2018

Mathematics in Contemporary Society - Chapter 6 (Spring 2018)

Patrick J. Wallach
CUNY Queensborough Community College

Chapter 6

Lab Assignment #5–Using Your Own Data Set
Due _____

1) Collect a data set of 30 **related** scores from the Internet. **(Print out the data set.)** Some sources you might try are:
Weather sites: www.weather.com/ , www.weather.yahoo.com/ , http://www.cnn.com/WEATHER/ for temperatures, precipitation, etc.
Financial sites: www.nasdaq.com/ , www.morningstar.com , finance.yahoo.com/ for stock prices and business information
Sports sites: www.espn.com/ , sports.yahoo.com/, www.si.com for tons of sports related values.

2) Enter the 30 values into Excel in column A in cells A1 to A30.
3) Calculate measures of average using Excel's Paste Function button:
   a) Type the word Mean in a cell, such as D1.
   b) Click on the cell right next to it, such as E1.
   c) Click on the f_x button.
   d) Under **Select a Category**, click on **Statistical**:
   e) Under **Select a Function**, click on **AVERAGE**. Click OK.
   f) For the first field in the box (Number 1), use the mouse to highlight all 30 values in A1 to A30.
   g) Click on OK. The value that appears is the mean!
   h) Repeat the process in cells D2 and E2 using the **MEDIAN** function.
   i) Repeat the process in cells D3 and E3 using the **MODE** or **MODE.MULT** function. Be careful! Excel will only give you one value for the mode even if there are multiple modes. Note any other modes in F3. If there is no mode, #N/A will appear.
4) Create a simple frequency table (as we had in Lab#3) from your data. Include at least 6 equally spaced categories.
5) Create a line chart from the frequency table.
6) Answer the following questions (you may refer to the notes):
   1) Describe your data set. What is the population you are examining? How did you obtain your sample?

2) Describe your distribution from the line chart. How many peaks do you have? Is it symmetric? Does it appear to be left-skewed, right-skewed, or normal?
3) How do the mean, median and mode compare to the peak of your line chart? That is, if your chart has a peak, does it match any of the three?

Variation

If we have three sets:

Set A: 78 79 80 80 81 82
Set B: 70 75 80 80 85 90
Set C: 60 70 80 80 90 100

If we were to calculate the mean, median and mode of each data set, we would see that all three values are the same (80 in every case) for each set. But obviously, each set is different. How are the sets different?

If Set A were a set of students' test scores, everyone would be getting roughly the same grade (B-/C+). In Set B, the values are spread further apart; one student is getting a C- while another is getting an A- (along with a C grade and B grade). Finally in Set C, one student is almost failing with a D- while another is getting an A+ grade (along with a C- and A- grade). In each set, there are two students getting B- grades, but these are certainly very different situations.

What makes each set different is the variation, which measures how close (or far apart) the values of a data set are. Just as there are different measures of average, there are also several measures of variation of interest to us.

Measures of Variation

Range

The **range** of a data set is the highest value minus the lowest value. That's it. If we examine our three data sets, we see that:

Range of Set A = 82-78 = 4
Range of Set B = 90-70 = 20
Range of Set C = 100 - 60 = 40

Clearly, the values in Set A must be close together if the range is so small. The values in Set B and Set C are clearly spread further apart.

The range tells us something about how close or far apart values are in a data set are. But the range has its limits. If we consider two more sets:

Set D: 40 84 86 87 88 91 92 94 97 98 99 100
Set E: 40 47 52 59 66 71 73 79 85 91 95 100

Each set has the same range (100-40=60), but the sets are obviously different. The value 40 is more of an **outlier** of the first set (all other values are between 84 and 100), but the values in Set E are spread throughout the set.

So we look to other measures of variation.

Five-number summary

The **five-number summary** describes a data set with five values: The **low value**, the **lower (or first) quartile**, the **median**, the **upper (or third) quartile** and the **high value**.

For example, we start with another data set:

Set F: 51 67 72 77 81 82 84 86 89 92 99

It's easy to get that the low value=51, the median=82 and the high value=99. What else do we do?

The lower quartile is the "median" of the values below the median. The values 51 67 72 77 81 are below the median; 72 is in the middle of them. So 72 is the lower quartile value.

The upper quartile is the "median" of the values above the median. The values 84 86 89 92 99 are above the median; 89 is in the middle of them. So 89 is the upper quartile value.

The five-number summary is then:

Low value = 51
Lower quartile = 72
Median = 82
Upper quartile = 89
High value = 99

By breaking a data set into five values, we can get a better sense of how the values of a data set are spread apart. Outliers are easier to spot; a low value would be much lower than the lower quartile or a high value would be much higher than the upper quartile.

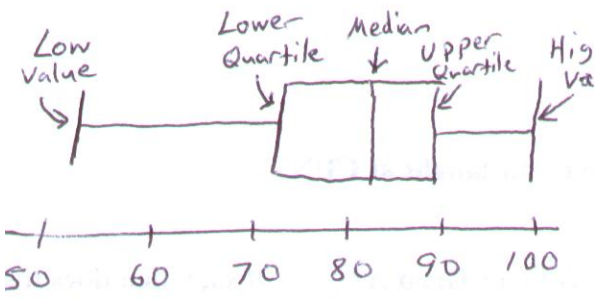
Question 1: Find the range and five-number summary of the following data set:
104 108 111 113 115 119 125 131 139 148 151 161 167

More on Variation

If the five-number summary of a data set is:

Low value = 51
Lower quartile = 72
Median = 82
Upper quartile = 89
High value = 99

You can get a visual sense of it using a **boxplot** (or **box and whisker plot**):



Each value of the five-number summary is represented by a vertical line. The five lines break the data set into quarters. Depending on how close or far apart each line is, you can get a sense of whether the numbers in the data set are close or far apart for that portion of the data set.

Standard Deviation

The **standard deviation** measures how close (or how far) the values in a data set are spread around the mean of the data set. It is calculated by measuring differences between the values of the data set and the mean itself.

The standard deviation is the most commonly used measure of variation, although it is considerably more difficult to calculate for even small data sets. The following is an example of how to calculate the standard deviation by hand. Don't worry—I won't expect you to be able to do it for an exam!

Example 1
Suppose we have a data set of 6 values: **78 85 92 81 83 91**

1) First, we calculate the mean of these values, which is 510/6=85 (See column 1 below)

2) Next, calculate the difference between each value and the mean. (See column 2. Note that the total should add up to zero.)
3) Next, square the values from column 2. Find the sum of these squares. (See column 3)
4) Finally, you take this sum, divide it by the numbers of values you have (6) minus 1, and take the square root of the result.
5) Hooray! This is the standard deviation!

Column	1	2	3
	Value	Value-Mean	(Value-Mean) squared
	78	-7	49
	85	0	0
	92	7	49
	81	-4	16
	83	-2	4
	91	6	36
Sum	510	0	154
Mean	85		

Standard Deviation = sqrt(154/5)= 5.54977477

For the future, you don't have to worry about calculating the standard deviation by yourself. Excel's **STDEV** formula will do it for you!

What do we do with the standard deviation?

We need to realize that, even if the mean is the same, the further spread out the values in a data set are, the larger the standard deviation will be.

Example 2

Consider the following set, which also has a mean of 85 (as in Example 1):

66 94 88 73 100 89

Column	1	2	3
	Value	Value-Mean	(Value-Mean) squared
	66	-19	361
	94	9	81
	88	3	9
	73	-12	144
	100	15	225
	89	4	16
Sum	510	0	836
Mean	85		

Standard Deviation = sqrt(836/5)= 12.9305839

The numbers are spread apart further away from the mean. This makes the differences larger, and the squares larger, and the sum of the squares larger. Therefore, the standard deviation is larger.

Example 3
Consider the following set, which also has a mean of 85 (as in Example 1):
82 84 89 86 88 81

Column	1	2	3
	Value	Value-Mean	(Value-Mean) squared
	82	-3	9
	84	-1	1
	89	4	16
	86	1	1
	88	3	9
	81	-4	16
Sum	510	0	52
Mean	85		

Standard Deviation = sqrt(52/5)= 3.224903099

The numbers are spread apart closer to the mean. This makes the differences smaller, and the squares smaller, and the sum of the squares smaller. Therefore, the standard deviation is smaller.

<u>Question 2:</u> Use the above technique to find the standard deviation of the values:
75 86 92 71 99 76 89

What else do I do with the standard deviation?

The standard deviation is particularly useful when working with normal distributions, which we will discuss in the next set of notes. It does give us a quick understanding of what the data set looks like, in terms of variation.

Range Rule of Thumb

The range of a data set (generally one without outliers) is approximately four times the standard deviation:

$$\text{Range} \approx \text{Standard Deviation} \bullet 4$$

In other words, the standard deviation is about ¼ of the range:

$$\text{Standard Deviation} \approx (\text{Range}/4)$$

With this information, we can estimate the range or standard deviation.

If the standard deviation of a data set is 4.5, then the range is approximately:

$$\text{Range} \approx 4.5 \bullet 4$$
$$\text{Range} \approx 18$$

If the range of a data set is 35, then the standard deviation is approximately:

$$\text{Standard Deviation} \approx (35/4)$$
$$\text{Standard Deviation} \approx 8.75$$

If we also know the mean of the data set, we can approximate the lowest and highest value of the data set using:

$$\text{Lowest Value} \approx \text{Mean} - (\text{Range}/2)$$
$$\text{Highest Value} \approx \text{Mean} + (\text{Range}/2)$$

But the range is just four standard deviations, we can also say:

$$\text{Lowest Value} \approx \text{Mean} - (\text{Standard Deviation} \bullet 2)$$
$$\text{Highest Value} \approx \text{Mean} + (\text{Standard Deviation} \bullet 2)$$

Example 4

Suppose we know that the mean of a recent exam was 83 and the standard deviation was 6.5. We can approximate the range of scores in the class.

$$\text{Range} \approx \text{Standard Deviation} \bullet 4$$
$$\text{Range} \approx 6.5 \bullet 4$$
$$\text{Range} \approx 26$$

We can also approximate the range in terms of lowest and highest value:

$$\text{Lowest Value} \approx \text{Mean} - (\text{Standard Deviation} \bullet 2)$$
$$\text{Lowest Value} \approx 83 - (6.5 \bullet 2)$$
$$\text{Lowest Value} \approx 83 - 13$$
$$\text{Lowest Value} \approx 70$$

$$\text{Highest Value} \approx \text{Mean} + (\text{Standard Deviation} \bullet 2)$$
$$\text{Highest Value} \approx 83 + (6.5 \bullet 2)$$
$$\text{Highest Value} \approx 83 + 13$$
$$\text{Highest Value} \approx 96$$

We notice the range of these values is $96 - 70 = 26$

Question 3: Suppose the mean score of the October SAT in Bayside High School was 950 and the standard deviation was 250. Approximate the range of scores for the school in terms of lowest value and highest value.

Statistical Project, Part I:

You are going to make your own statistical study. It will involve gathering your own data set; calculating measures of average and variation; making graphs; calculating z-scores and percentiles (defined in the next two weeks) and ultimately writing a report about your work.

You should have a data set ready by _____. You should be ready to discuss it.

Step 1: Choose the population you are going to analyze and the specific parameter (what do you care about) of interest. Examples of these are: student GPA, student Height, IBM Stock price, NYC winter temperatures, NYC family income, etc.
Step 2: Collect a sample of 50 values from some source (newspaper, book, internet, survey, etc.)

Ultimately, the goal is to perform statistical analysis on your raw data and create an analytical report. Try to choose a data set that is interesting and worth writing about!

For _____:
You will participate in a data set discussion. Be ready to identify the goal of your study, the population you are examining, and a "prediction" for your sample results.

You should also have your data set ready for the next lab assignment, to be given in the next chapter.