Lesson 19: Comparing Data Distributions

Student Outcomes

- Given box plots, students comment on similarities and differences in data distributions.

Classwork

As students work through this lesson, encourage them to think about the data they have collected for their projects and to think about how they might summarize that data using both numerical summaries and graphical displays.

Lesson 18 was about understanding data distributions and seeing how graphical displays and numerical summaries are used to learn about important characteristics of a data distribution. In this lesson, the focus is on comparing two or more data distributions. As students work through this lesson, each time they encounter a new example that presents them with data distributions, encourage them to think about how the distributions are similar and how they are different and what this means in the context of the example.

Have students read the following paragraphs.

Suppose that you are interested in comparing the weights of adult male polar bears and the weights of adult male grizzly bears. If data were available on the weights of these two types of bears, they could be used to answer questions such as the following:

- Do adult polar bears typically weigh less than adult grizzly bears?
- Are the weights of adult polar bears similar to each other, or do the weights tend to differ a lot from bear to bear?
- Are the weights of adult polar bears more consistent than the weights of adult grizzly bears?

These questions could be answered most easily by comparing the weight distributions for the two types of bears. Graphs of the data distributions (such as dot plots, box plots, or histograms) that are drawn side by side and that are drawn to the same scale make it easy to compare data distributions in terms of center, variability, and shape.

In this lesson, when two or more data distributions are presented, think about the following:

- How are the data distributions similar?
- How are the data distributions different?
- What do the similarities and differences tell you in the context of the data?

Example 1 (3 minutes): Comparing Groups Using Box Plots

Review box plots and five-number summaries. Remind students of the following important points.

- A box plot provides a useful summary of a distribution, and certain summary measures (such as median, quartiles, IQR) can be obtained or estimated from the box plot.
- When box plots are shown together (using the same scale), visual differences between the box plots correspond to quantitative differences between the corresponding summary measures of the distributions.
Example 1: Comparing Groups Using Box Plots

Recall that a box plot is a visual representation of a five-number summary. The box part of a box plot is drawn so that the width of the box represents the IQR. The distance from the far end of the line on the left to the far end of the line on the right represents the range.

If two box plots (each representing a different distribution) are drawn side by side using the same scale, it is easy to compare the values in the five-number summaries for the two distributions and to visually compare the IQRs and ranges.

Here is a data set of the ages of 43 participants in a 5-kilometer race (shown in a previous lesson).

<table>
<thead>
<tr>
<th>20</th>
<th>30</th>
<th>30</th>
<th>35</th>
<th>36</th>
<th>34</th>
<th>38</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>16</td>
<td>43</td>
<td>23</td>
<td>47</td>
<td>27</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>32</td>
<td>32</td>
<td>31</td>
<td>32</td>
<td>36</td>
<td>74</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>51</td>
<td>61</td>
<td>50</td>
<td>34</td>
<td>34</td>
<td>34</td>
<td>35</td>
<td>28</td>
</tr>
<tr>
<td>57</td>
<td>26</td>
<td>29</td>
<td>49</td>
<td>41</td>
<td>36</td>
<td>37</td>
<td>41</td>
</tr>
<tr>
<td>38</td>
<td>30</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Here is the five-number summary for the data: Minimum = 18, Q1 = 30, Median = 35, Q3 = 41, Maximum = 74.

There was also a 15-kilometer race. The ages of the 55 participants in that race appear below.

<table>
<thead>
<tr>
<th>47</th>
<th>19</th>
<th>30</th>
<th>30</th>
<th>36</th>
<th>37</th>
<th>35</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>49</td>
<td>47</td>
<td>16</td>
<td>45</td>
<td>22</td>
<td>50</td>
<td>27</td>
</tr>
<tr>
<td>19</td>
<td>20</td>
<td>30</td>
<td>32</td>
<td>32</td>
<td>31</td>
<td>32</td>
<td>37</td>
</tr>
<tr>
<td>22</td>
<td>81</td>
<td>43</td>
<td>43</td>
<td>54</td>
<td>66</td>
<td>53</td>
<td>35</td>
</tr>
<tr>
<td>22</td>
<td>35</td>
<td>35</td>
<td>36</td>
<td>28</td>
<td>61</td>
<td>26</td>
<td>29</td>
</tr>
<tr>
<td>38</td>
<td>52</td>
<td>43</td>
<td>37</td>
<td>38</td>
<td>43</td>
<td>39</td>
<td>30</td>
</tr>
<tr>
<td>58</td>
<td>30</td>
<td>48</td>
<td>49</td>
<td>54</td>
<td>56</td>
<td>58</td>
<td></td>
</tr>
</tbody>
</table>

Does the longer race appear to attract different runners in terms of age? Below 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. What do you notice about the two box plots?

Both the range and the IQR are larger for the longer race, which means there is more variability in the ages for the long race than in the ages for the short race.
Exercises 1–6 (10 minutes)

In some cases, the questions have more than one correct answer. Answers may not be exact when students are estimating values of summary measures from the box plots.

Students work individually. If time allows, discuss answers as a class.

Exercises 1–6

1. Based on the box plots, estimate the values in the five-number summary for the ages in the 15-kilometer race data set.
   - Minimum = 16, Q1 = 30, Median = 37, Q3 = 48, and Maximum = 81

2. Do the two data sets have the same median? If not, which race had the higher median age?
   - The two data sets do not have the same median; 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?
   - The two data sets do not have the same IQR; the 15 km race has a slightly higher IQR: 18 years of age compared to 11 years of age for the 5 km race. The ages of runners in the middle 50% for the 15 km race are more spread out.

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 runner? About how old was this runner?
   - The 15 km race had the oldest runner at 81 years of age. The oldest runner for the 5 km race was only 74.

6. Now, consider just the youngest 25% of the runners 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?
   - For the 15 km race, the youngest 25% of the runners had ages between the minimum and Q1. The youngest runner was 16 years old because this was the minimum for this distribution. The oldest runner in this group was 30 years old or younger because Q1 was 30. For the 5 km race, this youngest runner was 18 years old, and the oldest runner in the bottom 25% was 30 years old or younger (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 (two-minute) discussion.

In some cases, the questions have more than one correct answer. Answers may not be exact when students are estimating values of summary measures from the box plots. Note: Non-baseball-related questions with similar objectives appear in the Problem Set.
Exercises 7–12: Comparing Box Plots

In 2012, Major League Baseball had two leagues: an American League of 14 teams and a National League of 16 teams. Jesse wondered if American League teams have higher batting averages and on-base percentages. (Higher values are better.) Use the following box plots to investigate. (Source: [http://mlb.mlb.com/stats/sortable.jsp](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? Approximately how large was the difference, and which league had the higher maximum value?

The highest batting averages for both leagues appear to be around 0.274. (This value must be estimated from the box plot, so other answers that are close to 0.274 are acceptable.)

8. Was the range of the American League team batting averages very different or only slightly different from the range of the National League team batting averages?

They appear to be only slightly different, with the American League range being slightly higher. The American League minimum (0.234) is slightly lower than the National League minimum (0.236), and 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 American League has the higher median batting average at about 0.258, while the median batting average for the National League median is about 0.252. In answering the second part of this question, make sure that students look at the difference in the medians. Students might say that they think this is a small difference because 0.006 is a small number. However, a better answer would consider the scale and the range of the data. The difference of 0.006 is roughly $\frac{1}{6}$ of the range for the National League, which suggests that the difference may be large enough to be considered meaningful.
10. Based on the box plots below for on-base percentage, which three summary values (from the five-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 National League data set appears to have less variability, as it has a smaller IQR and smaller range.

12. Recall that Jesse wondered if American League teams have higher batting averages and on-base percentages. Based on the box plots given above, what would you tell Jesse?

Answers will vary.

It does look like the American League teams tend to have higher team batting averages. The median and upper quartile for the American League team batting average box plot are greater than the median and upper quartile for the National League. However, there is at least one team in the American League with a team batting average that is lower than every team in the National League, and there is at least one team in the National League that has a team batting average that is about the same as the American League team with the highest batting average. For on-base percentage, the median and lower quartile and the maximum are about the same for both leagues. So it does not really look like it would be reasonable to say that the American League teams tend to have higher team on-base percentages than National League teams.
Closing (5 minutes)

- 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, and number of observations
- What other kinds of graphs might be placed side by side to visually communicate the similarities and differences between data distributions?
  - Side-by-side dot plots would be effective for this, 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 the groups’ distributions side by side using graphs drawn to the same scale. This makes it easier to describe the similarities and differences in the distributions of the groups.

Exit Ticket (7 minutes)
Lesson 19: Comparing Data Distributions

Exit Ticket

Hay is used to feed animals such as cows, horses, and goats. Almost \( \frac{1}{3} \) of the hay grown in the United States comes from just five states. Is this because these states have more acres planted in hay, or could it be because these states produce more hay per acre than other states? The following box plots show the distribution of hay produced (in tons) per acre planted in hay for three different regions: 22 eastern states, 14 midwestern states, and 12 western states.


1. 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 (such as median and IQR) or aspects of the graphical displays (such as shape and variability).
2. 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 you know this is false.

3. 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 aspects of the graphical displays.
Exit Ticket Sample Solutions

Hay is used to feed animals such as cows, horses, and goats. Almost $\frac{1}{3}$ of the hay grown in the United States comes from just five states. Is this because these states have more acres planted in hay, or could it be because these states produce more hay per acre than other states? The following box plots show the distribution of hay produced (in tons) per acre planted in hay for three different regions: 22 eastern states, 14 midwestern states, and 12 western states.


1. 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 (such as median and IQR) or aspects of the graphical displays (such as shape and variability).

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.

2. 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 you know 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.

3. 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 aspects of the graphical displays.

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 more than two times the largest eastern and midwestern yields.
Problem Set Sample Solutions

Before students begin the Problem Set, consider allowing time for them 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 create 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 the time is needed for completion of the project.

1. College athletic programs are separated into divisions based on school size, available athletic scholarships, and other factors. A researcher wondered if members of swimming and diving programs in Division I (usually large schools that offer athletic scholarships) tend to be taller than the swimmers and divers in Division III programs (usually smaller schools that do not offer athletic scholarships). To begin the investigation, the researcher creates side-by-side box plots for the heights (in inches) of 41 male swimmers and divers at Mountain Vista University (a Division I program) and the heights (in inches) of 10 male swimmers and divers at Eaglecrest College (a Division III program).

   ![Box plots](image)

   a. Which data set has the smaller range?

   *Eaglecrest College has the smaller range.*

   b. True or false: A swimmer who had a height equal to the median for the Mountain Vista University would be taller than the median height of swimmers and divers at Eaglecrest College.

   *True*

   c. To be thorough, the researcher will examine many other colleges’ sports programs to further investigate the 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 the claim might be valid? Carefully support your answer using summary measures or graphical attributes.

   *Based on just these two teams, it looks like the claim may be correct. A large portion of the Mountain Vista University distribution is higher than the maximum value of the Eaglecrest College distribution. The median value for the Mountain Vista University appears to be 4 inches higher than the median value of the Eaglecrest College distribution.*
2. Data on the weights (in pounds) of 100 polar bears and 50 grizzly bears are summarized in the box plots shown below.

- **Polar Bears**:
- **Grizzly Bears**:

![Box plots for polar and grizzly bears](image)

Weight (pounds)

420 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600

a. True or false: At least one of the polar bears weighed more than the heaviest grizzly bear. Explain how you know.

True. The lower quartile and the median weight for the polar bears were greater than the maximum weight for the grizzly bears, so at least 75% of the polar bears weighed more than the heaviest grizzly bear.

b. True or false: Weight differs more from bear to bear for polar bears than for grizzly bears. Explain how you know.

True. The weights of the grizzly bears are more compact than the weights of the polar bears. The grizzly bear weight distribution has a smaller range and IQR than the polar bear weight distribution.

c. Which type of bear tends to weigh more? Explain.

Polar bears tend to have greater weights. The median weight for grizzly bears is much smaller than the median polar bear weight, and about half of the grizzly bear weight distribution is lower than the minimum weight for the polar bears.
Many movie studios rely heavily on viewer data to determine how a movie will be marketed and distributed. Recently, previews of a soon-to-be-released movie 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 have ever seen” and 0 representing “worst movie I have ever seen.”

Below are some side-by-side box plots that summarize the ratings by gender and by age.

For 150 women and 150 men:

For 3 age groups:

a. 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 variability.

It appears that the men and women rated the film in a very similar manner: The box plots show the same quartile values, the same medians, and the same maximums. The only difference is that the minimum rating from a man was slightly lower than the minimum rating from a woman.

b. It appears that the film tended to receive better ratings from the older members of the group. Write a few sentences using comparative measures of center and spread or aspects of the graphical displays 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).