Lesson 8: Distributions—Center, Shape, and Spread

Student Outcomes

- Students describe data distributions in terms of shape, center, and variability.
- Students use the mean and standard deviation to describe center and variability for a data distribution that is approximately symmetric.

Lesson Notes

In this lesson, students review key ideas developed in Grades 6 and 7 and Algebra I. In particular, this lesson revisits distribution shapes (approximately symmetric, mound shaped, and skewed) and the use of the mean and standard deviation to describe center and variability for distributions that are approximately symmetric. The steps to calculate standard deviation are not included in the student edition since it is recommended that students use technology to calculate standard deviation. If students are not familiar with standard deviation, the steps to calculate this value are included under the notes for Example 1. The major emphasis should be on the interpretation of the mean and standard deviation rather than on calculations.

Classwork

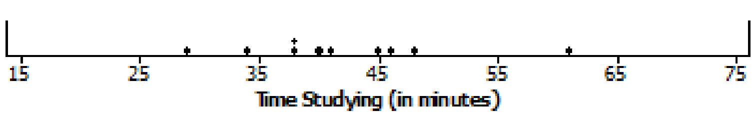
Opening (5 minutes): Review of Distribution Shapes and Standard Deviation

Prior to beginning Example 1, students may need to review the interpretation of the *mean* of a distribution. The mean is the "fair share" value. It also represents the balance point of the distribution (the point where the sum of the deviations to the left of the mean is equal to the sum of the deviations to the right of the mean). The mean can be interpreted as an average or a typical value for a data distribution.

Show students the following data on the amount of time (in minutes) spent studying for a math test for 10 students:

34	38	48	41	38	61	29	46	45	40

For this data set, the mean is 42 minutes, and this would be interpreted as a typical amount of time spent studying for this group of students. Also, consider showing the dot plot of the data shown below, and point out that the data values are centered at 42 minutes.

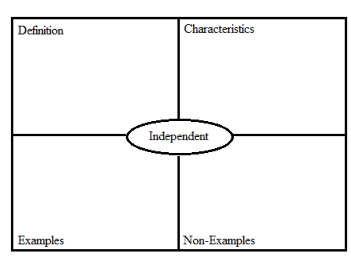


In addition to discussing the definition of *mean*, it may be necessary to review the definition of *standard deviation*.

Scaffolding:

The word *mean* has multiple definitions (including from different parts of speech). Although students have been exposed to the term *mean* in Grades 6 and 7 and Algebra I, those new to the curriculum or English language learners may struggle with its meaning. Teachers should clarify the term for these students.

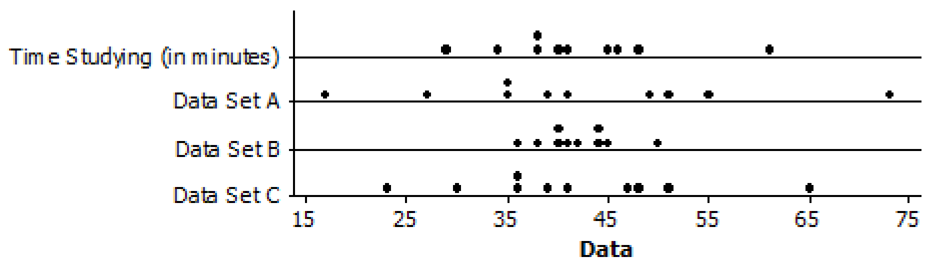
Consider using visual displays and repeated choral readings to reinforce the mathematical meaning of this word. A Frayer diagram may be used.



Remind students that the *standard deviation* is one measure of *spread* or *variability* in a data distribution. The standard deviation describes variation in terms of deviation from the mean. The standard deviation can be interpreted as a typical deviation from the mean. A small standard deviation indicates that the data points tend to be very close to the mean, and a large standard deviation indicates that the data points are spread out over a large range of values.

The formula for the standard deviation of a sample is $s = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$, where $\overline{x}$ is the sample mean.

The standard deviation, s, of the study time data is approximately 8.77, which would be interpreted as a typical difference from the mean number of minutes this group of students spent studying. Consider showing the following dot plots, and ask students if they think that the standard deviation would be less than or greater than 8.77 for each of Data Sets A, B, and C. (The standard deviation is less than 8.77 for Data Set B and greater for Data Sets A and C.)



To calculate the standard deviation,

1. Find the mean, $\overline{x}$.
2. Find the difference between each data point and the mean. These values are called *deviations* from the mean, and a deviation is denoted by $(x - \overline{x})$.
3. Square each of the deviations.
4. Find the sum of the squared deviations.
5. Divide the sum of the squared deviations by $n - 1$.
6. Determine the square root.

To find the standard deviation of data from an entire population (as opposed to a sample), divide by n instead of $n - 1$.

Review the different distribution shapes: *symmetric* and *skewed*. A distribution is approximately *symmetric* if the left side and the right side of the distribution are roughly mirror images of each other (if students were to fold the distribution in half, the two sides would be a close match). A distribution is *skewed* if it has a noticeably longer tail on one side than the other. A distribution with a longer tail to the right is considered *skewed to the right*, whereas a distribution with a longer tail on the left side is *skewed to the left*. Display the distributions shown in the Lesson Summary, and discuss the shape of each distribution. Also, point out that a distribution that is approximately symmetric with a single peak is often described as mound shaped.

Example 1 (10 minutes): Center, Shape, and Spread

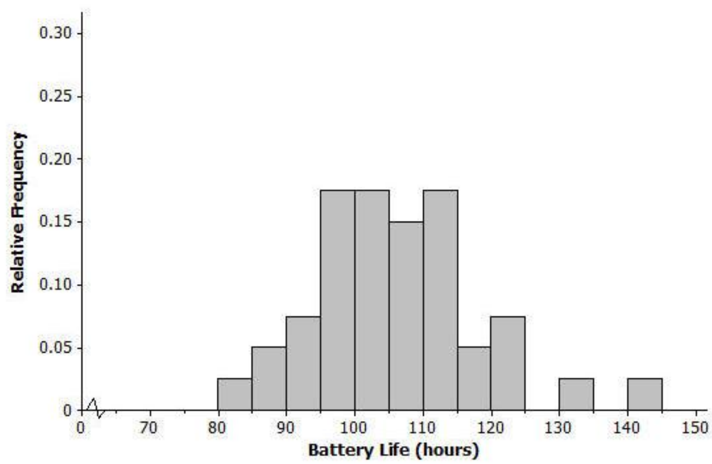
All of the histograms in this lesson are relative frequency histograms. Relative frequency histograms were introduced in earlier grades. If necessary, remind students that the heights of the bars in a relative frequency histogram represent the proportion of the observations within each interval and not the number of observations (the frequency) within the interval. Prior to students beginning to work on Exercises 1 through 6, discuss the scales on the battery life histogram. Address the following points using the histogram of battery life:

- What is the width of each bar?
  - 5 *hours*
- What does the height of each bar represent?
  - *The proportion of all batteries with a life in the interval corresponding to the bar. For example, approximately* 5% *of the batteries lasted between* 85 *and* 90 *hours.*

Example 1: Center, Shape, and Spread

Have you ever noticed how sometimes batteries seem to last a long time, and other times the batteries seem to last only a short time?

The histogram below shows the distribution of battery life (hours) for a sample of 40 batteries of the same brand. When studying a distribution, it is important to think about the shape, center, and spread of the data.



Exercises 1–6 (10 minutes)

Have students work independently and confirm answers with a partner or in a small group. Discuss answers as needed.

MP.2 Before beginning Exercise 7, review the interpretation of standard deviation. Have students describe how they made their estimates for the standard deviation in Exercises 3 and 6. This allows students to practice reasoning abstractly and quantitatively.

In the next several exercises, students are asked to estimate and interpret the standard deviation. Standard deviation can be challenging for students. Often in the process of focusing on how it is calculated, students lose sight of what it indicates about the data distribution. The calculation of the standard deviation can be done using technology or by following a precise sequence of steps; understanding what it indicates about the data should be the focus of the exercises.

The standard deviation is a measure of variability that is based on how far observations in a data set fall from the mean. It can be interpreted as a typical or an average distance from the mean. Various rules and shortcuts are often used to estimate a standard deviation, but for this lesson, keep the focus on understanding the standard deviation as a value that describes a typical distance from the mean. Students should observe that a typical distance is one for which some distances would be less than this value and some would be greater. When students estimate a standard deviation, ask them whether that value is representative of the collection of distances. Is the estimate a reasonable value for the average distance of observations from the mean? If the estimate is a value that is less than most of the distances, then it is not a good estimate and is probably too small. If the estimate is a value that is greater than most of the distances, then the estimate is probably too large. This understanding is developed in several of the following exercises:

Exercises 1–9

1. **Would you describe the distribution of battery life as approximately symmetric or as skewed? Explain your answer.**

   The distribution is approximately symmetric. The right and left halves of the distribution are similar.

Indicate that because this distribution is approximately symmetric, the standard deviation is a reasonable way to describe the variability of the data. From students' previous work in Grade 6 and Algebra I, they should recall that for a data distribution that is skewed rather than symmetric, the interquartile range (IQR) would be used to describe variability.

2. **Is the mean of the battery life distribution closer to 95, 105, or 115 hours? Explain your answer.**

   The mean of the battery life distribution is closer to 105 hours because the data appear to center around 105.

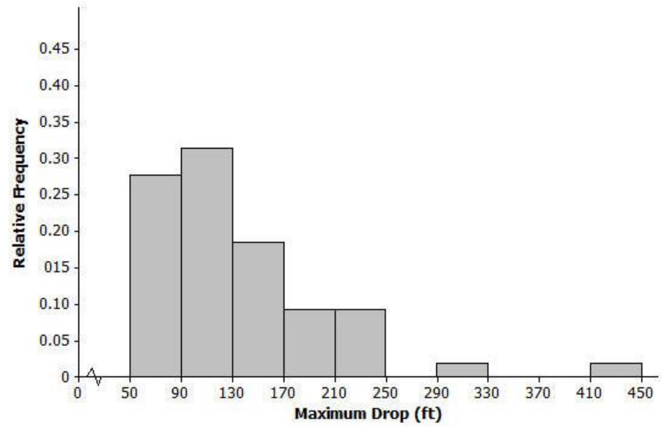
The next exercise provides an opportunity to discuss a typical distance from the mean. If students struggle with understanding this question, for each estimate of the standard deviation, ask if it would be a good estimate of a typical or an average distance of observations from the mean. If 5 was the standard deviation, how many of the distances would be greater than and how many less than this value? Students should be able to see that most of the data values are more than 5 units from the mean. If 25 was the standard deviation, how many of the distances would be greater than and how many less than this value? Again, students should see that most of the data values are less than 25 units away from the mean. The estimate of 10 is more reasonable as an estimate of the average distance from the mean and is a better estimate of the standard deviation.

3. **Consider 5, 10, or 25 hours as an estimate of the standard deviation for the battery life distribution.**
   a. **Consider 5 hours as an estimate of the standard deviation. Is it a reasonable description of a typical distance from the mean? Explain your answer.**
      Most of the distances from the mean are greater than 5 hours. It is not a good estimate of the standard deviation.

   b. **Consider 10 hours as an estimate of the standard deviation. Is it a reasonable description of a typical distance from the mean? Explain your answer.**
      It looks like 10 is a reasonable estimate of a typical distance from the mean. It is a reasonable estimate of the standard deviation.

   c. **Consider 25 hours as an estimate of the standard deviation. Is it a reasonable description of a typical distance from the mean? Explain your answer.**
      Nearly all of the data values are less than 25 hours from the mean of 105. It is not a good estimate of the standard deviation.

The histogram below shows the distribution of the greatest drop (in feet) for 55 major roller coasters in the United States.



4. **Would you describe this distribution of roller coaster maximum drop as approximately symmetric or as skewed? Explain your answer.**

   The distribution is skewed to the right because there is a long tail on the right side.

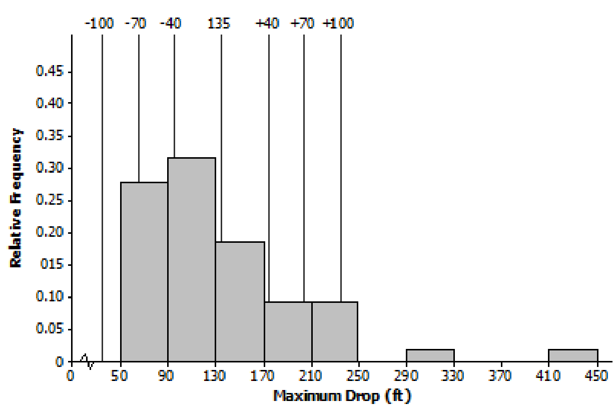
5. **Is the mean of the maximum drop distribution closer to 90, 135, or 240 feet? Explain your answer.**

   The mean is closer to 135 feet because 90 is too small and 240 is too large to be considered a typical value for this data set.

In the same way that students estimated the standard deviation for battery life, the following exercise asks students to select an estimate for the drop data. Here again, students should consider each estimate separately and determine which one is the most typical of the distances from the estimated mean. As this is a skewed distribution, estimating a typical distance from the mean is a challenge and requires careful thought.

6. **Is the standard deviation of the maximum drop distribution closer to 40, 70, or 100 hours? Explain your answer.**

   It seems that 70 is about right for a typical distance from the mean. A deviation of 40 would be too small, and 100 would be too large to be considered a typical distance from the mean for this data set. Most of the data values differ from the estimated mean of 135 by more than 40, which means that 40 is not a reasonable estimate of the standard deviation. Most of the data values differ from the estimated mean of 135 by less than 100, which means that 100 is not a reasonable estimate of the standard deviation. This can be illustrated for students using the following picture:



Exercises 7–9 (10 minutes)

Encourage students to work in pairs on Exercises 7, 8, and 9. Then, discuss and confirm the answers.

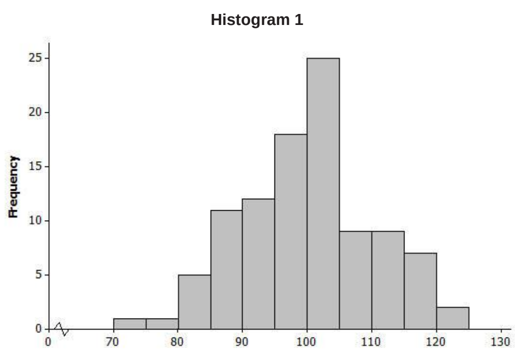
7. Consider the following histograms: Histogram 1, Histogram 2, Histogram 3, and Histogram 4. Descriptions of four distributions are also given. Match the description of a distribution with the appropriate histogram.

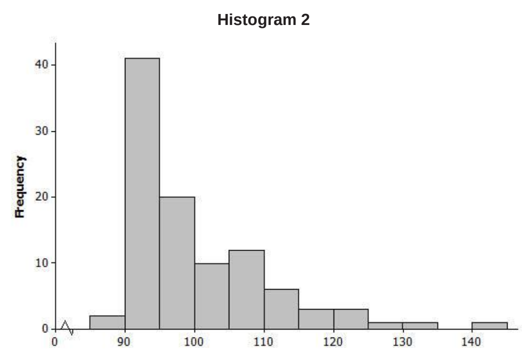
Histogram	Distribution
1	*B*
2	*A*
3	*C*
4	*D*

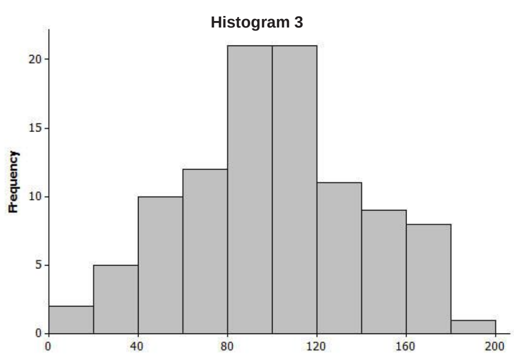
Description of distributions:

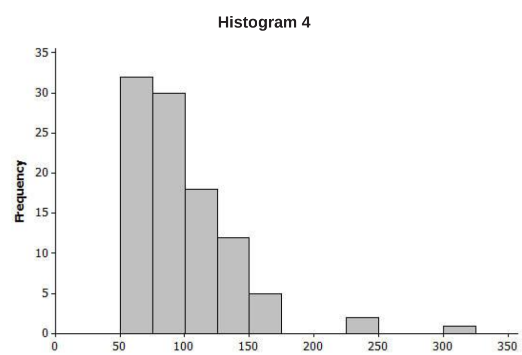
Distribution	Shape	Mean	Standard Deviation
A	Skewed to the right	100	10
B	Approximately symmetric, mound shaped	100	10
C	Approximately symmetric, mound shaped	100	40
D	Skewed to the right	100	40

Histograms:

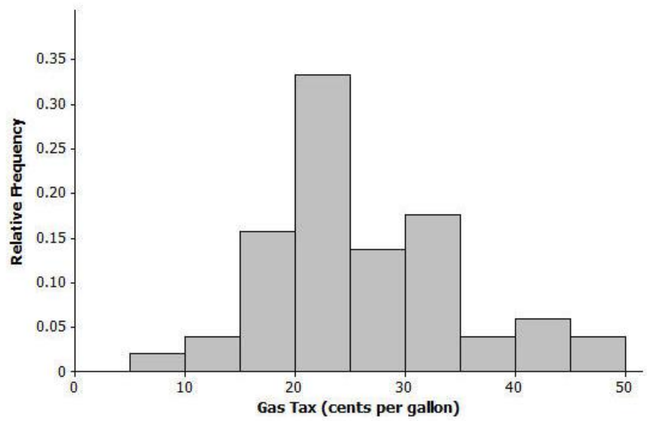






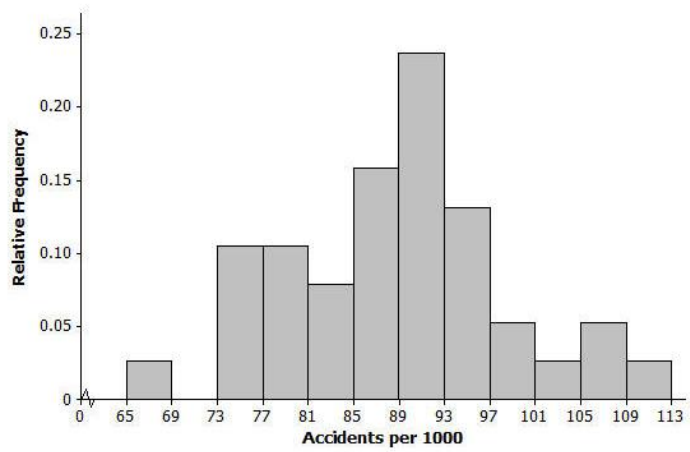


8. **The histogram below shows the distribution of gasoline tax per gallon for the 50 states and the District of Columbia in 2010. Describe the shape, center, and spread of this distribution.**



The distribution shape is skewed to the right. Answers for center and spread will vary, but the center is approximately 25, and the standard deviation is approximately 10.

9. **The histogram below shows the distribution of the number of automobile accidents per year for every $1,000$ people in different occupations. Describe the shape, center, and spread of this distribution.**



The shape of the distribution is approximately symmetric. Answers for center and spread will vary, but the center is approximately 89, and the standard deviation is approximately 10.

Closing (5 minutes)

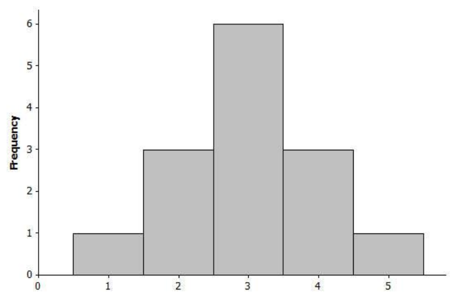
Ask students to summarize one of the histograms presented in class in terms of center, shape, and spread. Allow them to select any one of the many examples presented in this lesson. Call on a representative group of students to present descriptions of the histogram they selected.

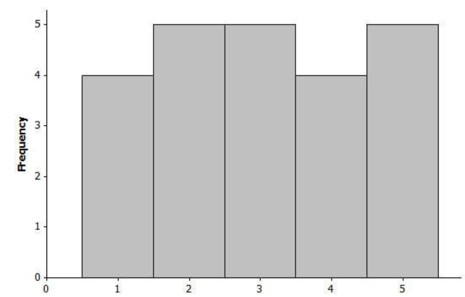
Ask students to summarize the main concepts of the lesson in writing or with a neighbor. Use this as an opportunity to informally assess comprehension of the lesson. The Lesson Summary below offers some important concepts that should be included.

Lesson Summary

Distributions are described by the shape (symmetric or skewed), the center, and the spread (variability) of the distribution.

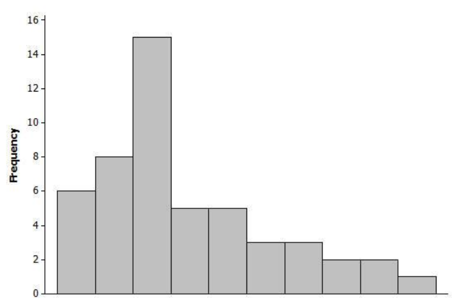
A distribution that is approximately symmetric can take different forms.



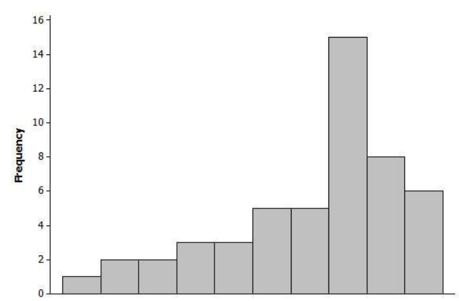


A distribution is described as *mound shaped* if it is approximately symmetric and has a single peak.

A distribution is *skewed to the right* or *skewed to the left* if one of its tails is longer than the other.



Skewed to the Right



Skewed to the Left

The *mean of a distribution* is interpreted as a typical value and is the average of the data values that make up the distribution.

The *standard deviation* is a value that describes a typical distance from the mean.

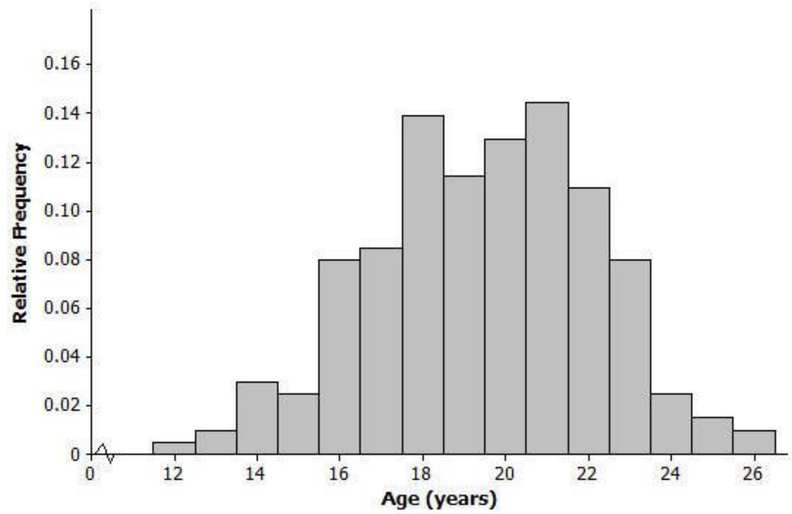
Exit Ticket (5 minutes)

Name ____ Date ____________________

Lesson 8: Distributions—Center, Shape, and Spread

Exit Ticket

A local utility company wanted to gather data on the age of air conditioners that people have in their homes. The company took a random sample of 200 residents of a large city and asked if the residents had an air conditioner, and if they did, how old it was. Below is the distribution in the reported ages of the air conditioners.



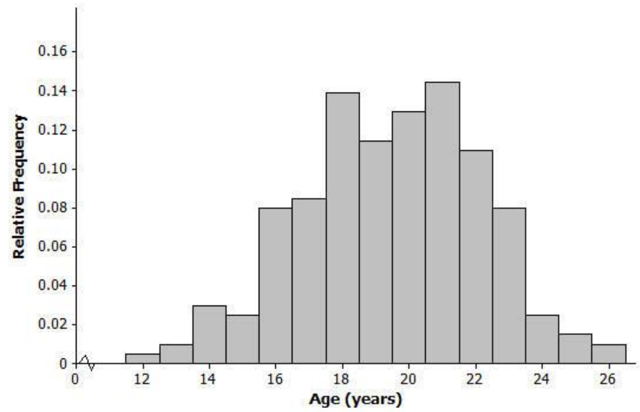
1. Would you describe this distribution of air conditioner ages as approximately symmetric or as skewed? Explain your answer.

2. Is the mean of the age distribution closer to 15, 20, or 25 years? Explain your answer.

3. Is the standard deviation of the age distribution closer to 3, 6, or 9 years? Explain your answer.

Exit Ticket Sample Solutions

A local utility company wanted to gather data on the age of air conditioners that people have in their homes. The company took a random sample of 200 residents of a large city and asked if the residents had an air conditioner, and if they did, how old it was. Below is the distribution in the reported ages of the air conditioners.



1. **Would you describe this distribution of air conditioner ages as approximately symmetric or as skewed? Explain your answer.**

   The distribution is approximately symmetric. The left and right sides of the distribution are similar. This distribution would also be described as mound shaped.

2. **Is the mean of the age distribution closer to 15, 20, or 25 years? Explain your answer.**

   The mean of the age distribution is closer to 20 years because the distribution is centered at about 20.

3. **Is the standard deviation of the age distribution closer to 3, 6, or 9 years? Explain your answer.**

   A reasonable estimate of an average distance from the mean would be 3, so the standard deviation of the age distribution would be about 3.

Problem Set Sample Solutions

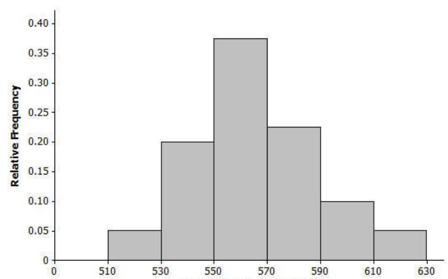
1. **For each of the following histograms, describe the shape, and give estimates of the mean and standard deviation of the distributions:**

   a. **Distribution of head circumferences (mm)**

   Shape: Approximately symmetric and mound shaped

   Mean: Approximately 560

   Standard Deviation: Approximately 25

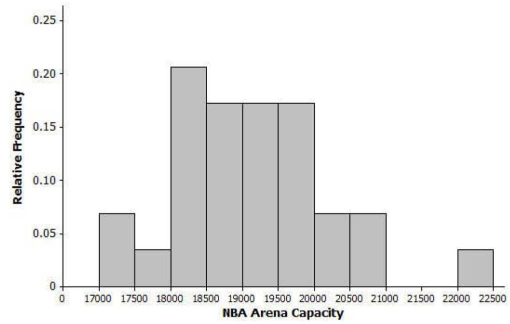


b. **Distribution of NBA arena seating capacity**

Shape: Approximately symmetric

Center: Approximately $19,000$

Spread: The standard deviation is approximately $1,000$



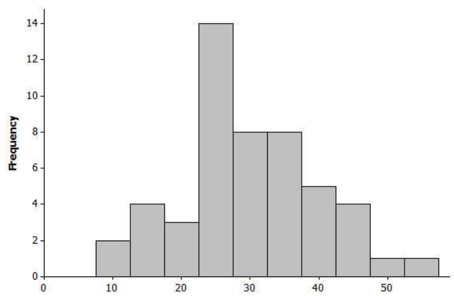
2. **For the each of the following, match the description of each distribution with the appropriate histogram:**

Histogram	Distribution
1	C
2	B
3	A
4	D

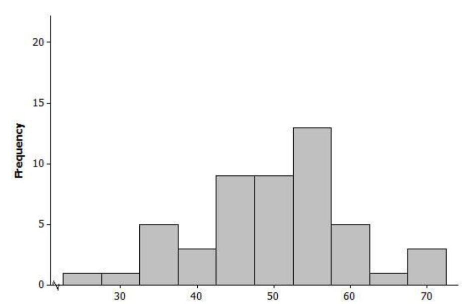
Description of distributions:

Distribution	Shape	Mean	Standard Deviation
A	Approximately symmetric, mound shaped	50	5
B	Approximately symmetric, mound shaped	50	10
C	Approximately symmetric, mound shaped	30	10
D	Approximately symmetric, mound shaped	30	5

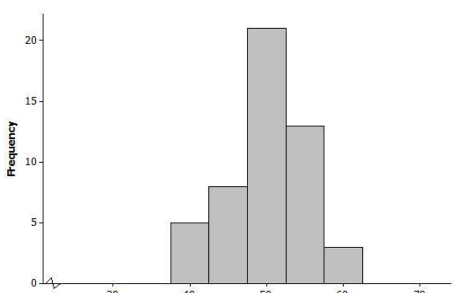
Histogram 1



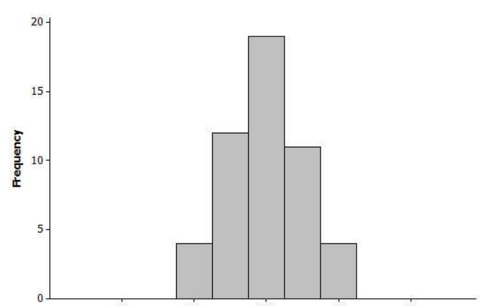
Histogram 2



Histogram 3



Histogram 4

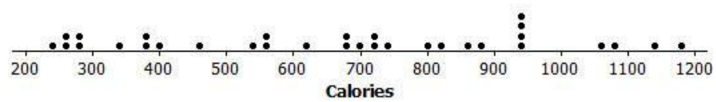


3. **Following are the number of calories in a basic hamburger (one meat patty with no cheese) at various fast food restaurants around the country:**

$380,\ 790,\ 680,\ 460,\ 725,\ 1130,\ 240,\ 260,\ 930,\ 331,\ 710,\ 680,\ 1080,\ 612,\ 1180,\ 400,\ 866,\ 700,\ 1060,\ 270,\ 550,\ 380,\ 940,\ 280,\ 940,\ 550,\ 549,\ 937,\ 820,\ 870,\ 250,\ 740$

a. **Draw a dot plot on the scale below.**

Completed dot plot:



b. **Describe the shape of the calorie distribution.**

Approximately symmetric

c. **Using technology, find the mean and standard deviation of the calorie data.**

Mean: 665

Standard deviation: 283

d. **Why do you think there is a lot of variability in the calorie data?**

Answers will vary. Possible answers include that the size (number of ounces) of the hamburger patty may be different, the type of bun may vary, and burgers may have different types and amounts of various condiments.