



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).

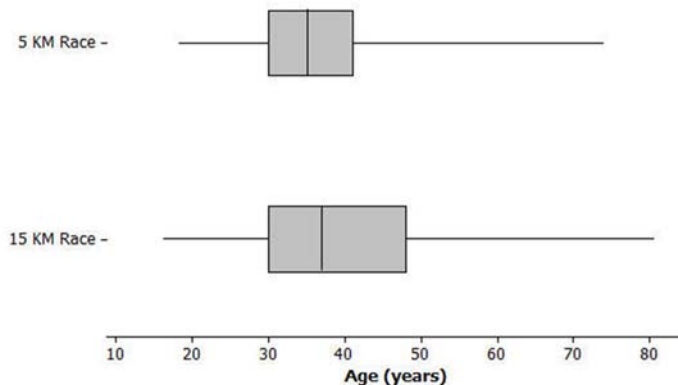
20	30	30	35	36	34	38	46
45	18	43	23	47	27	21	30
32	32	31	32	36	74	41	41
51	61	50	34	34	34	35	28
57	26	29	49	41	36	37	41
38	30	30					

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).

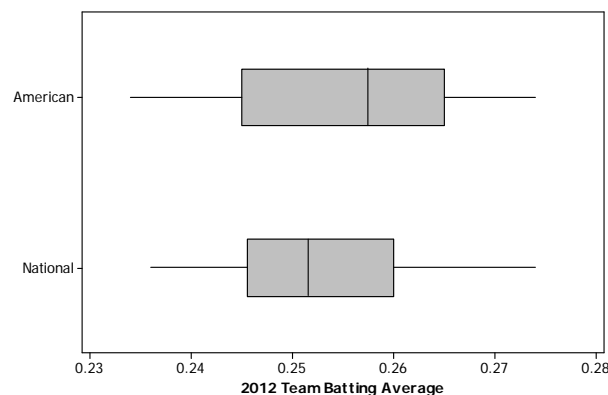
Exercises 7–12 (20 minutes): Comparing Box Plots

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.)

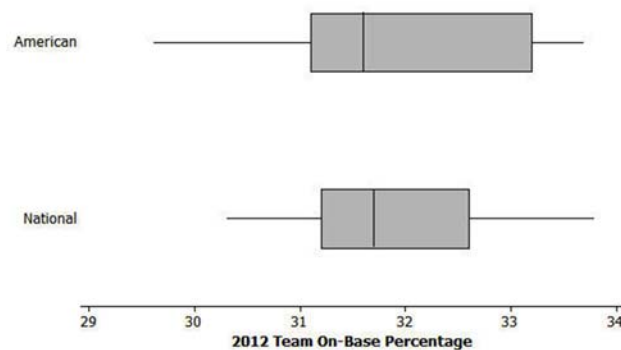
8. Was the range of American League team batting averages very different or only slightly different from the range of National League team batting averages?

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 _____

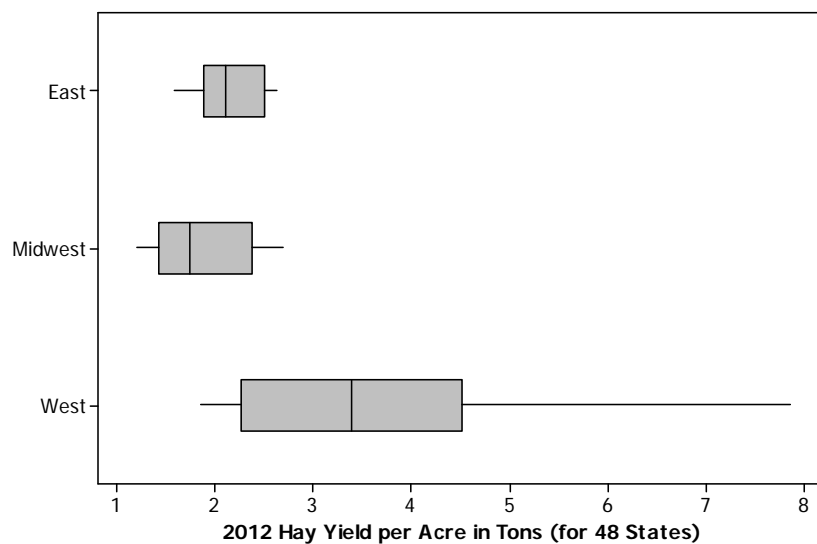
Date _____

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>.)

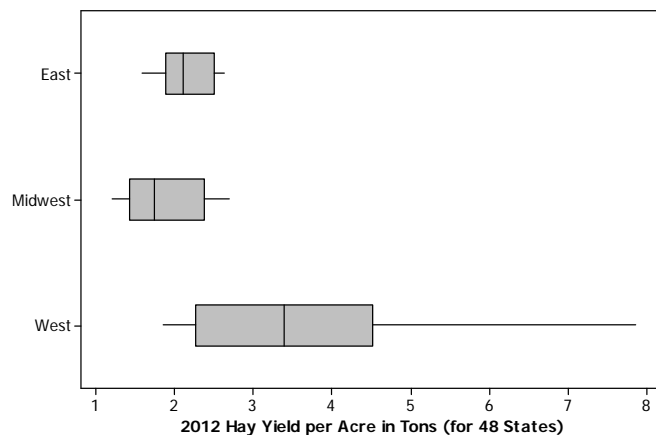


- Which of the three regions' data sets has the least variability? Which has the greatest variability? To explain how you chose your answers, write a sentence or two that supports your choices by comparing relevant summary measures (i.e., median, IQR, etc.) or graphical attributes (i.e., shape, variability, 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.

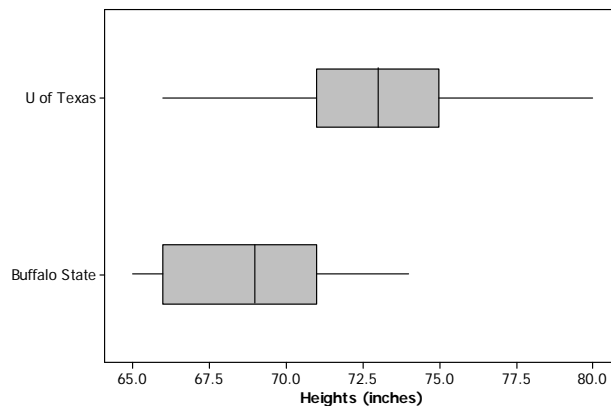
The West typically has states with the largest hay yield per acre. Over half of the Western states have hay yields that are higher than any yield in either of the other two regions. Also, some Western yields are up to two or three times the largest Eastern and Midwestern yields.

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.texasports.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)



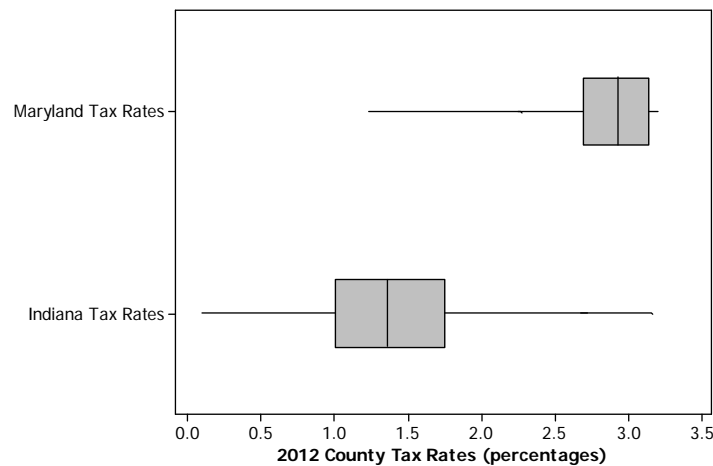
- Which data set has the smaller range?
Buffalo State
- True or False: A team member of median height on the University of Texas team would be taller than a team member of median height on the Buffalo State College team.
True.
- To be thorough, the researcher will examine many other college's sports 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.

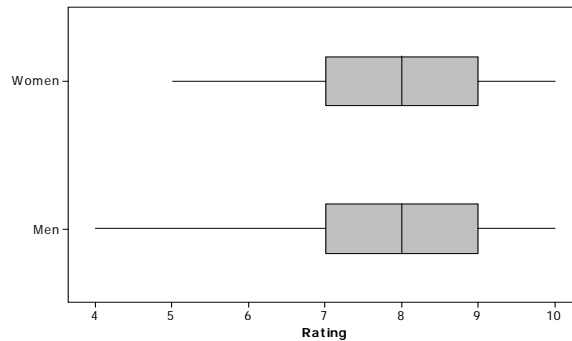
- c. Which state appears to have typically lower county income tax rates? Explain.

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.

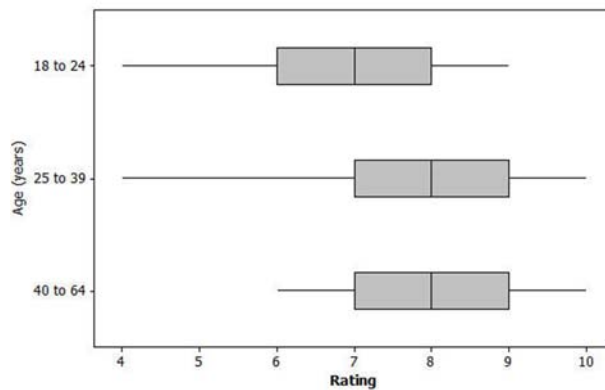
3. Many movie studios rely heavily on customer data in test markets to determine how a film will be marketed and distributed. Recently, previews of a soon to be released film were shown to 300 people. Each person was asked to rate the movie on a scale of 0 to 10, with 10 representing "best movie I've 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).