Math Lesson 3: Displaying Data Graphically

Hawaii DOE Content Standards:

Math standard: [Data Analysis, Statistics, and Probability]-Pose questions and collect, organize, and represent data to answer those questions.

Key concept:

Compare the graphs of two data sets.

Performance indicator: After completing this lesson, students will . . .

- construct dot plots, histograms, stem-and-leaf plots, box-and-whisker plots
- examine the characteristics of graphed data

Vocabulary:

Measures of center, mean, median, mode, spread, resistance to extremes, quartiles, five-number summary, inter-quartile range, outlier

Time:

Three or four class periods

Materials:

Butcher paper, tape, Post-It™ Notes, TI-84 calculators, View Screen

Discussion

There are many ways to represent data in a graph, and each method reveals interesting information about the data. Suppose the Hawaii Board of Realtors listed the prices of every house sold on Oahu during the past year. These pages of numbers, simply compiled in a list, would not tell us anything. But if we represented them in a graph, we would learn all sorts of things about the price of houses on Oahu. In this lesson we will learn how to construct dot plots, histograms, stem-and-leaf plots, and box-and-whisker plots. Here we go--*let's let the numbers speak to us!*

Now that you have collected your data for the importance of mathematics in society question, let's pool our data together for the entire class and see what it tells us. In order to do that, we'll start by displaying our data in a dot plot.

A. Dot (line) plots (See Appendix C.)

Sample Dot Plot:

```
                                x
                                x        x
                                x        x        x
                x         x     x        x        x         x
x               x    x    x     x        x        x         x
________________________________________________________________
1       2       3    4    5     6        7        8         9
```

Once the dot plot is completed, you can answer the following questions:

i. What is the shape of the distribution of the data?
ii. What is the range of the data?
iii. What seems to be the most common response?

B. Histograms

Let's now introduce the use of the calculator for graphical data display. On your TI-84 go to STAT, 1:Edit, and enter all the data into L1. Arrow down after each data entry. Then go to Stat Plot (2nd y=). Enter Plot 1 and turn ON. Arrow down to Type:. Arrow right to histogram and press ENTER. For Xlist, enter the number of the list in which the data are stored – for example, (L1). Then press ZOOM 9 to see the histogram. Note the similarities and/or differences between the dot plot on the board and the histogram on the screen. Press TRACE to see the actual number of data entries in each column.

Measures of Center

Now that you have the class data both listed in your calculator and on the board in the form of a dot plot, we are ready to think about what we mean by the "center" of the data. There are three common ways to measure the center, and the words all begin with the letter "m" – **mean, mode and median.**

Mean

The **mean** is the **arithmetic average.** To find the mean, add up all the values in the data set and then divide by the number of data entries in the set. For example, if you wanted to find the average number of people living in a house for all the students in your class, you would add up the number of people living in each student's house, then divide by the number of houses represented by the students in the class. The result would be the arithmetic mean (commonly called the average); that is, the average number of people per household for students in your class. To find the mean response for your activity question above (What is the importance of mathematics in society?), take the sum of all of the responses and divide by the number of persons who responded. With so many data entries, that would take a long time, but your calculator will do it for you instantly. Here's how:

With your data in List 1, go to STAT, arrow right to CALC, press 1 for 1-Var Stats. Now the cursor is asking where your data is, so enter L1 (2nd 1) and press ENTER. The symbol x-bar ("x" with a bar over it) represents the mean. What is the mean response for your data? (See Appendix C.)

Practice problems:

1. a) Mr. Sneed's math class earned the following scores on their midterm exam: 81, 67, 93, 71, 65, 61, 32, 81, 83, 92, 81, 80. What is the mean score for Mr. Sneed's class? (Enter the data for Mr. Sneed's class into List 2) **[73.9]**

   b) Mrs. Short wanted to compare her class's scores with Mr. Sneed's scores. Her students got the following exam grades: 98, 65, 77, 73, 70, 81, 89, 84, 64, 68, 95, 56, 81. What was the mean score for Mrs. Short's class? (Enter the data for Mrs. Short's class into List 3) **[77]**

   c) Whose class did better? Are you sure? (See Appendix C.)

   d) What was the most common score in each of the classes?

   (Make histograms for the data in List 2 and in List 3 and compare the shapes of the two histograms.)

Mode

mode

That last question (#1d) leads us to the definition of another measure of center, the **mode**. The mode is the value that occurs most frequently in a set of data values. A data set may have one mode, or more than one mode, or no mode at all! What is the mode for Mr. Sneed's exam scores? **[81]** Mrs. Short's exam scores? **[There is no mode]** What is the mode for the set of observations your class collected in response to the "importance of math" question? How do you know? **[The mode would correspond to the tallest column.]**

Median

median

One of the most useful measures of center is the **median**. The median is the middle value when all the data are arranged in numerical order. It is the value that divides the data in half. That is, half of the data entries are less than the median, and half are greater than the median. The median is often a better measure of “average” or “center” than the mean because it is determined by its position when all the data are arranged in numerical order, and it is not affected by extreme values in the data set.

The way to compute the median is as follows:

1. Arrange all the data in order, from smallest to largest.
2. If the number of data entries is odd, the median is the data value exactly in the middle, with the same number of data observations below the median as above the median.
3. If the number of data entries is even, take the two middle values and find their average. In other words, add them together and divide their sum by 2. The result is the median. NOTE: when there is an even number of data entries, the median of a data set may not necessarily be one of the data entries.

Practice problems:

2. a) What is the median score for Mr. Sneed’s math class in Practice Problem #1 above? **[80.5]**

   b) What is the median exam score for Mrs. Short’s class? **[77]** Now whose class do you think did better? Why do you think we got different results when we calculated the means and the medians? What scores might have had a big influence on the mean but not the median?

   c) Now let’s discover an easy way to find the median on our calculators. Under EDIT, press SortA to sort the data in List 2 in ascending order. Do the same for Mrs. Sneed’s scores in List 3. Since these are relatively small data sets, you can count to the middle value to find the median. That is easy to do for Mrs. Short’s class since there are an odd number (13) of students. For Mr. Sneed’s class, find the average of the two middle scores.

   d) Carlos, the smartest student in Mr. Sneed’s class, missed the midterm exam because he was taking his driver’s test that day. When he came back, he took the test and scored 100. By adding Carlos’ score to the others in the class, how did the mean change? [**It changed from 73.9 to 75.9]** How did the median change? **[It changed from 80.5 to 81, not much at all.]**

   e) What if Mr. Sneed allowed Carlos to tutor some of the students and let them retake their final exam. When they did, interestingly enough, only the three students who had scored the highest the first time around did better. In fact, they each scored 100. The other students' scores remained unchanged. The new set of scores is: 32, 61, 65, 67, 71, 80, 81, 81, 81, 100, 100, 100. (Carlos did not retake the test.) *Here you may copy the data in List 2 to appear in List 4 as follows: Highlight the cursor in the list name at the top of the column (L4). Press ENTER and note the cursor will be flashing at the bottom where it says "L3 =   ". Type in L2 and press ENTER. The data in L2 will be listed in L4. Delete the three highest entries (83, 92, 93) and replace them with 100,100,100. What is the new mean? **[76.5]** The new median? **[80.5]** Compare these with your answers in problems 1a and 2a. Why do you suppose the mean changed when the median did not change? What does this tell you about the median? **[The median is resistant to extreme scores on either the low or the high end while the mean is not. The mean is pulled in the direction of the lower or higher scores.]** What if the lowest score were changed to 0? Would the mean change? **[yes]** Would the median? **[no]** (See Appendix C.)

What is the mode of Mr. Sneed's new data set of scores? **[There are two modes – 81 and 100]**

C. Stem-and-Leaf Plots or Stemplots

An easy way to quickly find the median of a data set is to represent the data graphically in a stem-and-leaf plot. Use the data above in problem #1a to construct a stem-and-leaf plot for Mr. Sneed's scores as follows:

stem-and-leaf plot

Make a vertical line to separate the "stems" (the tens digits) from the "leaves" (the ones digits). Place the stem values to the left, and the corresponding leaf values to the right. A second run-through would enable you to put the data in numerical order by rearranging the order of the leaves.

3	2
4	
5	
6	7 5 1
7	1
8	1 1 3 1 0
9	3 2

3	2
4	
5	
6	1 5 7
7	1
8	0 1 1 1 3
9	2 3

1. Consider the shape, center and spread of the data reflected in your stem-and-leaf plot. Can you tell from the stem-and-leaf plot where the center is? Count from the lowest score to the "middle" score. Do you see how easy it is to find your median in a stem-and-leaf plot?

2. Both a dot plot and a histogram show the shape, center and spread for the distribution of the data, but neither of these graphical representations preserve the actual data values. The stem-and-leaf plot also shows the shape, center and spread of the data, but it has the additional advantage of also displaying the actual data values.

back-to-back

3. Suppose that Mrs. Short wanted to see how her students fared in comparison with Mr. Sneed's students on the midterm exam. Her students' scores were as follows: 98, 65, 77, 73, 70, 81, 89, 84, 64, 68, 95, 56, 81 Now add to your stem-and-leaf plot above, by putting the "leaves" for Mrs. Short's students' scores to the **left** of the stems shown in the vertical line. This is called a **back-to-back stem-and-leaf** plot. Note the order of the leaves on the left. In each stemplot, the leaves increase as they go away from the stems. With such a plot it is easy to compare two data sets by looking at shapes, centers and spreads. Can you tell by looking at the stemplot which class did better?

				3	2			
				4				
			6	5				
	8	5	4	6	1	5	7	
	7	3	0	7	1			
9	4	1	1	8	0	1	1	1
		8	5	9	2	3		

Practice problems:

3. Mrs. Poppington's class made the stem-and-leaf plot shown at right after they collected cans for a food drive. Use it to complete the parts below.

Cans collected by Mrs. Poppington's class

0	5	6	
1	1	1	4
2	3	4	7
3	5	5	5
4	7	9	9
5	3	7	

Key
2|3 means 23

a. List the number of cans collected by each of the students who collected 40 or more. **[47, 49, 49, 53, 57]**

b. Find the mode of the data. **[35]**

c. Find the median number of cans brought in by the students in Mrs. Poppington's class. **[31]**

d. What do you think the key in the box above is telling you about the stems and leaves? **[How to read the stem and leaf plot, that is, the left side of the line is the tens place and the right side is the ones place.]**

4. Jane recorded the number of minutes she spent talking on the phone per day: 15, 25, 60, 10, 120, 85, 35, 20, 60, and 30. Make a stem-and-leaf plot to organize her data.

1	0 5
2	0 5
3	0 5
4	
5	
6	0 0
7	
8	5
9	
10	
11	
12	0

   a. Find the mean, median, and mode. **[mean=46; median=32.5; mode=60]**

   b. If Jane wanted to convince her parents that she did not spend too much time on the phone, which of the measures of central tendency should she use? **[She would probably use median since it is the lowest number.]**

   c. If her parents argued that they disagree with her, what evidence could they use? **[They would be using mode since it is the highest number.]**

5. Shelley surveyed her classmates to find out how much money they had in their pockets, wallets, and backpacks. The dollar amounts were $1, $2, $1, $1, $8, $1, $7, $10, and $5.

   a. What is the mode of these dollar amounts? **[$1]**

   b. If the person with the most money had $100 instead of $10, would the mode change? Why or why not? **[No; mode measures frequency, not size or amount.]**

   c. Would changing the largest amount of money change the median? **[No; making the highest number larger does not change the middle number.]**

   d. What is the mean for the original dollar amounts? **[$4]**

   e. Would changing the largest amount to $100 change the mean? If your answer is yes, find the new mean. If not, explain why not. **[Yes; the mean changes from $4 to $14.]**

D. Box-and-Whisker Plots

Often we want to know more about how data are spread out. For example, if you were a teacher, you might anticipate that you would have some very bright and hard-working students in your class who always scored well, and you also might expect to have a few who consistently scored in the low range. But how were the majority of students in the middle doing? Could you find out? How would you measure the spread around the median? Once you have divided your data into two halves by finding the median, find the values that divide each half in half again. These two values are called the **lower quartile** or quartile 1 **(Q1)**, and the **upper quartile** or quartile 3 **(Q3)**. Together with the median they divide the data set into four equal parts. The distance between the upper and lower quartiles, called the inter-quartile range, or IQR, is a

measure of how the data are spread out. The data within the IQR comprise the middle 50% of your data.

$$\text{IQR} = Q_3 - Q_1$$

Let's investigate the inter-quartile range for the combined scores for both Mr. Sneed's and Mrs. Short's students. If we were to make a single stemplot for the combined data, it would look like this:

3	2								
4									
5	6								
6	1	4	5	5	7	8			
7	1	3	3	7					
8	0	1	1	1	1	1	3	4	9
9	2	3	5	8					

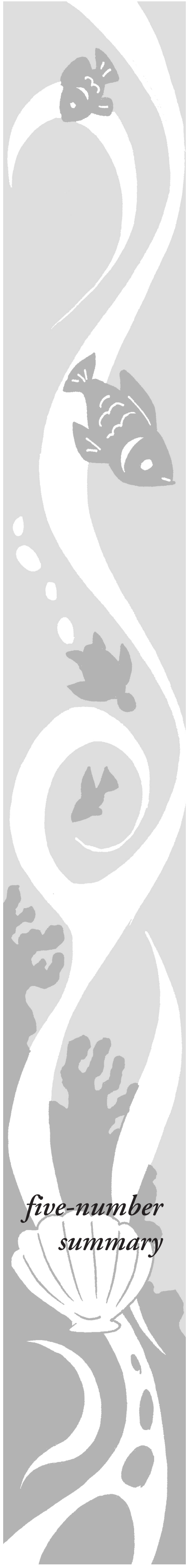
Practice problem:

6. Identify the mean, median, mode and range for the combined data set above. **[x bar= 75.6, median = 80, mode = 81, range = 98 – 32 = 66]**

   Now compute Q_1 by finding the median for the data in the bottom half of the data set. *Do not include the median in the bottom half or the top half when calculating the middle of either the bottom half or the top half of the data set.*

   Here's the bottom half of the data: 32, 56, 61, 64, 65, 65, 67, 68, 71, 73, 73, 77

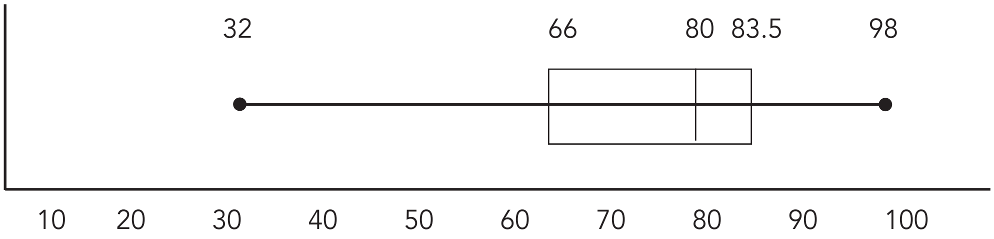
   The median of this half is 66, so Q_1 = 66. What is Q_3? **[Q3 = 83.5]** Now we know that 25% (one-fourth) of the students had scores ranging from 32 to 66; 50% of the students had scores between 66 and 83.5; and the top 25% of the students scored between 83.5 and 98. What would be another name for the median? **[Q2]** When we combine these five important dividing scores - the minimum score, the Q_1 score, the median, the Q_3 score and the maximum score - we have an important summary of the data called the **five-number summary:**



five-number summary

   Five-number summary: min - Q_1 – median – Q_3 – max

   For the data set above, the five-number summary is 32 – 66 – 80 – 83.5 – 98. A visual that helps to show how the scores are spread is found in a box-and-whisker plot, which is constructed from this five-number summary.



Can you see the spread of the bottom 50% of the scores? [**32 to 80]**
The top 50% of the scores? **[80 to 98]**
The middle 50% of the scores? **[66 to 83.5]**
The bottom 25% of the scores? **[32 to 66]**
The top 25% of the scores? **[32 to 66]**

The calculator computes the five-number-summary each time you go to STAT/CALC/1-Var Stats/ENTER. After entering the list number in which the data are stored, press ENTER. Scroll down past the n= and find the five-number summary.

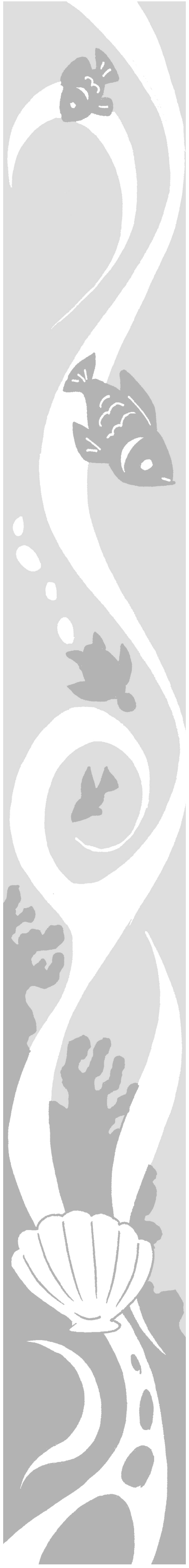
Your calculator will also graph a box-and-whisker plot for your data. Press 2nd/y= and turn one of the plots ON (arrow and press ENTER). Under Type: arrow to the middle plot of the second row and press ENTER. Your cursor will be next to Xlist. Enter there the name of the list in which you have your data stored (e.g. L1). Press ZOOM 9 to graph and you will see your box plot. Use the TRACE button to see the five-number summary as you arrow through the box plot.

E. Outliers

Did you notice the score of 32 in Mr. Sneed's class? Was this an unusual score for the class? In other words, was there perhaps some explanation for this low score (such as a long absence on the part of the student)? Maybe the score should not be included in the data set if it resulted from some unusual circumstance and if we want to get a truly reliable picture of how Mr. Sneed's class is doing. A score that differs significantly from most of the data entries is often considered an **outlier**. Sometimes it is best to eliminate it from the data set. For example, suppose you we entering heights in inches for students in your class into a list on your calculator, and instead of entering 65 you entered 6 by mistake. The inclusion of a data entry of 6 inches would greatly throw off your measure of center. (Which measure of center would it have the most effect on – mean or median?) It should be eliminated from the data set or be corrected.

outlier

But what about the score of 56 in Mrs. Short's class? Is that an outlier? Often an outlier stands away from the rest of the data in a dot plot or stem-and-leaf plot. But how do we know if a score is far enough away, or different enough from the rest of the data, to be considered a legitimate outlier? There is a mathematical rule that will allow you to make the determination so that you don't have to guess.



A data entry is an outlier if it is more than one-and-a-half times the inter-quartile range above Q3 or below Q1.

To find out if the score of 32 is really an outlier, first compute the IQR of the data set for the original scores in Mr. Sneed's class given in Practice Problem #1a.

$$\text{IQR} = Q_3 - Q_1 = 82 - 66 = 16$$
$$1.5 \times \text{IQR} = 1.5 \times 16 = 24$$

Is 32 far enough below Q_1 to be classified an outlier? Let's see:

$$Q_1 - 24 = 66 - 24 = 42$$

The score of 32 is below 42, so 32 is an outlier.

Is 98 an outlier in the other direction?

$$1.5 \times \text{IQR} = 1.5 \times 16 = 24$$
$$Q_3 + 24 = 82 + 24 = 106$$

The score of 98 is less than 106, so 98 is not an outlier.

Practice problem:

7. In "Ages of Oscar-Winning Best Actors and Actresses" (Mathematics Teacher magazine) by Richard Brown and Gretchen Davis, stem-and-leaf plots are used to compare the ages of actors and actresses at the time they won Oscars. Below are the ages for 34 recent Oscar winners in each gender category:

Actors: 32 37 36 51 53 33 61 35 45 55 39 76 37 42 40 32 32
60 38 56 48 48 40 43 62 43 42 44 41 56 39 46 31 47

Actresses: 50 44 35 80 26 28 41 21 61 38 49 33 74 30 33 41 31
35 41 42 37 26 34 34 35 26 61 60 34 24 30 37 31 27

a) Construct a back-to-back stem-and-leaf plot for the above data, using the tens digits for the stems. Discuss shape, center and spread for each data set, comparing the two sets and discussing any differences you see.

b) Construct two box-and-whisker plots on the same graph to compare ages of actors and actresses at the time they won Oscars. What do you observe? Write a short paragraph discussing the bottom 25% of ages, the middle 50% of ages and the top 25% of ages, and discuss difference you observe in the data sets. What conclusions can you draw?

c) Are there any outliers in the data sets? Verify your answer.

Your calculator will automatically show you if there are any outliers in your data set. Proceed as above to graph a box-and-whisker plot from your data, but instead of highlighting the middle graph in the second line of Type:, use the first graph in the second line. If there are any outliers in your data set, they will be identified by dots on your plot. Again, use the TRACE button to find their values.