Rising AP Stats Students

Hello AP Stats Class 2023-24!

This year, we will be watching videos and taking notes as homework and doing problems in class. This process will allow you to do the "easy stuff" on your own and work through the harder understanding in class when I am available to help you.

The first unit is relatively easy, and you will be doing the first four lessons over the summer. You will need to watch the videos, take notes, and do the homework. We will review this material the first few days of class and then take an assessment. If you find that you need more explanation than the review during the first few days, you will need to come for extra help.

Your summer math packet lists the videos. I will also send them to you via Renweb so that you can click on the video links more easily.

This is how you should go about doing your packet. It will probably take 8 – 12 hours to complete.

1. Watch the Notes 1 video and fill in the guided notes provided. You are welcome to use color and to write extra information in the margins.
2. Complete the homework which is just behind the guided notes.
3. Highlight anything that causes confusion so that you can ask questions once you return to school. Write in a specific question in the margin because you will forget why it confused you once you get back to school.
4. Continue this process for Lessons 2, 3 & 4.
5. You do not need to do Lesson 5 or 6
6. You will get eight homework credits for this process.

I hope you have a wonderful summer!

Sincerely,

Caren Sturgill

Unit 1: Exploring One Variable Data

Guided Student Notes with Video Lessons

If you flip your classroom, have distance learning needs, or have students that just need remediation or re-teaching, you can use the video lessons and the set of Guided Notes to help your AP Statistics students master the concepts for Unit 1 Exploring One Variable Data.

Using the Resource:
The purchaser of this resource is granted permission to post the video and the guided notes to a private password-protected learning management system for use by students (such as Google Classroom or Canvas).

Students will use the set of guided notes to follow along with the video lesson.

Video Lesson Links:

Lesson	YouTube Link
Notes 1 Representing Categorical Variables with Graphs *(20:04)*	https://youtu.be/Ct9kjnPrEh0
Notes 2 Representing Quantitative Variable with Graphs *(43:40)*	https://youtu.be/__hr8wTUEWk
Notes 3 Describing and Summarizing Quantitative Variables *(1:00:15)*	https://youtu.be/qHPgnPkDuas
Notes 4 Comparing Distributions *(35:34)*	https://youtu.be/EmU6ge3ICzg
Notes 5 Z-Scores and the Empirical Rule *(38:22)*	https://youtu.be/BDVI_0NYesw
Notes 6 The Standard Normal Distribution *(42:16)*	https://youtu.be/Vfsx_83P7o0

Unit 1 – Exploring One-Variable Data
Notes 1 – Representing Categorical Variables with Graphs

Statistics is ___________________________.

Data contains information about a group of *individuals*. The information is organized using *variables*.

Individuals are objects described by a set of data. Individuals may be people but may be animals or inanimate objects.

Variables are characteristics of individuals. A variable may take on different values of different variables. Variables can be split into two types: categorical or quantitative.

Categorical variables place individuals into specific groups.

Quantitative variables takes on numerical values for which it makes sense to do arithmetic operations like adding and averaging. Quantitative variables fall into two categories: discrete and continuous.

VARIABLES: CHARACTERISTICS WE MEASURE/OBSERVE

CATEGORICAL
-
-
-
-
-

QUANTITATIVE

DISCRETE
-
-
-

CONTINUOUS
-
-
-

**Be careful, just because it is a number DOES NOT automatically make it quantitative. Can you think of an example of a number that is categorical?

Discrete variables are numerical values where counting makes sense; in other words, decimals would not be an appropriate way to record the data.

Continuous variables are numerical values where decimals are appropriate; it usually involves some form of measuring.

**Note: The difference between discrete and continuous isn't always clear: For instance, is age continuous or discrete? You use whole numbers to describe it, but it can be measured with a decimal.

**If it is not obvious if it is discrete or continuous, like our age example, it depends on how you use the data. For age, we rarely say "I'm 16.4 years old", so we treat age like a discrete variable.

Representing Categorical Data: Tables

One of the easiest ways to display categorical data is with a table. Here is a list of the first 10 US presidents, their political party and their state of birth.

George Washington	Federalist	Virginia
John Adams	Federalist	Massachusetts
Thomas Jefferson	Democratic-Republican	Virginia
James Madison	Democratic-Republican	Virginia
James Monroe	Democratic-Republican	Virginia
John Quincy Adams	Democratic-Republican	Massachusetts
Andrew Jackson	Democrat	South Carolina
Martin Van Buren	Democrat	New York
William H. Harrison	Whig	Virginia
John Tyler	Whig	Virginia

Let's start with a few **one-way tables**:

Category	*Count*	*Relative Count*
Federalist		
Democratic-Republican		
Democrat		
Whig		

Category	*Frequency*	*Relative Frequency*
Virginia		
New York		
South Carolina		
Massachusetts		

If we wanted to display two categorical variables at a time, we would make a **two way table**:

		Political Party				
		Federalist	Democratic-Republican	Democrat	Whig	Total
State of Birth	Virginia					
	New York					
	South Carolina					
	Massachusetts					
	Total					

To better visualize the data, we can also make various graphs from our data. We want to visualize the graphs to get a better idea of the *distribution*.

The **distribution** of a variable tells us what values the variable takes and how often it takes these values.

Bar Graphs

Bar Graphs have the following important characteristics:

- Label each axis clearly (this goes for all graphs).
- The x-axis will contain the categorical variable and the y-axis will display the counts (or percentages).
- Each category has its own bar and the bars CANNOT touch.
- Order is not important when creating the x-axis.

Create a bar graph showing the distribution of the four political parties of the first 10 US presidents.

Questions we can answer from a bar graph:

1) Of the first 10 presidents, what political party was the most affiliated with?

2) What political party was the least affiliated with?

3) What are the individuals in this data set? What are the variables of the data set?

*We can also use Pie Graphs for categorical data, but those are not covered on the AP exam.

*To graph two way tables, we will explore segmented bars graphs and mosaic plots in Unit 2.

Unit 1 – Exploring One Variable Statistics
HW 1 – Categorical Data

Name: ________________________

1) Below is a list of unorganized data from the 2020 election. This gives the state, their geographical location, and the voting majority.

State	*Location*	*Electoral College Vote in 2020*
AL	South	Republican
AK	West	Republican
AZ	West	Democrat
AR	South	Republican
CA	West	Democrat
CO	West	Democrat
CT	Northeast	Democrat
DE	South	Democrat
FL	South	Republican
GA	South	Democrat
HI	West	Democrat
ID	West	Republican
IL	Midwest	Democrat
IN	Midwest	Republican
IA	Midwest	Republican
KS	Midwest	Republican
KY	South	Republican
LA	South	Republican
ME	Northeast	Democrat*
MD	South	Democrat
MA	Northeast	Democrat
MI	Midwest	Democrat
MN	Midwest	Democrat
MS	South	Republican
MO	Midwest	Republican
MT	West	Republican
NE	Midwest	Republican*
NV	West	Democrat
NH	Northeast	Democrat
NJ	Northeast	Democrat
NM	West	Democrat
NY	Northeast	Democrat
NC	South	Republican
ND	Midwest	Republican
OH	Midwest	Republican
OK	South	Republican
OR	West	Democrat
PA	Northeast	Democrat
RI	Northeast	Democrat
SC	South	Republican
SD	Midwest	Republican
TN	South	Republican
TX	South	Republican
UT	West	Republican
VT	Northeast	Democrat
VA	South	Democrat
WA	West	Democrat
WV	South	Republican
WI	Midwest	Democrat
WY	West	Republican

*Split Votes

a) Construct a bar graph of the amount of states in each region of the US.

b) Construct a bar graph of the amount of states who voted either republican or democrat in the 2020 election.

c) Create a two way table showing the distribution of region and political party.

2) You measure the age, marital status and earned income of a random sample of 1463 women. The number and type of variables you have measured is

(A) 1463; all quantitative.
(B) four; two categorical and two quantitative.
(C) four; one categorical and three quantitative.
(D) three; two categorical and one quantitative.
(E) three; one categorical and two quantitative.

3) As part of a survey of college students, a researcher is interested in the variable class standing. She records a 1 if the student is a freshman, a 2 if the student is a sophomore, a 3 if the student is a junior, and a 4 if the student is a senior. The variable class standing is

(A) Categorical
(B) Quantitative
(C) Quantitatively Categorical
(D) All of the above

Unit 1 – Exploring One-Variable Data
Notes 2 – Representing Quantitative Variables with Graphs

Graphing Quantitative Variables: Histograms

The table below shows the average income level for a household in the respective state.

State	Average Income	State	Average Income
Alabama	$78,871.29	Montana	$81,638.21
Alaska	$100,086.45	Nebraska	$93,878.49
Arizona	$96,364.72	Nevada	$92,457.31
Arkansas	$77,637.36	New Hampshire	$114,680.66
California	$111,632.93	New Jersey	$119,305.58
Colorado	$107,936.13	New Mexico	$71,531.93
Connecticut	$117,303.22	New York	$105,571.94
Delaware	$102,639.68	North Carolina	$84,727.74
Florida	$85,581.31	North Dakota	$90,647.13
Georgia	$84,224.69	Ohio	$90,396.68
Hawaii	$105,978.87	Oklahoma	$84,974.15
Idaho	$91,721.86	Oregon	$107,795.68
Illinois	$103,958.50	Pennsylvania	$99,681.52
Indiana	$87,139.09	Rhode Island	$98,980.03
Iowa	$86,536.71	South Carolina	$83,649.63
Kansas	$96,719.61	South Dakota	$83,574.96
Kentucky	$87,474.11	Tennessee	$81,911.63
Louisiana	$78,124.94	Texas	$98,362.04
Maine	$84,312.62	Utah	$112,799.70
Maryland	$125,053.40	Vermont	$95,683.34
Massachusetts	$127,460.73	Virginia	$114,127.44
Michigan	$92,073.46	Washington	$110,680.46
Minnesota	$104,195.36	West Virginia	$72,857.68
Mississippi	$65,648.61	Wisconsin	$90,695.26
Missouri	$83,396.09	Wyoming	$84,500.23

To make a histogram, we need to put the data into "bins" (even intervals that capture our data). We will do this first by hand by counting how many data scores are in each bin.

Lowest Income: ____________ Highest Income: _____________ Bin Width: ________________

Interval	Count
$65,000 - <$77,600	
$77,600 - <$90,200	
$90,200 - <$102,800	
$102,800 - <$115,400	
$115,400 - <$128,000	

To make the histogram:

- Draw rectangles for each bin, and their height extends up to the highest count
- Bars MUST touch
- Label with each boundary with the "boarder value"

You can also use a graphing calculator to create a histogram, which we will do a little later!

Graphing Quantitative Variables: Stem and Leaf

Gather Data: Determine your pulse rate. Hold the fingers of one hand on the artery on the inside of your wrist. Count the number of pulse beats in 20 seconds and then triple it. You will do this three times. Find the average of your tripled rate:

Pulse rate 1:	Pulse rate 1 tripled:
Pulse rate 2:	Pulse rate 2 tripled:
Pulse rate 3:	Pulse rate 3 tripled:
	Average:

Come to the board and write your average pulse along with your gender (M or F). When everyone has done so, write down all the class's data in the chart below.

Gender	Average Pulse Rate	Gender	Average Pulse Rate	Gender	Average Plus Rate

<u>Stemplots</u> (sometimes called a stem and leaf plot) are an alternate way of illustrating data using a semi-graph. It is similar to a histogram but, unlike a histogram, the data isn't lost. If the data has two digits, the stem is the first digit and the leaf is the second. If the data has 3 digits, the stem is the first two digits and the leaf is the 3rd.

Create a stemplot for the average pulse rates you collected (ignore the M/F for this graph). ALWAYS add a key to your graph when you are done.

Stem	Leaf

Leaves must go in order.

Line up the leaves as best you can!

Think of this as a sideways histogram.

<u>Back-to-back stemplots</u> are created when you can separate the data into two categories. Let's create a back-to-back stemplot using the same data, but split the sides into males and females. The stem is still the tens digit and the leaf is the ones digit. DON'T FORGET A KEY!

Male Leaf	Stem	Female Leaf

*Did you know that male hearts are typically bigger than female hearts, resulting in the male heart needing to beat at a lower rate per minute.

Split Stem and Leaf

The last type of stem and leaf plot is called split stem and leaf. When you have too many data values in a single stem, it can be helpful to split the stem; the same way we would create more bins on a histogram if our bin width resulted in a skyscraper.

Stem	Leaf

→

Stem	Leaf

Here are the scores from last year on the Unit 1 Test:

80	88	78	93	69	80	92	90	88	84	82	92	88	62	84
75	74	96	88	88	90	94	92	79						

Create a split stem and leaf plot of the data.

Graphing Quantitative Variables: Dot plot

A **dotplot** is a very simple type of graph that involves plotting the data values, with dots, above the corresponding values on a number line.

How to construct a dotplot:

Step 1: Label your axis and title your graph. Draw a horizontal line and label it with the variable.
Step 2: Scale the axis based on the values of the variable.
Step 3: Mark a dot above the number on the horizontal axis corresponding to each data value.

Here are the quiz scores (out of 50) on a past quiz. Construct a dotplot.

36 42 42 42 42 41 38 41 40 42 43 44 45 45 46
44 43 48 43 42 43 41 50 38 38 40 44

Graphing Quantitative Variables: Cumulative Relative Frequency Graphs

Cumulative relative frequency graphs (ogives) display percentiles.

A **percentile** will tell you what percent of data falls below a value.

You first must make a table of the cumulative relative frequencies in order to graph it. This can be done by finding the relative frequencies and then find the cumulative percentages.

Here is a frequency table summarizing the ages of the 46 U.S. presidents when they took office. (Biden was the 46th president and the oldest sworn in at 78 years old. Teddy Roosevelt, the 26th president, was the youngest at 42 years old.)

First Way: Find the relative frequencies and then find the cumulative percentages.

Age	40-<45	45-<50	50-<55	55-<60	60-<65	65-<70	70-<75	75-<80
Frequency	2	7	13	12	7	3	1	1
Relative Frequency								
Cumulative Relative Frequency								

Age	Cumu. Rel. Frequ
40-<45	
45-<50	
50-<55	
55-<60	
60-<65	
65-<70	
70-<75	
75-<80	

Creating an ogive:

- The first interval is 40-<45, with a cumulative relative frequency of 4%. Since 40 is the starting interval, that starts at 0 and 4% is graphed at 45.
- The next interval is 45-<50, with a cumulative relative frequency of 20%. At 45, 4% is graph, so at 50, the 20% will be graphed.
- This pattern continues until 100% is graphed at 80.

Questions we can answer with an ogive:

1) At approximately what age is the 30th percentile? What does this mean in the context of the problem?

2) Richard Nixon was the 37th president and was 56 years old when he was inaugurated. Approximately what percentile is this and what does it mean in the context of the problem?

Unit 1 – Exploring One Variable Statistics Name: ______________________
HW 2 – Quantitative Data

1) The following are the temperature highs for an Illinois town in August, 2021.

78 75 83 85 87 91 78 91 85 84 80 79 83 85 86
80 88 94 89 86 88 84 81 85 83 80 79 87 88 91
95

a) Create a dot plot of this data.

b) Create a histogram of this data (include your table!).

c) Create a stem and leaf plot for this data.

d) Which graph do you think displays the data the best? Why?

2) 30 people were asked how many miles, to the nearest mile, they commute to work each day:

2	5	7	3	2	10	18	15	20	7	10	18	5	12	13
12	4	5	10	7	10	12	7	3	4	11	13	14	5	6

a) Create a cumulative relative frequency graph (include your table!)

b) What percent of people traveled less than 12 miles?

c) What percent of people traveled over 15 miles?

d) What percent of people traveled between 10 and 15 miles?

Unit 1 – Exploring One-Variable Data
Notes 3 – Describing and Summarizing Quantitative Variables

Describing Distributions

In AP Statistics, to describe a distribution, we use the acronym SOCS:

S:

O:

C:

S:

After we graph a quantitative variable, describing what we see helps identify patterns, and answer questions we might have about the data.

**Note: We do NOT use SOCS to describe categorical variables. Why?

Categories can be placed in any order on the x-axis, so shape wouldn't make sense. Same thing for center and spread, we cannot find the average zip code (for example)

Shape

Once the distribution is graphed, the first thing we identify is the shape of the graph.

Unimodal	Bimodal	Uniform
Symmetric	Left Skewed	Right Skewed

You might only use one word to describe the shape or two words might be appropriate:

Draw a symmetric, bimodal distribution.	Draw a unimodal, right skewed distribution.

Outliers

An **outlier** is an individual piece of data that falls outside the overall pattern of the distribution.

When an outlier occurs, we must find out *why* it occurs. Many times it occurs because of a mistake. It may have been written down incorrectly or input into a computer or calculator incorrectly. That is why we always PLOT OUR DATA first because our eye will usually tell us when an outlier exists. Outliers can be eliminated from the data if there is a good reason.

It isn't that easy to determine whether or not to eliminate an outlier and different statisticians may handle the situation differently. But with outliers, there are several rules:

1) Plot the data
2) Determine whether there is an outlier. If so, be sure that the data is accurate.
3) Determine if there is a valid reason that the outlier can be removed. If there isn't, don't remove it just identify it.

Center

There are three measures of center in statistics: ________________, ________________, and ______________.

Let's gather some data to analyze the measures of center: How many siblings (step and half included, because they are!) do you have? Copy down the class data below.

I'm sure you know how to find the "average" (technically, average is a generic term for a measure of center. When you say average, you usually are referring to the **mean**. We do not use the term "average" in AP Statistics), but here is the formula and notation:

______: mean of a sample

______: each individual observation

______: number of observations

______: capital sigma = "add up"

Mean number of siblings in our class: ________________

The reason we do not always use the mean to describe the center has to do with the inclusion of outliers in a data set.

For instance, suppose a student took 7 exams and had these scores: 90, 92, 94, 98, 86, 88, 0.

- There is very clearly an outlier: the zero score. There might be legitimate reasons to eliminate the outlier. Why?

- If we cannot eliminate the outlier, we calculate the mean to be: ______________

- If we did eliminate the outlier, we calculate the mean to be: _______________

- What is the outlier doing to the mean? ____

The mean is called **non-resistant**. This means that the mean is strongly influenced by extreme values. If you do have extreme values in a data set (these extreme values may or may not be outliers), the mean is not reliable as a measure of center.

Luckily, we have another measure of center that is **resistant**: meaning that if there are extreme values in a data set, this measure of center will not be as affected by it.

The **median** is found by ordering the data and then finding the middle value in that list.

For example, order the number of siblings from lowest to highest:

____

Mean: ___________ Median: _____________ How do these compare? ____________________________

Let's take a look back at our exam score example: 0, 86, 88, 90, 92, 94, 98

Mean: ___________ Median: _____________ How do these compare? ____________________________

**Note: This isn't saying that the median is always a better measure of center, but whenever there are extreme values or outliers, the median is the better measure of center.

**Note: The median does not have a formula or special math symbol. (Use Med for short)

The last measure of center is perhaps the most useless: the mode. It finds the most occurring value in a data set. The idea of having this as a measure of center comes from having a symmetric, unimodal distribution, where the most occurring value happens in the middle. However, as we have seen, there are many shapes to different distributions and the most occurring doesn't always occur in the middle.

78, 79. 80, 80, 85, 91, 99	77, 77, 78, 79, 85, 100	73, 73, 88, 89, 90, 93, 98
Mean: ______ Median: ______ Mode: ______ Shape: ______	Mean: ______ Median: ______ Mode: ______ Shape: ______	Mean: ______ Median: ______ Mode: ______ Shape: ______

To summarize our mini exploration above:

Unimodal, Symmetric	*Left Skewed*	*Right Skewed*

Spread

There are three common measures of spread in statistics: ____________, ______________, and __________.

The **range** is the difference: ___________ - _____________= ________. The range is a number, not an interval!

The **standard deviation** is the average deviation of an observation from the mean of the data set. To understand where the formula comes from, let's walk through an example where we find the standard deviation by hand (which we will rarely have to do):

Example: A group of elementary school children was asked how many pets they have. Here are their responses, arranged from highest to lowest:

1 3 4 4 4 5 7 8 9

Calculate the mean: __________

Observations	1	3	4	4	4	5	7	8	9	
Deviations										
Square Deviations										

Now we will compute the average squared deviation – sort of. Instead of dividing by the total number of observations, we will divide by n-1 (the reason for this involves some calculus, so just trust me on this one).

This value, __________, is called the **variance** (S_x^2). This is the average squared deviation of the number of pets this group of elementary children have.

Because this is the average squared deviation, our units are in "squared pets", and that is pretty useless to us.

We have to take the square root to get back to the correct units – pets.

Variance = ___________ → _______________ → Standard Deviation (S_x) = _____________

The variance formula is:

	______: symbol for variance ______: mean of a sample ______: each individual observation ______: number of observations ______: capital sigma = "add up" ______: symbol for standard deviation

The standard deviation formula is very similar, and you can see the difference is just undoing the "square" by square rooting both sides of the formula:

Interpreting the standard deviation in context:

- The number of pets typically varies from the mean by______________.
- It measures the typical distance of the values in a distribution from the mean.

Standard deviation should only be used as your measure of spread when the mean is your chosen measure of center.

The standard deviation has the following additional properties

- The standard deviation is always _______.
- The standard deviation is 0 when _______________________________________.
- The standard deviation has the same units of measure as the original variable measured.
- The standard deviation is __________________, meaning that a few outliers will make it _______________.
- The greater the standard deviation, the greater the variation of the distribution.

The last measure of spread is **IQR**. IQR stands for *Inter-Quartile Range* and uses percentiles to describe the spread of the distribution.

Recall: **Percentiles** measure the percent of the observations that fall below a value.

If the median is the middle of a data set (50% below and 50% above), what percentile is it? ____________

Here are some other percentiles:

- 0th percentile: ___________ - lowest value in a data set.
- 25th percentile: ______________ or ______ - 25% of the data is below this value.
- 50th percentile: ______________ - middle value in a data set.
- 75th percentile: ______________ or ______ - 75% of the data is below this value.
- 100th percentile: __________ - highest value in a data set.

These 5 numbers make up what is called the Five Number Summary (clever, right?). To calculate these values by hand, place the observations in order and find the median. Then, find the middle value to the left and right of your median to identify your quartiles.

Example – Here is the data from a previous AP statistics class. They were asked how many hours of sleep they got before the first day of school.

2 4.5 5 5 6 6 6 6.5 7 7 7 7 7 7 7 7.5 8 8 8 8 8 8.5

Min : ________ Q1: ________ Median: ________ Q3: ________ Max: ________

The IQR is the range of the inter-quartiles: ____________ Just like range, this value is a SINGLE value. It is not defined as an interval.

What is the IQR of the amount of hours of sleep that group of AP Stat students got? ______________.

Boxplots

The five number summary can also be used to create another graph: the **boxplot**.

Example: Using the five number summary above, create a boxplot of the amount of hours of sleep the AP stat students got.

Using the five number summary and the IQR, we also have a numerical way of determining if a point is an outlier. For simplicity's sake, we call this the **outlier test**.

Steps:

1. Find the 5-number summary
2. Find the IQR
3. Compute Q1 - (1.5 * IQR). Any data below that number is an outlier.
4. Compute Q3 + (1.5 * IQR). Any data above that number is an outlier.

Using the sleep hours in the example above, is there an outlier? Show your work.

When we have outliers in our data set, and we cannot throw them out, a **modified boxplot** helps us create a boxplot that will identify the outliers. Using the example above, once we have determined that we have an outlier, we mark it's location with a star (or dot). Then, instead of the whisker extending to the outlier, it extends to the next observation that is within the acceptable range.

Example – create a modified boxplot for the sleep example above.

Shape of Boxplot

Distribution	Boxplot	Key Features
		Symmetric Uniform mean ≈ median
		Symmetric Bimodal mean ≈ median
		Symmetric Unimodal Bell-Shaped mean ≈ median
		Unimodal Left Skewed mean < median
		Unimodal Right Skewed mean > median

Unit 1 – Exploring One Variable Data
HW 3 – Summary Statistics and Boxplots

Name: ______________________________

1) Following is a list of calories and cholesterol in certain fast food burgers.

Company	Fast Food	Calories	Cholesterol (mg)
McDonald's	McLean Deluxe	320	60
Wendy's	Single	340	65
McDonald's	Quarter Pounder	410	85
McDonald's	Big Mac	500	100
Burger King	Hamburger Deluxe	344	43
Wendy's	Big Classic	570	90
Burger King	Whopper	614	90
Hardee's	Big Burger Deluxe	500	70
Burger King	Double Whopper w/Cheese	935	194
Hardee's	Grilled Chicken	310	60
Hardee's	Chicken Fillet	370	55
Wendy's	Grilled Chicken	340	60
Wendy's	Chicken	430	60
Burger King	BK Broiler Chicken	379	53
McDonald's	McChicken	415	42
McDonald's	Chicken McNuggets	270	56
Burger King	Chicken Sandwich	685	82
Kentucky Fried Chicken	Lite n'Crispy	198	60
Kentucky Fried Chicken	Original Recipe	248	90
Kentucky Fried Chicken	Extra Crispy	324	99

a) Complete the chart below for CALORIES.

	Mean	Min	Q1	Med	Q3	Max	Std. Dev
Burgers							
Chicken							

Which food has the greatest calorie variability in the distribution? ______________________

Are there any outliers in the "burger calories" category? Show work to support your answer.

Are there any outliers in the "chicken calories" category? Show work to support your answer.

Draw two modified boxplots comparing the calories for burgers and the calories for chicken.

b) Take out the chicken outlier and recalculate the data below.

	Mean	Min	Q1	Med	Q3	Max	Std. Dev
Chicken							

What values changes the most? Why?

c) Complete the chart below for CHOLESTEROL. Then answer the questions that follow.

	Mean	Min	Q1	Med	Q3	Max	Std. Dev
Burgers							
Chicken							

Which food has the greatest cholesterol variability in the distribution? ____________________

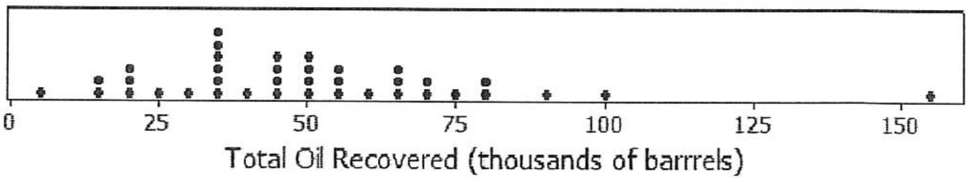
Are there any outliers in the "burger cholesterol" category? Show work to support your answer.

Are there any outliers in the "chicken cholesterol" category? Show work to support your answer.

Draw two modified boxplots comparing the cholesterol for burgers and the cholesterol for chicken.

2) How much oil wells, in a given field, will ultimately produce is key information in deciding whether to drill more wells. Here are the total amounts of oil recovered from 38 wells in the Michigan basin, in thousands of barrels.

3	31	38	50	65	92
13	33	43	50	66	98
15	35	43	53	70	157
19	35	45	56	70	
21	35	46	57	74	
22	37	48	59	80	
25	37	49	63	82	



a) What measures would you use to describe the center and spread of these data? Justify your answer.

b) Find the five number summary for these data.

c) Are there any outliers? Justify your answer.

d) Draw a boxplot of this distribution.

e) For this data, how can you will (without doing any calculations) that the mean of these data is larger than the median?

3) A student studying the sleeping habits of seniors at his school asked 36 randomly-selected seniors how many hours of sleep they got the previous night. The data, rounded to the nearest half-hour, is given in the table below.

8	7.5	9	7.5	9	6	5	9	7.5	7	8	7
6.5	8.5	8	6.5	8.5	6	7	7.5	7	6	8.5	7
7	8	7	7.5	7	6	7	8	7.5	6	7	5

a) Using your calculator, find the mean and median of this data. Without graphing it, what shape do you suspect, give the values of the mean and median?

b) Find and interpret the standard deviation.

c) Suppose 4 more values were added to the data, each exactly equal to the mean. Would this have any impact on the standard deviation? Explain, without using any calculations.

Multiple Choice Practice

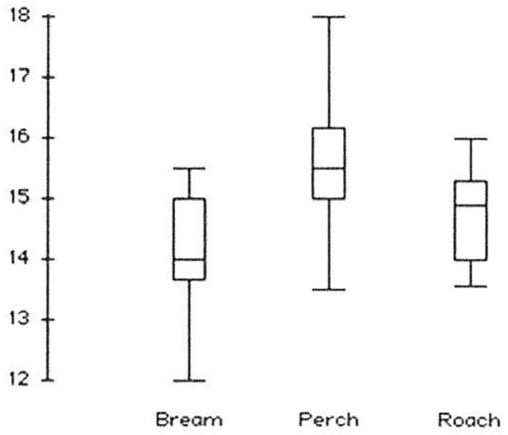
______ 4) In a distribution of 3000 scores, the mean is 78 and the median is 95. One would expect this distribution to be:

(A) skewed to the right
(B) skewed to the left
(C) symmetrical and mound-shaped
(D) symmetrical and uniform
(E) bimodal

_______ 5) The five-number summary for a one-variable dataset is {5, 18, 20, 40, 75}. If you wanted to construct a modified boxplot for the dataset (that is, one that would show outliers if there are any), what would be the maximum possible length of the right side "whisker"?

(A) 35
(B) 33
(C) 5
(D) 55
(E) 73

_______ 6) A study was conducted on the weights of three different species of fish found in a lake in Finland. These three fish (bream, perch and roach) are commercial fish. Their weights are displayed in the boxplots below. Which of the following statements comparing these boxplots is **NOT** correct?



(A) The median weights of the three species differ.

(B) The spread of roach is less than the spread of the other two species.

(C) The distributions of weights are approximately symmetric for all three species.

(D) There are no outliers in weight for the three species.

(E) The variability in the weights for the three species exceeds the variation in the three species' means.

_____ 7) Which of the following are true statements?

I. The standard deviation is the square root of the variance.
II. The standard deviation is zero only when all values are the same.
III. The standard deviation is strongly affected by outliers.

(A) I and II
(B) I and III
(C) II and III
(D) I, II, and III
(E) None of the above gives the complete set of true responses.

_____ 8) In order to rate TV shows, phone surveys are sometimes used. Such a survey might record several variables, some of which are listed below. Which of these variables are categorical?

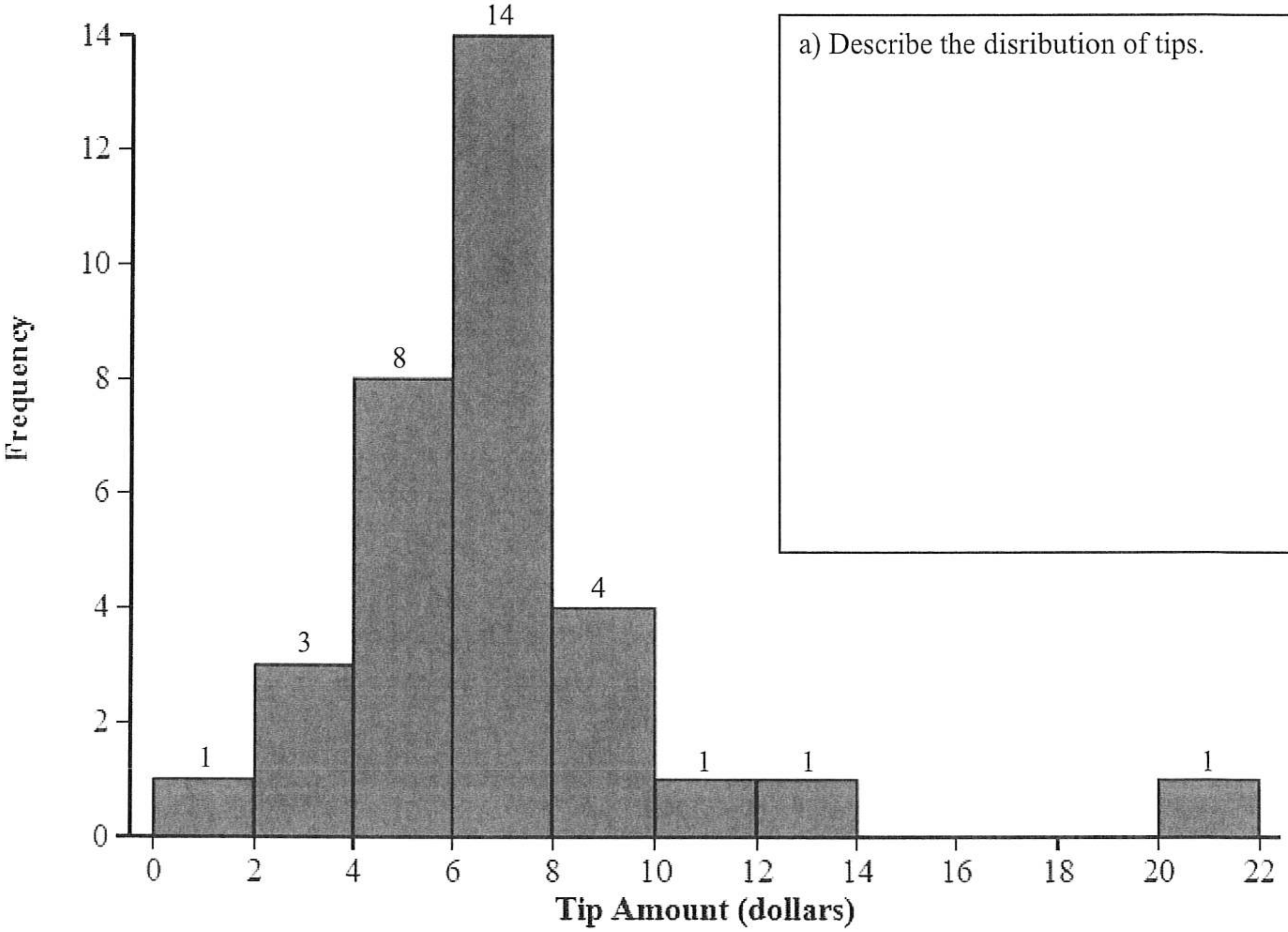
I. The type of show being watched
II. The number of persons watching the show
III. The ages of persons watching the show
IV. The name of the show being watched
V. The number of times the show has been watched in the last month

(A) II, III, and V
(B) I only
(C) I and V
(D) I and IV
(E) None of the above describes the complete set of correct responses

Unit 1 – Exploring One Variable Data
HW 4 – SOCS and Comparing

Name: ______________________

1) Karley works as a Dominoes delivery driver and on Friday night, she recorded how many tips she got from each of her deliveries. The histogram below shows the distribution of the 33 deliveries.



a) Describe the disribution of tips.

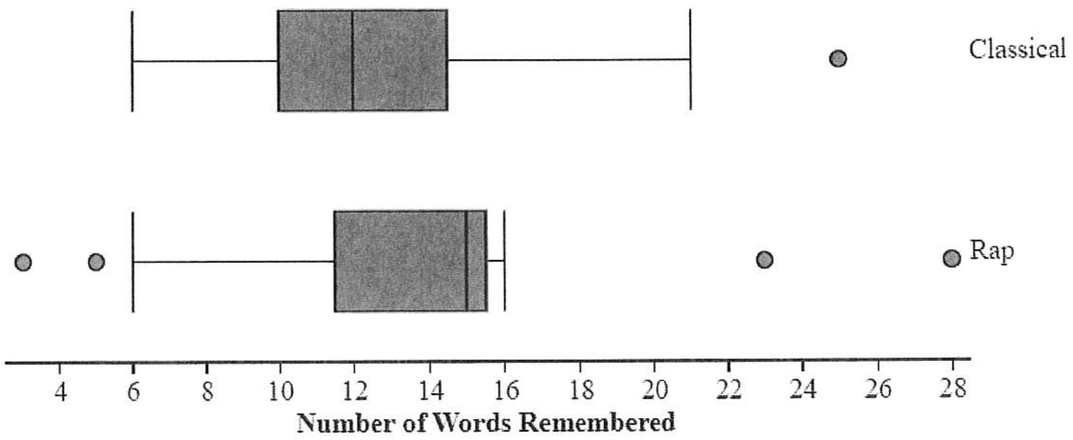
b) One of the tip amounts was $9. If this tip had been $14 instead of $9, describe what effect this would have on the following measures of center. Justify your answer.

Mean:

Median:

c) Describe how you would estimate the median of this distribution.

2) Jordan and Alex want to know if listening to different types of music while studying will help you remember the material better. They randomly assigned a group of students to two different groups: one group studied a list of words while listening to classical music while the second group studied the same list of words while listening to rap music. There was a total of 30 words on the list and each group studied the list while listening to the music for 5 minutes. They were then asked to write down as many words as they could remember. The data is displayed in the boxplots below.

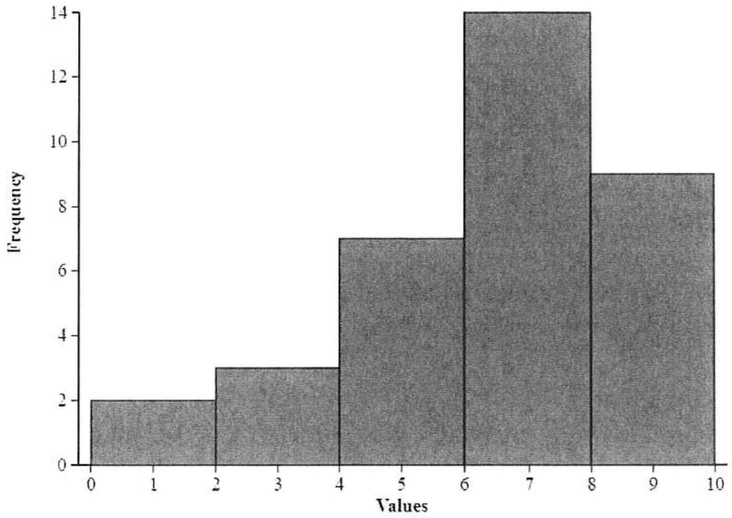


a) Approximate the interquartile range for each set of data. Why is this the appropriate measure of spread to use for these two data sets?

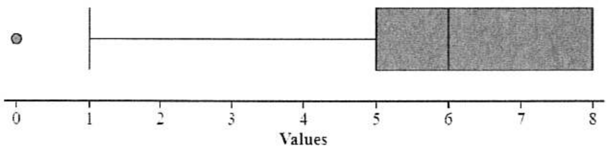
b) Write three sentences comparing the number of words remembered between each of the two groups.

Multiple Choice Practice

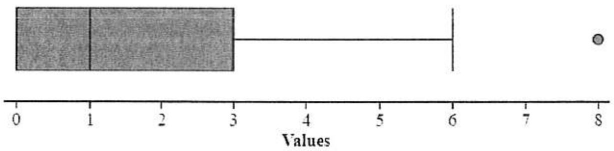
______ 3) The histogram below displays a set of measurements. Which of the boxplots below displays the same set of measurements?



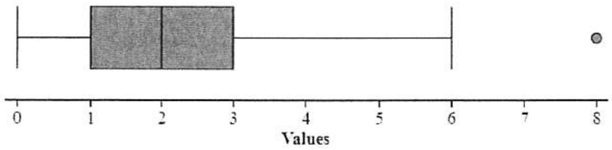
(A)



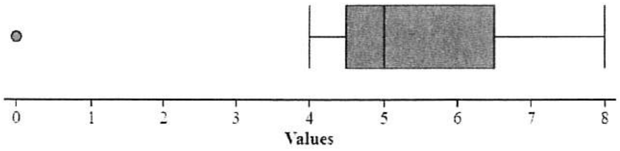
(B)



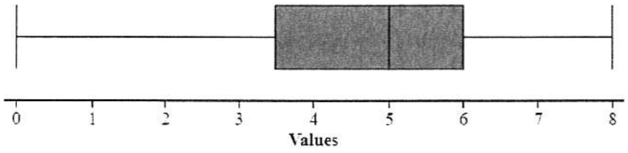
(C)



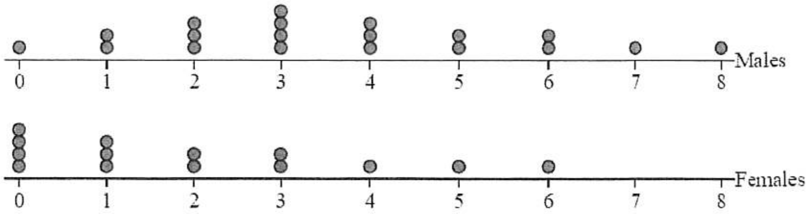
(D)



(E)



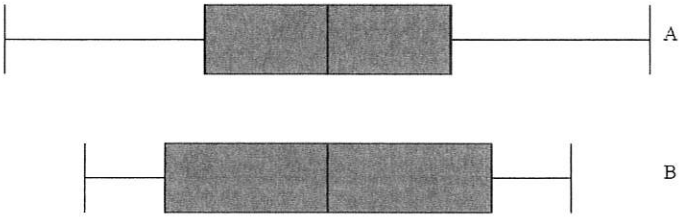
_____ 4) These dotplots for randomly selected male and female students at a particular high school show the number of times per week they eat at fast food restaurants.



Which of the following is a true statement?

(A) One distribution is roughly symmetric; the other is skewed left.
(B) The males' median is larger than the females' mean.
(C) The ranges are both 8.
(D) The standard deviations are equal.
(E) Combining the male and female times into one set of student times will increase the range to 16.

______ 5) The boxplots below summarize two sets of data, A and B. Which of the following must be true?



I. Set A contains more data than Set B.
II. Set A has a larger range than Set B.
III. Set A and Set B have the same median.

(A) I only
(B) III only
(C) I and II only
(D) II and III only
(E) I, II, and III

_______ 6) Which of the following statements is incorrect?

(A) Both dotplots and stemplots can show symmetry, gaps, clusters, and outliers.
(B) Sets with different distribution shapes can have identical boxplots.
(C) Boxplots, dotplots, stemplots, and histograms can all show skewness.
(D) In histograms, relative areas correspond to relative frequencies.
(E) In histograms, frequencies can be determines from relative heights.

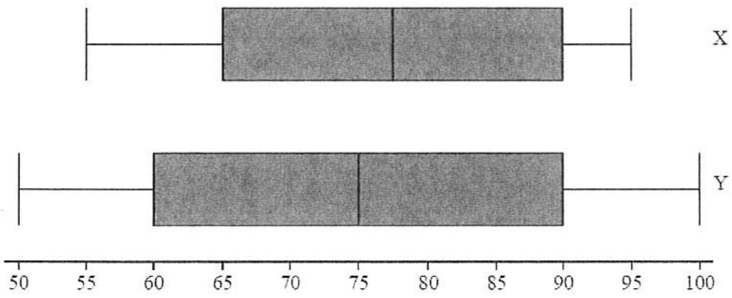
_______ 7) Consider the following test scores:

Class A: 65, 72, 73, 73, 74, 82, 91   Class B: 60, 65, 79, 81, 85, 86, 86

Which of the following is a true statement?

(A) Class A has the greater range, while Class B has the greater standard deviation.
(B) Class B has the greater range, while Class A has the greater standard deviation.
(C) The ranges and standard deviations are the same for both classes.
(D) The ranges are the same, but Class A has the higher standard deviation.
(E) The ranges are the same, but Class B has the higher standard deviation.

_______ 8) The boxplots shown below summarize two sets of data, X and Y.



Which of the following must be true?

(A) 50 is an outlier for Set Y.
(B) Set Y contains more data than Set X.
(C) The IQR of Set X is smaller than the IQR of Set Y.
(D) The range of Set Y is smaller than the range of Set X.
(E) The means of the two data sets are equal.

Unit 1 – Exploring One-Variable Data
Notes 4 – Comparing Distributions

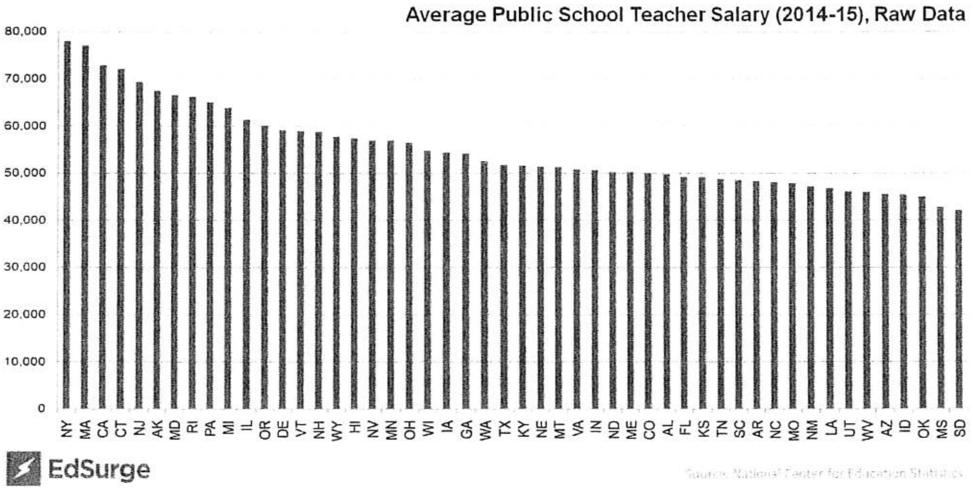
Describing and Comparing Distributions

In the world of statistics, it is not enough to just report the graph or just report the summary statistics.

Why?

The graph alone might not give us all the information we need to make a valid conclusion.

Ex. Do teachers in New York get paid too much? The graph below shows the average teacher salary for each state (from 2014/15).



This graph shows that teachers in New York state make almost double than teachers in South Dakota. Is this a valid conclusion or is the graph not telling the whole story?

The summary statistics might not give us all the information we need to make a valid conclusion.

Ex. The mean teacher salary in the state of California (from 2019/20 data) is $84,531. This lets us know that teachers are making above the median income for everyone in California, which is $78,672. Is this a valid conclusion or are the summary statistics not telling the whole story?

Being the great statistics students you are, you must report both a visual display of the data as well as detailed summary statistics when asked to "describe the distribution":

First: Create your graph (if one is not given)

- Is your data categorical? Use a bar graph! (With two variables next unit, we might also use a segmented bar graph or a mosaic plot)
- Is your data quantitative? (Two variables will require a scatterplot!)
  - Discrete? Use a dotplot, stemplot, or boxplot.
  - Continuous? Use a histogram or a boxplot.

Cumulative relative frequency graphs will usually take too long to create on the AP exam, but remember if you are interested in percentile positions, that you can create these graphs as well!

Second: Summarize your findings
We've discussed the acronym "SOCS" as a way to remember the 4 things to comment on once your graph is created.

- Shape – Skewed, Symmetric, Bimodal, etc.
  - Be as specific as possible and combine shape terms if you can.
- Outliers – an individual observation that falls outside the overall pattern of the graph.
  - You will just have to comment on if there are any visual outliers. You do not have to do the outlier test unless they instruct you to.
- Center – Mean and Median
  - Use the mean, unless the distribution is skewed or has outliers.
  - Unless asked for otherwise, a verbal description of the center will do.
- Spread – Range, Standard Deviation, IQR
  - If you describe the center with the mean, comment on the standard deviation.
  - If you describe the center with the median, comment on the range or IQR.

When the problem asks you to "compare the distributions", you want to follow the above steps, but you want to make sure that you are using comparison words to compare the distributions.

Do not do…	What you should do…
The median teacher salary in California is $61,595 and the median teacher salary in Arizona is $47,606.	The median teacher salary of $61,595 in California is higher than the median teacher salary of $47,606 in Arizona.

Example: The following data are for two popular songs on the Billboard Top 100 (occurring during different decades). The length of each word in the song was recorded and below shows the number of words with the corresponding number of letters.

Sweet Child O'Mine by Guns and Roses

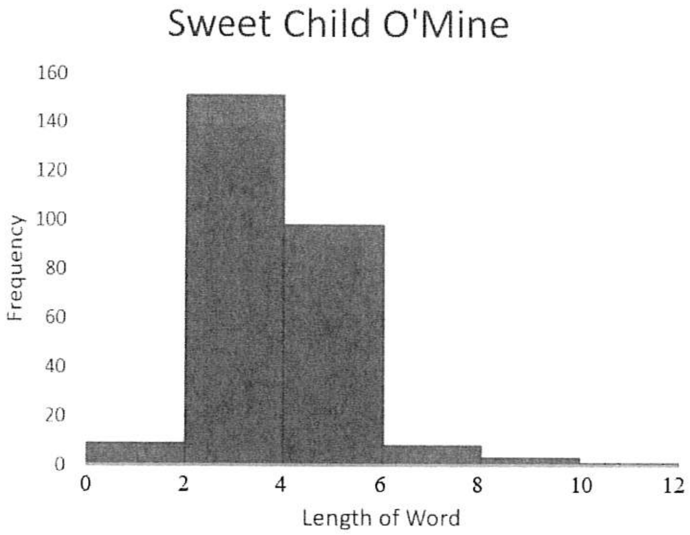
Length of Word	1	2	3	4	5	6	7	8	9	10
Number of Words	9	109	42	51	47	2	6	2	1	1

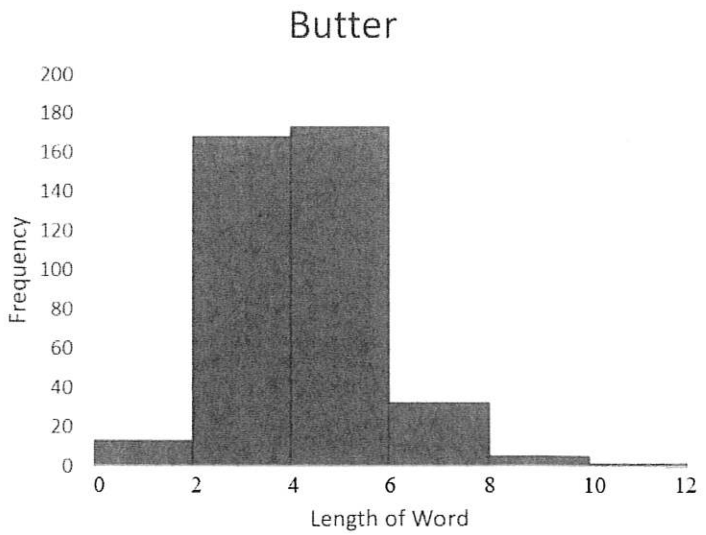
Butter by BTS

Length of Word	1	2	3	4	5	6	7	8	9	10
Number of Words	13	77	91	137	36	26	6	3	2	1

(a) Determine the five number summary for each data set.

(b) Below are the graphs of the two distributions. Write a few sentences comparing the distributions.





(c) There are two “rules” we use to mathematically determine outliers. Method A is the 1.5 x IQR rule and Method B is the two standard deviation rule.

(i) Using method A, determine the outliers that are present in the Sweet Child O’Mine distribution. Justify your answer.

(ii) The mean number of letters in the Butter distribution is 3.62 and the standard deviation is 1.42. Using method B, determine the outliers that are present in the Butter distribution. Justify your answer.

(d) Explain why method A is better for determining outliers than method B in these distributions.