MSP

Grade 6 Module 6

Lesson Refreshers
&
Homework Starters

NOTE:
Lesson 17, 21, + 22
are projects

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").

measured or counted

words, categories or labels

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.

*Statistical questions have more than one answer.

Problem Set

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; only one answer.
   b. How many letters are in the last names of the students in my 6th grade class? yes; many answers.
   c. What are the colors of the shoes worn by the students in my school?
   d. What is the maximum number of feet that roller coasters drop during a ride?
   e. What are the heart rates of the students in a 6th grade class?
   f. How many hours of sleep per night do 6th graders usually get when they have school the next day?
   g. How many miles per gallon do compact cars get?

2. Identify each of the following data sets as categorical (C) or numerical (N). Explain your answer.
   a. Arm spans of 12 6th graders Numerical; measurement
   b. Number of languages spoken by each of 20 adults
   c. Favorite sport of each person in a group of 20 adults categorical; non-numerical value
   d. Number of pets for each of 40 3rd graders
   e. Number of hours a week spent reading a book for a group of middle school students

3. Rewrite each of the following questions as a statistical question.
   a. How many pets does your teacher have?
   b. How many points did the high school soccer team score in its last game?
   c. How many pages are in our math book?
   d. Can I do a handstand?

Melissa Farrell

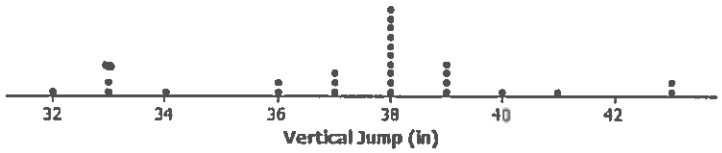
Lesson Summary

In this lesson, numerical data collected to answer a statistical question were shown in a *dot plot*. In a dot plot, a data value is represented by a dot over a number line. The number of dots over the number line at a particular value tells how many of the data points have that value. A dot plot can help you find the smallest and largest values, see how spread out the data are, and see where the center of the data is.

Problem Set

1. 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



a. What statistical question do you think could be answered using these data? How high can an NBA player jump from a standstill?

b. What was the highest vertical jump by a player? 43 in. because that's the highest number with a dot above it.

c. What was the lowest vertical jump by a player? 32 in.; the lowest number with a dot.

d. What is the most common vertical jump? 38 in.; has the most dots

e. How many players jumped that high? 10 players, because there are 10 dots above 38 in

f. How many players jumped higher than 40 inches? 3 players; there are 3 dots that are past 40 in

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 cannot jump as high as most of the other players.

Howie Rim

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:

11. What are the ages of 4th graders in our school?

A- Many 4th graders are around 9 or 10 years old.

12. What are the heights of the players on the 8th grade boys' basketball team?

D- There is a tall player (73 inches), while most others are around 5 feet or 60 inches.

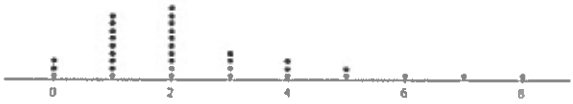
13. How many hours do 6th graders in our class watch TV on a school night?

14. How many different languages do students in our class speak?

Dot plot A - Age

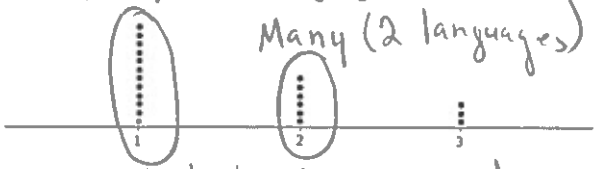
Age →

Dot plot B - TV on school night



Dot plot C - languages

Most (1 language)
Many (2 languages)



- Most students know one language.
- Many students in our class speak another language

Dot plot D - Height

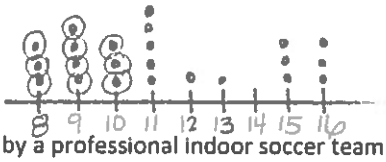
tall ↖

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.

Problem Set



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

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

   a. Make a dot plot of the number of goals scored. (rows (across) of dots must be in line)
   b. What number of goals describes the center of the data? 11 (count # of dots, pick middle one)
   c. What is the least and most number of goals scored by the team? least = 8 ; most = 16
   d. 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. The First 10 dots because they are the lowest ten.

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		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		

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

Melissa Farrell

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.

Tally (|) = 1
𝍸 = 5

Hat Sizes	Interval of Head Circumferences (mm)	Tally	Frequency
XS	510–< 530	||	2
S	530–< 550	𝍸 |||	8
M	550–< 570	𝍸 𝍸 𝍸	15
L	570–< 590	𝍸 |||	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: 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.

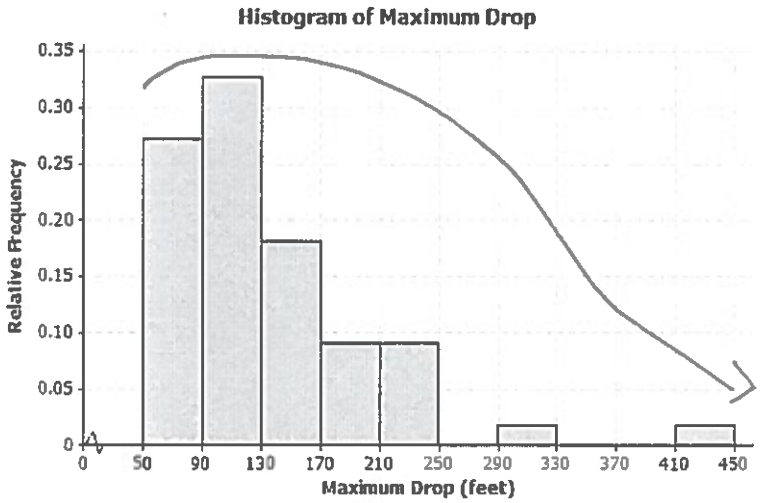
Lesson Summary

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.

Problem Set

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



a. 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?

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?

Answers

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

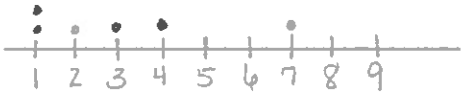
c. Yes, that span has 60% of the data.

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.

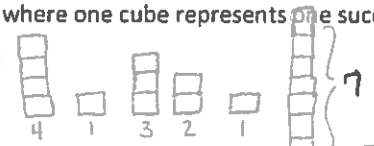
Problem Set



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.



every one is the same

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.

add up
divide by # of values

mean = 4+1+3+2+1+7
= 18 ÷ 6 = (3) ← want all to have 3 cubes.

Step described in words	"Fair Share" cube representation	Dot plot
Share 2 cubes from 7 with one of the 1 cubes		
Share 2 cubes from 5 with the 1 cube		
Share 1 cube from 4 with the 2 cubes		

* The cubes needed to be moved around so that they were all the same. They needed to be 3 because that was the mean (average) of the data.

Melissa Farrell

Howard Lin

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.

Problem Set

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

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.
   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 pockets.
   c. Sabina asks you to find the mean by using the balancing process. Demonstrate that 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)

negative deviations

370 - 380 = (-10)
375 - 380 = (-5)
total -15

positive deviations

384 - 380 = (+4)
384 - 380 = (+4)
390 - 380 = (+10)
total +18

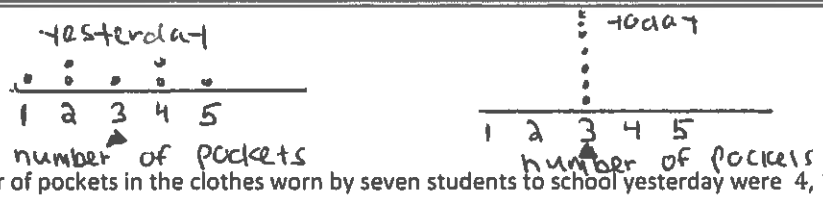
In order to balance properly, the mean sum of negative deviations and positive deviations need to be opposites, and equal 0 when added together.

The deviation from 380 for the sixth station must be -3. Therefore the price of gasoline must be 377 cents at the 6th station.

Lesson Summary

We can compare distributions based on their means, but variability must also be considered. The mean of a distribution with small variability (not a lot of spread) is considered to be a better indication of a typical value than the mean of a distribution with greater variability (wide spread).

Problem Set



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

yesterday 1+2+2+3+4+4+5 = 21
21 ÷ 7 = (3) mean
today 3+3+3+3+3+3+3 = 21
21 ÷ 7 = (3) mean

The mean of 3 is a better indicator for "today" because there is no variability and all students have 3 pockets.

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

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

Howard Pain

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	
0	−3	
3	3 − 3 = 0	
3		
5		
7		
Sum		

Girls' Team		
#Goals	Deviations	Absolute Deviations
2		
2		
3		
3		
3		
5		
Sum		

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

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:

how far each number is from the mean.

add

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 deviations of data around the mean is zero.

-24 -21 -16 -13 -8 -2 2 8 13 18 21 22

39 42 47 50 55 61 Δ 65 71 76 81 84 85

-24 + -21 + -16 + -13 + -8 + -2 = -84 2+8+13+18+21+22 = 84

-84+84 = 0

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. Add the deviations. (all are positive) 84+84 = 168

Absolute deviation (make all positive) then divide by total # of data

+24 +21 +16 +13 +8 +2 2 8 13 18 21 22

39 42 47 50 55 61 Δ 65 71 76 81 84 85

168 ÷ 12 = 14

on average the temperatures differ from the mean of 63° by 14 degrees.

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

MAD = average of absolute deviations

mean

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

7+4+2+1+0+3+3+4 +6+5+1+6 = 42

42 ÷ 12 = 3.5

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.

The MAD in NYC is 14° and in San Francisco it is 3.5°. If he wants more variability he should choose NYC because it has more changes.

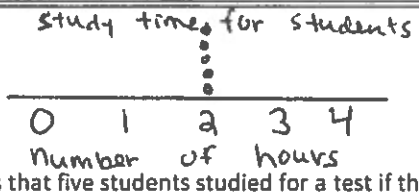
Melissa Lamm

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.

Problem Set



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.

If the MAD is zero, all values must be the same. Since the mean was 2, all values must be 2.

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?

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.

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

d. Describe the shape of this data distribution.

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.

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.

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.

*Calculate MAD:
Step 1: Find average (mean)
Step 2: Determine how far away each data # is from the mean.
Step 3: Find the mean (average) from Step 2.

Problem Set

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.

b. Suppose that your score was below the mean score. In which class would you prefer to have been? Explain your reasoning.

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. add & divide by total # of plants

b. Calculate the mean number of tomatoes produced for each variety. Which one produces more tomatoes on average? Love Em = 28 Wonderful = 32 On average, Wonderful produces more.

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.

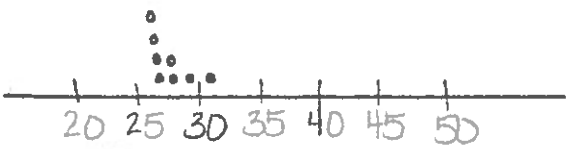
The one with the smaller MAD because it has less variability.

d. Calculate the MAD of each plant variety.

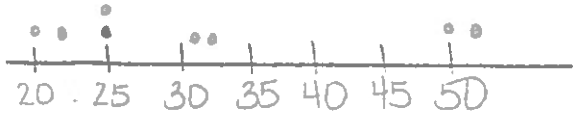
Love Em
1+1+1+0+3+1+0+1 = 8÷8
= 1

Wonderful
1+12+7+18+0+7+10+19 = 74÷8
= 9.25

Love Em



Wonderful

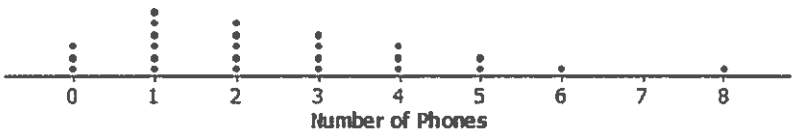


Melissa Farrell

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 0, 1, 5, and 6, 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. If the set was 10, 50, 6 the median of 50 isn't halfway between 10 and 6
   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 also increase by 10. If the set is 1, 2, 3, 4, 5, the median is 3 If you add 10, the set becomes 11, 12, 13, 14, 15 and the median would be 13 (10 more than before)

3. Make up a data set such that the following is true:
   a. The set has 11 different values and the median is 5.
   b. The set has 10 values and the median is 25.
   c. The set has 7 values and the median is the same as the smallest value.

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?
   b. Why do you think three people have no landline phones in their homes?
   c. Find the median number of phones for the people in the sample.
   d. Use the median and the range (maximum-minimum) to describe the distribution of the number of phones.

Howard Pinn

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.

Problem Set

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	58	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? ans. will vary.

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, the mean for July is 89. There are 31 days, so $\frac{1}{4}$ is 8 days. 95° for 5 days 100° for 3 days 87° for the rest 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 is smaller because it has less variability.

d. Find the IQR for the average monthly high temperature for each city. How do the results compare to your conjecture?

Step 1: order data least to greatest.
Step 2: Find the middle number.
Step 3: Find the middle of the first half and middle of the second half
Step 4: Subtract the two from step 3

St. Louis: 40, 43, [45, 55] 56, 67 | 69, 77, [81, 85], 88, 89
middle

San Fran: 57, 57 [60, 62] 63, 63 | 64, 67 [67, 68] 69, 70
middle

61 (middle) 67.5 (middle)

IQR → 67.5 − 61 = 6.5

45 + 55 = 100 ÷ 2 = 50

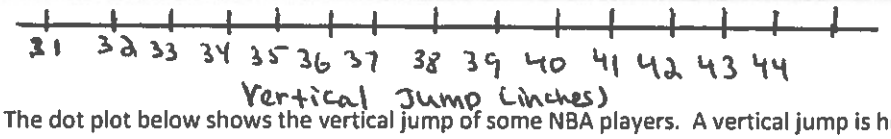
81 + 85 = 166 ÷ 2 = 83

IQR → 83 − 50 = 33

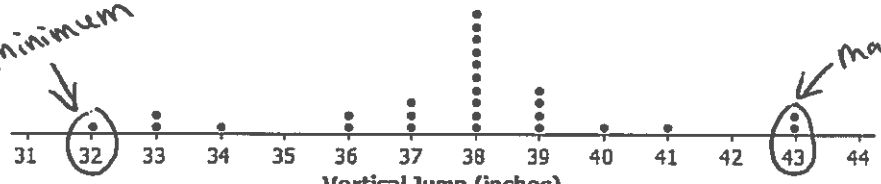
Melissa Kramer

Box Plot

Five summary values are 32, 37, 38, 39, and 43.



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.



Median of lower half of data

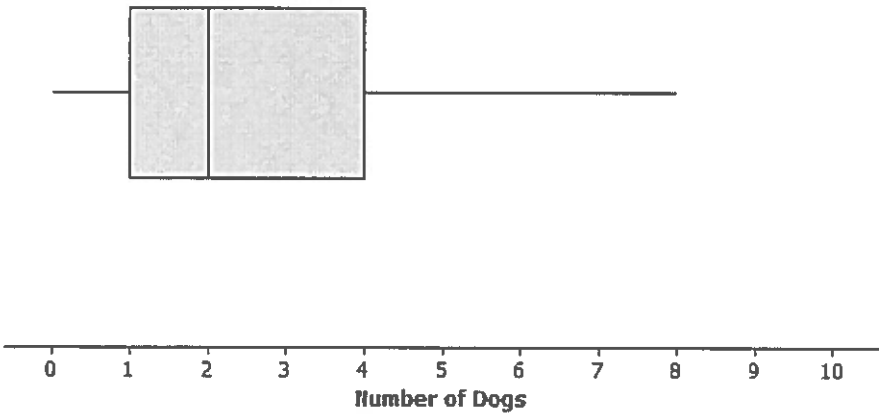
Median of upper half of data

32 33 33 33 34 36 36 37 37 37 38 38 38 38 38 38 38 38 38 39 39 39 39 40 41 43 43

Q1 Median Q3

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.
   b. Make a prediction about the part of the United States you think the city might be located. Explain your reasoning.
   c. Describe the data distribution of temperature. Include a description of the center and spread.

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.
   b. At least $\frac{1}{2}$ of the houses have 2 or more dogs.
   c. All of the houses have dogs.
   d. Half of the houses surveyed have between 2 and 4 dogs.
   e. Most of the houses surveyed have no dogs.

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.
   b. Sketch a box plot.
   c. Are there more data values above or below the median? Explain your reasoning.

4. The speeds for the fastest dogs are given in the table below.

maximum

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

Q3 (middle of upper half)

Q1 (middle of lower half)

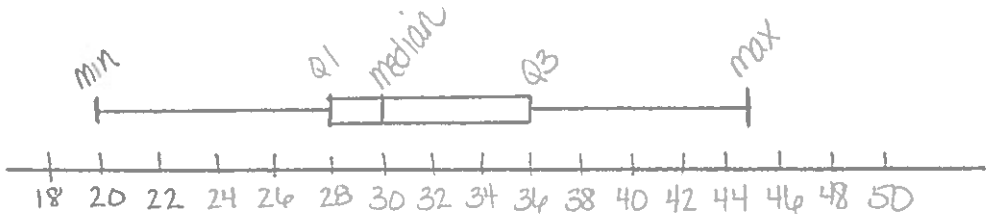
median (middle)

minimum

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.
   b. Why is the median not in the center of the box?
   c. Write a few sentences telling your brother or sister about the speed of the fastest dogs.

a.



b. The median is not in the center of the box because the data in the lower half (before the median) is close together. The data in the upper half (after the median) is spread out.

c. The fastest dog runs 45 mph. The slowest dog, the Australian cattle dog, runs 20 mph. Half of the speeds are between 28 mph and 36 mph. (Inside the box)

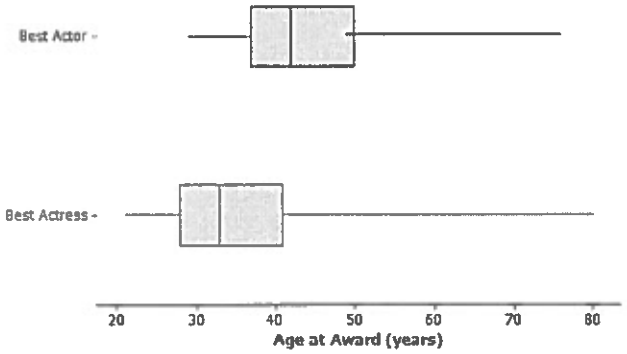
Melissa Farrell

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.

Problem Set

1. The box plots below summarize the ages at the time of the award for leading actress and leading actor Academy Award winners.



Data Source: http://en.wikipedia.org/wiki/List_of_Best_Actor_winners_by_age_at_win
http://en.wikipedia.org/wiki/List_of_Best_Actress_winners_by_age_at_win

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. You may answer yes or no, as long as you explain why.

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. You cannot tell from the box plot if they were a lot older than others.

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 older actors because he was past Q3. Less than 25% of the male winners 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 winning actresses is 33, and for actors it is 42. The upper quartile is also lower for actresses, 41, compared to 49 for actors. About $\frac{3}{4}$, or 75%, of actresses who won were younger than the median for men.

Herman Rim

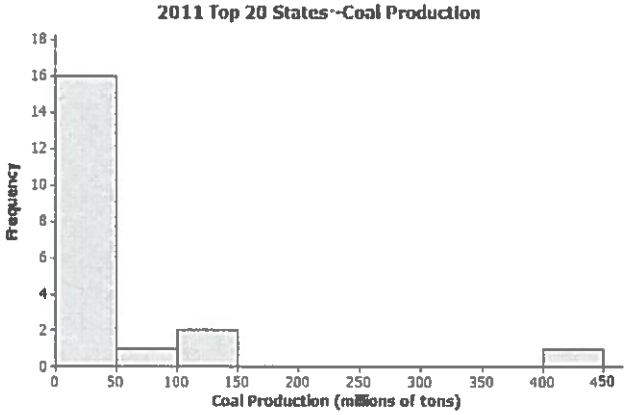
Lesson Summary

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.

Problem Set

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

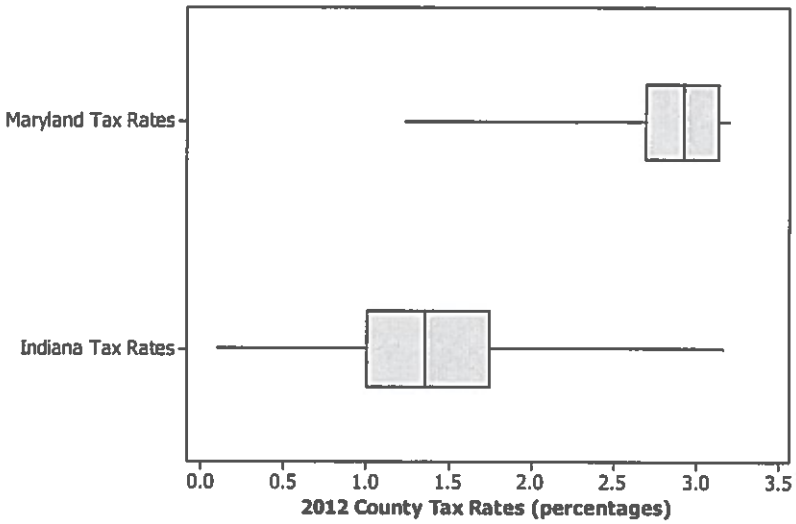
c. Minimum = 10, Q1 = 37.5, Median = 62, Q3 = 105, Maximum = 439; Mean = 54.6, MAD = 52.36

- Choice a is eliminated because the MAD will be larger than 2.76. This is because of the outlier bar between 400 and 450. That would make the mean higher.
- Choice c is eliminated because of the outlier the mean is higher than the median.
- This leaves choice b as the correct sets of summary.

Melissa Farrell

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.

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.

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

a. True, the median in Maryland is 3%. The maximum in Indiana is just over 3%.

b. True, the data in Maryland is more compacted (closer together) The data in Indiana is more spread out. Maryland has a smaller range and IQR.

c. Indiana appears to have typically lower county income tax rates. Indiana has a much smaller median and much of its data is below Marylands minimum.

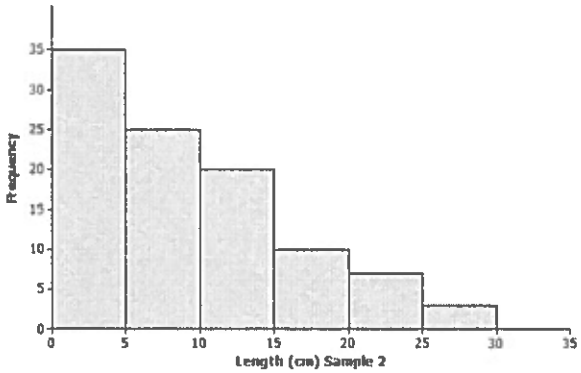
Melissa Farrell

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.

Problem Set

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.

2. Does this histogram represent a data distribution that is skewed or that is nearly symmetrical?

3. What measure of center would you use to describe a typical length of a yellow perch in this second sample? Explain your answer.

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:

   Min or minimum value = 2 cm (smallest)

   Q1 value = 4 cm (value greater than 2, but within the interval of 0 to 5 cm)

   Median = 7 cm (a value within the interval of 5 to 10 cm)

   Q3 value = 12 cm (a value within the interval of 10 to 15 cm)

   Max or maximum value = 29 cm (largest)

answers will vary for Q1, median and Q3.

Howard Pine