Math Virtual Learning

Algebra IIB

The Data Unit - Describing Numerical Data

May 4, 2020

Algebra IIB
Lesson: May 4, 2020

Objective/Learning Target: Students will describe the key features of data sets.

Let's Get Started:

Review: The ages of employees in a department are listed.   Compute the mean and the median for the following data set.

24, 27, 30, 35, 31, 27, 24, 37, 30, 28, 62, 34, 29, 30

If mean and median are measures of center, why do you think they are they different?

Let's Get Started Answers:

You should have computed the mean as 32 and the median as 30. One employee was much older at 62 years old than the other employees. This will make the mean higher and leave the median the same.

The SHAPE of the data and any OUTLIERS impact the statistics. Some statistics are RESISTANT to outliers and some are NONRESISTANT to outliers.

Outliers can change RANGE, MEAN, and STANDARD DEVIATION so they are not resistant and should not be used if there are outliers or the graph is obviously skewed.

If there are outliers or the graph is obviously skewed, use the IQR and MEDIAN.

Describe your SOCS

There are 4 characteristics of data sets/graphs that help you find meaning in data sets:

Shape: Is the graph of the data symmetrical or skewed?

Outliers: Is there any unusual data?

Center: Where is most of the data centered at?

Spread: How spread out is the data?

Shape

The shape of the data is the first attribute to look at because it guides how you should look at the rest of the attributes. If it obviously skewed you need to use statistics that are RESISTANT to outliers.

Outliers

Outliers will skew the graph even if the rest of the data is symmetrical. You can choose to omit the outlier as long as you document WHY you are leaving it out. If you use the outlier you must use statistics that are RESISTANT to outliers.
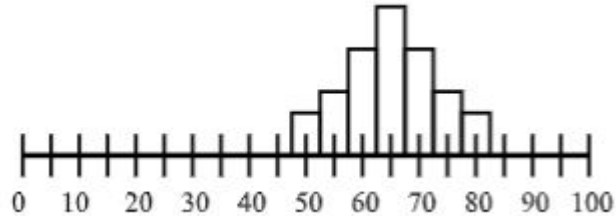
Center

For the center you can select either the mean or the median. Select the one that is appropriate for the graph.
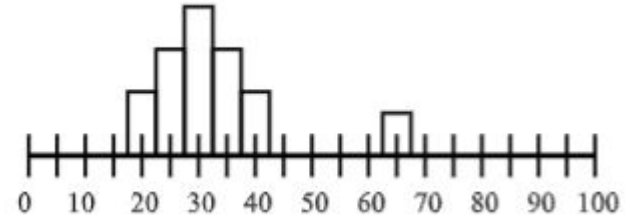
Spread

For spread you can use the range or the IQR depending on the shape of the data.

Putting it together:

The data is symmetrical with no outliers.
The mean of the data is approximately 65.
The range of the data is approximately
36(83-47).

The data is skewed to the right because of
the outlier at 65. This means it is more
appropriate to use the median and the IQR.
The median is approximately 30 and the
IQR is approximately 10 (35-25).

Practice Continued...

Here are the numbers of home runs that Babe Ruth hit in his 15 years with the New York Yankees, 1920 to 1934:

54   59   35   41   46   25   47   60   54   46   49   46   41   34   22

Babe Ruth's home run record for a single year was broken by another Yankee, Roger Maris, who hit 61 home runs in 1961. Here are Mari's home run totals for his 10 years in the American league:

13   23   26   16   33   61   28   39   14   8

Compare these distributions by constructing a back-to-back stemplot. Comment on the similarities and differences of the Shape, Outliers, Center and Spread.

Which baseball player do you think was better, based on the number of home runs hit? Why?

Answer to Practice

Back to Back Stemplots

Maris	Stem	Ruth
1	6	0
	5	4 9 4
	4	1 6 7 6 9 6 1
9 3	3	5 4
8 6 3	2	5 2
4 6 3	1	
8	0	

Key: 4 | 1 means 41
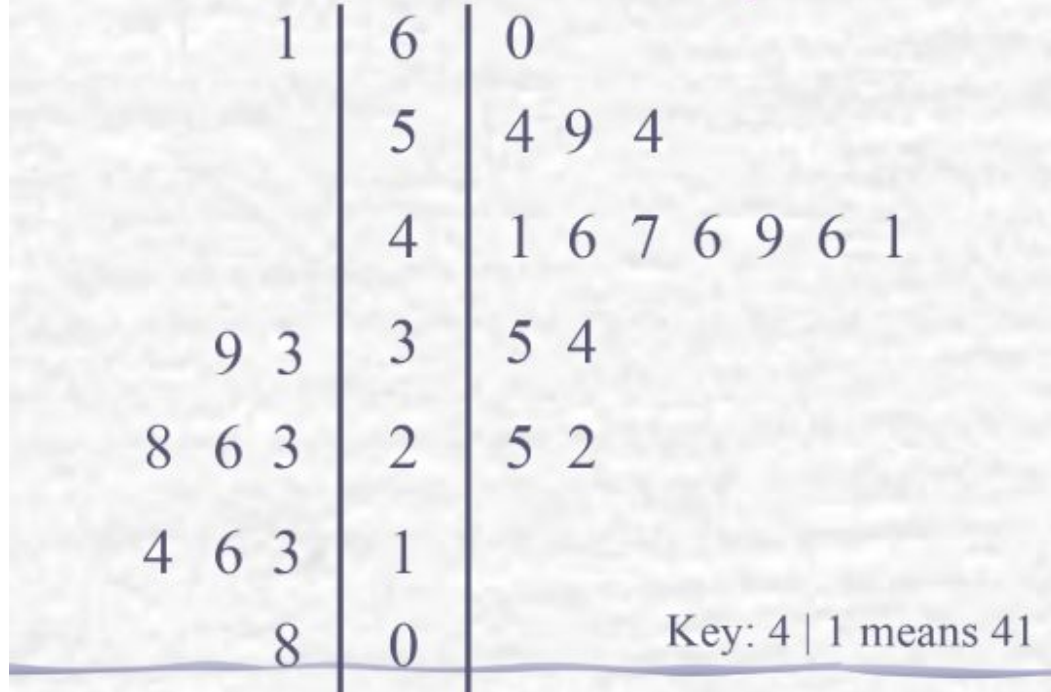
Maris   Ruth

Babe Ruth vs. Roger Maris

- Generally, we can see that Babe Ruth hit more home runs than Roger Maris.
- The center of Babe Ruth is higher at 46 than Roger Maris at 24.5 home runs.
- Roger Maris has an outlier at 61 while Ruth has no outliers.
- Ruth has a higher spread from 22 to 60 than Maris who's from 8 to 39.
- Both distributions are fairly symmetrical.

Who is better? The debate is ongoing…..

The easiest way to describe a data set is to look at a graph of the data set. Here are some online statistical calculators:

Box Plots: Online Box Plot Generator
Histograms: Easy Histogram Maker
Stem-and-Leaf: Stem and Leaf Maker
Statistics (mean, median, standard deviation, …):
Statistics Calculator

Practice

Ater you graph the data, describe the data using the four main characteristics (SOCS)

1. **Exam Scores**   Use a stem-and-leaf plot to display the data. The data represent the scores of a biology class on a midterm exam.

   75   85   90   80   87   67   82   88   95   91   73   80   83   92   94   68   75
   91   79   95   87   76   91   85

2. **Highest Paid CEOs**   Use a boxplot to display the data. The data represent the ages of the top 30 highest paid CEOs.

   64   74   55   55   62   63   50   67   51   59   50   52   50   59   62   64   57
   61   49   63   62   60   55   56   48   58   64   60   60   57

3. **Ice Thickness**   Use a histogram to display the data. The data represent the thickness (in centimeters) of ice measured at 20 different locations on a frozen lake.

   5.8   6.4   6.9   7.2   5.1   4.9   4.3   5.8   7.0   6.8   8.1   7.5   7.2   6.9   5.8   7.2   8.0
   7.0   6.9   5.9

4. **Apple Prices**   Use a box plot to display the data. The data represent the prices (in cents per pound) paid to 28 farmers for apples.

   19.2 19.6 17.1 19.0 17.4 17.3 20.1 19.0 17.5 17.6 18.6 18.4 17.7 19.5 18.4 18.9 17.5
   19.3 20.8 19.3 18.6 18.6 18.3 17.1 18.1 16.8 17.9

Practice Answers

PROBLEM 1

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

Basic Statistics:

Minimum:	67
Maximum:	95
Range:	28
Count:	24
Sum:	2009
Mean:	83.71
Median:	85
Mode:	91
Standard Deviation:	8.28
Variance:	68.56

The data is slightly skewed to the left. Because there are no obvious outliers and the mean and the median are approximately the same, it is safe to use the mean and standard deviation.
The mean score on the biology test about 84% with a standard deviation of 8.28 points. This means that most students scored a C or higher.

Practice Answers
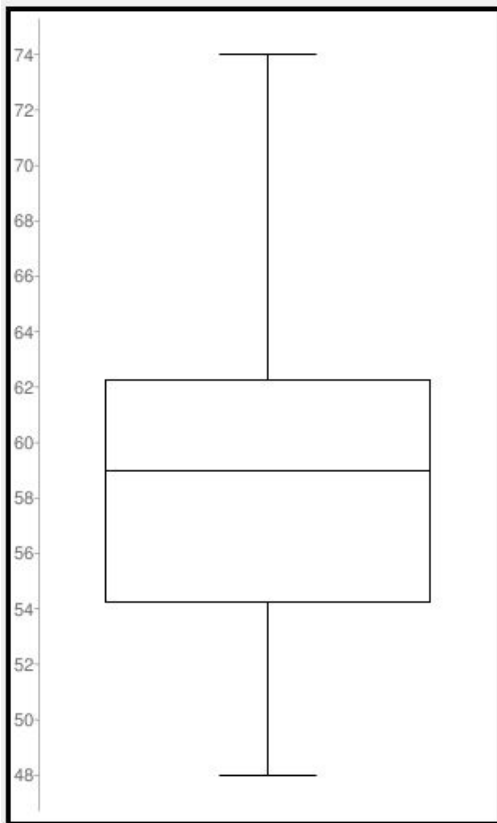
Population size: 30
Median: 59
Minimum: 48
Maximum: 74
First quartile: 54.25
Third quartile: 62.25
Interquartile Range: 8
Outliers: none

The graph is obviously skewed to the right so even though there are no outliers it is safer to use the median and the IQR. The median age of the CEOs was 59 years old. There was an IQR of 8 years so 50% of the CEOS were between the ages of 54 and 62.

Practice Answers

PROBLEM 3
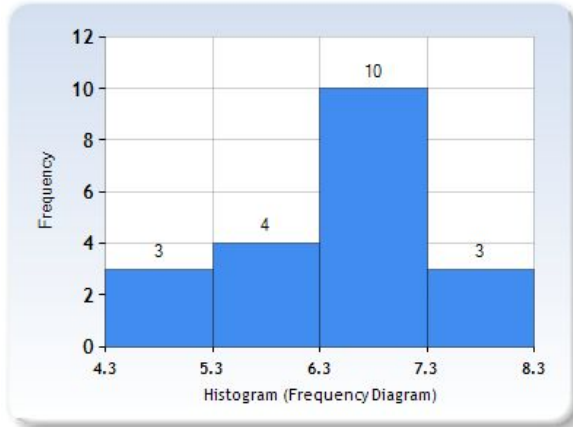
Your Histogram	
Mean	6.535
Standard Deviation (s)	1.01373
Skewness	-0.60715
Kurtosis	-0.14725
Lowest Score	4.3
Highest Score	8.1
Distribution Range	3.8
Total Number of Scores	20
Number of Distinct Scores	13
Lowest Class Value	4.3
Highest Class Value	8.2
Number of Classes	4
Class Range	1

The graph appears to be skewed to the left with no outliers. When I change the number of classes the graph does not get more symmetrical so it is safer to use the median and IQR. A box plot would probably show us this better but when I put the original data into the online statistics calculate the median ice depth is 6.3 cm and the IQR 1.4 cm which means 50% of the ice depths are between 5.8 cm and 7.2 cm.

Practice Answers

PROBLEM 4
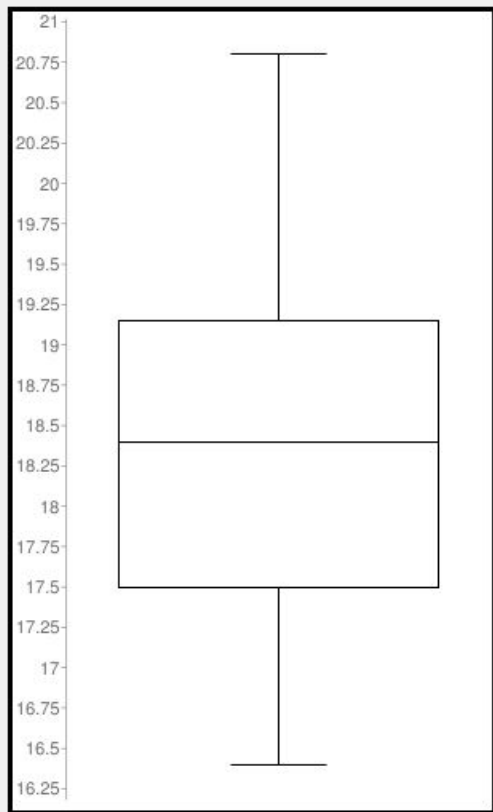
Population size: 28
Median: 18.4
Minimum: 16.4
Maximum: 20.8
First quartile: 17.5
Third quartile: 19.15
Interquartile Range: 1.65
Outliers: none

Arithmetic mean (μ): 18.357142857143
Standard deviation (σ): 1.0324826430324

This graph is approximately symmetrical with no outliers so it is appropriate to use the mean and standard deviation. The mean price of apples was 18.35 cents per bushel. There was a standard deviation of 1.03 so 95% of all apple farmers were paid between 16.29 cents and 20.41 cents per bushel.