84, 79, 78, 80, 81, 83, 80, 79, 81, 84, 82, 85, 85, 79, 79, 83

Restaurant C: 75, 75, 77, 85, 85, 80, 80, 80, 80, 81, 82, 84, 84, 84, 85, 77, 77, 86, 78, 78, 79, 79, 79, 79

Exercises 1–4

1. In Lesson 12, you thought about the claim made by a chain restaurant that the typical number of french fries in a large bag was 82. Then you looked at data on the number of fries in a bag from three of the restaurants.
 - a. How do you think the data was collected and what problems might have come up in collecting the data?

Answers will vary: They probably went to the restaurants and ordered a bunch of bags of french fries. Sometimes the fries are broken, so they might have to figure out what to do with those - either count them as a whole, discard them, or put them together to make whole fries.
 - b. What scenario(s) would give counts that might not be representative of typical bags?

Answers will vary: Different workers might put different amounts in a bag, so if you took the sample at lunch, you might have different numbers than if you did it in the evening. The restaurants might weigh the bags to see that the weight was constant despite the size of the fries, so you could have the same weight of fries even though you had different counts for the bags.
2. In Exercise 7 of Lesson 12 you found the median of the top half and the median of the bottom half of the counts for each of the three restaurants. These were the numbers you found: Restaurant A – 87.5 and 77; Restaurant B – 82 and 79; Restaurant C – 84 and 78. The difference between the medians of the two halves is called the interquartile range or IQR.
 - a. What is the IQR for each of the three restaurants?

Restaurant A is 10.5; Restaurant B is 3; Restaurant C is 6.
 - b. Which of the restaurants had the smallest IQR, and what does that tell you?

Restaurant B had the smallest IQR. This indicates that the spread around the median number of fries is smaller than for either of the other two restaurants. About half of the data are within 3 fries of the median, so the median is a pretty good estimate of what is typical.

c. About what fraction of the counts would be between the quartiles? Explain your thinking.

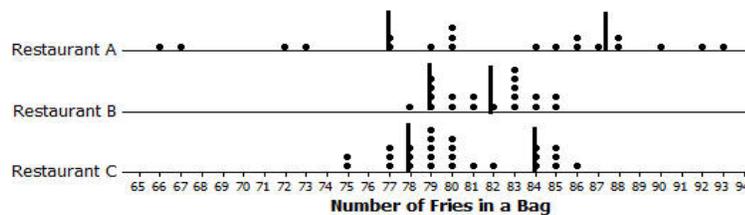
About $\frac{1}{2}$ or 50% of the counts would be between the quartiles because about $\frac{1}{4}$ of the counts are between the median and the lower quartile, and $\frac{1}{4}$ of the counts are between the median and upper quartile.

3. The medians of the lower and upper half of a data set are called quartiles. The median of the top half of the data is called the upper quartile; the median of the bottom half of the data is called the lower quartile. Do these names make sense? Why or why not?

Answers will vary: Students might say that quartile is related to quarter, and the lower quartile, the median, and the upper quartile divide the data into four sections with one fourth or a quarter of the data values in each section.

4.

a. Mark the quartiles for each restaurant on the graphs below.



b. Does the IQR help you decide which of the three restaurants seems most likely to really have 82 fries in a typical bag? Explain your thinking.

Restaurant B has the smallest IQR, which means that the middle half of the counts of the number of fries in a bag is really close to the median. Restaurant B also has the smallest range.

Example 1 (5 minutes): Finding the IQR

The example is intended as a reference for students and not to be reproduced during class unless students need more clarity. If that is the case, have students explain the diagram to each other rather than recreating it on the board. Make sure that students understand that the lower quartile is the number marking the median of the lower half.

Example 1: Finding the IQR

Read through the following steps. If something does not make sense to you, make a note and raise it during class discussion. Consider the data: 1, 1, 3, 4, 6, 6, 7, 8, 10, 11, 11, 12, 15, 15, 17, 17, 17

Creating an IQR:

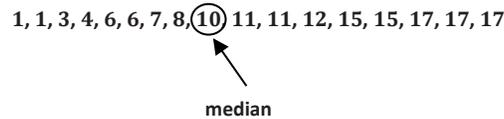
I. Order the data: The data is already ordered.

1, 1, 3, 4, 6, 6, 7, 8, 10, 11, 11, 12, 15, 15, 17, 17, 17

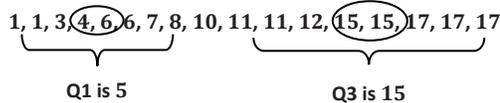
II. Find the minimum and maximum: The minimum data point is 1, and the maximum is 17.

① 1, 3, 4, 6, 6, 7, 8, 10, 11, 11, 12, 15, 15, 17, 17, ① 17

III. Find the median: There are 17 data points so the 9th one from the smallest or from the largest will be the median.



IV. Find the lower quartile and upper quartile: The lower quartile (Q1) will be half way between (the mean) the 4th and 5th data points (4 and 6), or 5 and the upper quartile (Q3) will be half way between the 13th and the 14th data points (15 and 15), or 15.



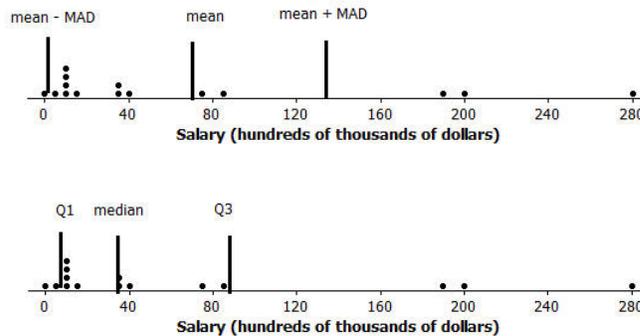
V. Find the difference between Q3 and Q1: The $IQR = 15 - 5 = 10$.

Exercise 5 (5 minutes)

This exercise points out how a skewed distribution might not be adequately summarized by the mean and its corresponding measure of spread, the MAD.

Exercises 5–6

5. When should you use the IQR? The data for the 2012 salaries for the Lakers basketball team are given in the two plots below (see problem 5 in the Problem Set from Lesson 12).



a. The data are given in hundreds of thousands of dollars. What would a salary of 40 hundred thousand dollars be?

\$4,000,000

b. The vertical lines on the top plot show the mean and the mean \pm the MAD. The bottom plot shows the median and the IQR. Which interval is a better picture of the typical salaries? Explain your thinking.

The median and the IQR seem to represent the typical salaries better than the mean \pm the MAD. The mean salary is above all but five of the salaries.

Exercise 6 (15–20 minutes): On Your Own with IQRs

Students should work together in pairs or groups of three on this exercise. After they develop three examples, encourage them to select one of their examples and explain it more fully by creating a simple poster. If a poster is used for this exercise, indicate to students that they should do the following:

- Draw or write out a specific context for each example that explains the data and how it would be collected.
- Explain on the poster how to find the median, the upper quartile, and the lower quartile.
- Explain what the IQR would mean for the context of the selected example.

Either discuss student examples or design a presentation of the posters.

Encourage students to use their imaginations and identify contexts for which the IQR might provide useful information. If time permits, have students explain their poster to other students. Display posters as possible examples of problems for future lessons or discussions.

6. Create three different contexts for which a set of data collected related to those contexts could have an IQR of 20. Define a median for each context. Be specific about how the data might have been collected and the units involved. Be ready to describe what the median and IQR mean in each case.

Examples that could be included in this exercise are as follows: number of books read by students during a school year (some students read a lot of books, while other students may not read as many), number of movies viewed at a theater during the last year by students in a class, number of text messages students receive during a specific day (for example, on Monday), number of commercials on TV during a specific time period that are about buying a car, number of different states students have visited, number of healthy trees on certain blocks of a city, and number of students in each classroom of a school during a specific time period. Remind students that the goal is to have them think of data that if collected might have an IQR of approximately 20. These ideas also allow students to start thinking of the process of actually collecting data that is needed later.

Closing (2–3 minutes)**Lesson Summary**

One of our goals in statistics is to summarize a whole set of data in a short concise way. We do this by thinking about some measure of what is typical and how the data are spread relative to what is typical.

In earlier lessons, you learned about the MAD as a way to measure the spread of data about the mean. In this lesson, you learned about the IQR as a way to measure the spread of data around the median.

To find the IQR, you order the data, find the median of the data, and then find the median of the lower half of the data (the lower quartile) and the median of the upper half of the data (the upper quartile). The IQR is the difference between the upper quartile and the lower quartile, which is the length of the interval that includes the middle half of the data, because the median and the two quartiles divide the data into four sections, with about $\frac{1}{4}$ of the data in each section. Two of the sections are between the quartiles, so the interval between the quartiles would contain about 50% of the data.

Small IQRs indicate that the middle half of the data are close to the median; a larger IQR would indicate that the middle half of the data is spread over a wider interval relative to the median.

Exit Ticket (5 minutes)



Name _____

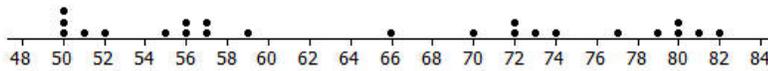
Date _____

Lesson 13: Describing Variability Using the Interquartile Range (IQR)

Exit Ticket

1. On the graph below, insert the following words in approximately the correct position.

Maximum Minimum IQR Median Lower Quartile (Q1) Upper Quartile (Q3)



2. Estimate the IQR based on the data set above.

Exit Ticket Sample Solutions

1. On the graph below, insert the following words in approximately the correct position.

Maximum Minimum IQR Median Lower Quartile (Q1) Upper Quartile (Q3)

2. Estimate the IQR based on the data set above.

The IQR is approximately 22.

Problem Set Sample Solutions

1. The average monthly high temperatures (in °F) for St. Louis and San Francisco are given in the table below.

	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
St. Louis	40	45	55	67	77	85	89	88	81	69	56	43
San Francisco	57	60	62	63	64	67	67	68	70	69	63	57

Data Source: www.weather.com/weather/wxclimatology/monthly/graph/USCA0987
www.weather.com/weather/wxclimatology/monthly/graph/USMO0787

a. How do you think the data might have been collected?

Someone at a park or the airport or someplace probably records the temperature every hour of every day and then takes all of the highest ones and finds the mean.

b. Do you think it would be possible for $\frac{1}{4}$ of the temperatures in the month of July for St. Louis to be 95° or above? Why or why not?

Yes, it is possible. The mean temperature in St. Louis for July is 89°. There are 31 days in July, so $\frac{1}{4}$ of the days would be about 8 days. If the temperature was 95° for 5 days, 100° for 3 days, and 87° for all of the rest of the days, it would work.

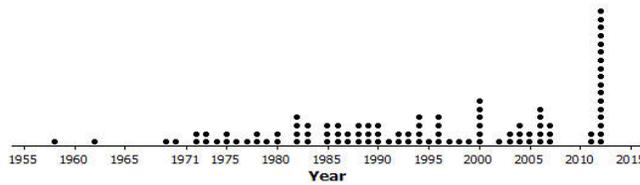
c. Make a prediction about how the sizes of the IQR for the temperatures for each city compare. Explain your thinking.

San Francisco probably has the smaller IQR because those temperatures don't seem to vary as much as the St. Louis temperatures.

d. Find the IQR for the average monthly high temperature for each city. How do the results compare to your conjecture?

For San Francisco the IQR is 6.5°; St. Louis is 33°.

2. The plot below shows the years in which each of 100 pennies were made.



a. What does the stack of 17 dots at 2012 representing 17 pennies tell you about the “age” of the pennies in 2014?

17 pennies were made in 2012, and they would be 2 years old in 2014.

b. Here is some information about the sample of pennies. The mean year they were made is 1994; the first year any of the pennies were made was 1958; the newest pennies were made in 2012; Q1 is 1984, the median is 1994, and Q3 is 2006; the MAD is 11.5 years. Use the information to indicate the years in which the middle half of the pennies was made.

In this case, the IQR is 22 years and the mean ± the MAD gives an interval of 23, so the middle half of the pennies were made over an interval of 22 years.

3. Create a data set with at least 6 elements such that it has the following:

a. A small IQR and a big range (maximum-minimum).

Answers will vary: {0, 100, 50, 50, 50, 50} where the range is 100 and the IQR is 0.

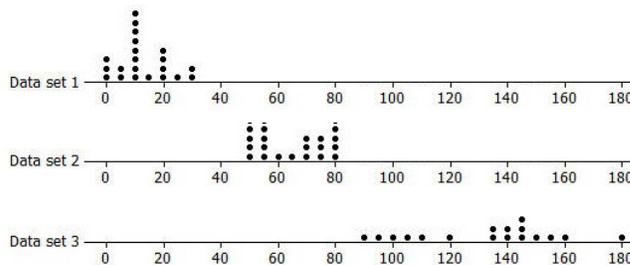
b. An IQR equal to the range.

Answers will vary: {10, 10, 10, 15, 20, 20, 20}

c. The lower quartile is the same as the median.

Answers will vary: {1, 1, 1, 1, 1, 5, 6, 7}

4. Rank the following three data sets by the value of the IQR.



Data set 1 has the smallest IQR at about 14, data set 2 the next smallest at about 22, and data set 3 the largest at about 41. (Be sure students do not confuse data set with the larger quartiles as having the larger IQR.)

5. Here are the counts of the fries in each of the bags from Restaurant A:

80, 72, 77, 80, 90, 85, 93, 79, 84, 73, 87, 67, 80, 86, 92, 88, 86, 88, 66, and 77.

a. Suppose one bag of fries had been overlooked in the sample and that bag had only 50 fries. Would the IQR change? Explain your reasoning.

The IQR would be larger, 12.5, because the median number of fries would be at 80 now instead of 82, which would make the lower quartile at 75 instead of 77.

b. Will adding another data value always change the IQR? Give an example to support your answer.

No, it depends on how many values you have in the data set. For example, if the set of data is {2, 2, 2, 6, 9, 9, 9}, the IQR is 9 – 2 = 7. If you add another 6, the IQR would stay at 7.



Lesson 14: Summarizing a Distribution Using a Box Plot

Student Outcome

- Students construct a box plot from a given set of data.

Lesson Overview

In this lesson, students transition from using dot plots to display data to using box plots. The lesson begins with exercises that lay the foundation for the development of a box plot. Students inspect dot plots of several sets of data and think about how to group or section the plots to get a sense of the span of data values in each of the sections. When individual students determine how to make the sections, the results differ and the process seems arbitrary and inconsistent. Thus, there is a need for a standard procedure for making a box plot. Using the median and the quartiles introduced in the previous lesson seems like a good strategy.

The lesson begins and ends with interactive activities. If time allows, students will create a “human box plot” of the time it took them to come to school. Supplies will be needed for this exercise and are listed in the teacher notes.

Classwork

A box plot is a graph that is used to summarize a data distribution. What does the box plot tell us about the data distribution? How does the box plot indicate the variability of the data distribution?

Example 1 (5–7 minutes): Time to Get to School

MP.3

The questions are designed to help students begin to think about grouping data in order to get a sense of the spread of the data values within sections of the data. Let students write an estimate of the time it took them to come to school on a post-it note. Teachers may want the individual students to place their post-it note on a dot plot that is displayed on the classroom board at the beginning of class or graph the dot plot as a class.

Example 1: Time to Get to School

What is the typical amount of time it takes for a person in your class to get to school? The amount of time it takes to get to school in the morning varies for each person in your class. Take a minute to answer the following questions. Your class will use this information to create a dot plot.

Write your name and an estimate of the number of minutes it took you to get to school today on a post-it note.

Answers will vary.

What were some of the things you had to think about when you made your estimate?

Answers will vary: Does it count when you have to wait in the car for your sister? Usually I walk, but today I got a ride. Does it matter that we had to go a different way because the road was closed? The bus was late.

As students discuss and complete the questions in this example; additionally, ask the following questions:

- What does a dot on the dot plot represent?
- What is an estimate of the median? What is the typical amount of time it takes someone to get to school?
- What are the minimum and maximum values?

Exercises 1–4 (7–10 minutes)

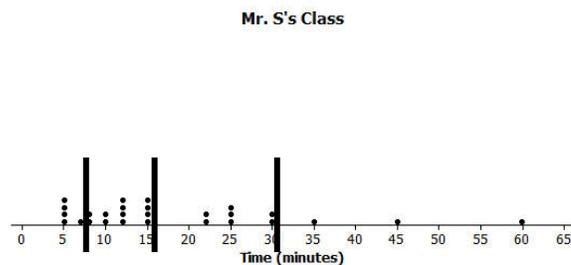
Let students work individually on the exercises and compare their plots with their neighbors. Students should recognize that their divisions of the data into four sections are close but not the same.

If time permits, bring them together for a discussion of their answers and stress the idea that it is useful to see how the data values group together in intervals across the entire distribution. Ask the students the following questions:

- Are there a lot of values in the middle or at one end?
- What was the shortest time a student in Mr. S's class got to school? The longest?
- Looking at the plot, how long does it seem to take a typical student in Mr. S's class to get to school?

Exercises 1–4

Here is a dot plot of the estimates of the times it took students in Mr. S's class to get to school one morning.



1. Put a line in the dot plot that seems to separate the shortest times and the longest times.

Some might put the dividing line between 15 and 20.

2. Put another line in the plot that separates those who seem to live really close to school and one that marks off those who took a long time to get to school.

Responses will be different. Some might put a line at 30 and a line at 10.

3. Your plot should be divided into four sections. Record the number of times in each of the four sections.

Answers will vary: Depending on the divisions, 7 or 8 in the lower one, 9 in the next, 5 in the next, and 5 in the upper section.

4. Share your marked up dot plot with some of your classmates. Compare how each of you divided the plot into four sections.

Different responses; students should recognize that the divisions might be close but are different.

Exercises 5–7 (10 minutes): Time to Get to School

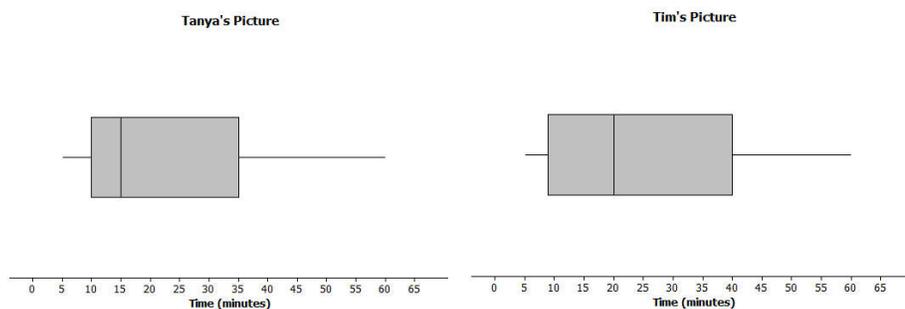
Let students work individually on the questions. Then discuss and confirm answers. Start to get students thinking about using the *5-number summary* (or the *minimum*, lower quartile or *Q1*, *median*, upper quartile or *Q3*, and the *maximum*) and how it relates to the construction of the box plot. Ask students the following questions:

- How could the quartiles and the median be helpful in making a plot that summarized the data?
- Where do you think the median is on Tanya’s and Tim’s plots?
- Where are the lower and upper quartiles? (*Q1* and *Q3*)?
- Why are their values different?

Exercises 5–7: Time to Get to School

The teacher asked the class to make a representation that would summarize the times it took students in Mr. S’s class to get to school and how they are spread out. Tim decided to get rid of the dots and just use a picture of the divisions he made of the shortest times and the longest times. He put a box around the two middle sections.

Tanya thought that was a good idea and made a picture of the way she had divided the times. Here are their pictures.



5. What do the pictures tell you about the length of time it takes the students to get to school?

Answers will vary: The fastest times for getting to school were from 3 or 4 minutes to 10 minutes, and the longest time was around 62 minutes. A bunch of students took from 10 to 35 or 40 minutes to get to school. It kind of looks like most students live closer to school – or at least got to school in a short time.

6. What don't the pictures tell you about the length of time it takes the students to get to school?

The pictures don't tell how many students were in the class or even how many students were in any of the sections.

7. How do the two pictures compare?

The pictures are pretty close, but Tim had a longer box with a group from 10 to 20, and Tanya had a shorter box with a group from 10 to 15. Their next groups were different too.

Example 2 (7–10 minutes): Making a Box Plot

This example defines the procedure for finding a box plot, referring back to Lesson 13 on quartiles and the IQR. You may want students to read through the process themselves and then ask them what the directions mean, or have the whole class work through the process together.

Ask the students the following when the box plot is complete:

- Why is it important to have a standard way to make a box plot?
- What does a box plot tell you about the *story* in the data? What does it not tell you?
- What proportion (percent) of the data is in each of the sections of the box plot? How do you know?

Example 2: Making a Box Plot

Mr. S suggested that to be sure everyone had the same picture, statisticians developed a standard procedure for making the cut marks for the sections.

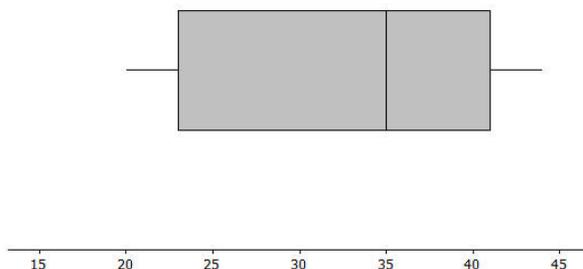
Mr. S. wrote the following on the board:

To make a box plot

- Find the median of all of the data
- Find Q1, the median of the bottom half of the data, and Q3, the median of the top half of the data.
- Draw a box that goes from Q1 to Q3, the two middle sections.
- Draw a line segment connecting the minimum value to the box and one that connects the maximum value to the box.

Now use the given number line to make a box plot of the data below.

20, 21, 25, 31, 35, 38, 40, 42, 44



The 5-number summary is as follows:

Min = 20
 Q1 = 23
 Median = 35
 Q3 = 41
 Max = 44

Exercises 8–11 (10–12 minutes): A Human Box Plot

Depending on the amount of time left in class, this exercise can either be completed on the board as a class or by using selected students to create a Human Box Plot. If possible, try to involve approximately 12 students. This would put 3 students in each quartile. If time is running short and the problem is completed as a class, use the focus questions listed below.

Preparation for Human Box Plot:

The data is already ordered from the post-it notes on the dot plot on the classroom board. Call out students' names to have them form an ordered line of data. Find a place in the classroom (or hall) that will allow all of the students to line up and that can accommodate a number line that will go from 0 to as large as 60 minutes. Have some props that can locate a number line – large cards with a scale in 5-minute intervals can work. Do not start with the number line, however. A ball of yarn or rope can be used to mark off the *box* part of the box plot.

Once students are in line, have them identify the median and the two quartiles. Give the signs for the five summary values to the appropriate students and ask them to step out with their signs. Ask each of the quartiles to hold one end of the rope marking off the *box* that extends from Q1 to Q3. Students may not recognize at first that the plot has no scale and thus does not really tell the story. Try to get them to see how important the scale is by asking questions such as the following:

- You all are a human box plot of the times it took you to come to school. Did it take most of you a short time or a longer time?
- Did it take anyone a really long time?
- Can you tell from our plot?

When students realize that it looks like the times were all evenly spaced because of how they are standing, bring out the props for the number line (be sure to make the intervals wide enough to accommodate several students). Then have students rearrange themselves using the scale and recreate the box plot with the five number summary values and the rope to represent the box. Then, ask the following:

- How many people are in each of the sections? Which section has the most people? The fewest?
- Were there any sections where the people were all crowded together? How did this show up in the box plot?
- Why do we need a scale to make a box plot?

Exercises 8–11: A Human Box Plot

Consider again your post-it note that you used to write down the number of minutes it takes you to get to school. If possible, you and your classmates will form a human box plot of the number of minutes it takes your class to get to school.

8. Find the median of the group. Does someone represent the median? If not, who is the closest to the median?

Answers will vary.

9. Find the maximum and minimum of the group. Who are they?

Answers will vary.

10. Find Q1 and Q3 of the group. Does anyone represent Q1 or Q3? If not, who is the closest to Q1? Who is closest to Q3?

Answers will vary.

11. Sketch the box plot for this data set below.

Answers will vary.

Closing (2–3 minutes)

Lesson Summary

The focus of this lesson is moving from a plot that shows all of the data values (dot plot) to one that summarizes the data with five points (box plot).

You learned how to make a box plot by doing the following:

- Finding the median of all of the data
- Finding Q1, the median of the bottom half of the data, and Q3, the median of the top half of the data.
- Drawing a box that goes from Q1 to Q3, the two middle sections.
- Drawing a line segment connecting the minimum value to the box and one that connects the maximum value to the box.

You also learned important characteristics of box plots:

- $\frac{1}{4}$ of the data are in each of the sections of the plot.
- The length of the interval for a section does not indicate either how the data are grouped in that interval or how many values are in the interval.

Exit Ticket (5 minutes)

Name _____

Date _____

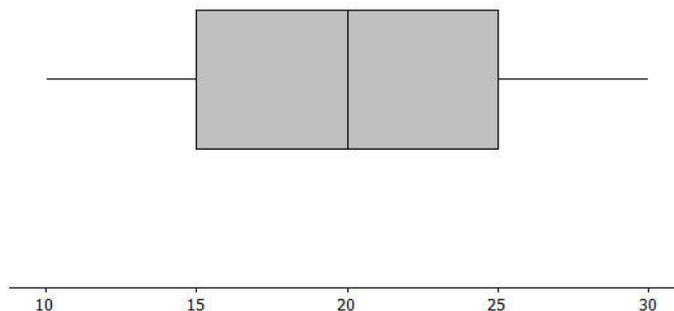
Lesson 14: Summarizing a Distribution Using a Box Plot

Exit Ticket

Sulee explained how to make a box plot to her sister as follows:

“First you find the smallest and largest values and put a mark halfway between them, and then put a mark halfway between that mark and each end. So, if 10 is the smallest value and 30 is the largest value, you would put a mark at 20. Then another mark belongs half way between 20 and 10, which would be at 15. And then one more mark belongs half way between 20 and 30, which would be at 25. Now, you put a box around the three middle marks and draw lines from the box to the smallest and largest values.”

Here is her box plot. What would you say to Sulee?

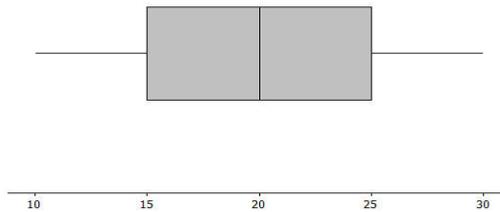


Exit Ticket Sample Solutions

Sulee explained how to make a box plot to her sister as follows:

“First you find the smallest and largest values and put a mark halfway between them, and then put a mark halfway between that mark and each end. So, if 10 is the smallest value and 30 is the largest value, you would put a mark at 20. Then another mark belongs half way between 20 and 10, which would be at 15. And then one more mark belongs half way between 20 and 30, which would be at 25. Now, you put a box around the three middle marks and draw lines from the box to the smallest and largest values.”

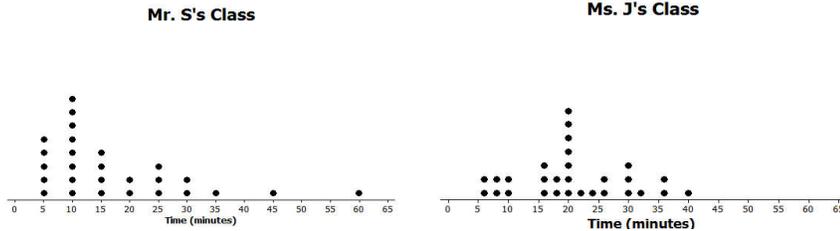
Here is her box plot. What would you say to Sulee?



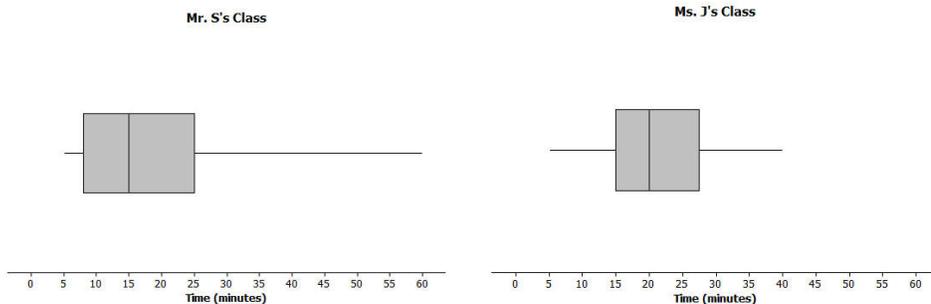
Sulee is wrong. This is not the correct way to create a box plot. Sulee did not find the median or the quartiles using the data values; she just divided up the length between the smallest and largest numbers.

Problem Set Sample Solutions

1. Dot plots for the amount of time it took students in Mr. S’s and Ms. J’s classes to get to school are below.



a. Make a box plot of the times for each class.

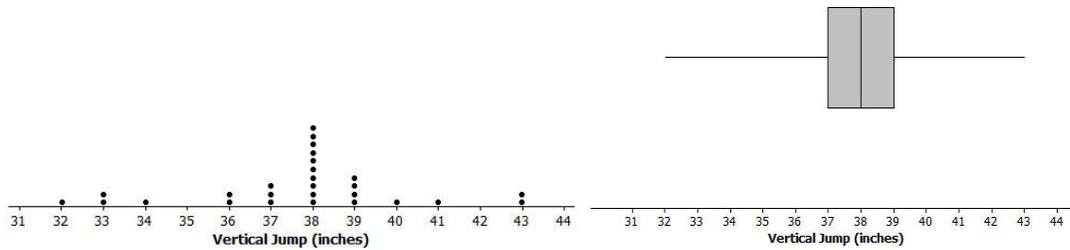


Mr. S five summary values: 5, 10, 15, 25, 60 and Ms. J five summary values: 5, 15, 20, 27.5, 40

- b. What is one thing you can see in the dot plot that you cannot see in the box plot? What is something that is easier to see in the box plot than in the dot plot?

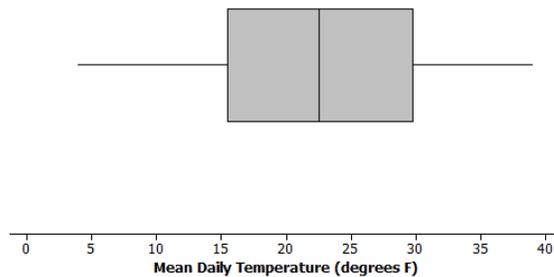
The dot plot shows individual times which you cannot see in the box plot. The box plot shows the location of the median and of the lower and upper quartiles.

2. The dot plot below shows the vertical jump of some NBA players. A vertical jump is how high a player can jump from a standstill. Draw a box plot of the heights for the vertical jumps of the NBA players above the dot plot.



Five summary values: 32, 37, 38, 39, 43

3. The mean daily temperatures in °F for the month of February for a certain city are as follows:
 4, 11, 14, 15, 17, 20, 30, 23, 20, 35, 35, 31, 34, 23, 15, 19, 39, 22, 15, 15, 19, 39, 22, 23, 29, 26, 29, 29
- a. Make a box plot of the temperatures.



5-summary values: 4, 16, 22.5, 29.5, 39

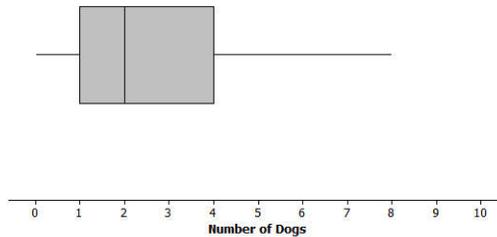
- b. Make a prediction about the part of the United States you think the city might be located. Explain your reasoning.

Answers will vary: The city was probably somewhere in the northern states, either Midwest or northeast or maybe Montana or Wyoming, because the temperatures are pretty cold.

- c. Describe the data distribution of temperature. Include a description of the center and spread.

The IQR is 29.5° – 16°, or 13.5°. Half of temperatures were near the middle between 16° and 29.5°. The median is 22.5°. A quarter of the temperatures are less than 16 but greater than or equal to 4°. A quarter of the temperatures are greater than 29.5° and less than or equal to 39°.

4. The plot below shows the results of a survey of households about the number of dogs they have. Identify the following statements as true or false. Explain your reasoning in each case.



- a. The maximum number of dogs per house is 8.
True because the line segment at the top goes to 8.
- b. At least $\frac{1}{2}$ of the houses have 2 or more dogs.
True because 2 is the median.
- c. All of the houses have dogs.
False because the lower line segment starts at 0 so at least one household does not have a dog as a pet.
- d. Half of the houses surveyed have between 2 and 4 dogs.
False because only about 25% of the houses would have between 2 and 4 dogs.
- e. Most of the houses surveyed have no dogs.
False because at least $\frac{3}{4}$ of those surveyed had 1 or more dogs.



Lesson 15: More Practice with Box Plots

Student Outcomes

- Given a box plot, students summarize the data by the 5-number summary (Min, Q1, Median, Q3, Max.)
- Students describe a set of data using the 5-number summary and the interquartile range.
- Students construct a box plot from a 5-number summary.

Lesson Overview

In this lesson, students are expected to summarize and describe distributions. They consider data displayed in dot plots and box plots and summarize data sets in relation to their context, describing center and spread using 5-number summaries (Minimum, Q1, Median, Q3, and Maximum) and the interquartile range. The questions in this lesson focus students on how box plots provide information about variability in a distribution.

Students begin by looking at a box plot, finding the 5-number summary, and using the 5-number summary to describe the data. They then consider the variability in two different data sets, the maximum speeds of selected birds and of land animals, with very different spreads. They create box plots using the 5-number summary for each set. In the last example, students interpret the IQR for different data sets.

To help students make sense of box plots and to confront typical misconceptions they may have, it would be very valuable to engage students with an interactive dynamic file that allows them to explore the relation between box plots and dot plots. One such example is the activity *Introduction to Boxplots* available at <http://education.ti.com/calculators/timathnspired/US/Activities/?sa=5026&t=1190>.

The file can be played using TI™-Nspire handhelds, TI™-Nspire software or on the TI™-Nspire Player, which is free and can be downloaded from <http://education.ti.com/calculators/products/US/document-player/>

Classwork

You reach into a jar of Tootsie Pops. How many Tootsie Pops do you think you could hold in one hand? Do you think the number you could hold is greater than or less than what other students can hold? Is the number you could hold a typical number of Tootsie Pops? This lesson examines these questions.

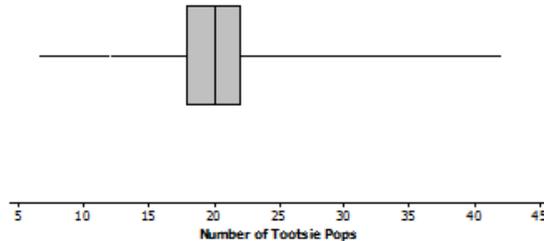
Example 1 (2–3 minutes): Tootsie Pops

You may want to actually do this experiment with students. Have them see how many Tootsie Pops they can grab and then replace the data in the example with the class data. (The data in the example were not collected from sixth grade students, so results might be different as hand sizes might be larger as students get older.) In later grades, data from the size of hand spans could be used to see whether any correlation exists between hand span and the number of Tootsie Pops someone can hold.

Example 1: Tootsie Pops

As you learned earlier, the five numbers that you need to make a box plot are the minimum, the lower quartile, the median, the upper quartile, and the maximum. These numbers are called the 5-number summary of the data.

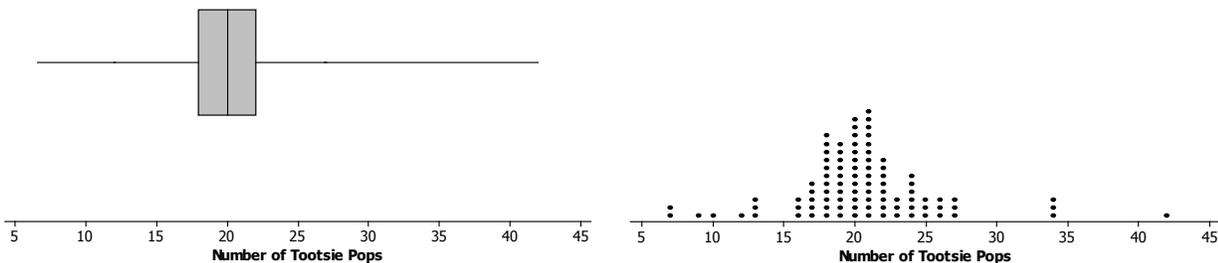
Ninety-four people were asked to grab as many Tootsie Pops as they could hold. Here is a box plot for these data. Are you surprised?



Exercises 1–5 (6–8 minutes)

As students work with the exercises, they should recognize that at least one person was able to hold a lot of Tootsie Pops because the upper segment extends to 42. They should also note that the typical number was pretty close to the median; the box that contains about half of the values only spans about four numbers. The second plot of the data informally introduces the concept of outliers as data values that seem far away from all of the others; you may want to use the term without giving a formal definition.

Technology can make the transition very visible from a dot plot to a box plot.



Note: An example on software that provides this visual transition is TI™ Nspire software. With this software, the dot plot is changed to a box plot as the dots “climb” into their place in the box plot. Going back and forth between plots several times will give students a very visual impression of what a box plot represents. Being able to simultaneously look at the two plots allows teachers to ask, “Why is the segment on the right so long? What do you think would happen if the point at 42 were removed or had been a 35?” and other questions that probe at student understanding. Interactive dynamic technology allows students to make conjectures and actually test them out by moving points and observing the consequences (See the *Mathematical Education of Teachers, Edition 2* from the Conference Board of Mathematical Sciences).

As students work through the exercises in small groups, ask them the following questions:

- How many Tootsie Pops do you think people can hold in one hand? Make a prediction.
 - Record students’ estimates for this question. If possible, demonstrate for students.

- How do you find the upper and lower quartiles?
 - Summarize with students the process they addressed previously as they order the data, find the median of the ordered data, and then find the middle of the upper half and the lower half as the upper and lower quartiles.
- What do we mean by a 5-number summary?
 - The 5-number summary refers to the following: the lowest or minimum data value, the lower quartile (Q1), the median, the upper quartile (Q3), and the maximum data value.
- About what fraction of the data values should be in each section of the box plot?
 - Approximately $\frac{1}{4}$ or 0.25 or 25% should be found in each section.

Exercises 1–5

1. What might explain the variability in how many Tootsie Pops those 94 people were able to hold?

Answers will vary: Size of people’s hands, hand span, whether they are flexible in moving their fingers.

2. Estimate the values in the 5-number summary from the box plot.

Min = 7, Q1 = 18, Median = 20, Q3 = 22, Max = 42

3. Describe how the box plot can help you understand the difference in the number of Tootsie Pops people could hold.

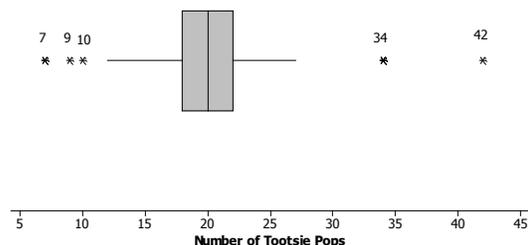
The maximum of about 42 and minimum of about 7 show you the range of 35 Tootsie Pops. The box shows that about half of the people can hold about 2 more or 2 fewer Tootsie Pops than the median, which was 20 Tootsie Pops. The box plot shows the overall spread, the bottom half, and the middle half of the number of Tootsie Pops people can hold.

4. Here is Jayne’s description of what she sees in the plot. Do you agree or disagree with her description? Explain your reasoning.

“One person could hold as many as 42 Tootsie Pops. The number of Tootsie Pops people could hold was really different and spread about equally from 7 to 42. About one half of the people could hold more than 20 Tootsie Pops.”

You cannot really tell that they are evenly spread – the box contains about half of the numbers of Tootsie Pops. However, the box is only four units long. That means half of the people were bunched over those four numbers.

5. Here is a different plot of the same data on the number of Tootsie Pops 94 people could hold.



a. Why do you suppose the five values are separate points and are labeled?

Maybe because they are far away from most of the other values. It shows that more than half of the data is from about 12 to 27 Tootsie Pops.

MP.6

MP.3

b. Does knowing these data values change anything about your responses to Exercises 1 to 4 above?
Not really, except maybe to say that only two of the people could hold a 34 and 42 Tootsie Pops; the rest held less than that.

Exercises 6–10 (15 minutes): Maximum Speeds

The intention in these exercises is not to compare the two data sets but rather to think about how the variability is different for birds and land animals. Note that two of the speeds are accurate to the hundredths place; the speed for the horse could have been clocked at a race track, but it is not clear how researchers were able to record such an accurate speed for the hummingbird. If you prefer, you could round the values for your students.

In working with any data set, a good habit is to start by looking over the values to see what might be unusual, different, or in some way interesting, which is the reason for the first question. When students describe the plots, encourage them to use fractions or percentages to talk about each of the four sections rather than “most” or “lots”, i.e., $\frac{1}{4}$ or 25% of the speeds were less than 76 mph.

Exercises 6–10: Maximum Speeds

The maximum speeds of selected birds and land animals are given in the tables below.

Bird	Speed (mph)
Peregrine falcon	242
Swift bird	120
Spine-tailed swift	106
White-throated needletail	105
Eurasian hobby	100
Pigeon	100
Frigate bird	95
Spur-winged goose	88
Red-breasted merganser	80
Canvasback duck	72
Anna's Hummingbird	61.06
Ostrich	60

Land Animal	Speed (mph)
Cheetah	75
Free-tailed bat (in flight)	60
Pronghorn antelope	55
Lion	50
Wildebeest	50
Jackrabbit	44
African wild dog	44
Kangaroo	45
Horse	43.97
Thomson's gazelle	43
Greyhound	43
Coyote	40
Mule deer	35
Grizzly bear	30
Cat	30
Elephant	25
Pig	9

Data Source: Natural History Magazine, March 1974, copyright 1974; The American Museum of Natural History; and James G. Doherty, general curator, The Wildlife Conservation Society; <http://www.thetravelalmanac.com/lists/animals-speed.htm>; http://en.wikipedia.org/wiki/Fastest_animals

As students answer the exercises, ask the following questions individually or in small groups to help students connect their work to the outcomes:

- The top recorded speed for a human is 27.79 mph for Usain Bolt during a 100-meter sprint in 2009. How does the human compare to the other land animals?
 - *One of the fastest human speeds would be similar to the fastest speeds of elephants and wild cats.*

6. As you look at the speeds, what strikes you as interesting?

Some might suggest birds are really fast, especially the falcon. Others may notice that only two of the speeds have decimals. The speeds of specific animals might strike students as interesting.

7. Do birds or land animals seem to have the greatest variability in speeds? Explain your reasoning.

It looks like the speeds of the birds vary a lot as they go from 60 mph for some birds to 242 mph for others. The speeds of the land animals vary, but not as much; they go from 9 mph to 75 mph.

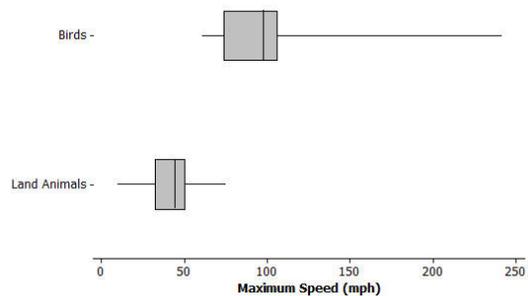
8. Find the 5-number summary for the speeds in each data set. What do the 5-number summaries tell you about the distribution of speeds for each data set?

Land Animal 5-number summary – Min = 9, Q1 = 32.5, Median = 43.97, Q3 = 50, Max = 75

Bird 5-number summary – Min = 60, Q1 = 76, Median = 97.5, Q3 = 105.5, Max = 242

The summaries give me a sense of the range or span of the speeds (Maximum – minimum speed) and how the speeds are grouped around the median.

9. Use the 5-number summaries to make a box plot for each of the two data sets.



10. Write several sentences to tell someone about the speeds of birds and land animals.

At least one bird flies really fast, the falcon at 242 mph. Three fourths of the birds fly less than 106 mph, and the slowest bird flies at 60 mph. The land animals' running speeds are slower going from 9 mph to 75 mph. The middle half of the speeds for land animals is between 32.5 mph and 50 mph.

Exercises 11–15: What is the Same and What is Different? (10 minutes)

The focus in thinking about the three box plots should be on the IQR for each, noting that the minimum, median, and maximum for each plot are the same. The spread of the middle half of the data is across the entire range (minimum to maximum) for the second plot. This could happen because the distribution is bimodal with the lower quartile and the minimum the same value and the upper quartile and the maximum the same value. The spread of the middle half of the data is much more tightly packed around the median in the third plot. Students estimate the quartiles from the plots; if their answers vary a bit that is okay because the emphasis is on the concept.

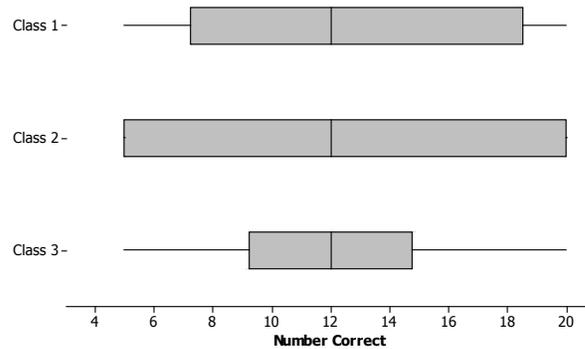
As students answer the questions for this exercise, ask the following questions individually or in small groups to help students connect their work to the outcomes:

- Are the IQR and the range enough to make a box plot? If not, what else do you need to know?
 - *No. We need to also know the median.*

- How many data values will be between the quartiles? Will there always be the same number in each quartile? Why or why not?
 - *Four data values are within each quartile. The number of data values within the quartiles will generally be the same as boundaries may include a data value. Consider discussing with students how many animals would be within each quartile if there were 18 animals or 19 animals or 20 animals. If one of the data values is the median, or two of the data values are the first quartile (Q1) and the third quartile (Q3), the values within the quartiles are the same. There are exceptions that students begin to see with the following examples.*

Exercises 11–15: What is the Same and What is Different?

Consider the following box plots, which show the number of questions different students in three different classes got correct on a 20-question quiz.



11. Describe the variability in the scores of the three classes.

The range (Max–Min) is the same for all three classes and so is the median, but the boxes that contain the middle half of the scores are spread very differently about the median. The third class has a small box so the scores are close together. In Class 2 the minimum and lower quartile are the same score, and the maximum and upper quartile are also the same score, so lots of scores are piled at the ends of the range. The middle half of the scores in Class 1 are spread out more than Class 3 but not as much as Class 2.

12.

a. Estimate the interquartile range for each of the three sets of scores.

Class 1 IQR = 10; Class 2 IQR = 15; Class 3 IQR = 5

b. What fraction of students does the interquartile range represent?

About one half.

c. What does the value of the IQR tell you about how the scores are distributed?

For Class 1, half of the scores are spread over an interval of width 10; for Class 2, half of the scores are spread out over an interval of width 15; and for Class 3, half of the scores are bunched together over an interval of width 5.

13. The teacher asked students to draw a box plot with a minimum value at 34 and a maximum value at 64 that had an interquartile range of 10. Jeremy said he could not draw just one because he did not know where to put the box on the number line. Do you agree with Jeremy? Why or why not?

The box could go anywhere from 34 to 44 all the way to from 54 to 64 and any width of 10 in between so Jeremy is correct.

14. Which class do you believe performed the best? Be sure to use the data from the box plots to back up your answer.

Class 3 as it has the smallest IQR. About half of the students scored close to the median score. Scores were more consistent for this class. Students may select Class 3 based on the smallest variability. Students might also make a case that although the variability is greater, approximately 25% of the students in Class 1 scored 18 or higher compared to 25% of the students in Class 3 scored 15 or higher. In Class 2, several students must have scored near the top for the Q3 and maximum to be the same. (Allow students to select the box plot they think answers the question and to describe why they selected the box plot.)

15.

- a. Find the IQR for the three data sets in the first two examples: maximum speed of birds, maximum speed of land animals, and number of Tootsie Pops.

Land Animals: 50 – 32.5 for an IQR of 17.5

Birds: 105.5 – 76 for an IQR of 29.5

Tootsie Pops: 22 – 18 for an IQR of 4

- b. Which data set had the highest percentage of data values between the lower quartile and the upper quartile? Explain your thinking.

All of the data sets should have about half of the data values between the quartiles.

Closing (2 minutes)

Lesson Summary

In this lesson, you learned about the 5-number summary for a set of data: minimum, lower quartile, median, upper quartile, and maximum. You made box plots after finding the 5-number summary for two sets of data (speeds of birds and land animals), and you estimated the 5-number summary from box plots (number of Tootsie Pops people can hold, class scores). You also found the interquartile range (IQR), which is the difference between the upper quartile and lower quartile. The IQR, the length of the box in the box plot, indicates how closely the middle half of the data is bunched around the median. (Note that because sometimes data values repeat and the same numerical value may fall in two sections of the plot, it is not always exactly half. This happened with the two speeds of 50 mph – one went into the top quarter of the data and the other into the third quarter – the upper quartile was 50.)

You also practiced describing a set of data using the 5-number summary, making sure to be as precise as possible—avoiding words like “a lot” and “most” and instead saying about one half or three fourths.

Exit Ticket (3 minutes)



Name _____

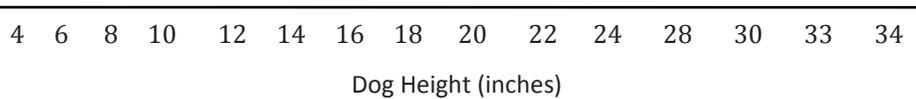
Date _____

Lesson 15: More Practice with Box Plots

Exit Ticket

Given the following information, create a box plot and find the IQR.

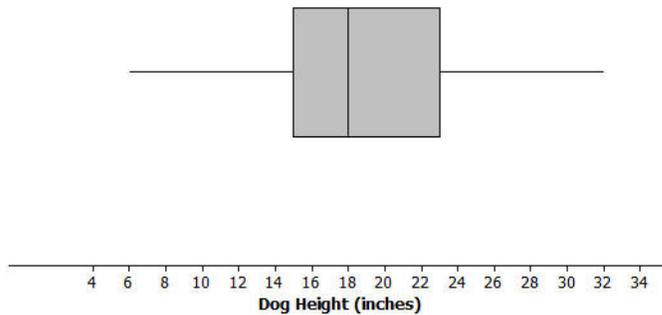
For a large group of dogs, the shortest dog was 6 inches, and the tallest was 32 inches. One half of the dogs were taller than 18 inches. One fourth of the dogs were shorter than 15 inches. The upper quartile of the dog heights was 23 inches.



Exit Ticket Sample Solutions

Given the following information, create a box plot and find the IQR.

For a large group of dogs, the shortest dog was 6 inches, and the tallest was 32 inches. One half of the dogs were taller than 18 inches. One fourth of the dogs were shorter than 15 inches. The upper quartile of the dog heights was 23 inches.

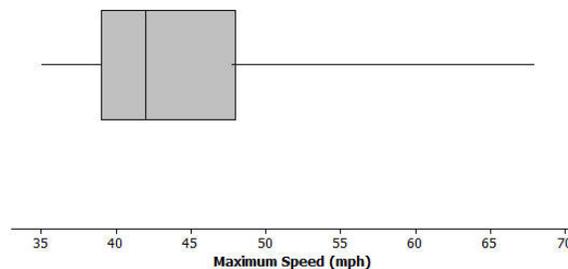


The IQR is $23 - 15 = 8$.

Problem Set Sample Solutions

All students should be encouraged to do problems 1 and 2 to be sure they understand the concepts developed in the lesson. Problem 4b should be discussed in some way as a whole class to raise awareness that medians are about counts and relative position of ordered data and not about distance or location.

1. The box plot below summarizes the maximum speeds of certain kinds of fish.



a. Estimate the 5-number summary from the box plot.

Answers will vary: Min – 35 mph; Q1 – 39 mph; Median – 42 mph; Q3 – 48 mph; Max – 68 mph.

b. The fastest fish is the sailfish at 68 mph followed by the marlin at 50 mph. What does this tell you about the spread of the fish speeds in the top quarter of the plot?

The Q3 is about at 48, so all but one of the top quarters are bunched between 48 and 50 mph.

c. Use the 5-number summary and the IQR to describe the speeds of the fish.

The speeds of fish vary from 35 mph to 68 mph. The IQR is 9 mph; the middle half of the speeds is between 39 mph and 48 mph. Half of the speeds are less than 42 mph.

Note: Data for box plot is provided below.

Fish	Maximum speed (mph)
Sailfish	68
Marlin	50
Wahoo	48
Tunny	46
Bluefin tuna	44
Great blue shark	43
Bonefish	40
Swordfish	40
Bonito	40
Four-winged flying fish	35
Tarpon	35

Data Source: <http://www.thetravelalmanac.com/lists/fish-speed.htm>

2. Suppose you knew that the interquartile range for the number of hours students spent playing video games during the school week was 10. What do you think about each of the following statements? Explain your reasoning.

a. About half of the students played video games for 10 hours during a school week.

This may not be correct as you know the width of the interval was 10, but you do not know where it starts or stops. You do not know the lower or upper quartile.

b. All of the students played at least 10 hours of video games during the school week.

This may not be correct for the same reason as in part (a).

c. About half of the class could have played video games from 10 to 20 hours a week or from 15 to 25 hours.

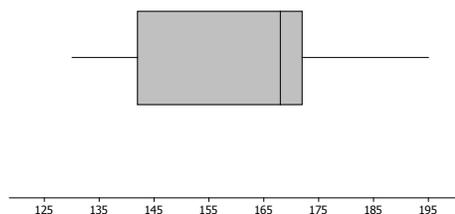
Either could be correct as the only information you have is the width of 10, and the statement says “could” not “is”.

3. Suppose you know the following for a data set: minimum value is 130, the lower quartile is 142, the IQR is 30, half of the data are less than 168, and the maximum value is 195.

a. Think of a context for which these numbers might make sense.

Answers will vary: The number of calories in a serving of fruit.

b. Sketch a box plot.



c. Are there more data values above or below the median? Explain your reasoning.

The number of data values on either side of the median should be about the same, one half of all of the data.

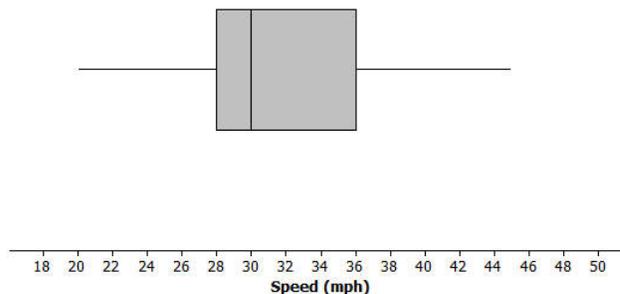
4. The speeds for the fastest dogs are given in the table below.

Breed	Speed (mph)
Greyhound	45
African Wild Dog	44
Saluki	43
Whippet	36
Basanji	35
German Shepherd	32
Vizsla	32
Doberman Pinscher	30

Breed	Speed (mph)
Irish Wolfhound	30
Dalmatian	30
Border Collie	30
Alaskan Husky	28
Giant Schnauzer	28
Jack Russell Terrier	25
Australian Cattle Dog	20

Data Source: <http://www.vetstreet.com/our-pet-experts/meet-eight-of-the-fastest-dogs-on-the-planet>;
<http://canidaepetfood.blogspot.com/2012/08/which-dog-breeds-are-fastest.html>

a. Find the 5-number summary for this data set and use it to create a box plot of the speeds.



Min = 20, Q1 = 28, Median = 30, Q3 = 36, Max = 45

b. Why is the median not in the center of the box?

The median is not in the center of the box because about $\frac{1}{4}$ of the speeds are between 30 and 36, and another $\frac{1}{4}$ are closer together, between 28 and 30. The data are skewed with lots of them at the lower speeds.

c. Write a few sentences telling your brother or sister about the speed of the fastest dogs.

Half of the dogs run faster than 30 mph; the fastest dog in the list is the greyhound with a speed of 45 mph. The slowest dog in the list is the Australian cattle dog. The middle 50% of the speeds are between 28 and 36 mph.



Lesson 16: Understanding Box Plots

Student Outcomes

- Students summarize a data set using box plots, the median, and the interquartile range.
- Students use box plots to compare two data distributions.

Lesson Notes

The activities in this lesson engage students in thinking about everything they learned in the last several lessons about summarizing and describing distributions of data using counts and position relative to the other data values. They consider displays of numerical data in plots on the number line – box plots and dot plots. They use quantitative measures of center (median) and variability (interquartile range) and describe overall patterns in the data with reference to the context.

If possible, students should have technology that allows them to construct the box plots so they can focus on what can be learned from analyzing the plots as a way to summarize the data, rather than on figuring out the scale and plotting points. If students can use technology, it would be important to transfer the data sets to the students in order to reduce the time spent entering the data and also the time spent tracking down entry errors.

Classwork

Exercise 1 (7–10 minutes): Supreme Court Chief Justices

This example should take students a short time to do if they understood the concepts from the prior lessons. One error might be forgetting to order the data before finding the 5-number summary. If time permits, it is also an opportunity to make connections to social studies. Ask students if they know what cases are before the current Supreme Court, whether they think any data would be involved in those cases, and whether any of the analysis techniques might involve what they have been learning about statistics. Note that this discussion might extend the time necessary for the activity.

Ask the following questions to students as they discuss the answers to the example questions in small groups:

- Why is it important to order the data before you find a median?
 - *The middle value, or median, is based on the order of the data.*
- What other mistakes do you think most people make when thinking about a box plot?
 - *Allow students to indicate their own problems when they work with a box plot (for example, not counting correctly to locate the median, or not figuring out the median correctly based on whether there is an even number or odd number of data).*

Exercise 1: Supreme Court Chief Justices

The Supreme Court is the highest court of law in the United States, and it makes decisions that affect the whole country. The Chief Justice is appointed to the Court and will be a justice the rest of his or her life unless he or she resigns or becomes ill. Some people think that this gives the Chief Justice a very long time to be on the Supreme Court. The first Chief Justice was appointed in 1789.

The table shows the years in office for each of the Chief Justices of the Supreme Court as of 2013:

Name	Years	Appointed in
John Jay	6	1789
John Rutledge	1	1795
Oliver Ellsworth	4	1796
John Marshall	34	1801
Roger Brooke Taney	28	1836
Salmon P. Chase	9	1864
Morrison R. Waite	14	1874
Melville W. Fuller	22	1888
Edward D. White	11	1910
William Howard Taft	9	1921
Charles Evens Hughes	11	1930
Harlan Fiske Stone	5	1941
Fred M. Vinson	7	1946
Earl Warren	16	1953
Warren E. Burger	17	1969
William H. Rehnquist	19	1986
John G. Roberts	8	2005

Data Source: http://en.wikipedia.org/wiki/List_of_Justices_of_the_Supreme_Court_of_the_United_States

1. Use the table to answer the following:

- a. Which Chief Justice served the longest term and which served the shortest term? How many years did each of these Chief Justices serve?

John Marshall had the longest term, which was 34 years. He served from 1801 to 1835. John Rutledge served the shortest term, which was one year in 1795.

- b. What is the median number of years these Chief Justices have served on the Supreme Court? Explain how you found the median and what it means in terms of the data.

First, you have to put the data in order. There are 17 justices so the median would fall at the 9th value (11 years) counting from the top or from the bottom. The median is 11. Half of the justices served less than or equal to 11 years as chief, and half served greater than or equal to 11 years.

- c. Make a box plot of the years the justices served. Describe the shape of the distribution and how the median and IQR relate to the box plot.

The distribution seems to have more justices serving a small number of years (on the lower end). The range (max – min) is 33 years, from 1 year to 34 years. The IQR is: $18 - 6.5 = 11.5$, so about half of the Chief Justices had terms in the 11.5-year interval from 6.5 to 18.

- d. Is the median half way between the least and the most number of years served? Why or why not?

The halfway point on the number line between the lowest number of years served, 1, and the highest number of years served, 34, is 16.5, but because the data are clustered in the lower end of the distribution, the median, 11, is to the left of (smaller than) 16.5. The middle of the interval from the smallest to the largest data value has no connection to the median. The median depends on how the data are spread out over the interval.

MP.1

MP.3

Exercises 2–3 (8–10 minutes): Downloading Songs

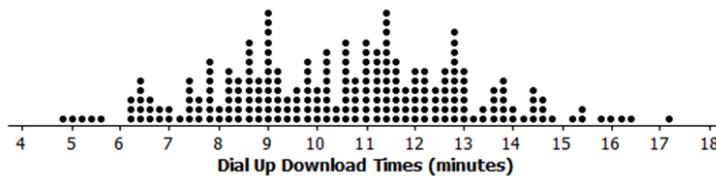
This exercise illustrates how box plots can be useful for large data sets. The 5-number summary visible in a box plot gives quantifiable information about the distribution and provides a way to think about the location of the lower 25% of the data, the middle 50% of the data, and the top 25% of the data. The questions ask students to think about these percentages as well as their fraction equivalents.

Ask students the following questions as they develop their answers for this exercise:

- Why is a distribution with a lot of data values harder to quantify than one with few values?
 - *It is more difficult to locate the median, or Q1 or Q3. Ask students if they can easily find the median from the data distribution in Exercise 2.*
- In what situations might box plots be really useful?
 - *Box plots are particularly useful when comparing two or more data sets.*

Exercises 2–3: Downloading Songs

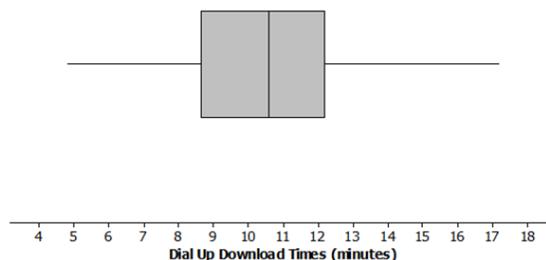
2. A broadband company timed how long it took to download 232 four-minute songs on a dial up connection. The dot plot below shows their results.



- a. What can you observe about the download times from the dot plot?

The smallest time was a little bit less than 5 minutes and the largest a bit more than 17 minutes. Most of the times seem to be between 8 to 13 minutes.
- b. Is it easy to tell whether or not 12.5 minutes is in the top quarter of the download times?

You cannot easily tell from the dot plot.
- c. The box plot of the data is shown below. Now answer parts (a) and (b) above using the box plot.



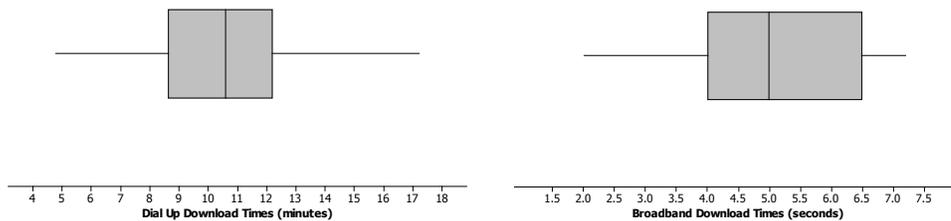
Answer for part (a): About half of the times are above 10.6 minutes. The distribution is roughly symmetric around the median. About half of the times are between 8.6 minutes and 12.1 minutes.

Answer for part (b): 12.5 is above Q3, so it was in the top quarter of the data.

d. What are the advantages of using a box plot to display a large set of data? What are the disadvantages?

With lots of data, the dots in a dot plot overlap, and while you can see general patterns, it is hard to really get anything quantifiable. The box plot shows at least an approximate value for each of the 5-number summary measures and gives a pretty good idea of how the data are spread out.

3. Molly presented the plots below to argue that using a dial up connection would be better than using a broadband connection. She argued that the dial up connection seems to have less variability around the median even though the overall range seems to be about the same for the download times using broadband. What would you say?



The scales are different for the two plots and so are the units, so you cannot just look at the box plots. The time using broadband is centered near 5 seconds to download the song while the median for dial up is almost 11 minutes for a song. This suggests that broadband is going to be faster than dial up.

Exercises 4–5 (12 minutes): Rainfall

Students are asked to compare the variability that can be observed in two different graphs relating to the same topic. Students then use the data from the two graphs to make box plots and think about the difference in comparing the bar graphs and comparing the box plots, in particular using box plots to estimate the percent of data values between the minimum, Q1, median, Q3, and maximum values. Working in pairs might help students sort out the ideas involved in the work and help them learn to communicate their thinking.

Ask students the following questions individually or in small groups as they answer the questions to this exercise:

- Before looking at the graphs carefully, which city would you expect to have the most variability in the amount of precipitation? Explain your thinking.
 - *The answer to this question is based on students having some background of the two cities. If they are not aware of the cities, a short discussion about the location of each city and the general weather patterns of these cities might be considered. If time permits, locate each city on a map and talk about what might influence the amount of precipitation in each city based on location, and what type of precipitation (rain or snow) each city would have. Understanding data often is connected to the background of the data.*
- Notice that the horizontal scales are the same in both dot plots. Is this important? Why or why not?
 - *Having the same scales is important if the two distributions are to be accurately compared. It is also important to have the same scales when comparing the box plots of each city.*

MP.6

- b. Compare the percent of months that have above 2 inches of precipitation for the two cities. Explain your thinking.

In St. Louis the average amount of precipitation each month is always over 2 inches, while this happens at most half of the time in San Francisco because the median amount of precipitation is just above 1 inch.

- c. How do the top fourths of the average monthly precipitation in the two cities compare?

The highest 25% of the precipitation amounts in the two cities are spread over about the same interval (about 4 to 4.5 inches). St. Louis has a bit more spread; the highest 25% in St. Louis are between 4.2 to 4.8 inches; the top 25% in San Francisco are between 4.3 to 4.61 inches.

- d. Describe the intervals that contain the smallest 25% of the average monthly precipitation amounts for each city.

In St. Louis, the smallest 25% of the monthly averages are between about 2.5 inches to 3.2 inches; in San Francisco, the smallest averages are much lower ranging from 0 to 0.16 inches.

- e. Think about the dot plots and the box plots. Which representation do you think helps you the most in understanding how the data vary?

Answers will vary: Some students may say the dot plot because they like seeing individual values; others may say the box plot because it just shows how the data are spread out in each of the four sections made by finding the medians.

Note: The data used in this problem are displayed in the table below.

Average Precipitation (inches)

	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
St. Louis	2.45	2.48	3.36	4.10	4.80	4.34	4.19	3.41	3.38	3.43	4.22	2.96
San Francisco	4.5	4.61	3.76	1.46	0.70	0.16	0	0.06	0.21	1.12	3.16	4.56

Data Source: www.weather.com/weather/wxclimatology/monthly/graph/USCA0987

www.weather.com/weather/wxclimatology/monthly/graph/USMO0787

Closing (1–2 minutes)

Lesson Summary

In this lesson, you reviewed what you know about box plots, the 5-number summary of the data used to construct a box plot, and the IQR. Box plots are very useful for comparing data sets and for working with large amounts of data. When you compare two or more data sets using box plots; however, you have to be sure that the scales and units are the same.

Exit Ticket (5 minutes)

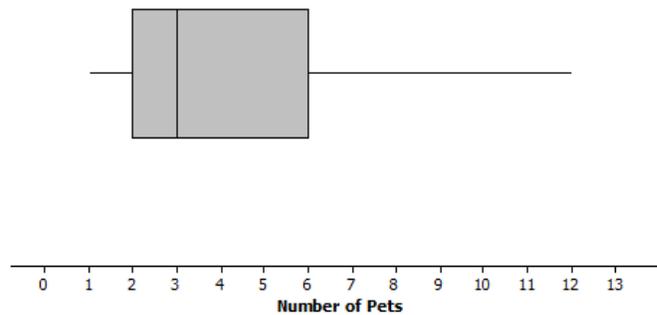
Name _____

Date _____

Lesson 16: Understanding Box Plots

Exit Ticket

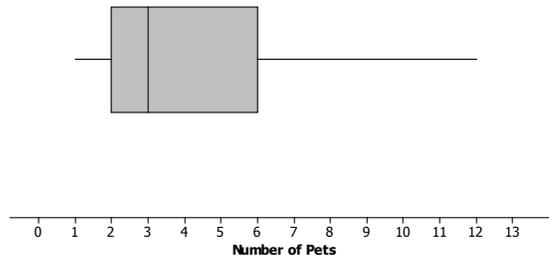
The number of pets per family for students in a sixth grade class is below:



- Can you tell how many families have two pets? Explain why or why not.
- Given the plot above, which of the following statements are true? If the statement is false, modify it to make the statement true.
 - Every family had at least one pet.
 - About one fourth of the families had six or more pets.
 - Most of the families had three pets.
 - Half of the families had five or fewer pets.
 - Three fourths of the families had two or more pets.

Exit Ticket Sample Solutions

The number of pets per family for students in a sixth grade class is below:



1. Can you tell how many families have two pets? Explain why or why not.

You cannot tell from the box plot. You only know that the lower quartile (Q1) is 2 pets. You do not know how many families are included in the data set.

2. Given the plot above, which of the following statements are true? If the statement is false, modify it to make the statement true.

- a. Every family had at least one pet.

True.

- b. About one fourth of the families had six or more pets.

True.

- c. Most of the families had three pets.

False because you cannot determine the number of any specific data value. Revise to "You cannot determine the number of pets most families had."

- d. Half of the families had five or fewer pets.

False. Revise to "More than half of the families had five or fewer pets."

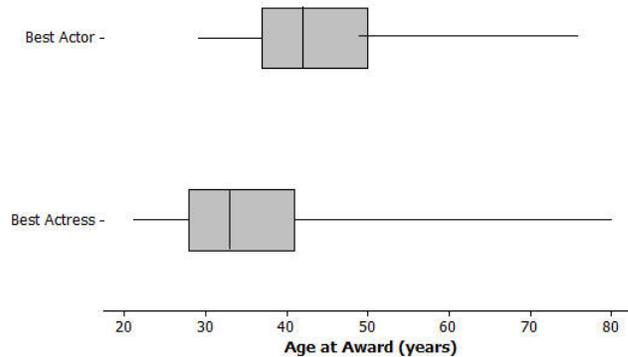
- e. Three fourths of the families had two or more pets.

True.

Problem Set Sample Solutions

All students should do problems 1 and 2. Problem 4 could be an extension, making connections to previous work on the mean.

1. The box plots below summarize the ages at the time of the award for leading actress and leading actor Academy Award winners.



- a. Do you think it is harder for an older woman to win an academy award for best actress than it is for an older man to win a best actor award? Why or why not?

Answers will vary: Students might take either side as long as they given an explanation for why they made the choice they did.

- b. The oldest female to win an academy award was Jessica Tandy in 1990 for *Driving Miss Daisy*. The oldest actor was Henry Fonda for *On Golden Pond* in 1982. How old were they when they won the award? How can you tell? Were they a lot older than most of the other winners?

Henry Fonda was 76 and Jessica Tandy was 80. Those are the maximum values. But there might have been some that were nearly as old—you cannot tell from the box plot.

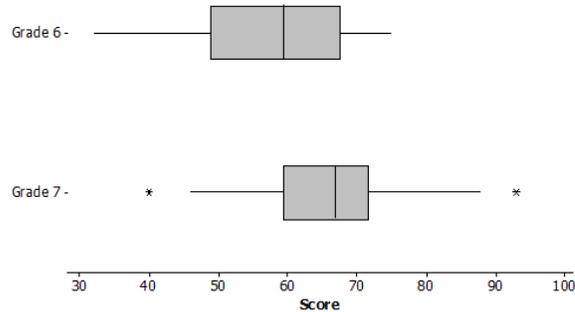
- c. The 2013 winning actor was Daniel Day-Lewis for *Lincoln*. He was 55 years old at that time. What can you say about the percent of male award winners who were older than Daniel Day-Lewis when they won their Oscar?

He was in the upper quarter as one of the older actors. There were less than 25% of the male award winners who were older than Daniel Day-Lewis.

- d. Use the information you can see in the box plots to write a paragraph supporting or refuting the claim that fewer older actresses than actors win academy awards.

Overall, the box plot for actresses starts about 10 years younger than actors and is centered around a lower age than for actors: the median age for actresses who won the award is 33, and for actors it was 42. The upper quartile is also lower for actresses, 41, compared to 49 for actors. The range for actresses' ages was larger, $80 - 21 = 59$, compared to $76 - 29 = 47$ for actors. About $\frac{3}{4}$ of the actresses who won the award were younger than the median for the men.

2. The scores of sixth and seventh graders on a test about polygons and their characteristics are summarized in the box plots below.



- a. In which grade did the students do the best? Explain how you can tell.
Three fourths of the seventh grade students did better than half of the sixth graders. You can tell by comparing Q1 for grade seven to the median of grade six.
- b. Why do you think two of the data values in grade seven are not part of the line segments?
The highest and lowest scores were pretty far away from the other scores so they were marked separately.
- c. How do the median scores for the two grades compare? Is this surprising? Why or why not?
The median score in grade seven was higher than the median in grade six. This makes sense because the seventh graders should know more than the sixth graders.
- d. How do the IQRs compare for the two grades?
The middle half of the Grade 7 scores were close together in a span of about 11 with the median around 66. The middle half of the Grade 6 scores were spread over a larger span, about 17 points from about 50 to 67.
3. A formula for IQR could be written as $Q3 - Q1 = IQR$. Suppose you knew the IQR and the Q1. How could you find the Q3?
 $Q3 = IQR + Q1$. Add the lower quartile to the IQR.
4. Consider the statement, “Historically, the average length of service as Chief Justice on the Supreme Court has been less than 15 years; however, since 1970 the average length of service has increased.” Use the data given in Exercise 1 to answer the following questions.
- a. Do you agree or disagree with the statement? Explain your thinking.
The mean number of years as Chief Justice overall is about 13. The mean number of years since 1969 is about 14.7. Even though the mean has increased, it does not seem really like a big difference because there have only been three justices since then to cover a span of 43 years (and three times 13 is 39, so not enough to really show an increasing trend).
- b. Would your answer change if you used the median number of years rather than the mean?
The median overall was 11 years; the median since 1970 was 17 years, which is considerably larger. This seems to justify the statement.



Topic D:

Summarizing and Describing Distributions

6.SP.B.4, 6.SP.B.5

Focus Standard:	6.SP.B.4	Display numerical data in plots on a number line, including dot plots, histograms, and box plots.
	6.SP.B.5	Summarize numerical data sets in relation to their context, such as by: <ol style="list-style-type: none"> Reporting the number of observations. Describing the nature of the attribute under investigation, including how it was measured and its units of measurement. Giving quantitative measures of center (median and/or mean) and variability (interquartile range and/or mean absolute deviation), as well as describing any overall pattern and any striking deviations from the overall pattern with reference to the context in which the data were gathered. Relating the choice of measures of center and variability to the shape of the data distribution and the context in which the data were gathered.
Instructional Days:	6	
	Lesson 17:	Developing a Statistical Project (E) ¹
	Lesson 18:	Connecting Graphical Representations and Numerical Summaries (P)
	Lesson 19:	Comparing Data Distributions (P)
	Lesson 20:	Describing Center, Variability, and Shape of a Data Distribution from a Graphical Representation (P)
	Lesson 21:	Summarizing a Data Distribution by Describing Center, Variability, and Shape (E)
	Lesson 22:	Presenting a Summary of a Statistical Project (E)

In Topic D, students integrate what they have learned about graphical and numerical data summaries in the previous topics. They match dot plots and histograms to numerical measures of center and variability. Students estimate means and medians from graphical representations of data distributions. They also estimate mean absolute deviation (MAD) and interquartile range (IQR) from graphical representations based

¹ Lesson Structure Key: P-Problem Set Lesson, M-Modeling Cycle Lesson, E-Exploration Lesson, S-Socratic Lesson

on an understanding of data distributions in terms of shape, center, and variability. Two of the lessons in this topic (Lessons 17 and 22) allow students to experience the four-step process described at the beginning of this module through completion of a project. In this project, students experience the four-step investigative process by (1) formulating a statistical question, (2) designing and implementing a plan to collect data, (3) summarizing collected data graphically and numerically, and (4) using the data to answer the question posed.



Lesson 17: Developing a Statistical Project

Student Outcomes

- Students construct a statistical question and a plan to collect data to answer the question.
- Given a statistical question, students use data to construct appropriate graphical and numerical summaries.
- Students use graphical and numerical summaries to answer a statistical question.

Lesson Notes

This lesson is an exploratory lesson. The following agenda provides an overview of the events of this lesson:

- Part 1: Review statistical questions posed in this module. (Approximately 10 minutes)
- Part 2: Summarize the four-step process that was used to begin this module. (Approximately 10 minutes)
- Part 3: Organize students in small groups to evaluate statistical questions. (Approximately 10 minutes)
- Part 4: Complete the four-step outline that is provided in the lesson. (Approximately 15 minutes)

In this lesson, students implement the four-step investigative process based on their own statistical question and data they collect. Students provide a plan to collect data to answer their statistical question. After their plans have been approved, students collect and organize their data and proceed to summarize the data using graphical and numerical summaries. The activity in this lesson is designed to provide a review of the four-step process that outlines for students a statistical study. This activity will require between 30 to 45 minutes of class time.

Students are expected to provide a daily update regarding their progress in collecting and summarizing the data they propose in this lesson. Students should have completed the numerical and graphical summaries before they start Lesson 21. Lesson 21 reviews the four steps and directs students to create a poster (or an outline for a presentation) based on their question, the data collected, and the numerical and graphical summaries. Lesson 22 outlines what is expected in their posters or presentations. Students conclude Lesson 22 by explaining their statistical project to their classmates or invited guests.

Several statistical questions from previous lessons launch this lesson. Direct students to read the list of statistical questions that they studied in this module. After they read this list and the two questions that follow, have a discussion based on the two summary questions.

Classwork

Statistical questions you investigated in this module included the following:

- How many hours of sleep do 6th graders typically get on a night when there is school the next day?
- What is the typical number of books read over the course of 6 months by a 6th grader?
- What is the typical heart rate of a student in a 6th grade class?
- How many hours does a 6th grader typically spend playing a sport or a game outdoors?
- What is the head circumference of adults interested in buying baseball hats?
- How long is the battery life of a certain brand of batteries?
- How many pets do students have?
- How long does it take a student to get to school?
- What is a typical daily temperature of New York City?
- What is the typical weight of a backpack for students at a certain school?
- What is the typical number of french fries in a large order from a fast food restaurant?
- What is the typical number of minutes a student spends on homework each day?
- What is the typical height of a vertical jump for a player in the NBA?

What do these questions have in common?

Why do several of these questions include the word “typical”?

Two discussion questions are posed to highlight the characteristics of a statistical question that was initially developed in Lesson 1 and continued throughout the lessons of this module.

- What is common about these questions?

If necessary, remind students that answering each of these questions require data. It is also anticipated that the data would vary, making them statistical questions.

- Why do several of these questions include the word “typical”?

Answering statistical questions often involves describing a data distribution. Often we are interested in finding a single value that describes what is typical of the values in a data set. Remind students that asking what value is typical for a group is a statistical question, whereas asking about the value of a single observation or about a single individual is not a statistical question because it is not a question that would be answered by collecting data that vary.

A Review of a Statistical Study

MP.4

To provide students an opportunity to see the entire four-step process, review a statistical study based on one or more of the questions they studied in the previous lessons and listed in the introduction of the lesson. A table to structure the discussion of the first three steps of the process is provided in the student’s material. First, encourage students to select at least one of the statistical questions given and write it in the first row of the table. Next, students should recall the data collected to answer this question and to think about how the data might have been collected. Third, review the types of numerical summaries and graphs that students were either given or were expected to construct. Generally, students are expected to start with a graphical summary of the data. Then, based on the shape of the data distribution, the mean and MAD or the median and IQR would be used to describe center and spread. Finally, have students use the summaries to answer the statistical question. Again, reviewing one of the investigations from the lessons provides an opportunity to connect a conclusion to the other three steps.

The following table is a completed example that uses the question, “How many hours of sleep do 6th graders typically get on a night when there is school the next day?”

Recall from the very first lesson in this module that a statistical question is a question answered by data that you anticipate will vary.

Let’s review the steps of a statistical investigation.

- Step 1: Pose a question that can be answered by data.
- Step 2: Collect appropriate data.
- Step 3: Summarize the data with graphs and numerical summaries.
- Step 4: Answer the question posed in Step 1 using the numerical summaries and graphs.

The first step is to pose a statistical question. Select one of the above questions and write it in the following Statistical Study Review Template.

The second step is to collect the data. In all of these investigations, you were given data. How do you think the data for the question you selected in Step 1 was collected? Write your answer in the summary below for Step 2.

The third step involves the various ways you summarize the data. List the various ways you summarized data for Step 3.

<p>Step 1: Statistical question.</p> <p><i>How many hours of sleep do 6th graders typically get on a night when there is school the next day? (Lesson 3)</i></p>
<p>Step 2: Collect data.</p> <p><i>Students from a 6th grade class might have been asked to indicate how many hours they slept. The data would consist of the answers from all of the students.</i></p>
<p>Step 3: Summarize the data.</p> <p><i>The first summary was to organize the data in a dot plot. This data set indicated a nearly symmetrical data distribution. Numerical summaries of the mean and the MAD would provide a description of the typical number of hours of sleep for 6th grade students and a measure of how much variability there was in the sleep times.</i></p>

Finally, the fourth step is to answer the statistical question. The answer to the statistical question was the focus of the investigation in each of the lessons. Describing a data distribution in terms of shape, center, and spread or, depending on the shape of the data distribution, calculating the mean or the median of the data often answer statistical questions.

MP.3 The 4th and final step is the conclusion based on the summaries of Step 3. In the above example, students indicated that a typical number of hours of sleep for 6th graders were approximately 8.5 hours (the mean) and the mean absolute deviation was approximately one hour. Using the opening questions, complete as many of these summaries as necessary in order for students to understand what is expected of them in completing a statistical project.

Make sure students understand that the focus of each lesson was built around a question and that data were then summarized using graphs and the measures of center and variability. The process of collecting the data, however, was not a step students implemented in previous lessons. This lesson adds this new and important step into the investigative process. In general, collecting data will be done by asking students in your class to respond to the questions. In some cases, students may pose a question that could involve collecting data from other students in the school or from other friends or family connections. You might want to encourage some students to obtain data from an appropriate website. Monitor website choice by asking students to clearly indicate the site they plan to use. Check out the site to make sure students are getting appropriate data. A good site that was referenced in several lessons is the website of the American Statistical Association and their Census at School project (www.amstat.org/education/posterprojects/index.cfm).

This lesson should involve students collecting numerical data rather than categorical data. Indicate that students are expected to collect data that is numerical. Most of the questions that were used to start this lesson are possible questions that students could use to collect their own data. If necessary, review with students the difference between numerical and categorical data.

Often, students at this grade level do not appreciate data. They may construct a question that is unclear for those who are expected to answer it. For example, “How many hours of sleep do you get?” Several students answering that question would want to know what day of the week they should use. Help students clearly state their questions and avoid possibly collecting inaccurate data. Finally, make sure you screen the questions to make sure they are appropriate. If there are problems due to the immaturity of students in your class, remind them of the importance of getting their collecting plan approved. If problems continue, assist students in getting data from the Census at School project mentioned above or from other appropriate data sites.

Ideas for students to consider are provided in the opening questions of this lesson. Students could select one of these questions in their own statistical project. Review the winning posters from American Statistical Association’s poster competition (www.amstat.org/education/posterprojects/index.cfm) for additional ideas. Other ideas you may want to suggest during a discussion with students include the following:

- Number of languages spoken by teachers at your school.
- Height of students at your school (would require a sample of students as not every student attending the school could be included in the data collection).
- Number of words in the sentences of a Dr. Seuss book.
- Number of siblings for students in the school (would again require a sample of students).

Note: As indicated with some of the above suggestions, students may select a question that would require selecting a sample. Obtaining a good sample is an important part of statistical investigation. Students are introduced to random sampling and random samples in the seventh grade. For this project, however, challenge students to make sure that they consider ways in which a good representation of the population is collected. Although introducing students to a definition of random samples is not expected, students should be challenged to think about collecting a representative sample in order to answer their statistical question.

A formal problem set has not been added to this lesson. However, teachers are encouraged to design a problem set based on students’ progress during this lesson. The following options are possible ideas for setting up a problem set. Teacher discretion in organizing the project is important.

Option 1: Students who struggled with completing the three-step table developed around one of the questions launched with this lesson should be encouraged to select a different question and complete the table for this second question. The first three steps provide students a structure for connecting a question to a plan of collecting data and then summarizing the data. These steps were involved in the previous lessons; however, in this lesson, students need to bring the steps together. Once students organize these steps for a given question, they are ready to formulate their own question and data collection plan.

Option 2: Students provide a question and a data collection plan to teacher for review as outlined in the lesson. Direct students to complete the three-step table for their question and plan. Using the table provided in this lesson is an excellent way for students to organize their progress, plus it provides a good record for the teacher of how students are thinking at the beginning of this project. For students ready to begin this process, direct them to provide a summary of their statistical question, what plan they have for collecting the data, and what summaries of the data they anticipate will be done. Although these steps were discussed in the lesson, organizing this into a table similar to the one presented in the lesson provides you a summary of their progress. Periodically ask students during the next several days to update you on their progress by providing a summary of the table used in this lesson.

Option 3: For students going beyond the scope of questions outlined in the launch, they need to provide specific descriptions of what they plan to research and how they plan to collect the data. For example, if students explore research on honeybees, make sure students clearly indicate the statistical question (or what is the question that will be answered by data that is anticipated to vary), where they plan to obtain data to summarize the question, and how they plan to summarize the data. A brief written report or summary of their progress might constitute a workable problem set option for students at this level.

Project (Exploratory Challenge)

Now it is your turn to answer a statistical question based on data you collect. Before you collect data, explore possible statistical questions. For each question, indicate data that you would collect and summarize to answer the question. Also indicate how you plan to collect the data.

Think of questions that could be answered by data collected from members of your class or school or data that could be collected from recognized websites (e.g., The American Statistical Association and the project Census at Schools). Check with your teacher if you are planning to work with data from an outside source such as one of the above websites. Your teacher will need to approve both your question and your plan to collect data before data are collected.

As a class, explore possibilities of a statistical investigation. Record some of the ideas discussed by your class using the following table.

Possible statistical questions	What data would be collected and how would the data be collected?

After discussing several of the above possibilities of a statistical project, prepare a statistical question and a plan to collect data to present to your teacher. After your teacher approves your question and data collection plan, begin collecting the data. Carefully organize your data as you begin developing the summaries to answer your statistical question. In future lessons, you will be directed to begin creating a poster or an outline of a presentation that will be shared with your teacher and other members of your class.

For this lesson, complete the following to present to your teacher:

1. The statistical question for my investigation is as follows:
2. Here is the plan I propose to collect my data. (Include the exact questions you may ask an individual or a clear description of what you plan to measure or count.)

Lesson Summary

A statistical study involves a four-step investigative process:

- Pose questions that can be answered by data.
- Design a plan for collecting appropriate data and then use the plan to collect data.
- Analyze the data.
- Interpret results and draw valid conclusions from the data to the question posed.

Problem Set Sample Solutions

Your teacher will outline steps you are expected to complete in the next several days to develop this project. Keep in mind that the first step in developing your project is a statistical question. With one of the statistical questions posed in this lesson or with a new one developed in this lesson, organize your question and plan to collect and summarize data. Complete the process as outlined by your teacher.



Lesson 18: Connecting Graphical Representations and Numerical Summaries

Student Outcomes

- Students match the graphical representations and numerical summaries of a distribution. Matches involve dot plots, histograms, and summary statistics.

Lesson Notes

It can be difficult to understand a data set by just looking at raw data. Imagine that Joaquin's project is on finding the typical weight of bears. He finds an article about bears that provides an unsorted list of the weights of 250 bears with no numerical summaries or graphs of these data. It would be very difficult to quickly draw some conclusions from this data. Joaquin decides to design his project using this data.

Next Joaquin finds an article that states, "The median weight for the bears studied was 305 pounds." This is useful, but sometimes even numerical summaries alone cannot completely convey interesting aspects of a data distribution. Often, readers like Joaquin want to have a concise and useful summary of the information that is both numerical *and* visual.

In the next couple of lessons, students will begin to take the graphical representations and numerical summaries they learned and apply them to different situations. While working through these lessons, students should keep in mind their own statistical question. They should think about which graphs will best showcase their data and which numerical summaries will represent the data they are collecting.

Classwork

It can be difficult to understand a data set by just looking at raw data. Imagine that Joaquin's project is on finding the typical weight of bears. He found an article about bears that provided an unsorted list of the weights of 250 bears with no numerical summaries or graphs of these data. It would be very difficult to quickly draw some conclusions. Joaquin decided to design his project using this data.

Now consider the case where the article provides you with a statement, "the median weight for the bears studied was 305 pounds." This is useful, but sometimes even numerical summaries alone cannot completely convey interesting aspects of a data distribution. Often, readers want to have a concise and useful summary of the information that is both numerical and visual.

In the next couple of lessons, you will begin to take the graphical representations and numerical summaries you learned and apply them to different situations. While working through these lessons, keep in mind your own statistical question. Think about which graphs will best showcase your data and which numerical summaries will represent the data you are collecting.

Example 1 (3 minutes): Summary Information from Graphs

MP.4

Review dot plots and histograms. Important points are as follows:

- Each picture reveals a great deal about what is occurring.
- Some pictures may look different from one another and highlight different aspects of the same data set, but graphs of the same data set can still have several similarities and impart similar information.
- Summary measures can be obtained or estimated from dot plots and histograms, and these measures complement the graphical information.

Example 1: Summary Information from Graphs

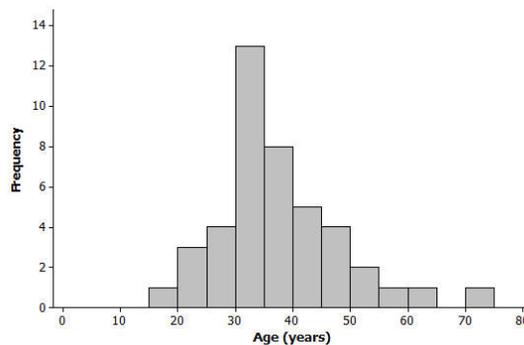
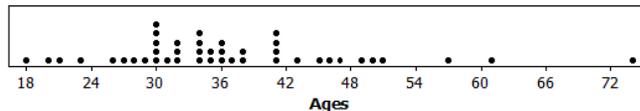
Recall that a *dot plot* includes a dot on a scale or number line for each observation in a data set. Dots are stacked on top of one another when there are multiple occurrences of a data value. Recall also that a *histogram* similarly uses a scale or number line to present the frequency or relative frequency of groups of data based on intervals of equal width. For each interval, the height of the bar is proportional to the number of observations in the interval; the taller the bar, the greater the number of observations in that interval. This means that when both graphs are generated for a given data set, the two graphs will display some similarities.

Here is a data set of the ages (in years) of 43 participants in a recent local 5-kilometer race.

20	30	30	35	36	34	38	46
45	18	43	23	47	27	21	30
32	32	31	32	36	74	41	41
51	61	50	34	34	34	35	28
57	26	29	49	41	36	37	41
38	30	30					

Here are some summary statistics, a histogram, and a dot plot for the data:

Minimum = 18, Q1 = 30, Median = 35, Q3 = 41, Maximum = 74; Mean = 36.81, MAD = 8.1



Exercises 1–7 (7–10 minutes)

Pose these questions to students one at a time.

Exercises 1–7

1. Based on the histogram, would you describe the shape of the data distribution as approximately symmetric or as skewed? Would you have reached this same conclusion looking at the dot plot?

Both graphs show a slightly skewed right data distribution.

2. Is it easier to see the shape of the data distribution from the histogram or the dot plot?

Generally, it is easier to see the shape of the data distribution from a histogram. In this case, the clustering and high frequency of ages in the 30s is more evident in the histogram.

3. What is something you can see in the dot plot that is not as easy to see in the histogram?

When using the histogram, we cannot determine the exact minimum or maximum age—for example, we only know that the minimum age is between 15 and 19 years of age. Also, we can only approximate the median (we generally cannot figure out the exact median value from a histogram).

Since the dot plot provides us with a dot for each observation, we can obtain specific (or sometimes rounded) values for a 5-number summary—and the entire data set—from the dot plot. With the dot plot, we see that the minimum is specifically 18. The median is the 22nd observation (since there are 43 observations) and the 22nd dot counting from left to right is 35 (we cannot be that precise with the histogram). The oldest runner (74) also appears to be a more extreme departure from the rest of the data in the dot plot as compared to the histogram.

4. Do the dot plot and the histogram seem to be centered in about the same place?

Yes, as both graphs are based on the same data, they should generally communicate the same information regarding center.

5. Do both the dot plot and the histogram convey information about the variability in the age distribution?

Yes, as both graphs are based on the same data, they should generally communicate the same information regarding variability. However, as mentioned earlier, the oldest runner (74) appears to be a more extreme departure from the rest of the data in the dot plot as compared to the histogram.

6. If you did not have the original data set and only had the dot plot and the histogram, would you be able to find the value of the median age from the dot plot?

Yes, see response to Exercise 3.

7. Explain why you would only be able to estimate the value of the median if you only had a histogram of the data.

The median is the 22nd ordered observation in this data set since there are 43 observations. Counting from left to right, we know that the first 21 observations are in the first 4 classes: 15–19 (1 value), 20–24 (3 more values), 25–29 (4 more values), and 30–34 (13 more values). Cumulatively, we have encountered the lowest 21 observations by the time we are finished with the 30–34 class. So, the 22nd value must be in the next class, which is 35–39 years of age. We just cannot determine the exact value from the histogram.

MP.6

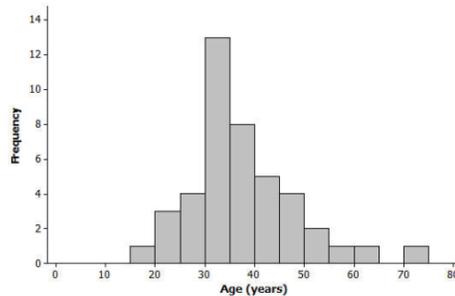
Exercises 8–13 (25 minutes): Graphs and Numerical Summaries

Pose the questions to students one at a time. Allow for more than one student to offer an answer for each question encouraging a brief (2 minute) discussion.

Note: In some cases, the questions have multiple and/or inexact answers.

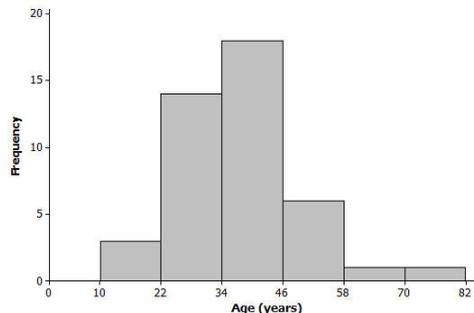
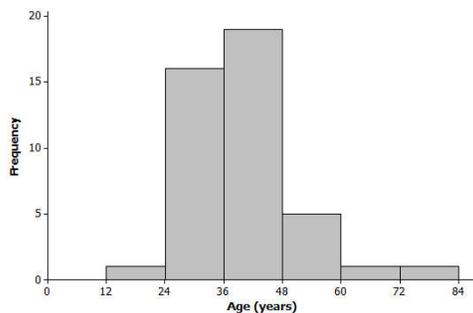
Exercises 8–13: Graphs and Numerical Summaries

8. Suppose that a newspaper article was written about the race and the histogram of the ages from Example 1 was shown in the article. The writer stated, “The race attracted many older runners this year; the median age was 45.” Explain how we would know that this is an incorrect statement based on just the histogram.



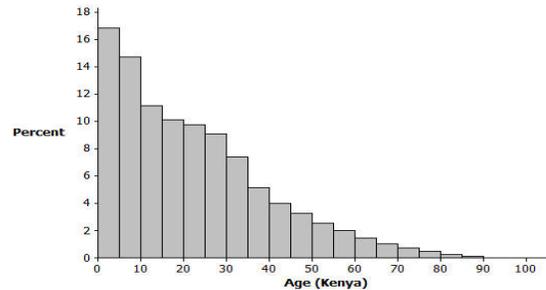
Several answers are possible, but students should concentrate on the shaded area (distribution) shown and the definition of a median. Specifically, at 45, it appears that less than half of the data are (or area is) at or above that value (or alternatively, more than half of the data are at or below that value). Another approach would be to state that the value at which the data (area) might be split, 50% below/50% above appears to be at a lower value than 45 (or in the interval 35–39).

9. One of the histograms below is another valid histogram for the runners’ ages. Select the correct histogram, and explain how you determined which graph is valid (and which one is incorrect) based on the summary measures and dot plot.



One of the objectives is to reinforce that there is more than one way to draw a proper histogram for a distribution. This question is especially detail-oriented because students need to carefully reconcile components of the histogram with the data set (either as shown in raw form or in the dot plot). The histogram on the right is the correct graph because it is consistent with the dot plot/data. Most notably, the histogram on the right correctly shows there are 3 runners in the 10–21 age group, while the left histogram shows only 1 runner in the 12–23 age group (and there are actually 4 runners in that class). Other classes in the left histogram do not match the dot plot/data (e.g., the 48–59 group), so several answers are possible.

10. The histogram below represents the age distribution of the population of Kenya in 2010.



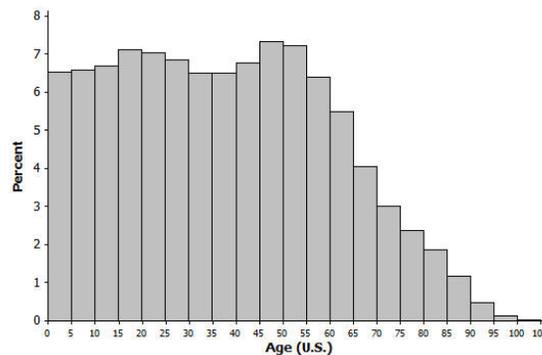
- a. How do we know from the graph above that the first quartile (Q1) of this age distribution is between 5 and 9 years of age?

Since a histogram should display information that is consistent with summary measures, we are seeking a data value such that 25% of the distribution is at or below that value. While the 0–4 age group represents the lowest 17% approximately, the next group (age 5–9) appears to account for the next approximately 15% of the distribution. This means that cumulatively this second group (ages 5–9) roughly represents the lowest 17%–32%, thus the first quartile would be in that group.

- b. Someone believes that the median age is near 30. Explain how the graph supports this belief, OR explain why the graph does not support this belief.

The median does NOT appear to be 30 years of age. See answers for Exercises 1 and 3 for guidance in determining this. Specifically, the 50th percentile estimated by adding approximate percentages (and/or visually assessing the point at which the area seems split evenly) appears to be in the 15–19 age group.

11. The histogram below represents the age distribution of the population of the United States in 2010. Based on the histogram, which of the following ranges do you think includes the median age for the United States: 20–29, 30–39, or 40–49? Why?



Using similar arguments as described in the response to Exercise 10 part (b), the median appears to be in the 30–39 age group, most likely in the 35–39 class.

12. Use the histograms from Exercises 10 and 11 to answer the following:

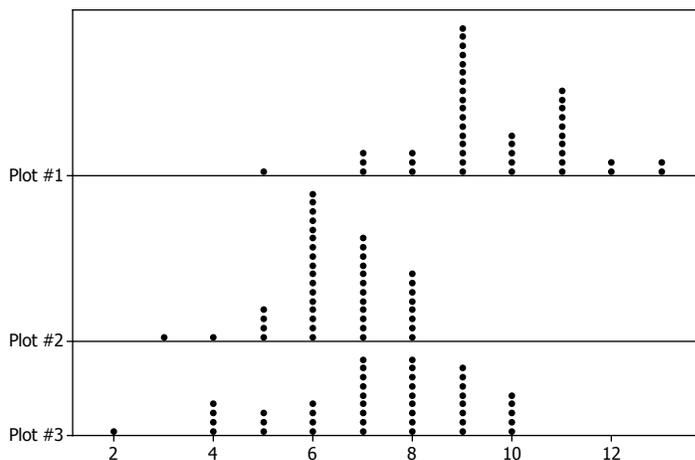
- a. Which country's age distribution (Kenya or United States) has a third quartile in the 50s? How did you decide?

The third quartile of the U.S. is in the 50s, and this can be determined using methods of area visualization or cumulative percentage counting as described above. Note also that for the value of 50 in the Kenya distribution, there appears to be far less than 25% of the distribution shown above that value.

- b. If someone believed that the average age of a person living in the United States was greater than the average age of a person living in Kenya, how could you support that claim by comparing the histograms?

There are a few ways to support this remark. First, there is a considerably higher percentage of high ages in the U.S. distribution. Secondly, as skewed right distributions tend to have a mean that is higher than the median, the fact that the U.S. has a higher median age than Kenya (Exercises 4 and 5) would support the idea that the U.S. would have a higher mean age than Kenya. Lastly, using a balance point argument, the balance point for the U.S. would be much further up the number line than the balance point for Kenya.

13. Match the following sets of summary measures with the corresponding dot plot. Only ONE dot plot matches each group of summary measures. Explain why you selected the dot plot or why the other dot plots would not represent the summary measures. Note: the same scale is used in each dot plot.



- a. Median = 8 and IQR = 3 Plot # _____

Plot #3 – It is the only distribution visually centered near 8 and one can tell that the 22nd ordered observation (the median in this case) is 8. Also, plot #3 is the only distribution with an IQR of 3.

- b. Mean = 9.6 and MAD = 1.28 Plot # _____

Plot #1 – This plot is the only plot for which one could assume a mean value as high as 9.6. It is the only plot with values of 11, 12, and 13 – and there are several of these values.

- c. Median = 6 and Range = 5 Plot # _____

Plot #2 – It is the only plot which appears to have a central value of 6. It is also the only plot with a range of 5 (each of the other plots has a range of 8).

Closing (5 minutes)

Consider posing the following questions; allow a few student responses for each:

- What kinds of information about a quantitative data distribution might not be presented well if we only use summary measures?
 - *Clustering, aspects of shape, extremeness of certain values, etc.*
- If dot plots can provide us with a way of figuring out exact (or nearly exact) observation values, why don't we always use dot plots to show a data distribution? What are some cases where a histogram might provide a better visual summary of the distribution or a dot plot might not work well?
 - *Clustering and gaps might be more easily shown in a histogram. A dot plot may be cumbersome for large data sets -- like the population distribution of an entire country!*

Lesson Summary

Generally, we can compute or approximate many values in a numerical summary for a data set by looking at a histogram or a dot plot for the data set. Thus, we can generally match a histogram or a dot plot to summary measures provided.

When making a histogram and a dot plot for the same data set, the two graphs will have similarities. However, some information may be more easily communicated by one graph as opposed to the other.

Exit Ticket (5–8 minutes)



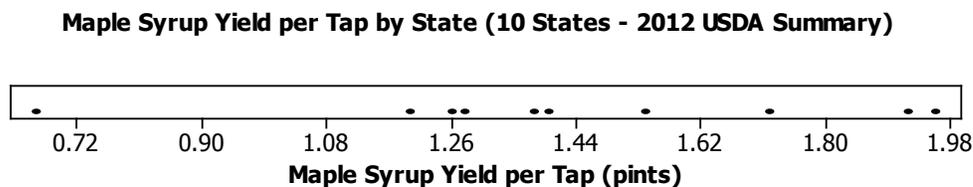
Name _____

Date _____

Lesson 18: Connecting Graphical Representations and Numerical Summaries

Exit Ticket

1. Many states produce maple syrup, which requires tapping sap from a maple tree. However, some states produce more pints of maple syrup per tap than other states. The following dot plot shows the pints of maple syrup yielded per tap in each of the 10 maple syrup producing states as listed in the *US Department of Agriculture's 2012 Crop Production Summary*. For the dot plot, which ONE of the three sets of summary measures could match the graph? For each choice that you eliminate, list at least one reason for eliminating the choice.

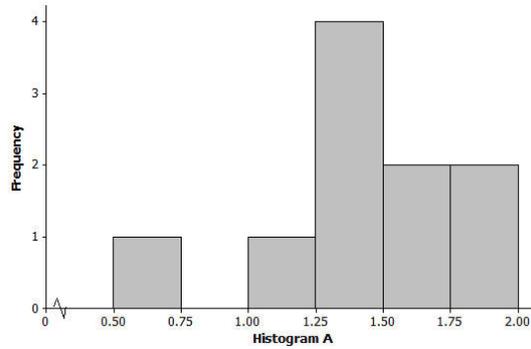


(From the *United States Department of Agriculture National Agricultural Statistics Service Crop Production 2012 Summary*, ISSN: 1057-7823, p. 75, accessed May 5, 2013 available at <http://usda01.library.cornell.edu/usda/current/CropProdSu/CropProdSu-01-11-2013.pdf>.)

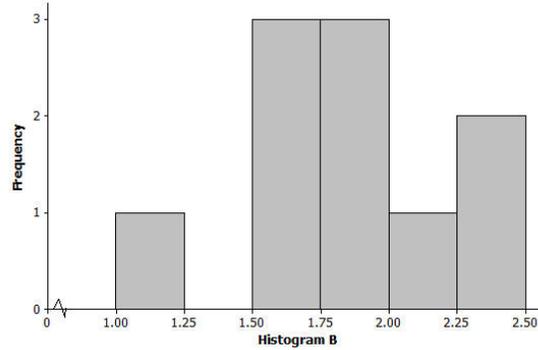
- Minimum = 0.66, Q1 = 1.26, Median = 1.385, Q3 = 1.71, Maximum = 1.95, Range = 2.4; Mean = 1.95, MAD = 0.28
- Minimum = 0.66, Q1 = 1.26, Median = 1.71, Q3 = 1.92, Maximum = 1.95, Range = 1.29; Mean = 1.43, MAD = 2.27
- Minimum = 0.66, Q1 = 1.26, Median = 1.385, Q3 = 1.71, Maximum = 1.95, Range = 1.29; Mean = 1.43, MAD = 0.28

2. For the dot plot in problem 1, which ONE of the three histograms below could be a match? For each choice that you eliminate, list at least one reason for eliminating the choice.

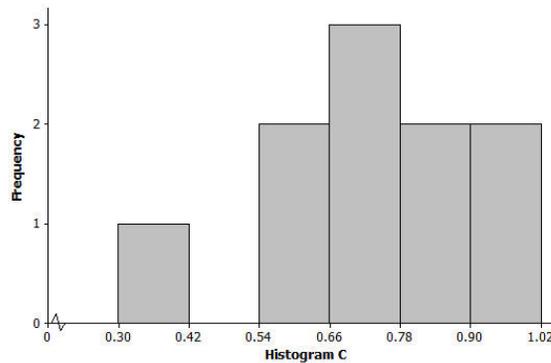
a.



b.



c.

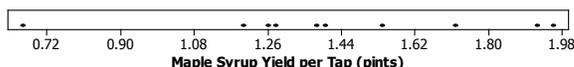


Exit Ticket Sample Solutions

Note: Students do not need to list all elimination reasons listed below. The instructions asked, "For each choice that you eliminate, list *at least one* reason for eliminating the choice."

- Many states produce maple syrup, which requires tapping sap from a maple tree. However, some states produce more pints of maple syrup per tap than other states. The following dot plot shows the pints of maple syrup yielded per tap in each of the 10 maple syrup producing states as listed in the *US Department of Agriculture's 2012 Crop Production Summary*. For the dot plot, which ONE of the three sets of summary measures could match the graph? For each choice that you eliminate, list at least one reason for eliminating the choice.

Maple Syrup Yield per Tap by State (10 States - 2012 USDA Summary)



(From the *United States Department of Agriculture National Agricultural Statistics Service Crop Production 2012 Summary*, ISSN: 1057-7823, p. 75, accessed May 5, 2013 available at <http://usda01.library.cornell.edu/usda/current/CropProdSu/CropProdSu-01-11-2013.pdf>.)

- Minimum = 0.66, Q1 = 1.26, Median = 1.385, Q3 = 1.71, Maximum = 1.95, Range = 2.4; Mean = 1.95, MAD = 0.28
- Minimum = 0.66, Q1 = 1.26, Median = 1.71, Q3 = 1.92, Maximum = 1.95, Range = 1.29; Mean = 1.43, MAD = 2.27
- Minimum = 0.66, Q1 = 1.26, Median = 1.385, Q3 = 1.71, Maximum = 1.95, Range = 1.29; Mean = 1.43, MAD = 0.28

The correct answer is (c).

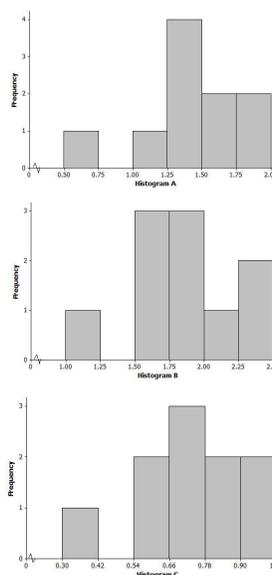
Choice (a) would not work because the range is too large as the difference between maximum and minimum is only 1.29 pints. Also, the mean would not be that close to (or the same as) the maximum value in this case.

Choice (b) would not work because a median value of 1.71 would be too high. Estimating the dot values, the 5th and 6th ordered observations (the median for a dataset of 10 items) are near 1.4. Also, the MAD is much too large as the range of the data is only 1.29 pints.

- For the dot plot in problem 1, which ONE of the three histograms below could be a match? For each choice that you eliminate, list at least one reason for eliminating the choice.

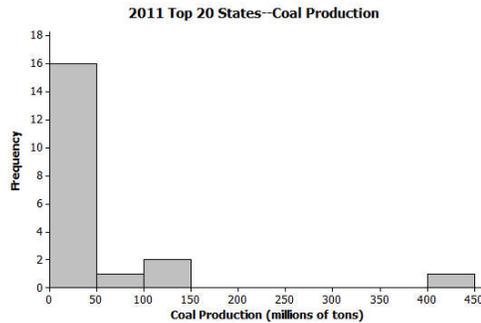
The correct answer is (a).

Graphs (b) and (c) are developed from data sets with similar shape features to the correct graph (graph (a)), but the range and distribution of values do not match. For example, graph (b) would not be valid as it is based on 3 observation values of 2 pints or more and there are no values that large in the original dot plot. Also, the smallest value in graph (b) is at least 1 pint, and the actual data set contains a value less than 1 pint. Graph (c) is based on values that are smaller than many of those presented in the dot plot; in fact all of the values in graph (c) are less than 1.02 pints, and nearly all of the 10 observations in the actual data set are greater than 1.02.



Problem Set Sample Solutions

1. The following histogram shows the amount of coal produced (by state) for the 20 largest coal producing states in 2011. Many of these states produced less than 50 million tons of coal, but one state produced over 400 million tons (Wyoming). For the histogram, which ONE of the three sets of summary measures could match the graph? For each choice that you eliminate, list at least one reason for eliminating the choice.



(U.S. Coal Production by State data as reported by the National Mining Association from http://www.nma.org/pdf/c_production_state_rank.pdf accessed May 5, 2013)

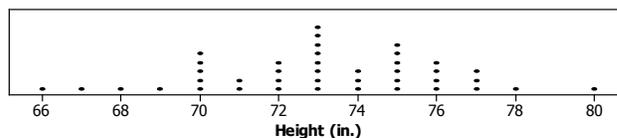
- a. Minimum = 1, Q1 = 12, Median = 36, Q3 = 57, Maximum = 410; Mean = 33, MAD = 2.76
- b. Minimum = 2, Q1 = 13.5, Median = 27.5, Q3 = 44, Maximum = 439; Mean = 54.6, MAD = 52.36
- c. Minimum = 10, Q1 = 37.5, Median = 62, Q3 = 105, Maximum = 439; Mean = 54.6, MAD = 52.36

The correct answer is (b).

Choice (a) would not work because Q3 (the average of the 15th and 16th ordered observations) must be less than 50 since both the 15th and 16th ordered observations are less than 50. The mean is most likely greater than (not less than) the median given the skewed right nature of the distribution and the large outlier. The MAD value is most likely much larger than 2.76 given the presence of the outlier and its distance from the cluster of remaining observations.

Choice (c) would not work because since there are 20 observations, the median (the average of the 10th and 11th ordered observations) must be less than 50 since both the 10th and 11th ordered observations are less than 50. Likewise, the Q3 (the average of the 15th and 16th ordered observations) must be less than 50 since both the 15th and 16th observations are less than 50. The mean is most likely greater than (not less than) the median given the skewed right nature of the distribution and the large outlier.

2. The heights (rounded to the nearest inch) of the 41 members of the 2012–2013 University of Texas Men's Swimming and Diving Team are shown in the dot plot below.



Data Source: <http://www.texassports.com/sports/m-swim/mtt/tex-m-swim-mtt.html> accessed April 30, 2013

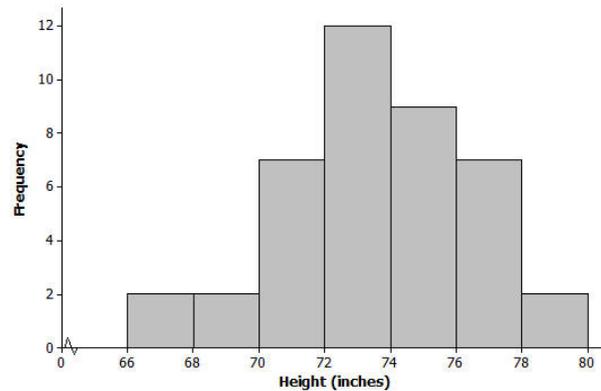
- a. Use the dot plot to determine the 5-number summary (minimum, lower quartile, median, upper quartile, and maximum) for the data set.

The 5-number summary values for an ordered data set of 41 observations would be:

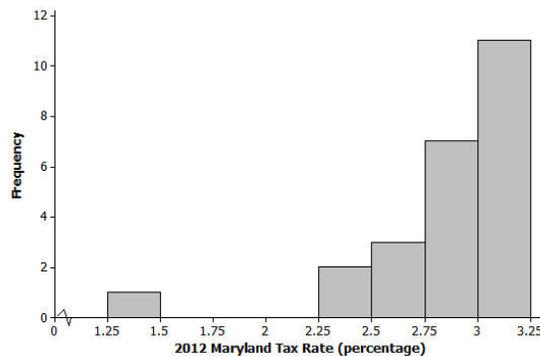
Min = 1st, Q1 = Average of 10th and 11th, Median = 21st, Q3 = Average of 31st and 32nd, Max = 41st

Summary: 66, 71, 73, 75, 80

- b. Based on this dot plot, make a histogram of the heights using the following classes: 66–< 68 inches, 68–< 70 inches, and so on.

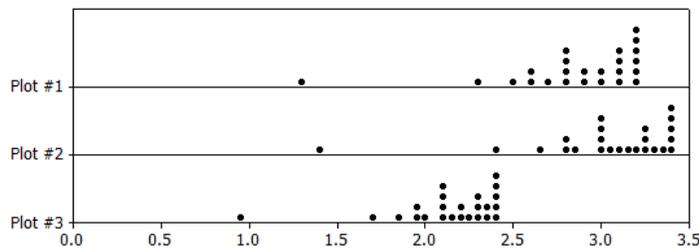


3. According to the website of the Comptroller of Maryland, “Maryland’s 23 counties and Baltimore City levy a local income tax ... Local officials set the rates, which range between 1.25% and 3.20% for the current tax year (2012).” A histogram of the 24 tax rates (in percentages) appears below.



Data Source: <http://taxes.marylandtaxes.com> accessed May 5, 2013

Which ONE of the three dot plots below matches the “2012 Maryland Tax Rates” histogram above? Explain how you determined the correct dot plot.



The correct response is “Plot #1.” Plot #3 is eliminated as both its minimum and maximum values are too low (along with other reasons). Plot #2 is eliminated because several of its large values exceed the histogram’s maximum possible value.

4. For each of the following five sets of summary measures, indicate if the set of summary measures could match the “2012 Maryland Tax Rates” histogram above. For each set of summary measures that you eliminate, explain why you eliminated that choice.

a. Mean = 1.01, MAD = 5.4

b. Median = 2.93, IQR = 0.45

c. Mean = 3.5, MAD = 1.1

d. Median = 3.10, IQR = 2.15

e. Minimum = 1.25, Maximum = 3.20

Options (b) and (e) could match the picture (and option (e) matches the text introducing the context). Options (a) and (c) are eliminated as the mean values of 1.01 and 3.5 are not supported by the histogram (these values are more extreme, respectively, than the minimum and maximum values shown in the histogram). Option (d) is eliminated since 13 of the observations (that's more than half) are less than 3.0, so a median of 3.1 would be too large. (The IQR in option (d) is also too large.)



Lesson 19: Comparing Data Distributions

Student Outcomes

- Given box plots of at least two data sets, students will comment on similarities and differences in the distributions.

Lesson Notes

As you have seen in previous lessons, it can be difficult to understand a data set just by looking at raw data. Often, readers want to have a concise and useful summary.

This becomes extremely important when data distributions are compared to one another. While a reader may be interested in knowing if a typical adult male polar bear in Alaska is larger than a typical adult male grizzly bear in British Columbia, it would also be useful to be able to compare the variability and shape of the distributions of these two groups of bears as well. With summary graphs of the two distributions placed side-by-side, you can more easily assess and compare the characteristics of one distribution to the other distribution.

By this point, you should have completed the collection of data for your statistical question. This lesson will provide graphical representations of data distributions that are part of the summaries expected in your project.

Classwork

Example 1 (3 minutes): Comparing Groups Using Box Plots

Review box plots and 5-number summaries. Important points are as follows:

MP.2

- Each box plot tells a great deal about the distribution as certain summary measures can be obtained or estimated from the plot.
- When two (or more) box plots are shown together (using the same scale), visual differences between the two box plots correspond to quantitative differences between the corresponding summary measures of the two (or more) distributions.

Pose the question to the class that is presented in the text:

Example 1: Comparing Groups Using Box Plots

Recall that a *box plot* is a visual representation of a 5-number summary. It is drawn with careful reference to a number line, so the difference between any two values in the 5-number summary is represented visually as a distance. For example, the box of a box plot is drawn so that width of the box represents the IQR. The whiskers (the lines that extend from the box) are drawn such that the distance from the far end of one whisker to the far end of the other whisker represents the range. If two box plots (each representing a different distribution) were drawn side-by-side using the same scale, one could quickly compare the IQRs and ranges of the two distributions while also gaining a sense of the 5-number summary values for each distribution.

Here is a data set of the ages of 43 participants in a local 5-kilometer race (shown in a previous lesson).

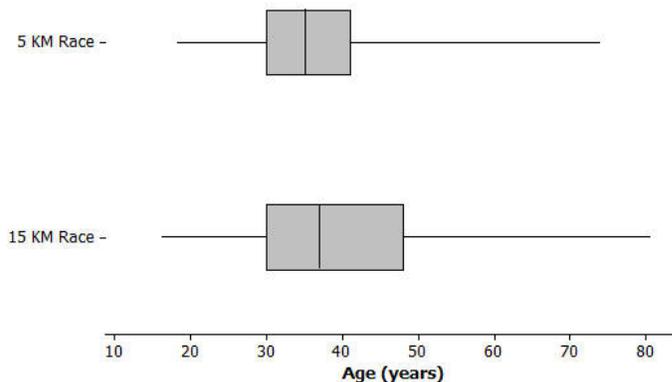
20	30	30	35	36	34	38	46
45	18	43	23	47	27	21	30
32	32	31	32	36	74	41	41
51	61	50	34	34	34	35	28
57	26	29	49	41	36	37	41
38	30	30					

Here is the 5-number summary for the data: Minimum = 18, Q1 = 30, Median = 35, Q3 = 41, Maximum = 74.

Later that year, the same town also held a 15-kilometer race. The ages of the 55 participants in that race appear below.

47	19	30	30	36	37	35	39
19	49	47	16	45	22	50	27
19	20	30	32	32	31	32	37
22	81	43	43	54	66	53	35
22	35	35	36	28	61	26	29
38	52	43	37	38	43	39	30
58	30	48	49	54	56	58	

Does the longer race appear to attract different runners in terms of age? Here are side-by-side box plots that may help answer that question. Side-by-side box plots are two or more box plots drawn using the same scale.



Exercises 1–6 (10 minutes)

In some cases, the questions have multiple and/or inexact answers. Also note that in some cases original data sets are not provided as the outcomes are based on analysis of box plots, and students are encouraged to estimate summary measures from the graph.

Exercises 1–6

- Based on the side-by-side box plots, estimate the 5-number summary for the 15-kilometer race data set.

Minimum = 16, Q1 = 30, Median = 37, Q3 = 48, Maximum = 81.

- Do the two data sets have the same median? If not, which race had the higher median age?

No, the 15-km race has a slightly higher median: 37 years of age compared to 35 years of age for the 5-km race.

3. Do the two data sets have the same IQR? If not, which distribution has the greater spread in the middle 50% of its distribution?

No, the 15-km race has a slightly higher IQR: 18 years of age compared to 11 years of age for the 5-km race.

4. Which race had the smaller overall range of ages? What do you think the range of ages is for the 15-kilometer race?

The 5-km race had the smaller range of ages: 56 compared to 65 for the 15-km race.

5. Which race had the oldest participant? About how old was this participant?

The 15-km race had the oldest participant at 81 years of age. The oldest participant for the 5-km race was 74.

6. Now consider just the youngest 25% of participants in the 15-kilometer race. How old was the youngest runner in this group? How old was the oldest runner in this group? How does that compare with the 5-kilometer race?

These values would be the minimum and Q1 respectively. For the 15-km race, this is 16 to 30 years of age. For the 5-km race, this is 18 to 30 years of age (both distributions have the same Q1).

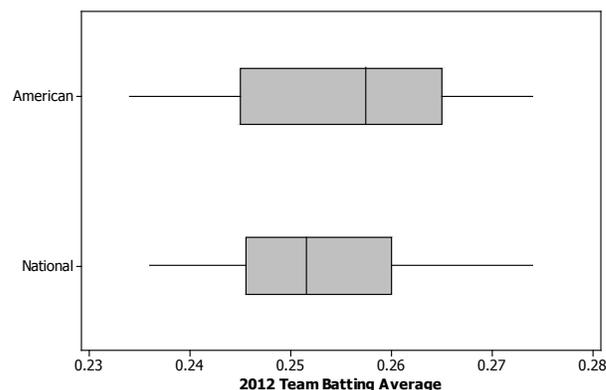
Exercises 7–12 (20 minutes): Comparing Box Plots

Pose the questions to students one at a time. Allow for more than one student to offer an answer for each question encouraging a brief (2 minute) discussion.

In some cases, the questions have multiple and/or inexact answers. Note: non-baseball related questions with similar objectives appear in the Problem Set.

Exercises 7–12: Comparing Box Plots

In 2012, Major League Baseball was comprised of two leagues: an American League of 14 teams and a National League of 16 teams. It is believed that the American League teams would generally have higher values of certain offensive statistics such as batting average and on-base percentage. (Teams want to have high values of these statistics.) Use the following side-by-side box plots to investigate these claims. (Source: <http://mlb.mlb.com/stats/sortable.jsp> accessed May 13, 2013)



7. Was the highest American League team batting average very different from the highest National League team batting average? If so, approximately how large was the difference and which league had the higher maximum value?

No, the highest batting averages for both leagues appear to be around 0.274. (Allow for estimation by students.)

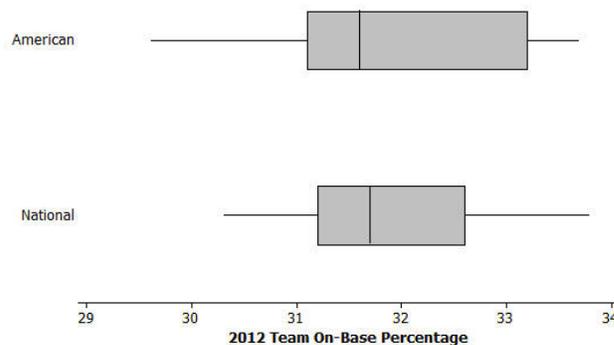
8. Was the range of American League team batting averages very different or only slightly different from the range of National League team batting averages?

They appear to be only slightly different with the AL range being slightly higher. AL minimum (0.234) is slightly lower than the NL minimum (0.236) and from above; both leagues appear to have the same maximum.

9. Which league had the higher median team batting average? Given the scale of the graph and the range of the data sets, does the difference between the median values for the two leagues seem to be small or large? Explain why you think it is small or large.

The AL has the higher median batting average at roughly 0.258 while the median batting average for the NL is roughly 0.252. Students could state that this 0.006 difference is significant based on several reasons, e.g., the difference of 0.006 is roughly $\frac{1}{6}$ of the NL range, the AL median is close to the NL Q3, visually, the difference appears to be about the same as the difference between Q1 and the median for the NL data set, and so on.

10. Based on the box plots below for on-base percentage, which 3 summary values (from the 5-number summary) appear to be the same or virtually the same for both leagues?



The Q1, median, and maximum appear to be roughly the same.

11. Which league's data set appears to have less variability? Explain.

The NL data set appears to have less variability as it has a smaller IQR and smaller range.

12. Respond to the original statement: "It is believed that the American League teams would generally have higher values of ... on-base percentage." Do you agree or disagree based on the graphs above? Explain.

A student might disagree with the statement given the similar medians and the other similar summary measures. Also the AL data set has a lower minimum. However, a student might agree with the statement in that the AL data set has a higher Q3 than the NL data set.

Closing (5 minutes)

Consider posing the following questions; allow a few student responses for each:

- What kinds of information about a quantitative data distribution might not be presented well if we only use box plots?
 - *Clustering, some aspects of shape, distribution within a quartile, number of observations, etc.*
- What other kinds of graphs might be graphed side-by-side to visually communicate the similarities and differences between data sets?
 - *Side-by-side dot plots would be effective for this, again assuming the same scale is used.*

Lesson Summary

When comparing the distribution of a quantitative variable for two or more distinct groups, it is useful to display graphs of the groups' distributions side-by-side using the same scale. Generally, you can more easily notice, quantify, and describe the similarities and differences in the distributions of the groups.

Exit Ticket (10 minutes)

Name _____

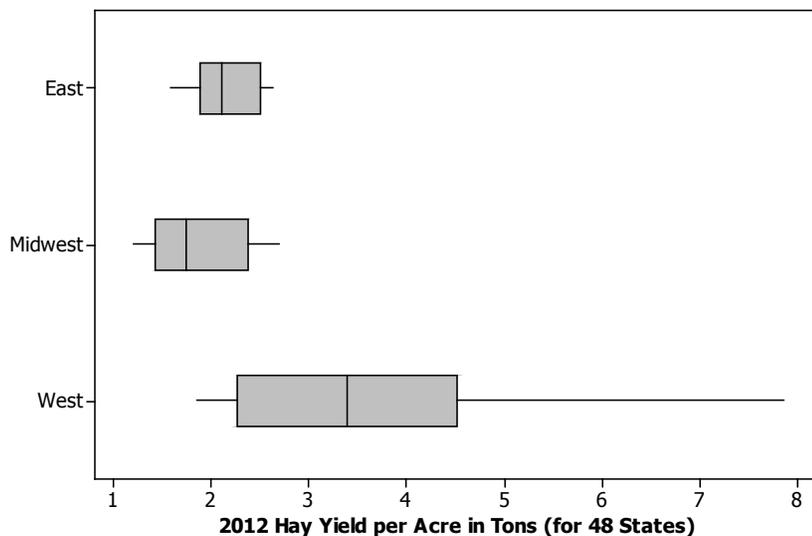
Date _____

Lesson 19: Comparing Data Distributions

Exit Ticket

According to the *United States Department of Agriculture National Agricultural Statistics Service Crop Production 2012 Summary*, in the contiguous 48 United States, there was a great deal of variability among states in terms of hay yield per acre. Do some regions of the United States generally have a higher hay yield per acre than other regions? The following box plots show the distribution of hay yield per acre (in tons) for 22 eastern states, 14 mid-western states, and 12 western states in 2012.

(From the *United States Department of Agriculture National Agricultural Statistics Service Crop Production 2012 Summary*, ISSN: 1057-7823, p. 75, accessed May 5, 2013 available at <http://usda01.library.cornell.edu/usda/current/CropProdSu/CropProdSu-01-11-2013.pdf>.)

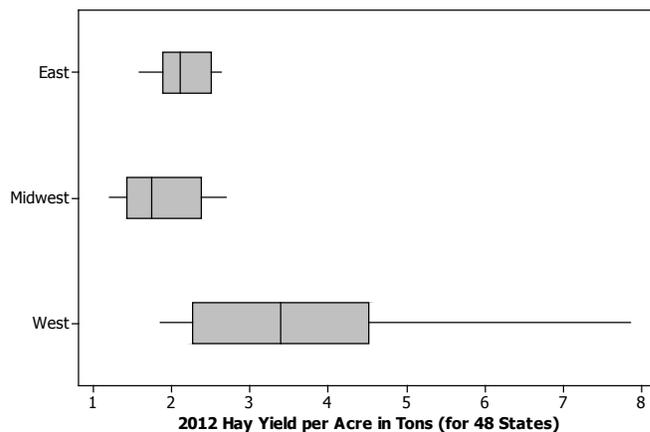


- Which of the three regions' data sets has the least variability? Which has the greatest variability? To explain how you chose your answers, write a sentence or two that supports your choices by comparing relevant summary measures (i.e., median, IQR, etc.) or graphical attributes (i.e., shape, variability, etc.) from the three groups.

Exit Ticket Sample Solutions

According to the *United States Department of Agriculture National Agricultural Statistics Service Crop Production 2012 Summary*, in the contiguous 48 United States, there was a great deal of variability among states in terms of hay yield per acre. Do some regions of the United States generally have a higher hay yield per acre than other regions? The following box plots show the distribution of hay yield per acre (in tons) for 22 eastern states, 14 mid-western states, and 12 western states in 2012.

(From the *United States Department of Agriculture National Agricultural Statistics Service Crop Production 2012 Summary*, ISSN: 1057-7823, p. 75, accessed May 5, 2013 available at <http://usda01.library.cornell.edu/usda/current/CropProdSu/CropProdSu-01-11-2013.pdf>.)



- Which of the three regions' data sets has the least variability? Which has the greatest variability? To explain how you chose your answers, write a sentence or two that supports your choices by comparing relevant summary measures (i.e., median, IQR, etc.) or graphical attributes (i.e., shape, variability, etc.) from the three groups.

The East data set has the least variability as it has the smallest range and the smallest IQR. The West data set has the greatest variability as it has the largest range and the largest IQR.

- True or False: The Western state with the smallest hay yield per acre has a higher hay yield per acre than at least half of the Midwestern states. Explain how you know this is true or how this is false.

This is true; the minimum value of the West data set is higher than the median value of the Midwest data set. Therefore, this minimum value for the West must be higher than at least half of the Midwestern states' values.

- Which region typically has states with the largest hay yield per acre? To explain how you chose your answer, write a sentence or two that supports your choice by comparing relevant summary measures or graphical attributes from the three groups.

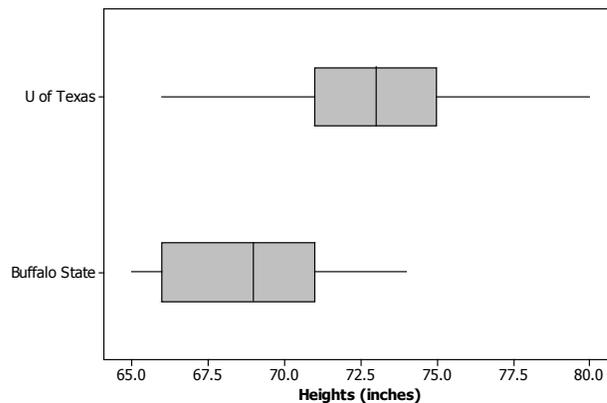
The West typically has states with the largest hay yield per acre. Over half of the Western states have hay yields that are higher than any yield in either of the other two regions. Also, some Western yields are up to two or three times the largest Eastern and Midwestern yields.

Problem Set Sample Solutions

Before students begin the problem set, consider providing them time to work on their projects. If students have not collected data, then provide assistance in completing that process. If students have collected data, then provide them time to develop numerical or graphical summaries of the data (dot plots, box plots, or histograms). Assign only one or two of the problems in the problem set if completion of the project needs to be addressed.

- College athletic programs are separated into divisions based on school size, available athletic scholarships, and other factors. A researcher is curious to know if members of swimming and diving programs in Division I (schools that offer athletic scholarships and tend to have large enrollment) are generally taller than the swimmers and divers in Division III programs (schools that do not offer athletic scholarships and tend to have smaller enrollment). To begin the investigation, the researcher creates side-by-side box plots for the heights (in inches) of members of the 2012–2013 University of Texas Men's Swimming and Diving Team (a Division I program) and the heights (in inches) of members of the 2012–2013 Buffalo State College Men's Swimming and Diving Team (a Division III program).

(From <http://www.texassports.com/sports/m-swim/mtt/tex-m-swim-mtt.html> accessed April 30, 2013, all 41 member heights listed, and <http://www.buffalostateathletics.com/roster.aspx?path=mswim&> accessed May 15, 2013, 11 members on roster; only 10 heights were listed)



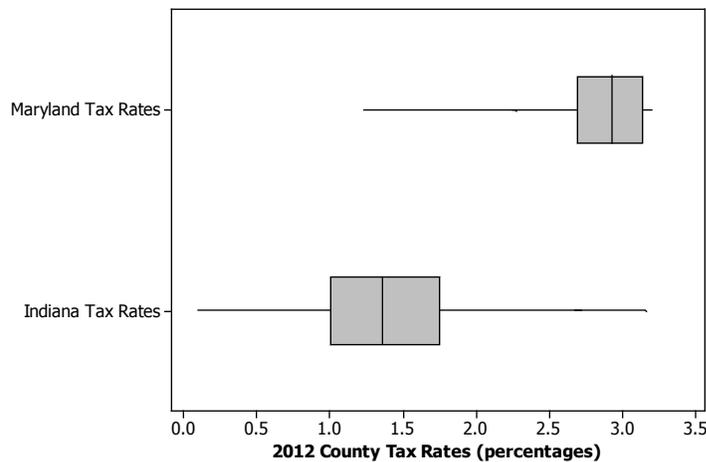
- Which data set has the smaller range?
Buffalo State
- True or False: A team member of median height on the University of Texas team would be taller than a team member of median height on the Buffalo State College team.
True.
- To be thorough, the researcher will examine many other college's sports programs to further investigate her claim that members of swimming and diving programs in Division I are generally taller than the swimmers and divers in Division III. But given the graph above, in this initial stage of her research, do you think that her claim might be valid? Carefully support your answer using comparative summary measures or graphical attributes.

Yes, a large portion of the University of Texas distribution is higher than the maximum value of the Buffalo State distribution. The median value for the University of Texas appears to be 4 inches higher than the median value of the Buffalo State distribution.

2. Different states use different methods for determining a person's income tax. However, Maryland and Indiana both have systems where a person pays a different income tax rate based on the county in which he/she lives. Box plots summarizing the 24 different county tax rates for Maryland's 23 counties and Baltimore City (taxed like a county in this case) and the resident tax rates for 91 counties in Indiana in 2012 are shown below.

(From http://taxes.marylandtaxes.com/Individual_Taxes/Individual_Tax_Types/Income_Tax/Tax_Information/Tax_Rates/Local_and_County_Tax_Rates.shtml accessed May 5, 2013 and www.in.gov/dor/files/12-county-rates.pdf accessed May 16, 2013)

- a. True or False: At least one Indiana county income tax rate is higher than the median county income tax rate



in Maryland. Explain how you know.

True. The median tax rate for Maryland appears to be a little under 3%, and the maximum tax rate in Indiana is over 3%.

- b. True or False: The 24 Maryland county income tax rates have less variability than the 91 Indiana county income tax rates. Explain how you know.

True. The tax rates in Maryland are more compact than for Indiana. Maryland has a smaller range and IQR compared to Indiana.

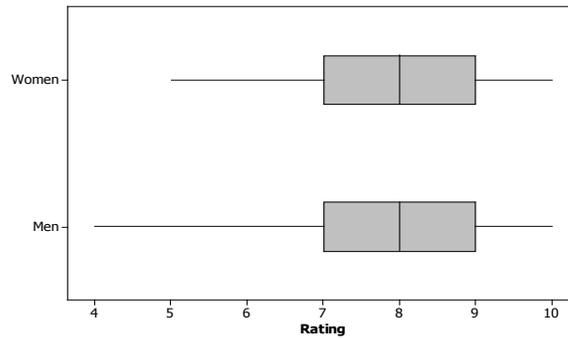
- c. Which state appears to have typically lower county income tax rates? Explain.

Indiana counties typically appear to have lower county income tax rates. The median Indiana tax rate is much lower than the median Maryland tax rate, and a large part of the Indiana distribution is lower than the minimum value of the Maryland distribution.

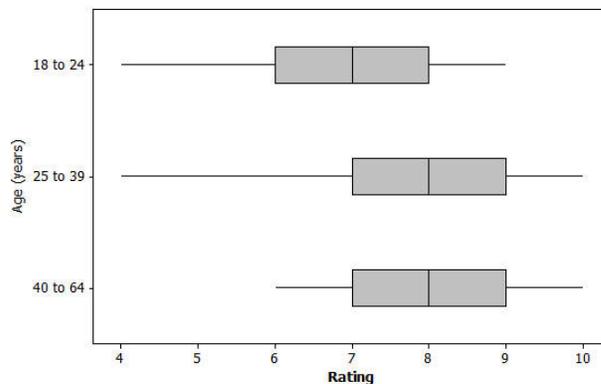
3. Many movie studios rely heavily on customer data in test markets to determine how a film will be marketed and distributed. Recently, previews of a soon to be released film were shown to 300 people. Each person was asked to rate the movie on a scale of 0 to 10, with 10 representing "best movie I've ever seen" and 0 representing "worst movie I've ever seen."

Below are some side-by-side box plots that summarize the ratings based on certain demographic characteristics.

For 150 women and 150 men:



For 3 distinct age groups:



Note: Students may be more likely to provide comparative values for this question given the discrete, integer nature of the data.

- a. Generally, does it appear that the men and women rated the film in a similar manner or in a very different manner? Write a few sentences explaining your answer using comparative information about center and spread from the graph.

It appears that the men and women rated the film in a very similar manner: same quartile values, same medians, and same maximums. The only difference is that the minimum rating from men was slightly lower than the minimum rating from women.

- b. Generally, it appears that the film typically received better ratings from the older members of the group. Write a few sentences using comparative measures of center and spread or graphical attributes to justify this claim.

For the two oldest age groups, the Q1, median, Q3, and maximum values are all higher than the 18–24 counterparts. In fact the Q1 value for each of these two older groups equals the median rating of the youngest group, and the median value for each of these two older groups equals the Q3 rating of the youngest group. Additionally, while the two oldest groups have similar distributions, the minimum score of the oldest group was much higher than the minimum value of the 25–39 group. This means that none of the 40–64 respondents rated the movie with a score as low as a 4 (as was the case in the 25–39 age group).



Lesson 20: Describing Center, Variability, and Shape of a Data Distribution from a Graphic Representation

Student Outcomes

- Given a frequency histogram, students are able to describe the data collected, including the number of responses, an estimate of the mean or median, and an estimate of the interquartile range (IQR) or the mean absolute deviation (MAD).

Lesson Notes

In each lesson of this module, students were either given graphic representations of a data distribution or they were expected to construct a graphic representation. The graphic representations used in this module are dot plots, box plots, and histograms. Each of these representations provides a summary of the data. If actual data are provided and the data set is not too large, a dot plot is generally a good way to display the data distribution. If the data set is large, a histogram is generally used to display the data distribution. Histograms are challenging for students at this grade level. A histogram provides a display of the data distribution, but the shape of a histogram can sometimes depend on the intervals used to construct the histogram. It is important to investigate the shape of the data distribution because the shape influences the choice of numerical summaries—the mean and MAD are used for data distributions that are approximately symmetric, and the median and IQR are used for data distributions that are skewed.

In this lesson, students consider a histogram of the length of yellow perch in the Great Lakes region. The data presented in this lesson are based on various scientific research studies of yellow perch during the 1990s. Before students analyze the data, share with them that the histogram is part of the yellow perch story that they will uncover in this lesson. The yellow perch is a valuable resource for the fishing industry, as well as a food source for several other species of fish and wildlife. Although the sample presented in this lesson has been simplified for students, it provides a graphic representation of data that was particularly disturbing to scientists researching the yellow perch. The histogram is a graphic representation of the length of the yellow perch. The length of the yellow perch is used to estimate the age of the fish. As a result, the histogram indicates that most of the fish were older or adult fish. It would be preferable for the younger fish to represent a larger proportion of the population of yellow perch. If the intervals representing the younger fish are less than the older intervals, then in time there will be fewer and fewer fish, possibly even indicating that the yellow perch will not survive.

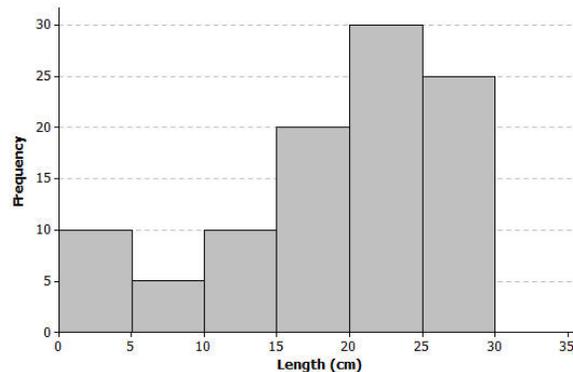
Classwork

Great Lakes yellow perch are fish that live in each of the five Great Lakes and many other lakes in the eastern and upper Great Lakes regions of the United States and Canada. Both countries are actively involved in efforts to maintain a healthy population of perch in these lakes.

Example 1 (10 minutes): The Great Lakes Yellow Perch**Example 1: The Great Lakes Yellow Perch**

Scientists collected data from many samples of yellow perch because they were concerned about the survival of the yellow perch. What data do you think researchers might want to collect about the perch?

Scientists captured yellow perch from a lake in this region. They recorded data on each fish, and then returned each fish to the lake. Consider the following histogram of data on the length (in centimeters) for a sample of yellow perch.



Discuss the following questions with students as you present the histogram.

MP.1

- What data do you think researchers might want to collect about the perch?
- How many fish captured had a length of 20 to 25 centimeters?
- Do you know how many fish had a length of 22 centimeters? Why or why not? (Remind students to understand that a histogram does not provide the frequency of a specific value, only the frequency for an interval of values.)
- Why were the scientists concerned about what they saw in the histogram of the lengths of yellow perch?

Exercises 1–11 (15 minutes)

Students should work individually or in small groups as they answer the questions in these exercises. Discuss as a group answers to these questions.

Exercises 1–11

Scientists were concerned about the survival of the yellow perch as they studied the histogram.

1. What statistical question could be answered based on this data distribution? How do you think the scientists collected these data?

Answers will vary: A possible statistical question would be, "What is a typical length of the Great Lakes yellow perch?" Remind students that a statistical question is a question that can be answered by data that you anticipate will vary.

2. Use the histogram to complete the following table:

Length of fish in centimeters (cm)	Number of fish
0 – < 5 cm	10
5 – < 10 cm	5
10 – < 15 cm	10
15 – < 20 cm	20
20 – < 25 cm	30
25 – < 30 cm	25

3. The length of each fish was measured and recorded before the fish was released back into the lake. How many yellow perch were measured in this sample?

100 fish were measured in this sample.

4. Would you describe the distribution of the lengths of the fish in the sample as a skewed data distribution or as a symmetrical data distribution? Explain your answer.

The data distribution is a skewed distribution, with the tail to the left.

5. What percentage of fish in the sample were less than 10 centimeters in length?

15 fish had a length of less than 10 centimeters, thus 15% are less than 10 centimeters.

6. If the smallest fish in this sample were 2 centimeters in length, what is your estimate of an interval of lengths that would contain the lengths of the shortest 25% of the fish? Explain how you determined your answer.

25% of the fish are represented in the first three intervals. If the smallest value in the first interval is known, then an estimate of the interval of the smallest 25% of the fish is 2 centimeters to 15 centimeters. Students would determine this by considering the histogram bars at the low end and looking for an interval that would represent 25 fish.

7. If the length of the largest yellow perch was 29 centimeters, what is your estimate of an interval of lengths that would contain the lengths of the longest 25% of the fish?

In a similar way, there are 25 fish in the interval 25 to 30 centimeters in length. If the longest fish were measured at 29 centimeters, then an estimate of the upper 25% would be 25 to 29 centimeters.

8. Estimate the median length of the yellow perch in the sample. Explain how you determined your estimate.

To estimate the median length, students would identify a length in which approximately 50% of the fish would be above and approximately 50% of the fish would be below the estimate. If a student would start from the smallest lengths, an estimate of the median would be located within the 20 to 25 centimeters interval. The same interval would be identified if students started with the largest lengths. As the actual values of the lengths of the fish are not known, any estimate within that interval would be a good estimate. For example, an estimate of 23 centimeters would be a good estimate.

9. Based on the shape of this data distribution, do you think the mean length of a yellow perch would be greater than, less than, or the same as your estimate of the median? Explain your answer.

Because the data distribution is skewed, the smaller lengths pull an estimate of the mean to the left of the median. Therefore, an estimate of the mean would be less than the estimate of the median.

10. Recall that the mean length is the balance point of the distribution of lengths. Estimate the mean length for this sample of yellow perch.

If students think of the mean as a balance point, they will estimate a length that is less than the median. Answers will vary, but an estimate of a length in the 15 to 20 centimeters interval would show an understanding of this idea. For example, 17 or 18 centimeters would be a good estimate of the mean.

11. The length of a yellow perch is used to estimate the age of the fish. Yellow perch typically grow throughout their lives. Adult yellow perch have lengths between 10 and 30 centimeters. How many of the yellow perch in this sample would be considered adult yellow perch? What percentage of the fish in the sample are adult fish?

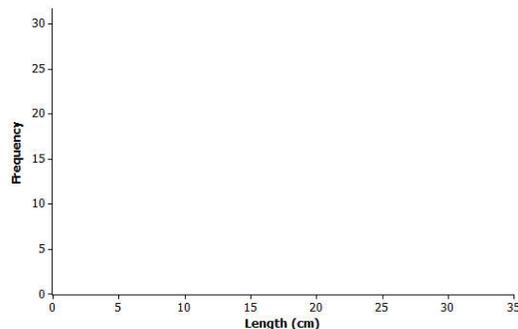
85 fish are counted in the intervals that represent 10 to 30 centimeters. Therefore, 85% of the fish in this sample were estimated to be adult fish.

Example 2 (5 minutes): What Would a Better Distribution Look Like?

Example 2: What Would a Better Distribution Look Like?

Yellow perch are part of the food supply of larger fish and other wild life in the Great Lakes region. Why do you think that the scientists worried when they saw the histogram of fish lengths given above?

Sketch a histogram representing a sample of 100 yellow perch lengths that you think would indicate the perch are not in danger of dying out?



Discuss the summary that was provided at the beginning of the teacher notes. A better distribution of fish would have more young fish. Because age is related to length, a better distribution would have more fish in the smaller length intervals. As the lengths (or ages) increased, you would expect the number of fish to decline. Allow students to sketch their own histogram shapes in response to the discussion question in this exercise. Point out that a histogram with the greater frequencies of fish in the smaller length intervals would be better.

MP.2

Exercises 12–17 (10 minutes): Estimating the Variability in Yellow Perch Lengths

Exercises 12–17: Estimating the Variability in Yellow Perch Lengths

You estimated the median length of yellow perch from the first sample in Exercise 8. It is also useful to describe variability in the length of yellow perch. Why might this be important? Consider the following questions:

12. In several previous lessons, you described a data distribution using the 5-number summary. Use the histogram and your answers to the questions in Exercise 2 to provide estimates of the values for the 5-number summary for this sample:

Min or minimum value = 2 centimeters

Q1 value = 15 centimeters

Median = 23 centimeters

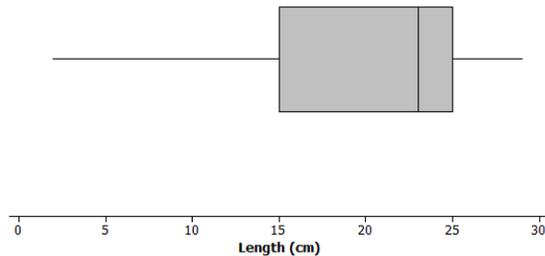
Q3 value = 25 centimeters

Max or maximum value = 29 centimeters

13. Based on the 5-number summary, what is an estimate of the value of the interquartile range (IQR) for this data distribution?

Based on the above estimates, an estimate of the interquartile range (IQR) would be as follows: 25 centimeters – 15 centimeters or 10 centimeters.

14. Sketch a box plot representing the lengths of the yellow perch in this sample.



15. Which measure of center, the median or the mean, is closer to where the lengths of yellow perch tend to cluster?

For a skewed distribution, the median is closer to where the lengths of yellow perch tend to cluster.

16. What value would you report as a typical length for the yellow perch in this sample?

Encourage students to use the median value they estimated in the previous questions as the typical value of the yellow perch.

17. The mean absolute deviation (or MAD) or the interquartile range (IQR) are used to describe the variability of a data distribution. Which measure of variability would you use for this sample of perch? Explain your answer.

When the median is selected as the measure of center for a typical value, then the interquartile range would be selected as the measure of variability. In this case, 10 centimeters, or the IQR determined in Exercise 2, would be the measure of the variability.

Closing (5 minutes)

Describe additional questions.

- What is the problem with the yellow perch length distribution shown in the opening histogram?
- What is a typical yellow perch length?
 - *(Basically, the answer to this question is the answer to students' statistical question.)*
- What would you use as a measure of the variability of yellow perch lengths?
 - *You would use the measurement of centimeters.*

Lesson Summary

Data distributions are usually described in terms of shape, center, and spread. Graphical displays, such as histograms, dot plots, and box plots, are used to assess the shape. Depending on the shape of a data distribution, different measures of center and variability are used to describe the distribution. For a distribution that is skewed, the median is used to describe a typical value, whereas the mean is used for distributions that are approximately symmetric. The IQR is used to describe variability for a skewed data distribution, while the MAD is used to describe variability for distributions that are approximately symmetric.

Exit Ticket (5 minutes)

Name _____

Date _____

Lesson 20: Describing Center, Variability, and Shape of a Data Distribution from a Graphic Representation

Exit Ticket

1. Great Lake yellow perch continue to grow until they die. What does the histogram in Example 1 indicate about the ages of the perch in the sample?
2. What feature of the histogram in Example 1 indicates that the values of the mean and the median of the data distribution will not be equal?
3. Adult yellow perch have lengths between 10 and 30 centimeters. Would a perch with a length equal to the median length be classified as an adult or pre-adult fish? Explain your answer.

Exit Ticket Sample Solutions

1. Great Lake yellow perch continue to grow until they die. What does the histogram in Example 1 indicate about the ages of the perch in the sample?

The histogram indicates that most of the perch are in the intervals corresponding to the longest lengths. Because length is related to age, the histogram indicates that there are more fish estimated as older fish.

2. What feature of the histogram in Example 1 indicates that the values of the mean and the median of the data distribution will not be equal?

The histogram indicates that the shape of the data distribution is skewed. For skewed distributions, the mean and the median are not equal.

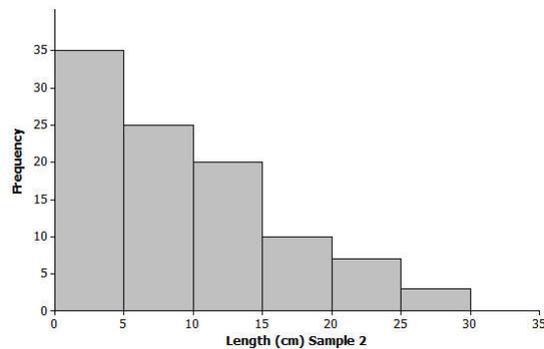
3. Adult yellow perch have lengths between 10 and 30 centimeters. Would a perch with a length equal to the median length be classified as an adult or pre-adult fish? Explain your answer.

A perch equal to the median length would be classified as an adult fish. The median is estimated to be between 20 and 25 centimeters in length. Adult fish are 10 centimeters or more in length.

Problem Set Sample Solutions

Consider again your students' progress on their projects. If needed, develop expectations for completing the four-steps of the project as part of the problem set. Indicate to students that you expect a brief written summary of their progress or a sample of the graphs or numerical summaries from their data. Assign four or five of the problems in the problem set if completion of the project needs to be addressed.

Another sample of Great Lake yellow perch from a different lake was collected. A histogram of the lengths for the fish in this sample is shown below:



1. If the length of a yellow perch is an indicator of its age, how does this second sample differ from the sample you investigated in the exercises? Explain your answer.

The second sample has more fish with lengths in the intervals corresponding to shorter lengths. Therefore, more of the fish are younger.

2. Does this histogram represent a data distribution that is skewed or that is nearly symmetrical?

This distribution is also skewed. However, the tail of this distribution is to the right, or toward the longer lengths.

3. What measure of center would you use to describe a typical length of a yellow perch in this second sample? Explain your answer.

Students should recommend the median of the data distribution as a description of a typical value of the length of the yellow perch because this distribution is also skewed.

4. Assume the smallest perch caught was 2 centimeters in length, and the largest perch caught was 29 centimeters in length. Estimate the values in the 5-number summary for this sample:

Q1, Q3, and median values are not as clear-cut in this distribution as in Exercise 4, so allow a wider range of acceptable answers.

Min or minimum value = 2 centimeters

Q1 value = 4 centimeters (value greater than 2, but within the interval of 0 to 5 centimeters)

Median value = 7 centimeters (a value within the interval of 5 to 10 centimeters)

Q3 value = 12 centimeters (a value within the interval of 10 to 15 centimeters)

Max or maximum value = 29 centimeters

5. Based on the shape of this data distribution, do you think the mean length of a yellow perch from this second sample would be greater than, less than, or the same as your estimate of the median? Explain your answer.

An estimate of the mean would be greater than the median length because the values in the tail or to the right of the median pull the mean in that direction. Consider estimating the mean as the balance point of this distribution. (If students have problems with estimating the balance point, consider providing them a representation similar to the representation used to introduce a balance point in earlier lessons. Use a ruler with coins (or weights) taped to locations that would represent a skewed distribution. This representation helps them sense the point of balance.)

6. Estimate the mean value of this data distribution.

An estimate of the mean would be a value slightly larger than the median value. For example, a mean of 10 or 11 centimeters would be a reasonable estimate of a balance point.

7. What is your estimate of a typical length of a yellow perch in this sample? Did you use the mean length from problem 5 for this estimate? Explain why or why not.

As the median was selected as the estimate of a measure of center, a value of 7 centimeters (or whatever students used to estimate the median) would be an estimate of a typical value for a yellow perch from this sample.

8. Would you use the MAD or the IQR to describe variability in the length of Great Lakes yellow perch in this sample? Estimate the value of the measure of variability that you selected.

Students should use the IQR to describe the variability because the data distribution is skewed and the median was used as a measure of a typical value. An estimate of the IQR based on the above estimates would be as follows: $12 - 4$ centimeters = 8 centimeters.



Lesson 21: Summarizing a Data Distribution by Describing Center, Variability, and Shape

Student Outcomes

- Given a data set, students are able to describe the data collected, including the number of responses, mean or median, and the MAD or the interquartile range (IQR).

Lesson Notes

This lesson provides an opportunity for students to summarize a given data set. In the problem set of this module, students are expected to summarize the data collected in Lesson 17 by constructing a poster or an outline of a presentation. This lesson guides students through the four steps used to carry out a statistical study in order to prepare them for the presentation in Lesson 22.

Classwork

Each of the lessons in this module is about data. What are data? What questions can be answered by data? How do you represent the data distribution so that you can understand and describe its shape? What does the shape tell us about how to summarize the data? What is a typical value of the data set? These questions, and many others, were part of your work in the exercises and investigations. There is still a lot to learn about what data tell us. You will continue to work with statistics and probability in grades seven and eight and throughout high school. You have already, however, started to learn how to uncover the stories behind data.

When you started this module, the four steps used to carry out a statistical study were introduced:

- Step 1:** Pose a question that can be answered by data.
- Step 2:** Collect appropriate data.
- Step 3:** Summarize the data with graphs and numerical summaries.
- Step 4:** Answer the question posed in step 1 using the numerical summaries and graphs.

In this lesson, you will carry out these steps using a given data set.

Exploratory Challenge: Annual Rainfall in the State of New York (25 minutes)

This is an exploration lesson. Students should be given approximately 25 minutes to work independently in completing the template that provides a structure for summarizing the rainfall data.

Students are given the annual rainfall in inches for New York from 1983 to 2012. The data were obtained from the National Climate Data Center. (If any students need data for their presentation discussed in the problem set, this site provides climate data for regions, cities, and states and could be a source to help students struggling to obtain data.)

Before students organize their summary, discuss the context explained in the lesson. Make sure the students understand what the data represent by highlighting the words *annual* and *rainfall*. Ask students to guess a value that they think represents the typical rainfall for New York in a year. Record these guesses and refer to them after students have summarized this data set and calculated a measure of center. Also, ask students why a statistical study of rainfall is important. For example, when a reporter says that a certain year was unusually rainy, on what basis was that claim made?

Direct students to study the template that is included with this lesson. Review the four steps involved in a statistical study. Indicate that during the next 25 minutes they are expected to complete the template that organizes their statistical summary of the data.

Exploratory Challenge: Annual Rainfall in the State of New York

The National Climate Data Center collects data throughout the United States that can be used to summarize the climate of a region. You can obtain climate data for a state, a city, a county, or a region. If you were interested in researching the climate in your area, what data would you collect? Explain why you think this data would be important as a statistical study of the climate in your area.

For this lesson, you will use yearly rainfall data for the state of New York that were compiled by the National Climate Data Center. The following data are the number of inches of rain (averaged over various locations in the state) for the years from 1983 to 2012 (30 years).

45	42	39	44	39	35	42	49	37	42	41	42	37	50	39
41	38	46	34	44	48	50	47	49	44	49	43	44	54	40

Use the four steps to carry out a statistical study using this data.

Step 1: Pose a question that can be answered by data.

What is a statistical question that you think can be answered with these data? Write your question in the template provided for this lesson.

Step 2: Collect appropriate data.

The data have already been collected for this lesson. How do you think these data were collected? Recall that the data are the number of inches of rain (averaged over various locations in the state) for the years from 1983 to 2012 (30 years). Write a summary of how you think the data were collected in the template for this lesson.

Step 3: Summarize the data with graphs and numerical summaries.

A good first step might be to summarize the data with a dot plot. What other graph might you construct? Construct a dot plot or another appropriate graph in the template for this lesson.

What numerical summaries will you calculate? What measure of center will you use to describe a typical value for these data? What measure of variability will you calculate and use to summarize the spread of the data? Calculate the numerical summaries and write them in the template for this lesson.

Step 4: Answer your statistical question using the numerical summaries and graphs.

Write a summary that answers the question you posed in the template for this lesson.

The following directions should be considered as students develop a statistical summary of this data. Work with students individually or in small groups as they complete the template.

Step 1: Pose a question that can be answered by data.

It is important that students are reminded of the two most important parts of the definition of a statistical question. A statistical question is (1) a question that is answered by data, and (2) a question that anticipates the data will vary. As students examine the data, point out to them that there is variability. Although students may vary the wording of their questions, it is anticipated most students will form a question that essentially asks, “What is the typical annual rainfall in New York?”

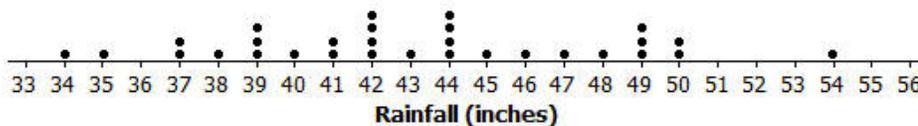
Step 2: How do you think the data were collected?

The data are given to students in this lesson. This step will be more challenging as they carry out their own statistical study because they need to explain the plan they developed to collect their data. For this lesson, allow students to speculate on how the National Climate Data Center probably collected this data. As the data represent the annual rainfall for the state of New York, the Center had to collect rainfall totals from several reporting weather centers around the state. They calculated an average of those levels for each day of the year. At the end of the year, the National Climate Data Center averaged those daily results. Students might be asked how a rainfall level is measured at a weather center. A rain gauge might be a good visual to share with students.

MP.4 Step 3: Construct graphs and calculate numerical summaries of the data.

This step represents most of the work students will be expected to do in this lesson. As a first step, encourage students who are not sure how to start summarizing the data to construct a dot plot. A blank grid is provided at the end of the Teacher Notes that can be duplicated for students who may need some structure in making a dot plot. This grid could also be used if any student decides to develop a box plot or a histogram of the data distribution.

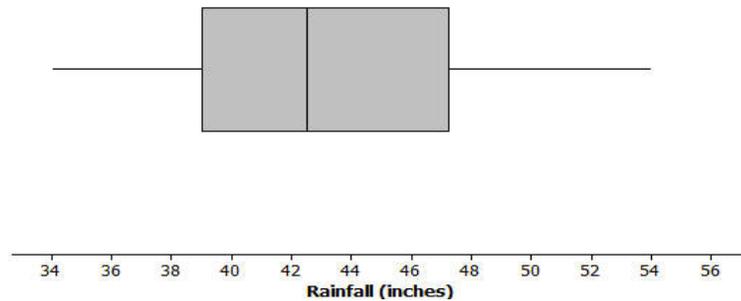
Dot plot of annual rainfall from 1983 to 2012



Students examine the dot plot and make decisions about the data distribution. For example, is the distribution approximately symmetric or is the distribution skewed? This dot plot shows a data distribution that is approximately symmetric.

Based on the decision that the distribution is approximately symmetric, students should proceed to calculate the mean as a measure of center and the MAD as a measure of variability. Some students might also choose to investigate this distribution with a box plot to answer the question about the symmetry.

Box plot of annual rainfall from 1983 to 2012



The box plot does not indicate a perfectly symmetrical distribution; however, it is approximately symmetric. The mean and the median of this data distribution are nearly equal to each other. Based on this decision, students should calculate the mean and the MAD.

The mean rainfall is 43 inches (to the nearest inch), and the mean absolute deviation (MAD) is 3.9 or 4 inches (to the nearest inch).

At the end of the Teacher Notes is a table that could be used for students who need structure in calculating the MAD. There are several steps in calculating the MAD that might need to be organized for some students.

Step 4: Answer your statistical question using the numerical summaries and graphs.

MP.3

This step asks students to write a short summary interpreting the graphs and numerical summaries. Students should connect this back to their statistical question. Students would indicate that the typical rainfall for New York is 43 inches per year. They would also indicate that a typical deviation from the mean is about 4 inches.

Template for Lesson 21

Step 1: What is your statistical question?

Step 2: How do you think the data were collected?

Step 3: Construct graphs and calculate numerical summaries of the data.

Construct at least one graph of the data distribution. Calculate appropriate numerical summaries of the data. Also indicate why you selected these summaries.

Step 4: Answer your statistical question using your graphs and numerical summaries.

Closing (5 minutes)

If time permits, look at the original guesses students made to this question. Did students have a pretty good idea of the annual rainfall in New York? Discuss this question with students.

Lesson Summary

Statistics is about using data to answer questions. The four steps used to carry out a statistical study include posing a question that can be answered by data, collecting appropriate data, summarizing the data with graphs and numerical summaries, and using the data, graphs, and summaries to answer the statistical question.

Exit Ticket

Consider a special type of Exit Ticket for this lesson. As students are expected to complete a summary of the four-step investigative study, use this opportunity to assess your students' understanding of this process as related to the question they formed and the data they collected. The Exit Ticket for this lesson is to complete the following direction (state this direction to the students):

Based on your current preparation, summarize the four steps you are expected to complete as part of presenting a statistical study.



Name _____

Date _____

Lesson 21: Summarizing a Data Distribution by Describing Center, Variability, and Shape

Exit Ticket

Based on your current preparation, summarize the four steps you are expected to complete as part of presenting a statistical study.

Exit Ticket Sample Solutions

Based on your current preparation, summarize the four steps you are expected to complete as part of presenting a statistical study.

Step 1: *State my statistical question. My question is based on collecting data that will vary.*

Step 2: *Devise a plan to collect data. I prepared a question to ask the students in my class. (Allow students to explain the question they asked, the responses they received, and the method they used for recording answers.)*

Step 3: *Summarize my data. I prepared a dot plot of the responses to the question. My dot plot indicated that the responses to my question were skewed to the left; therefore, I used the median of the data distribution to describe my center and the IQR to describe the variability. (Allow for a summary of the specific median or mean, and a specific summary of the variability as the MAD or IQR.)*

Step 4: *Based on my graphs and numerical summaries, I answered my question.*

Problem Set Sample Solutions

The problem set for this lesson involves creating a poster or an outline for a presentation using the data collected in Lesson 17. The directions in the lesson indicate that students are expected to carry out the four steps either on their poster or outlined for a presentation. If students provided an adequate summary of the four-step process in the exit ticket, they could use their summary as a guide in completing the poster. Highlight the following with students:

For Step 1, students are expected to have a question clearly identified as their statistical question. The question should involve the data they collected. Students should have anticipated variability in the data.

For Step 2, students should indicate how they collected the data based on the plan proposed in Lesson 17. For example, for a question that investigates a typical height of students in the class, did every student state his or her height in inches or was there a way to measure everyone's height? For a question that investigates how many books students read, did students ask members of their class how many books they read each month?

For Step 3, students include graphs and numerical summaries of the data. (Again, if students need more structure in constructing their graphs, provide them with the coordinate grids for this lesson.) It is anticipated that students begin with a dot plot. From the dot plot, students might construct a box plot or a histogram. Based on the shape of the distribution, students select appropriate numerical summaries—either the mean and the mean absolute deviation (MAD) or the median and the interquartile range (IQR). Posters or outlines should indicate what summaries were used and why.

For Step 4, students should have a concluding statement that answers the statistical question. Students should provide a brief description of their numerical summaries and graphs.

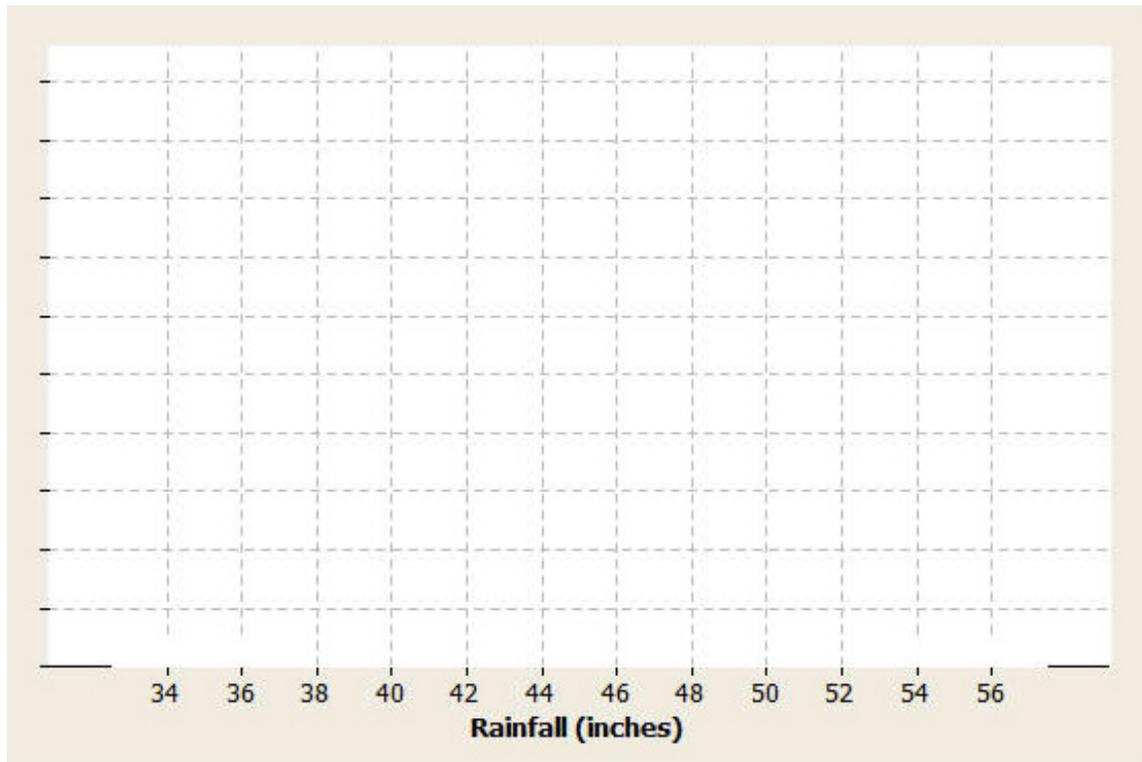
In Lesson 17, you posed a statistical question and a plan to collect data to answer your question. You also constructed graphs and calculated numerical summaries of your data. Review the data collected and your summaries.

Based on directions from your teacher, create a poster or an outline for a presentation using your own data. On your poster, indicate your statistical question. Also, indicate a brief summary of how you collected your data based on the plan you proposed in Lesson 17. Include a graph that shows the shape of your data distribution, along with summary measures of center and variability. Finally, answer your statistical question based on the graphs and the numerical summaries.

Share the poster you will present in Lesson 22 with your teacher. If you are instructed to prepare an outline of the presentation, share your outline with your teacher.

Additional Resource Materials

The following could be used to provide structure in constructing a dot plot, histogram, or box plot of the rainfall data. A similar type of grid (or graph paper) could be prepared for students as they complete the problem set. The grid provided for students should not include the units along the horizontal axis as that is part of what they are expected to do in preparing their summaries.



The following table could be used for students requiring some structure in calculating the mean absolute deviation, or MAD:

Data Value	Mean	Data Value – Mean	Absolute Value of Data Value – Mean
45			
42			
39			
44			
39			
35			
42			
49			
37			
42			
41			
42			
37			
50			
39			
41			
38			
46			
34			
44			
48			
50			
47			
49			
44			
49			
43			
44			
54			
40			



Lesson 22: Presenting a Summary of a Statistical Project

Student Outcomes

- Based on the data collected by students or on a sample set of data (for cases in which collecting data was not possible), students communicate conclusions based on the data distribution.

Lesson Notes

This is an exploratory lesson. As indicated in previous lessons, students build up to this lesson. In this lesson, each student has an opportunity to present a summary of his or her statistical study. Students should be reminded that their presentation should focus on the four-step investigative process. It is this process that defines a statistical study for students at this grade level.

MP.3

If students carried out the process outlined in previous lessons, this lesson is a formal presentation day in which they either display and explain their posters or are provided a few minutes to explain their statistical study. If there is not enough time for students to formally present their study, organize a gallery walk. Hang posters around a classroom and allow students to view as many as possible. Encourage students to take notes as they read the posters. Provide each student with a general template (see a suggested template below) to summarize at least one poster as part of a whole class discussion. Conclude the gallery walk with a short discussion of what they saw and what questions interested them. Ask students if there were any studies that surprised them. (Often a statistical study confirms a conjecture; there are times, however, that data lead to conclusions that were not expected.)

The audience for the presentations may vary. In most cases, the class is the audience. However, this type of project allows for other formats. It might be possible to use this day as an opportunity to invite parents to listen to the presentations, or school administrators or other available teachers.

Anticipate that problems will arise. In the event that there are students who did not complete Lesson 17 or were not able to collect data on their own, the posters or presentations can be based on data obtained from an outside source. It was pointed out in each of the lessons leading up to this presentation day that students were to advise their teachers about their progress. Students presenting a study based on data they did not collect must give proper credit to the source of that data on their poster or in their presentation.

Formal speaking is a comfortable and exciting experience for some students. For other students, it is an intimidating and possibly frightening experience. Teachers should use their best judgment in terms of organizing the formal presentations. If there are any students who need a little more structure in sharing their ideas, the following partially completed table could be provided to these students. Use it to help them organize their thoughts. The posters provide a format for students to present their ideas without formally presenting their studies.

Presentation Outline

A statistical study involves the following four-step investigative process:

- Step 1: Pose a question that can be answered by data.
- Step 2: Collect appropriate data.
- Step 3: Summarize the data with graphs and numerical summaries.
- Step 4: Answer the question posed in Step 1 using the numerical summaries and graphs.

Now it is your turn to be a researcher and to present your own statistical study. In Lesson 17, you posed a statistical question, proposed a plan to collect data to answer the question, and collected the data. In Lesson 21, you created a poster or an outline of a presentation that included the following: the statistical question, the plan you used to collect the data, graphs and numerical summaries of the data, and an answer to the statistical question based on your data. Use the following table to organize your presentation.

Points to consider:	Notes to include in your presentation:
(1) Describe your statistical question.	<i>"My statistical question is"</i>
(2) Explain to your audience why you were interested in this question.	<i>"I am interested in finding an answer to this question because"</i>
(3) Explain the plan you used to collect the data.	<i>"My plan for collecting data to answer my question was... "I was able to collect my data as planned." (If you were not able to collect the data, explain why.) Explain any challenges or unexpected reactions in collecting your data.</i>
(4) Explain how you organized the data you collected.	<i>"Let me explain how I organized my data and prepared my summaries." Students might use a table to summarize the data or organize data in a list that could be used to prepare a dot plot or a box plot.</i>
(5) Explain the graphs you prepared for your presentation and why you made this graph.	<i>"I developed a dot plot to start my statistical study because"</i>
(6) Explain what measure of center and what measure of variability you selected to summarize your study. Explain why you selected these values.	<i>"I selected as a measure of center the (mean or median). I selected this measure because"</i>
(7) Describe what you learned from the data. (Be sure to include an answer to the question from step (1) above.)	<i>"Let me tell you the answer to my statistical question..."</i>

Evaluation of Posters

Various evaluation techniques are possible. Given that students' work involves several steps, including how well students display and organize their work, it is recommended that a well-defined rubric be developed for this work. A sample rubric is available at the American Statistical Association's website:

<http://www.amstat.org/education/posterprojects/index.cfm>.

Rubric designs are highly dependent on the process used to complete this project; therefore, the final rubric design should be a teacher decision. Assessment of the project should provide students with feedback regarding the statistical question, the collection of the data, the summary of the data by graphs and numerical summaries, and the conclusions reached in answering the statistical question.

Closing Exercise (10 minutes)

Final thoughts on this module and the statistical study are encouraged with the closing questions provided to students.

Closing Exercise

After you have presented your study, consider what your next steps are by answering the following questions:

1. What questions still remain after you concluded your statistical study?
2. What statistical question would you like to answer next as a follow-up to this study?
3. How would you collect the data to answer the new question you posed in (2)?

Encourage a discussion around these questions, or if time is not available for a discussion, encourage students to write their responses.

Lesson Summary

Statistics is about using data to answer questions. The four steps used to carry out a statistical study include posing a question that can be answered by data, collecting appropriate data, summarizing the data with graphs and numerical summaries, and using the data, graphs, and summaries to answer the statistical question.

Template for Lesson 22: Summarizing a Poster

Step 1: What was the statistical question presented on this poster?

Step 2: How was the data collected?

Step 3: What graphs and calculations were used to summarize data?

Summarize at least one graph presented on the poster. (For example, was it a dot plot? What was represented on the scale?) Summarize any appropriate numerical summaries of the data (for example, the mean or the median). Also indicate why these summaries were selected.

Step 4: Summarize the answer to the statistical question.

Name _____

Date _____

1. A group of students was asked how many states they have visited in their lifetime. Below is a dot plot of their responses.



- a. How many observations are in this data set?

- b. In a few sentences, summarize this distribution in terms of shape, center, and variability.

- c. Based on the dot plot above and without doing any calculations, circle the best response below and then explain your reasoning.
 - A. I expect the mean to be larger than the median.
 - B. I expect the median to be larger than the mean.
 - C. The mean and median should be similar.

Explain:

d. To summarize the variability of this distribution, would you recommend reporting the interquartile range or the mean absolute deviation? Explain your choice.

e. Suppose everyone in the original data set visits one new state over summer vacation. Without doing any calculations, describe how the following values would change (i.e., larger by, smaller by, no change – be specific).

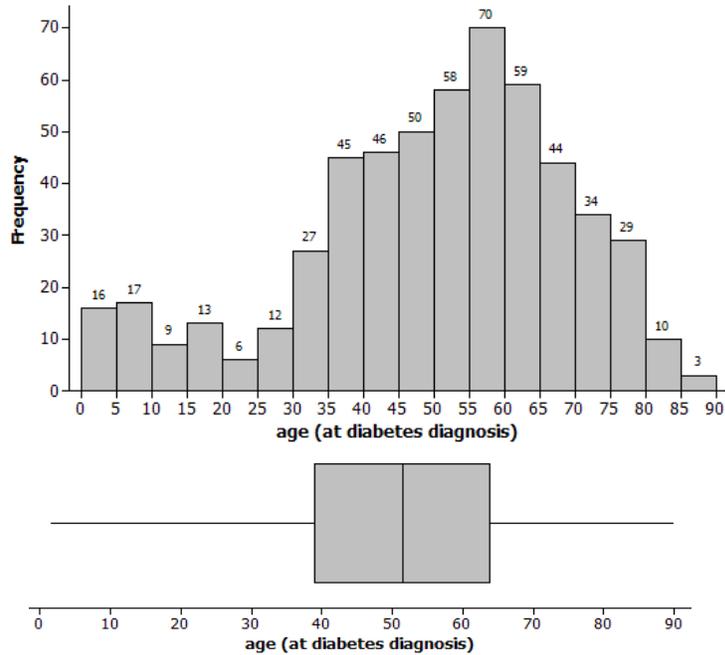
Mean:

Median:

Mean Absolute Deviation:

Interquartile Range:

2. Diabetes is a disease that occurs in both young and old people. The histogram and box plot below display the ages at which 548 people with diabetes first found out that they had this disease.



The American Diabetes Association has identified two types of diabetes:

- Type I diabetes is when the body does not produce insulin. Type I diabetes is usually first found in children and young adults (less than 20 years of age).
- Type II diabetes is when either the body does not produce enough insulin or the cells ignore the insulin. Type II diabetes is usually first found in older adults (50 years of age or older).

- a. Explain how the histogram reflects there being these two types of diabetes.

- b. The American Diabetes Association says that only about 5% of people with diabetes have type I diabetes. From the graphs, estimate the percentage of these 548 people who found out they had the disease before age 20. Clearly explain how you are doing so and which graph(s) you are using.
- c. Suggest a statistical question that the box plot of the age data would allow you to answer more quickly than the histogram would.
- d. The interquartile range for these data is reported to be 24. Write a sentence interpreting this value in the context of this study.

3. The following table lists the diameters (in miles) of the original nine planets.

Planet	Diameter (in miles)
Mercury	3030
Venus	7520
Earth	7926
Mars	4217
Jupiter	88838
Saturn	74896
Uranus	31762
Neptune	30774
Pluto	1428

- a. Calculate the 5-number summary (minimum, lower quartile, median, upper quartile, and maximum) of the planet diameters. Be sure to include measurement units with each value.

Minimum:

Lower quartile:

Median:

Upper quartile:

Maximum:

- b. Calculate the interquartile range (IQR) for the planet diameters.
- c. Draw a box plot of the planet diameters.

- d. Would you classify the distribution of planet diameters as roughly symmetric or skewed? Explain.
- e. Pluto was recently reclassified as a “dwarf planet” because it is too small to “clear other objects out of its path.” The mean diameter with all 9 planets is 27,821 miles with $MAD = 25,552$ miles. Use this information to argue whether or not Pluto is substantially smaller than the remaining eight planets.

A Progression Toward Mastery

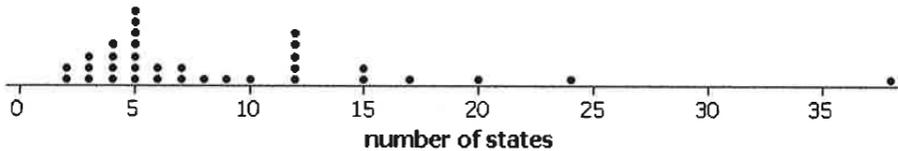
Assessment Task Item	STEP 1 Missing or incorrect answer and little evidence of reasoning or application of mathematics to solve the problem.	STEP 2 Missing or incorrect answer but evidence of some reasoning or application of mathematics to solve the problem.	STEP 3 A correct answer with some evidence of reasoning or application of mathematics to solve the problem <u>or</u> an incorrect answer with substantial evidence of solid reasoning or application of mathematics to solve the problem.	STEP 4 A correct answer supported by substantial evidence of solid reasoning or application of mathematics to solve the problem.	
1	a 6.SP.B.5a	Student does not use information in dot plot.	Student counts 15 unique outcomes for <i>number of states</i> .	Student counts 33 or 35 dots.	Student correctly counts 34 dots.
	b 6.SP.A.2	Student does not use information in dot plot.	Student's descriptions are not consistent with the graph.	Student only addresses two of the main components of shape, center, and variability.	Student gives complete description of each component, e.g., "The distribution is skewed with many students visiting around 5 states but another clump around 12 states. Almost all of the students have visited fewer than 15 states, but one outlier has visited more than 35."
	c 6.SP.B.5d	Student only circles an incorrect response with no justification.	Student circles A but does not justify choice or circles B or C but justification is inconsistent, e.g., "The median should be larger because there are some students who have visited a lot of states."	Student circles A but justification is incomplete or inconsistent, e.g., "The mean will be larger because most of the day is between 0 and 10."	Student circles A and comments on how the skewness and/or outliers will pull the value of the mean to the right of the value of the median.
	d 6.SP.B.5d	Student does not address the choice of measure of variability.	Student makes a choice but does not justify response.	Student makes a choice but justification is inconsistent (e.g., MAD because of the outlier).	Student chooses the interquartile range because of the presence of skewness and/or outliers.

	e 6.SP.A.3	Student only indicates that there is not enough information to answer the question.	Student indicates a correct change for each measure but provides no justification <u>QR</u> can only address two of the four measures.	Student answers and justifications are not completely consistent with mean and median measuring center and MAD and IQR measuring spread.	Mean and median increase by one because everyone shifts the same amount. MAD and IQR stay the same because everyone shifts the same amount.
2	a 6.SP.A.2	Student does not utilize information from the histogram.	Student discusses ages and the recommendation but does not connect to the histogram.	Student discusses observations below 20 and above 20 but does not address the distinctness of the two clumps of values around 20.	Student focuses on the two clumps, with the valley between the clumps around the cited 20 years of age.
	b 6.SP.B.5a	Student makes no use of distribution (e.g., $\frac{20}{90}$).	Student uses the box plot and assumes half of the 25% below 40 years of age are below 20 years of age.	Student uses the histogram but makes a minor calculation error. Result is still consistent with graph (e.g., below 20%).	Student uses the histogram and finds $\frac{16+17+9+13}{548} \approx 0.10$ or about 10%.
	c 6.SP.A.1	Student fails to distinguish between box plot and histogram.	Student suggests a feature of the box plot but does not relate to a statistical question, e.g., "How many people have diabetes?"	Student suggests a feature of the box plot but statistical question is vague or incomplete, e.g., "Where are most of the ages?"	Student suggests a statistical question that relates to quartiles or median, e.g., "What is the median age at diagnosis?"
	d 6.SP.A.3	Student only attempts to explain how IQR is calculated and does so incorrectly.	Student does not interpret the IQR as a measure of spread, e.g., "Most people are diagnosed around 24 years."	Student addresses the IQR as a measure of spread but does not provide an interpretation in context, e.g., "Q3 – Q1."	Student correctly answers that the width of the middle 50% of the diagnosed ages is 24 years.
3	a 6.SP.B.5	Student is not able to identify information correctly from table.	Student does not sort the observations before making computations: 3030, 7926, 88838, 31762, 1428.	Student sorts data but performs minor calculation errors or only correctly finds 3 of the values.	Student sorts data correctly: 1428, 3623.5, 7926, 53329, 88838 miles.
	b 6.SP.B.5c	Student does not perform calculation.	Student calculates Q3 – Q2 or Q2 – Q1 or reports a negative value.	Student reports correct values from (a), but as a range, and does not calculate the difference in values.	Student uses Q3 – Q1 using values from (a).
	c 6.SP.B.4	Student does not construct a graph.	Student constructs a dot plot or histogram.	Graph is well constructed and scaled but does not match values reported in (a).	Graph is correctly constructed and scaled using values from (a).

<p>d 6.SP.A.2</p>	<p>Student does not justify the choice.</p>	<p>Student gives a reasonable response but is not based on a graph of the data or on the quartiles, e.g., “A few planets are really large so the data are skewed.”</p>	<p>Student refers to box plot and/or quartiles but draws inconsistent conclusion (e.g., describes the correct box plot as symmetric or judges distance between quartiles as similar because values are so large).</p>	<p>Student uses the box plot and judges the relative widths of segments to answer question.</p>
<p>e 6.SP.B.5</p>	<p>Student only uses the context and does not use the mean and MAD values.</p>	<p>Student compares Pluto’s diameter to the mean without considering the MAD.</p>	<p>Student answers generically, e.g., “Not all observations will fall within the mean,” but does not relate to Pluto’s value explicitly.</p>	<p>Student compares Pluto’s diameter (1428) to mean – MAD = 27821 – 25552. Pluto is smaller but is not much more than one MAD from the mean.</p>

Name _____ Date _____

1. A group of students was asked how many states they have visited in their lifetime. Below is a dot plot of their responses.



- a. How many observations are in this data set? *34*
- b. In a few sentences, summarize this distribution in terms of shape, center, and variability.

A typical number of states is 5 but there is a lot of variability (eg., 2 states to over 35 states). There were also 5 people who visited 12 states. The distribution is skewed with one very high value.

- c. Based on the dot plot above and without doing any calculations, circle the best response below and then explain your reasoning.

- A. I expect the mean to be larger than the median.
- B. I expect the median to be larger than the mean.
- C. The mean and median should be similar.

Explain:

The mean will get pulled higher by the student who has been to a lot more states.

- d. To summarize the variability of this distribution, would you recommend reporting the interquartile range or the mean absolute deviation? Explain your choice.

The interquartile range because we do have a few extreme values which would enlarge the MAD or the mean absolute deviation.

- e. Suppose everyone in the original data set visits one new state over summer vacation. Without doing any calculations, describe how the following values would change (i.e., larger by, smaller by, no change – be specific).

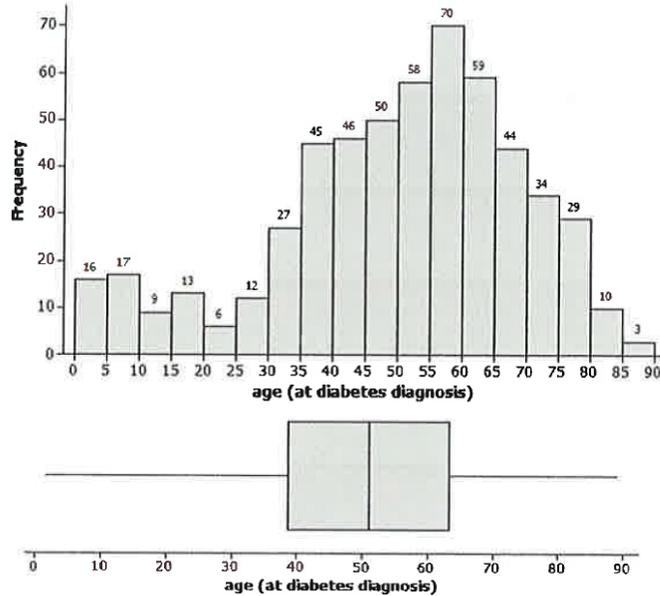
Mean: The mean would increase by one state.

Median: The median would increase by one state.

Mean Absolute Deviation: The MAD would not change.

Interquartile Range: The interquartile range would not change.

2. Diabetes is a disease that occurs in both young and old people. The histogram and box plot below display the ages at which 548 people with diabetes first found out that they had this disease.



The American Diabetes Association has identified two types of diabetes:

- Type I diabetes is when the body does not produce insulin. Type I diabetes is usually first found in children and young adults (less than 20 years of age).
- Type II diabetes is when either the body does not produce enough insulin or the cells ignore the insulin. Type II diabetes is usually first found in older adults (50 years of age or older).

- a. Explain how the histogram reflects there being these two types of diabetes.

There are two "humps" in the distribution - one to the left of 20 years and one around 50-55 years.

- b. The American Diabetes Association says that only about 5% of people with diabetes have type I diabetes. From the graphs, estimate the percentage of these 548 people who found out they had the disease before age 20. Clearly explain how you are doing so and which graph(s) you are using.

$$\frac{16 + 17 + 9 + 13}{548} = \frac{55}{548} \approx .10$$

About 10% of this sample
(found by finding the sum of the bars below 20 and dividing by the number of people in the study).

- c. Suggest a statistical question that the box plot of the age data would allow you to answer more quickly than the histogram would.

What is the median age at which people are diagnosed with diabetes?

- d. The interquartile range for these data is reported to be 24. Write a sentence interpreting this value in the context of this study.

The "length" of the ages for the middle 50% of ages is 24 years, or the distance between the top 25% and the bottom 25% of ages is 24 years.

3. The following table lists the diameters (in miles) of the original nine planets.

Planet	Diameter (in miles)
Mercury	3030
Venus	7520
Earth	7926
Mars	4217
Jupiter	88838
Saturn	74896
Uranus	31762
Neptune	30774
Pluto	1428

Handwritten notes showing the sorted diameters: 1428, 3030, 4217, 7520, 7926, 30774, 31762, 74896, 88838. The values 7926 and 88838 are circled, and an arrow points to 3030.

a. Calculate the five-number summary (minimum, lower quartile, median, upper quartile, and maximum) of the planet diameters. Be sure to include measurement units with each value.

Minimum: 1428 miles

Lower quartile: $(3030 + 4217) / 2 = 3623.5$ miles

Median: 7926 miles

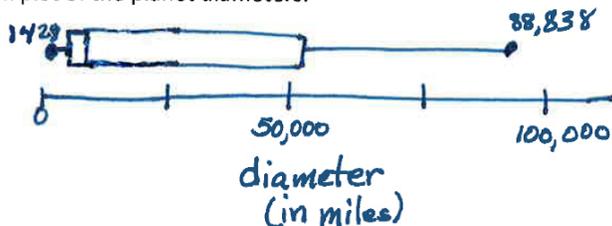
Upper quartile: $(31762 + 74896) / 2 = 53,329$ miles

Maximum: 88,838 miles

b. Calculate the interquartile range (IQR) for the planet diameters.

$$53,329 - 3,623.5 = 49,705.5 \text{ miles}$$

c. Draw a box plot of the planet diameters.



- d. Would you classify the distribution of planet diameters as roughly symmetric or skewed? Explain.

Skewed because the box to the right of the median stretches out much further than the box to the left of the median.

- e. Pluto was recently reclassified as a “dwarf planet” because it is too small to “clear other objects out of its path.” The mean diameter with all 9 planets is 27,821 miles with MAD = 25,552 miles. Use this information to argue whether or not Pluto is substantially smaller than the remaining eight planets.

Pluto = 1428 miles in diameter

$$\text{Mean} - \text{MAD} = 27,821 - 25,552 = 2,269 \text{ miles}$$

Pluto is a little more than one mean absolute deviation from the mean. As a result, Pluto is not substantially smaller.