THE SCHOOL DISTRICT OF PHILADELPHIA

Philadelphia High School for Girls

1400 West Olney Avenue
Philadelphia, PA 19141
215-400-3700 215-276-5738 (fax)

Summer 2019 Mathematics Assignment
For Students Entering
Statistics

Your summer assignment consists of statistics review content and practice problems. Please learn all the definitions from the review content and show your work (all steps) for each problem on a separate sheet of paper. The assignment is online. Write down all the necessary computations including those that you do on your calculator. If you are unable to complete a problem after reading the review, please research the topic online (Khan Academy is a great resource). No problem should be left blank.

You must do this assignment during the summer and complete it before the first day of class. Turn in only the practice problems (show your work neatly and staple them) Make sure your name is on every page of your work. Keep the notes section for reference during the school year

Summer Assignment: Available on Big Ideas Website Join by June 15th

Student Instructions for logging into Summer Assignment

1. Visit www.BigIdeasMath.com.

2. Click *New to Big Ideas Math?*

3. Enter your access code: **Access Code:** PPTS-RXYS-P88Y and click *Next.*

4. Fill out the required information and click *Next*

5. Write down your username and click *Next.* Make your password the same as your StudentNet password or something you will remember.

Select each of the summer assignments. Complete prior to the first day of school.

Questions: E-mail Ms. Baker vlbaker@philasd.org

Vocabulary & Problem Review of Statistics for Quantitative Variables

MEASURES OF CENTER: mean, median, and mode (the 3 m's) are called measures of center or central tendency. They tell us what the center or typical value of a data set is.

Mean = arithmetic average. Find the mean by adding up all the data points (numbers) and dividing by how many data points there are.

Median = the middle value. Find the median by listing the data set in order from smallest to largest and locating the middle value. If the data set has an odd number of data points, there will be one exact number in the middle. If the data set has an even number of data points, you need to take the average of the two middle values.

Mode = the number that shows up most often. A data set may have no mode (none of the values are repeated), one mode, two modes, or several modes.

MEASURES OF SPREAD OR RANGE: Range, interquartile range and standard deviation are all measures of spread

Range = maximum — minimum. The range is the difference between the largest number in the data set (maximum) and the smallest number (minimum).

Interquartile Range (IQR) = Q3 — Q1. The IQR is the range of the middle 50% of your data. We'll talk about this later, when we review box and whisker plots.

Standard Deviation (represented by the Greek letter sigma σ) is a measure that quantifies the amount of variation or dispersion of a set of data values. We'll save this for class.

Examples: Mean, Median & Mode

Data set #1: 3 5 6 6 8 9 12 13 16 19 20

Mean – There are 11 values, so add the numbers and divide the sum by 11.
(3+5+6+6+8+9+12+13+16+19+20) ÷ 11 = 10.64

Median – there are 11 values, the 6th value is in the center. Since the data set is already in order, count over to the 6th number, which is 9. The median is 9.

Mode – the only number that is repeated is 6, so 6 is the mode.

Range – the maximum is 20 and the minimum is 3. So, the range is 20 — 3 = 17.

Data set #2 3 7 11 8 5 6 9 9 5 10

Mean = (3+7+11+8+5+6+9+9+5+10) ÷ 10 = 7.3

Median – the data is not in order, so we must sort the data to find the median!
3 5 5 6 7 8 9 9 10 11

There are 10 numbers, so the 5th and 6th numbers are in the middle. The 5th number is 7 and the 6th number is 8, so average these. (7+8) ÷ 2 = 7.5. So, the median is 7.5.

Mode – there are two 5's and two 9's, so we have 2 modes in this data set: 5 and 9.

Range – the range is 8. The maximum is 11 and 3 is the minimum. So, 11 — 3 = 8.

STEM AND LEAF PLOTS – a stem and leaf plot is one way to display your data. Each data item is separated into a stem and a leaf. The leaf is the last digit in the number and the stem is the remaining digit or digits. For example, for the number 46, the stem is 4 and the leaf is 2. For the number 107, the leaf is 7 and the stem is 10.

Look at this stem and leaf plot. Notice how the leaves are in order from smallest to largest. The stem gives the hundreds and tens value for the number and the leaf gives the ones digit. So, 17|3 means 173. The first stem below gives the following values: 142, 146, and 148. There are no values in the 180s for this data set because there are no leaves on this stem. Never skip a stem – these gaps can be insightful.

14	2 6 8
15	0 2 5 8 9 9
16	0 3 5 6 7 7 7 8 9
17	0 3 4 5 7
18	
19	1

The minimum is 142 and the maximum is 191. So, the range is 49.

We can also find the median using a stem and leaf plot. There are 23 numbers displayed. The median (middle) is the 12th value, so, start counting from the minimum of 142. Be sure to start at the left of each row to count in order from smallest to largest! The median is 165.

To find the mean, add up all the numbers and divide by 23. The mean is 163.2

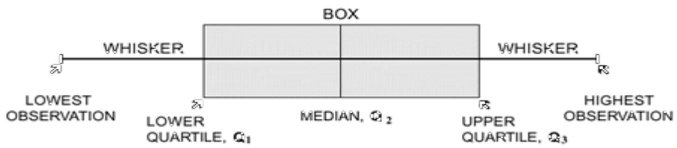
What is the mode? There are 2 modes in this data set -> 159 and 167.

Always give a title for your stem and leaf plot. Also give a key.

Example of a key: 6|2 = 62 years of age

BOX AND WHISKER PLOTS – a box and whisker plot breaks the data into quartiles (25% of the data in each part). It shows the minimum (also called the lower extreme), the maximum (upper extreme), the median, the **lower quartile** (also called Q1) and the **upper quartile** (also called Q3). The lower quartile is the median of the lower half of a data set. The upper quartile is the median of the upper half of a data set. A box and whisker plot looks like:

Figure 1. Box and whisker plot



The five numbers that show up on the box and whisker plot are called the **five-number summary** for the data set.

Look at the data set below. It shows the number of shots made from 50 free throws for 18 players. 12 32 42 50 39 38 27 34 42 35 48 36 13 15 23 16 25 43

What's the first thing we need to do? Sort it from lowest to highest! We've found the median – it's between the 9th and 10th value. Below, we've marked the median with a line to divide the data set into the lower half and the upper half. How many data values are in each half? The same number!

Median
= 34.5

12 13 15 16 23 25 27 32 34 | 35 36 38 39 42 42 43 48 50

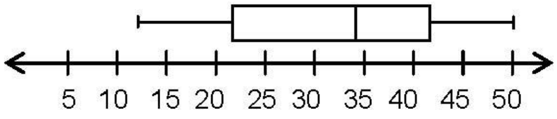
Lower Data Set Upper Data Set

Find the median of the lower data set -> 23. This is called the lower quartile, or Q1.
Find the median of the upper data set -> 42. This is called the upper quartile, or Q3.

Now, we have everything (the five-number summary) we need to make a box and whisker plot. Always use a scale to draw the box and whisker plot.

Number of shots made from 50 attempts

Min = 12
Q1 = 23
Median = 34.5
Q3 = 42
Max = 50

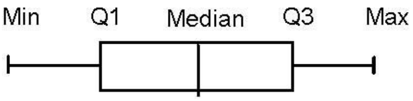


Always give a number line with a scale under your boxplot. You should also give a title!

So, if you are given a box and whisker plot, you can find the following: minimum, lower quartile, median, upper quartile, and maximum. You can find the range (maximum — minimum). But, you cannot find the mean or the mode if you are only given a box and whisker plot.

Remember, a box and whisker plot simply shows the following:

25% of the data is between the minimum and Q1.
25% of the data is between Q1 and the median.
25% of the data is between the median and Q3.
25% of the data is between Q3 and the maximum.



One more measure of range or spread that we mentioned way back on page 1 is the Interquartile Range (IQR). It is the range of the box on a box and whisker plot. It gives you the range of the middle 50% of your data. For the free throw data, IQR is:

IQR = Q3 — Q1 = 42 — 23 = 19

Below is another example with the median, Q1, and Q3 shown. Notice how sometimes the median is between 2 numbers (shown by a line |) and sometimes there is one middle number (shown by a box around it). If there is one middle number, it is not counted in either the lower or upper data set!

Q1 = 17.5 Median = 25 Q3 = 34.5

7 11 17 | 18 21 24 $\boxed{25}$ 29 32 33 | 36 36 38

OUTLIER TEST - Outliers are data values that are very small or very big in comparison to the other values in the data set. What we do with outliers (keep them or throw them out) depends on the techniques used to collect the data and on what we can find out about these data values (did something go wrong in the experiment that suggests we throw it away or does it reveal some anomaly about the pattern). For now, we want to look at the statistical test for identifying outliers.

To identify outliers, we essentially build a fence or boundary for our data. Anything outside of this fence is considered an outlier from a statistical perspective. Basically, we expand the size of the box by 1.5 times to give us a boundary for our data set. The boundary is found with the following equations:

Lower bound = Q1 - 1.5· IQR Upper bound = Q3 + 1.5· IQR

Let's look at this data set from before.

Q1 = 17.5 Median = 25 Q3 = 34.5

7 11 17 | 18 21 24 $\boxed{25}$ 29 32 33 | 36 36 38

Q1 = 17.5 Q3 = 34.5 IQR = 34.5 – 17.5 = 17

Lower bound = 17.5 – 1.5(17) = -8 Upper bound = 34.5 + 1.5(17) = 60

There are no outliers in the data set.

MODIFIED BOXPLOTS

Here are the annual deaths by tornado from 1990 to 2000. Are there any outliers in the data set? 53 39 39 33 69 30 25 67 130 94 40

<u>Sort the data and find the five - number summary:</u> (see if you get the same thing!)

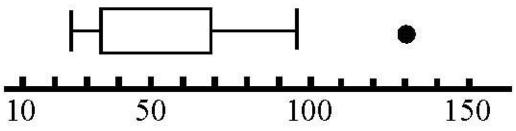
Min = 25
Q1 = 33
Median = 40
Q3 = 69
Max = 130

IQR = 69 – 33 = 36

<u>Conduct the outlier test</u>: Lower boundary = 33 – 1.5 · 36 = -21
Upper boundary = 69 + 1.5 · 36 = 123
Therefore, 130 deaths is an outlier.

<u>Create a Modified Boxplot</u> – for a modified boxplot, the outliers become dots and the whisker ends at the next data value in the ordered list. Since 130 is an outlier, the next highest value (94) becomes the end of the whisker.

Number of deaths by tornado (1990 – 2000)



HISTOGRAMS – before we create a histogram (a bar chart for quantitative data), we need to slice up or group the data into bins of equal width. Let's look at the free throw data again.

12 32 42 50 39 38 27 34 42 35 48 36 13 15 23 16 25 43

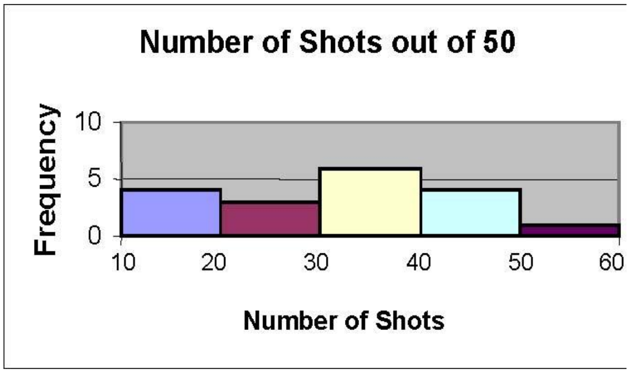
Possible groupings of the data would be 10-15, 15-20… (counting by fives) OR 10-20, 20-30… (counting by tens). Notice how groupings overlap. The first group would be 10 to less than 20. The second group would be 20 to less than 30.

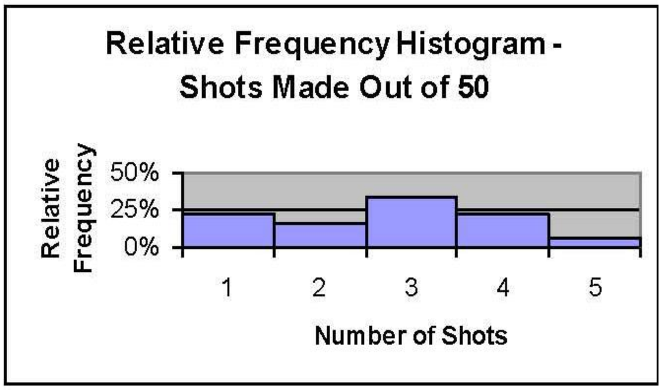
Here are the steps to creating a histogram:

1) Create a frequency chart. A frequency is just a count of the number of observations in each bin or group. Relative frequency is simply the percent in each group.

Group/Bin/Class	Tally	Frequency	Relative frequency
10-20	| | | |	4	22.2%
20-30	| | |	3	16.7%
30-40	| | | | |	6	33.3%
40-50	| | | |	4	22.2%
50-60	|	1	5.6%
TOTAL		18	

2) Create a histogram. You can create a frequency histogram or a relative frequency histogram. The groups are always the x-axis. Frequency or relative frequency the y-axis. **NOTE: there are no gaps between the bars in a histogram!**





You can see that there is not any difference in the shapes/peaks of the histograms when comparing a frequency to a relative frequency. Relative frequency is often used when comparing two different data sets that have different totals.

We create histograms using small sets of data to illustrate the technique. They are more useful to help describe large data sets. Technology is then used to create the histogram.

Always label your axes and give a title on your histogram!

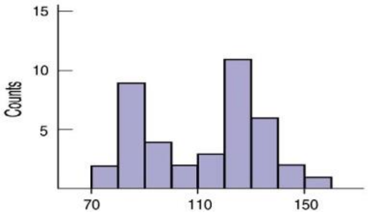
DESCRIBING A DATA SET FROM A PICTURE

You've heard the phrase a picture is worth a thousand words. Well, it is certainly true in statistics. Pictures of quantitative data include: stem and leaf plots, box and whisker plots, and histograms, which we've just reviewed. When you describe the picture of the data set, you should always address three things: shape, center, and spread.

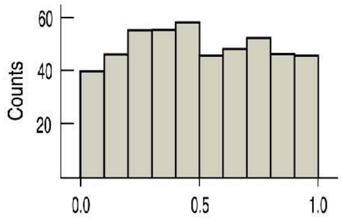
Shape:

1. Does the histogram have a single, central hump or several separated bumps?
    - Humps in a histogram are called modes.
    - A histogram with one main peak is dubbed unimodal; histograms with two peaks are bimodal; histograms with three or more peaks are called multimodal. Histograms with no peaks are called uniform.

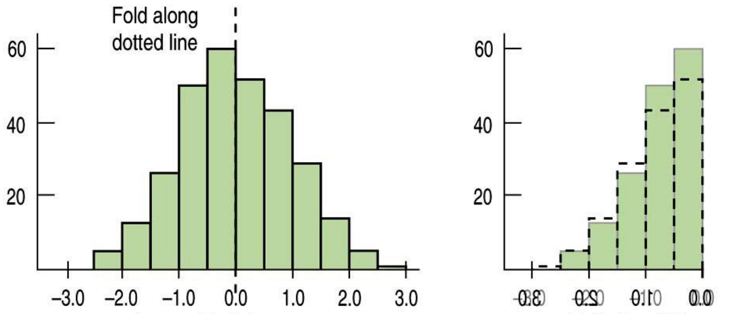
BIMODAL



UNIFORM



2. Is the histogram symmetric?
    - If you can fold the histogram along a vertical line through the middle and have the edges match closely, the histogram is symmetric.



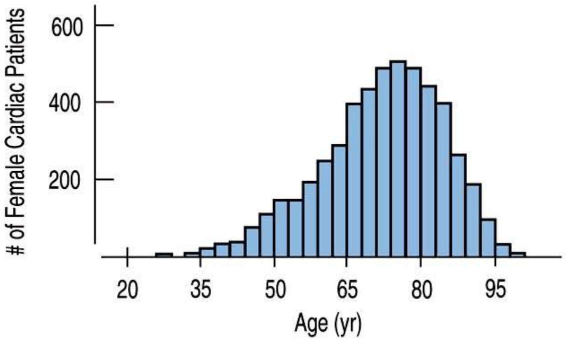
A symmetric histogram …

…can fold in the middle so that the two sides almost match.

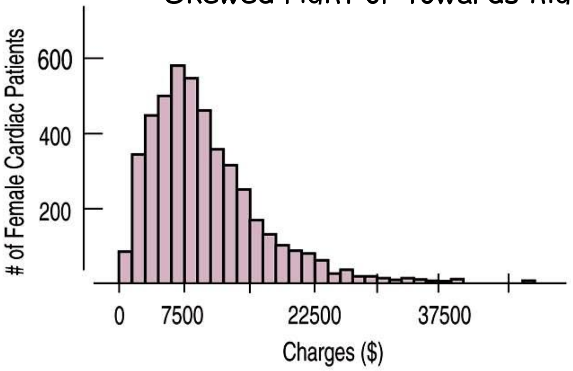
The distribution won't be perfecty symmetric like you see in geometric shapes. Perfection doesn't happen very often in statistics!

If the histogram is not symmetric, we decide whether it is **skewed**. A skewed histogram has a 'tail' to one side or the other. The tail is the direction of the skewness.

Skewed left or towards low values

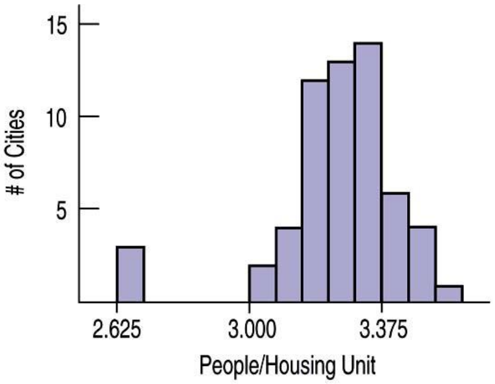


Skewed right or towards high values



3. Do any unusual features stick out?

- Sometimes it's the unusual features that tell us something interesting or exciting about the data.
- You should always mention any stragglers, or outliers, that stand off away from the body of the distribution.
- Are there any gaps in the distribution? If so, we might have data from more than one group.



This histogram has potential outliers. The bar to the left, which represents three cities, stands off by itself. These cities would warrant a closer look.

We call these potential outliers, because a mathematic test is needed to determine whether they are truly outliers.

<u>**Center**</u>**:** If you had to pick a single number to describe all the data what would you pick? It's easy to find the center when a histogram is unimodal and symmetric—it's right in the middle. On the other hand, it's not so easy to find the center of a skewed histogram or a histogram with more than one mode. If you are only given a picture of the data, we will "eyeball" the center of the distribution. If you have the actual data set, you can calculate the measures of center (median and mean).

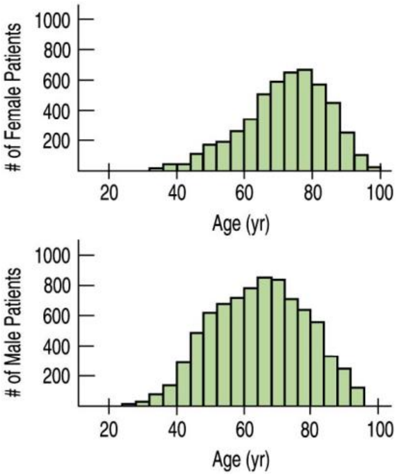
Spread: Variation and spread are very important concepts in Statistics and we will learn much more about them next year. In looking at a histogram or data set, for now you should look at the range and the IQR. For a histogram, you can note the range of the data (citing the minimum and maximum) and give a feel for the range of the middle 50% of the data as well. You are looking to see if the data are tightly clustered together or more spread out.

COMPARING DISTRIBUTIONS

When we compare distributions, it is first important to put them on the same scale! That way you can easily compare the centers and spread visually. Then address the big three: shape, center, and spread. Always answer in complete sentences!

Example: Compare these histograms of ages for female and male heart attack patients.

The age distribution for females appears skewed right while the distribution for men seems more symmetric. The center for female patients appears to be about 72 to 76 years of age compared to 64 to 68 for men. The ranges of the distributions are about the same, although the minimum for men appears to be about 24 years of age compared to 32 for women.



Note: Shape, center, and spread were all addressed!

Name: ______________________________________

Practice problems

1. For a class project, a student gathered data from her classmates. She made the request "Tomorrow morning, find out how many minutes it takes you to get ready – from the time you get up until the time you leave your house for school". Here are the responses:

47 28 78 47 58 93 34 76 35 72 45 53 23 43
75 27 23 87 33 43 25 35 49 35 48 37 28

A. Create a stem and leaf plot of the data:

B. Give the five-number summary for the data set and create a box and whisker plot. You need not check for outliers.

C. Now, answer the following questions:

a. How many students responded to the question?

b. What percent of student took longer than 60 minutes to get ready?

c. What are the mean and median times to get ready?

Name: ______________________________________

D. Complete the frequency table below and use it to create a histogram of the data.

Here’s the data list for you again:
47 28 78 47 58 93 34 76 35 72 45 53 23 40
75 27 23 87 33 43 25 35 49 35 48 37 28

Bin/Group *	Tally	Frequency	Relative Frequency
20-30			
30-40			
40-50			
50-60			
60-70			
70-80			
80-90			
90-100			

* refer to page 7 if you are unsure about bins

Name: ____

2. The following stem plot displays the distribution of the running times (minutes) for movies directed by Alfred Hitchcock.

8	1
9	
10	1 3 3 5 8 8 8
11	1 3 6 6 7 9
12	0 0 0 6 8
13	0 2 6

Key: 12|6 means 126 minutes.

a. What is the median running time? The mean running time?

b. Give the five-number summary. Conduct an outlier test. Create a modified boxplot.

3. In 1961 Roger Maris made baseball headlines by hitting 61 homeruns, breaking a famous record held by Babe Ruth. Here are Maris' home run totals for his 10 seasons in the American League. Would you consider his record-setting year to be an outlier?

8, 12, 14, 16, 23, 26, 28, 33, 39, 61

Name: ____

4. For a recent science project, you collected data regarding the distribution of aquatic life at various depths. Create and describe the histogram (shape, center, spread).

Depth (m)	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10
# fish	10	19	23	47	68	51	43	21	15	8

Name: ____

5. A Consumer Report on peanut butter reported the following scores for various brands. Scores are based on quality, taste, nutritional value, and price. Create boxplots for the two types of peanut butter and compare them. Note: when comparing boxplots, you should talk about the centers of the data sets, the ranges (both the range and the range of the middle 50%, represented by the IQR) and the minimum and maximum values. Answer in complete sentence form! Bulleted lists are ok.

Creamy:	56	44	62	36	39	53	50	65	45	22	40	56
	68	41	30	40	50	56	30					
Crunchy:	62	53	75	42	47	40	34	62	52	50	34	42
	36	42	36	75	80	47	56	62				

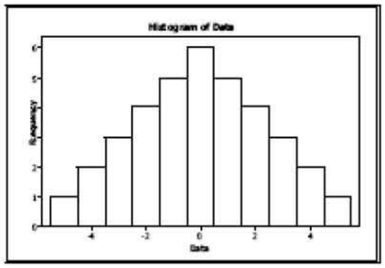
6. Create one data set that reflects all the following characteristics:

- The median of a set of 20 numbers is 24
- The range is 42
- To the nearest whole number, the mean is 24
- No more than three numbers are the same.

Show your strategy.

7. The mean age of 12 of the members attending a mathematics department faculty meeting is 37. Mr. Myers, who is 50, arrives late. What is the average of all 13 members?

8. Indicate the shape (symmetric, skewed left, skewed right) of each distribution.


Histogram of Data
Frequency
Data