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.

   MP.1

   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.

   MP.3

   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.

## 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.

Dial Up Download Times (minutes)
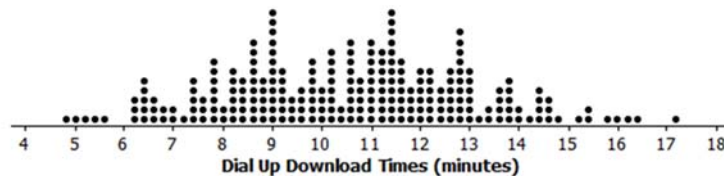
a. What can you observe about the download times from the dot plot?

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.
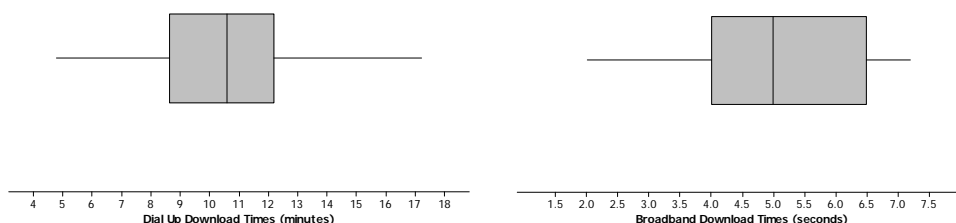

Dial Up Download Times (minutes)
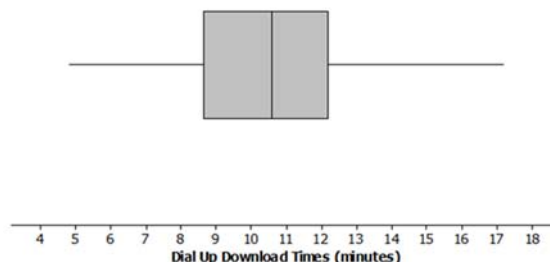
Answer for part (a): About half of the times are above 10.6 *minutes. The distribution is roughly symmetric around the median. About half of the times are between* 8.6 *minutes and* 12.1 *minutes.*

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?
>
>
>
> Dial Up Download Times (minutes)   Broadband Download Times (seconds)
>
> *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.*

MP.6

- 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.*

Exercises 4–5: Rainfall

4. Data on average rainfall for each of the twelve months of the year were used to construct the two dot plots below.
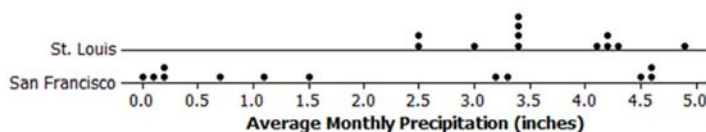
a. How many data points are in each dot plot? What does each data point represent?

There are 12 data points in St. Louis. There are also 12 data points in San Francisco. Each data point represents the average monthly precipitation in inches.

b. Make a conjecture about which city has the most variability in the average monthly amount of precipitation and how this would be reflected in the IQRs for the data from both cities.

San Francisco has the most variability in the average monthly amount of precipitation. It has the largest IQR of the two cities.

c. Based on the dot plots, what are the approximate values of the interquartile ranges (IQR) of the amount of average monthly precipitation in inches for each city? Use each IQR to compare the cities.

For St. Louis, the IQR is $4.205 - 3.16 = 1.045$; for San Francisco, the IQR is $4.13 - 0.185 = 3.945$. About the middle half of the precipitation amounts in St. Louis are within 1 inch of each other. In San Francisco, the middle half of the precipitation amounts is within about 4 inches of each other.

d. In an earlier lesson, the average monthly temperatures were rounded to the nearest degree Fahrenheit. Would it make sense to round the amount of precipitation to the nearest inch? Why or why not?

Answers will vary: It would not make sense because the numbers are pretty close together, or yes, it would make sense because you would still get a good idea of how the precipitation varied. If you rounded to the nearest inch, the IQR for San Francisco would be 4 because three of the values round to 0 and three round to 5. St. Louis would be 1 because most of the values round to 3 or 4. In both cases, that is pretty close to the IQR using the numbers to the hundredths place.

5. Use the data from Exercise 4 to answer the following.

a. Make a box plot of the amount of precipitation for each city.
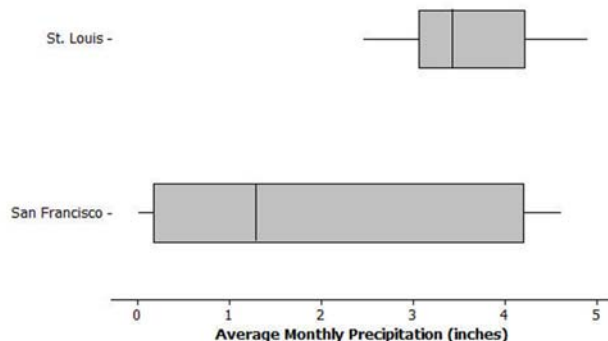
b.   Compare the percent of months that have above 2 inches of precipitation for the two cities. Explain your thinking.

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.
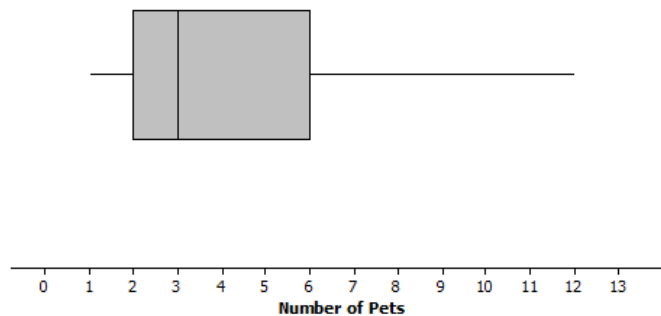
Exit Ticket (5 minutes)

Name _____   Date_____

Lesson 16: Understanding Box Plots
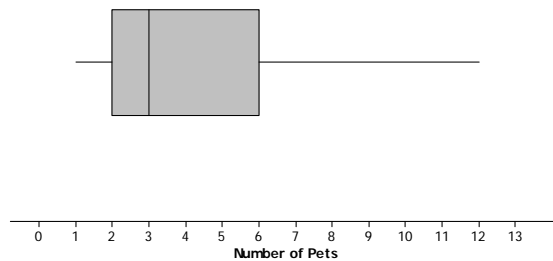
Exit Ticket

The number of pets per family for students in a sixth grade class is below:

1.   Can you tell how many families have two pets? Explain why or why not.

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.

   b.   About one fourth of the families had six or more pets.

   c.   Most of the families had three pets.

   d.   Half of the families had five or fewer pets.

   e.   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:

0 1 2 3 4 5 6 7 8 9 10 11 12 13
Number of Pets

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.

   Best Actor -

   Best Actress -

   20   30   40   50   60   70   80
   Age at Award (years)
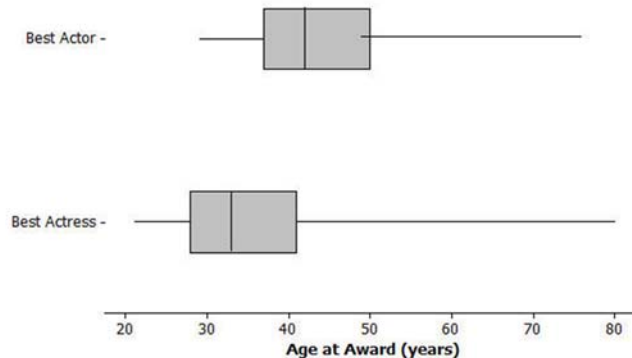
   a. Do you think it is harder for an older woman to win an academy award for best actress than it is for an older man to win a best actor award? Why or why not?

      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.
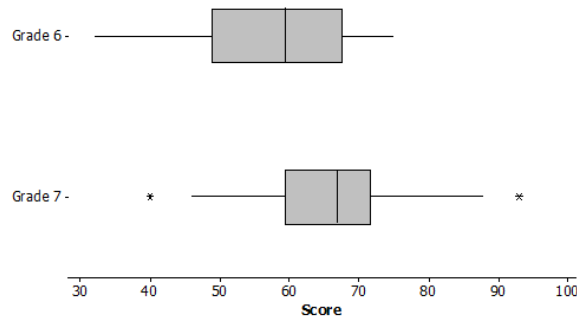
a. In which grade did the students do the best? Explain how you can tell.

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.