Chapter 4 – Analyzing Normal Quantitative Data

Introduction: In chapters 2 and 3, we focused on analyzing categorical data and exploring relationships between categorical data sets. We will now be doing the same for quantitative data. Let us start by reviewing the difference between quantitative and categorical data sets.

Categorical Data
Categorical data are generally labels that tell us something about the people or objects in the data set. For example, what country do they live in, what is the person's occupation, or what kind of pet they have. Usually categorical data is made up of words (do you smoke - yes or no), but occasionally a number can be used as a category. For example, a zip code can be used instead of the place a person lives. The numbers "1" and "2" can be used instead of female and male. Analyzing categorical data involved finding an comparing proportions and percentages.

Quantitative Data
Quantitative data are numbers that measure or count something. They usually have units and taking an average makes sense. For example: a list of people's heights in inches, or their weights in kilograms, or a list of how many dogs are there in various animal shelters across Los Angeles. Notice in each of these cases the data is numerical and an average seems appropriate in the context. We can find the average height, the average weight, or the average number of dogs in animal shelters in Los Angeles.

We are now moving into quantitative data analysis. Analyzing quantitative data is complex and involves shape, measures of center, averages, measures of spread, measures of position, finding typical values and finding unusual values. It is a very different approach than we took for categorical data.

Section 4A – Finding the Shape of Quantitative Data Sets with Dot Plots and Histograms

When analyzing numerical quantitative data, always start with finding the shape of the data set. Categorical data can be graphed, but does not have a shape. Categorical bar charts can be organized in a variety of ways depending on the order of the categories. Quantitative data is numerical measurement data and does have a shape.

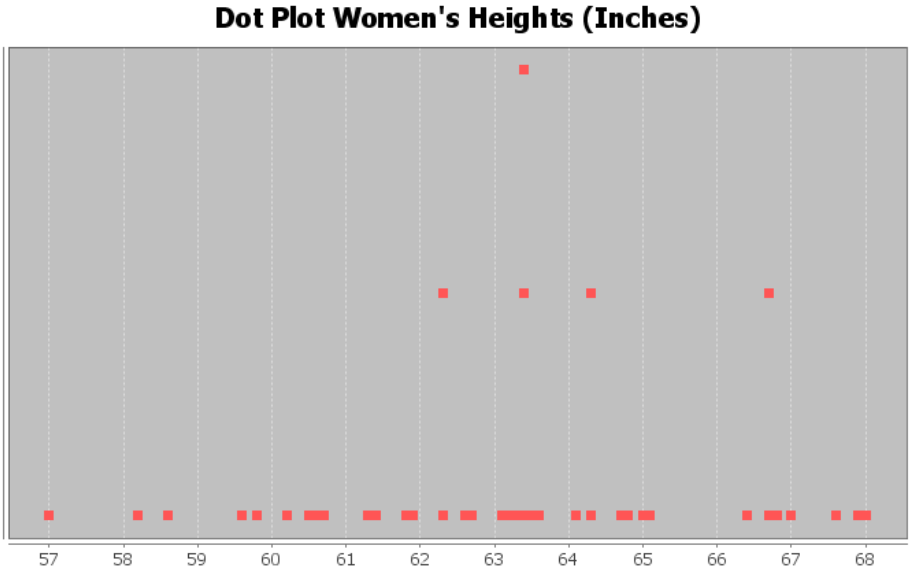
Why should we find the shape?

The goal in analyzing quantitative data is to find the average, spread, typical values and unusual values. In statistics, there are many types of averages, many types of spreads and different ways to find typical and unusual values. Shape helps us determine which averages, spreads and calculations are most accurate for the data.

Dot plots

The most basic kind of graph for quantitative data is the dot plot. The computer draws the numerical scale usually horizontally. It then draws a dot for every single number in the quantitative data set.

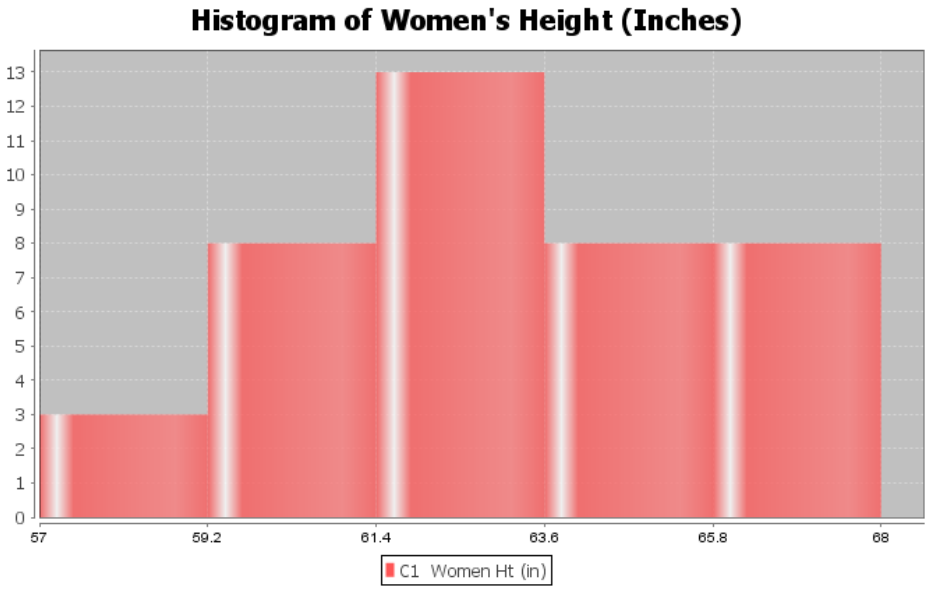
Here is the dot plot for the 40 women's heights created with Statcato.



Dot plots are very useful, especially when identifying unusual values in the data set. Most students find them a little difficult to determine shape from though.

When determining shape, it is better to make a histogram. Think of a histogram as braking the scale up into sections and counting how many dots are in each section. Then drawing a bar that represents the number of dots in that section (frequency).

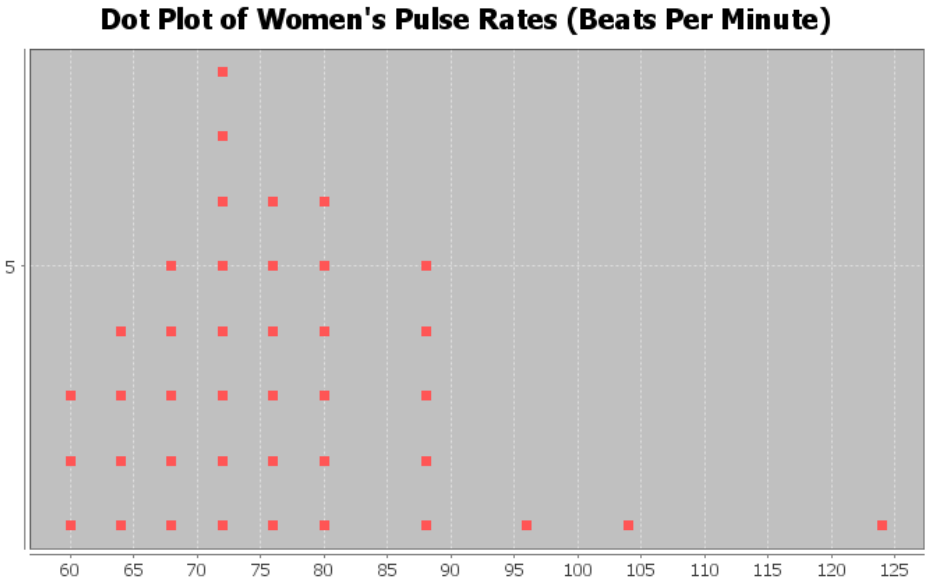
Here is a histogram made with Statcato for the same women's heights data as we used in the dot plot above.

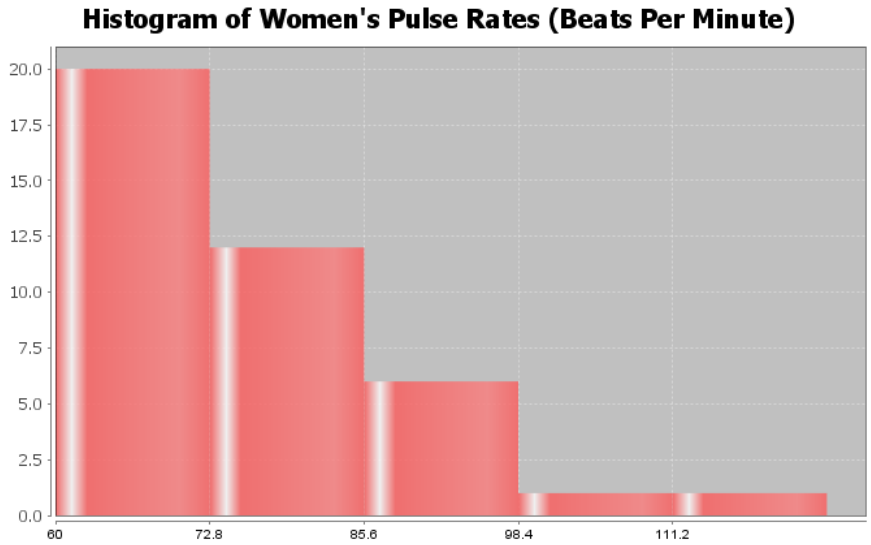


This is a very important shape in statistics. Notice the highest bar is close to the middle and the bars get smaller as we move away from the middle. This is often called "Bell Shaped" or "Normal Data". Some like to describe this shape as unimodal (1 hill) and symmetric (left and right side look about the same). Most people in statistics call this shape "normal" or "normally distributed".

Note: Histograms and Bar Charts are different graphs. A bar chart is a graph for categorical data where each bar gives the count or frequency for each categorical variable. A histogram is a graph used to see the shape of quantitative (numerical measurement) data. Do not confuse the two graphs.

Let us look at another example from the health data. This time we will look at women's pulse rates in beats per minute (BPM). Here is a dot plot and histogram for the data.

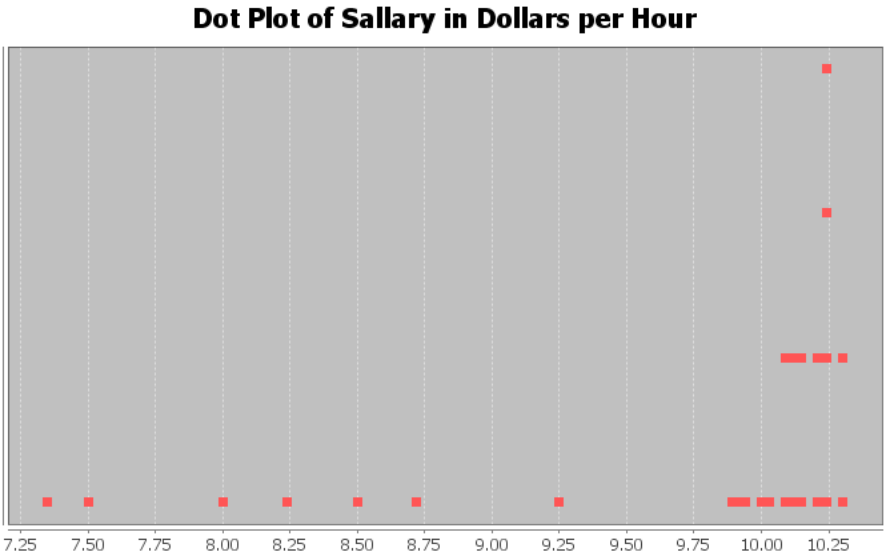


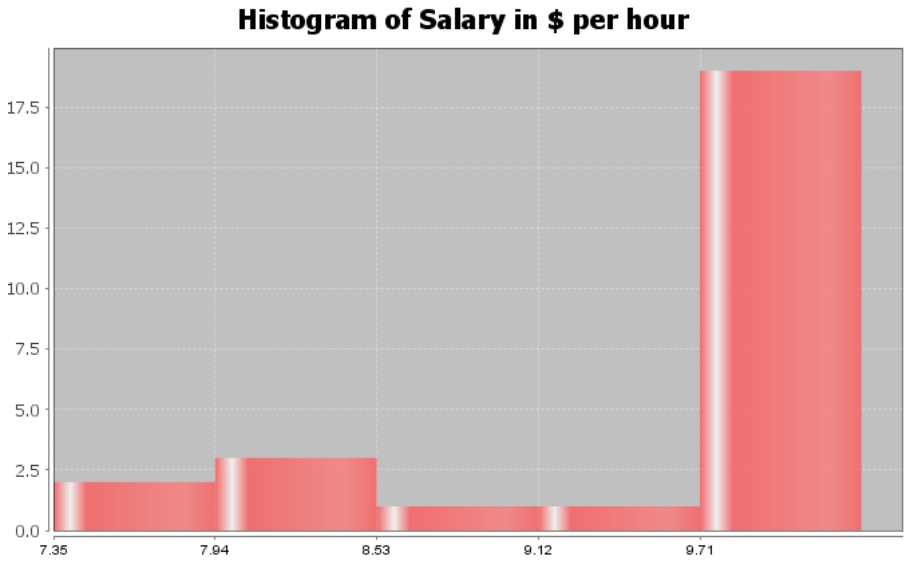


Notice this has a very different shape. There are more dots in the dot plot congregated on the far left. The highest bar in the histogram is on the far left and there are more bars to the right of the highest bar. There is a long tail to the right of the highest bar. This is called "Skewed Right". Some people also call this "Positively Skewed". Remember the skew is referring to the long tail. Look for the highest bar. If there is a significantly longer tail to the right, then it is skewed right. If there is a significantly longer tail to the left, then it is skewed left. If the highest hill is in the middle and the tails are approximately the same length, then it is closer to normal.

Let us look at another example.

Here is some salary data from a small company with 26 employees. The salaries are given in dollars per hour. We created a dot plot and histogram for this data.





What is the shape of these two graphs?

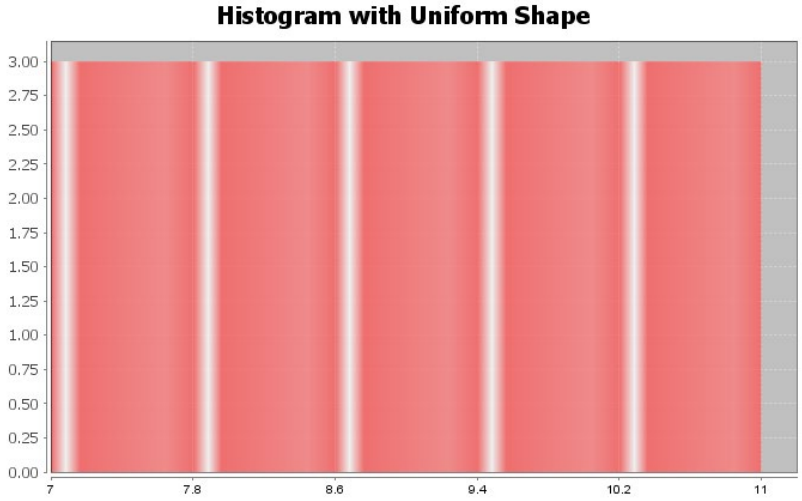
Notice the highest bar and most dots are on the far right, while there is a long tail to the left. Therefore, this is called skewed left.

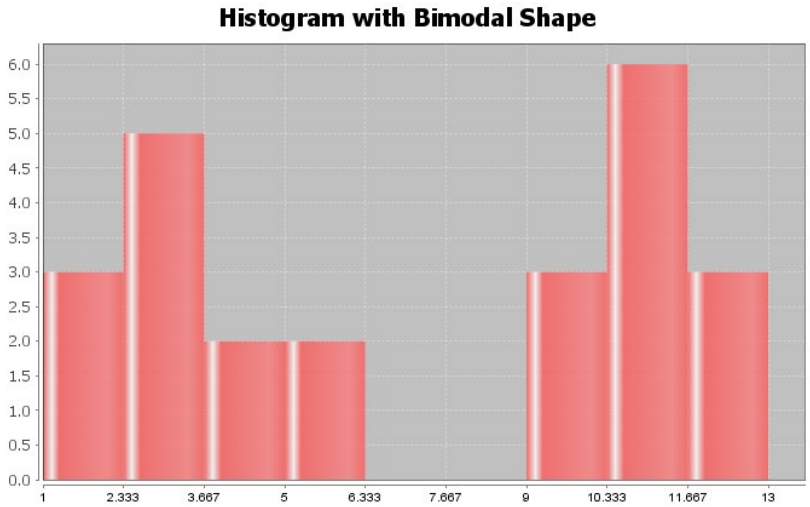
Note on Shape: *Real data rarely has a perfect shape. Most data has a shape somewhere in between bell shaped and skewed, and you will need to make a decision. Look for a significant difference in the length of the tail to classify something as skewed. If my highest hill is toward the middle and I had 2 bars to the right and 3 bars to the left of the highest bar, I would still classify that bell shaped or normal. Some say that is "nearly normal".*

If the highest hill is on the far right and I have 2 bars to the right of the highest hill and 7 bars to the left of the highest hill, I would classify that as skewed left. Some call this "negatively skewed" since negative numbers are to the left on the number line.

Here are a couple unusual shapes that sometimes appear.

A graph that looks like a rectangle is called "uniform". A graph with two distinct high bars is called "bimodal".





Note: In this chapter, we will be focusing on the how to analyze normal (bell shaped) quantitative data sets. We will discuss how to analyze skewed quantitative data sets in the next chapter.

Finding Quantitative Statistics and Creating Graphs with StatKey

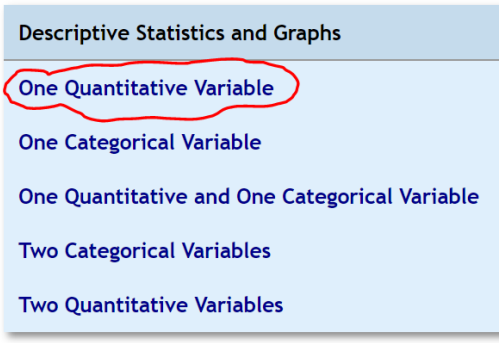
The most basic kind of graph for quantitative data is the dot plot. The computer draws the numerical scale usually horizontally. It then draws a dot for every single number in the data set. Another type of graph is a histogram. This graph counts the number of data values in certain sections and makes a bar telling us how many numbers are in that section. The number of bars is also called "bins" or "buckets".

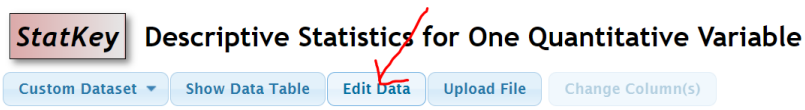
All of these graphs and statistics can be made with StatKey. The heights of women used earlier in this section, may be found in the "Health Data" on Canvas or at www.matt-teachout.org. Open the data set and copy the column of data that says women's heights data. Notice the data is quantitative. The data is made up of numbers that measure the height in inches of the women. It also seems reasonable to look for an average height for these women.

P	Q	R	S	T
Women Age (years)	Women Ht (in)	Women Wt (Lbs)	Women Waist (cm)	Women Pulse (Beats per min)
17	64.3	114.8	67.2	76
32	66.4	149.3	82.5	72
25	62.3	107.8	66.7	88
55	62.3	160.1	93	60
27	59.6	127.1	82.6	72
29	63.6	123.1	75.4	68
25	59.8	111.7	73.6	80
12	63.3	156.3	81.4	64
41	67.9	218.8	99.4	68
32	61.4	110.2	67.7	68
31	66.7	188.3	100.7	80
19	64.8	105.4	72.9	76

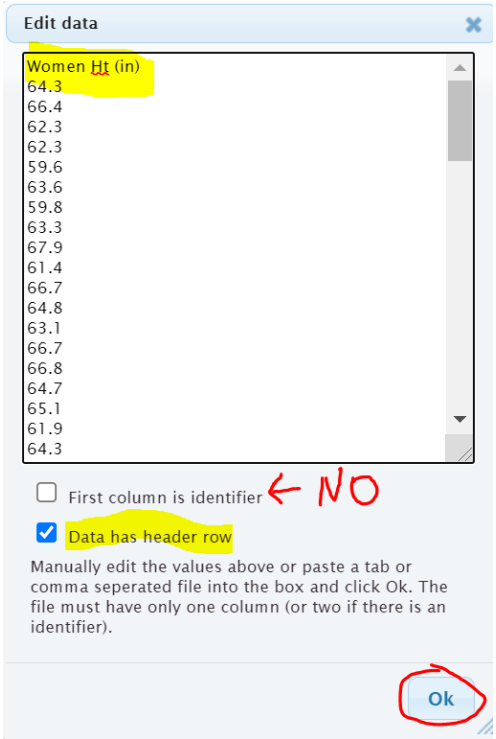
To copy the column of data, hold your cursor over the top of the column until it turns into a downward arrow "↓". Left click your mouse and the whole column will be highlighted. Then push "Contol C" on your keyboard to copy.

Go to www.lock5stat.com and click on the "StatKey" button. Under the "Descriptive Statistics and Graphs" menu, click on "One Quantitative Variable".

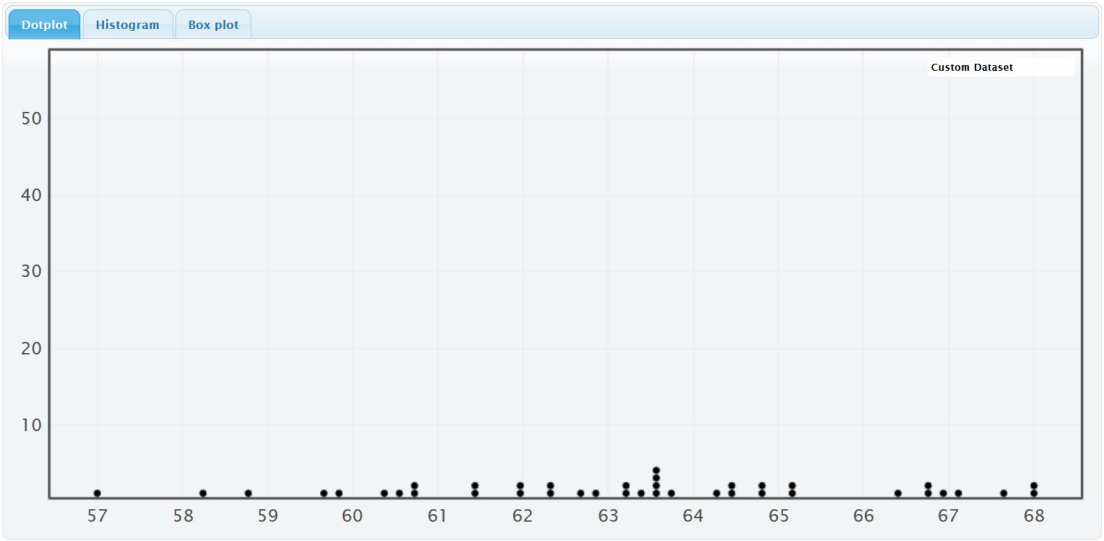


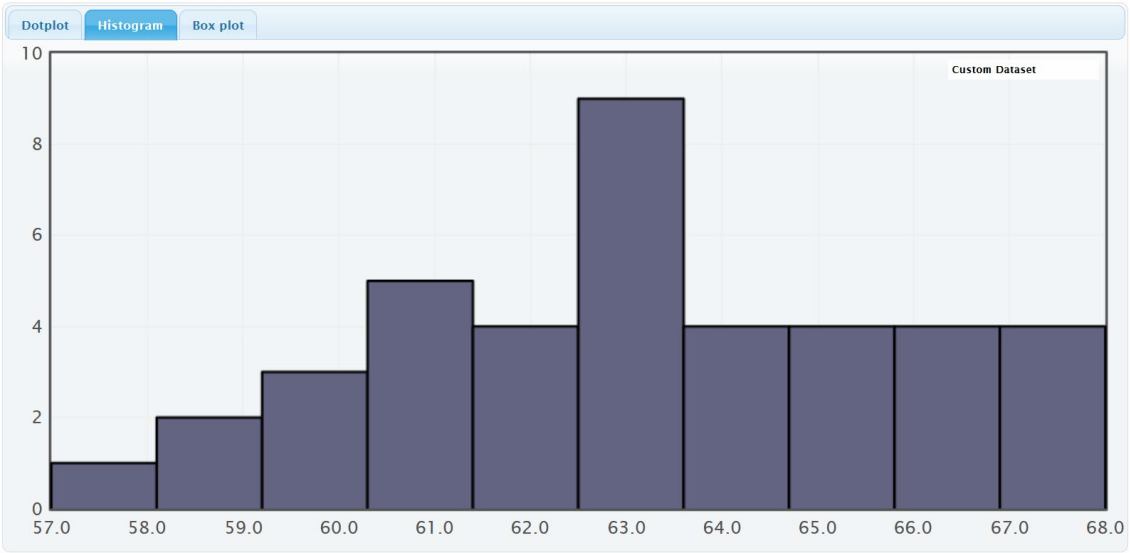


Click on the "Edit Data" button. Push "Control A" on your keyboard and then "delete" in order to get rid of any old data. Make sure your curser is at the top of the "edit data" field. Copy and paste the women's height data into StatKey. Do NOT check the box that says, "First column is an identifier". An identifier is a word next to every number. This data set does not have that. If your data has a title, then check the box that says, "Data has a header row". Now push "OK". Notice StatKey calculates many sample statistics and creates a dot plot and a histogram.

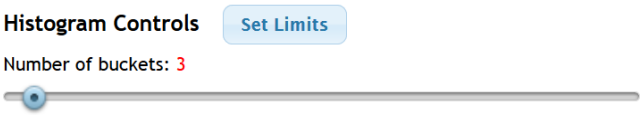


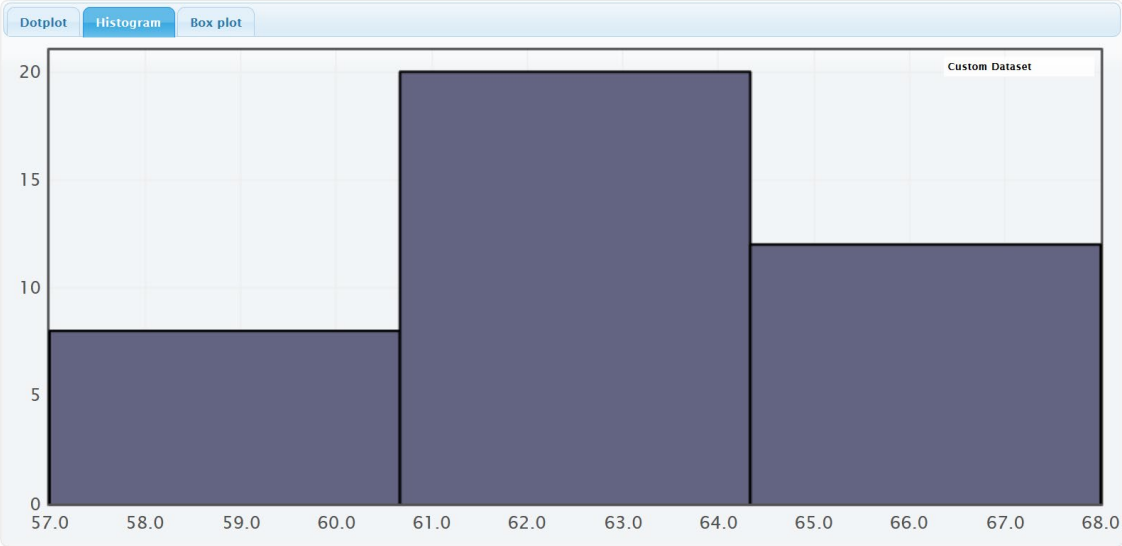
Here are the graphs that StatKey created. Notice there are buttons at the top to pick which graph you want to see. You can see a Dotplot, a Histogram or a Boxplot. We will focus on the dotplot and histogram. We will discuss boxplots in our next chapter.





On the right of this histogram, you will see a slider that can adjust the number of bars or "buckets" in your histogram. The smaller the data set the less bins you should have. Also the less bars you have the easier it is to see the shape. This data set only has 40 numbers, so we want only a few bars. If we slide it to 3 buckets (3 bars), we get the following Histogram.





We see that the highest bar is in the middle and the right and left tails are roughly symmetric. So this is "normal" data.

StatKey has also calculated many summary statistics. We will be discussing these statistics in future sections.

Summary Statistics

Statistic	Value
Sample Size	40
Mean	63.195
Standard Deviation	2.741
Minimum	57
Q_1	61.350
Median	63.350
Q_3	64.900
Maximum	68

Practice Problems Section 4A

Directions: Open the "Health" data in Canvas or at www.matt-teachout.org. You will be using the women's data (columns P – AB) and men's data (columns AD – AP). Go to www.lock5stat.com, click on StatKey and then "One Quantitative Variable". Paste the column of data under "edit data" in StatKey. Click on dotplot and histogram. Draw rough sketch of the dotplot and histogram on a sheet of paper. What is the shape of the data set?

1. Use a StatKey to create a dot plot and histogram of women's ages in years. Adjust the histogram to have three bars (3 buckets).

    a) Draw a rough sketch of the dotplot on a sheet of paper.
    b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
    c) What is the shape of the data?

2. Use a StatKey to create a dot plot and histogram of women's height in inches. Adjust the histogram to have three bars (3 buckets).

    a) Draw a rough sketch of the dotplot on a sheet of paper.
    b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
    c) What is the shape of the data?

3. Use a StatKey to create a dot plot and histogram of women's weight in pounds. Adjust the histogram to have three bars (3 buckets).

    a) Draw a rough sketch of the dotplot on a sheet of paper.
    b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
    c) What is the shape of the data?

4. Use a StatKey to create a dot plot and histogram of women's waist size in centimeters. Adjust the histogram to have three bars (3 buckets).

    a) Draw a rough sketch of the dotplot on a sheet of paper.
    b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
    c) What is the shape of the data?

5. Use a StatKey to create a dot plot and histogram of women's pulse rate in beats per minute. Adjust the histogram to have three bars (3 buckets).

    a) Draw a rough sketch of the dotplot on a sheet of paper.
    b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
    c) What is the shape of the data?

6. Use a StatKey to create a dot plot and histogram of women's systolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets).

    a) Draw a rough sketch of the dotplot on a sheet of paper.
    b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
    c) What is the shape of the data?

7. Use a StatKey to create a dot plot and histogram of women's diastolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets).

    a) Draw a rough sketch of the dotplot on a sheet of paper.
    b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
    c) What is the shape of the data?

8. Use a StatKey to create a dot plot and histogram of women's cholesterol in milligrams per deciliter. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

9. Use a StatKey to create a dot plot and histogram of women's body mass index (BMI) in kilograms per meters squared. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

10. Use a StatKey to create a dot plot and histogram of women's wrist circumference in inches. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

11. Use a StatKey to create a dot plot and histogram of men's ages in years. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

12. Use a StatKey to create a dot plot and histogram of men's height in inches. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

13. Use a StatKey to create a dot plot and histogram of men's weight in pounds. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

14. Use a StatKey to create a dot plot and histogram of men's waist size in centimeters. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

15. Use a StatKey to create a dot plot and histogram of men's pulse rate in beats per minute. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

16. Use a StatKey to create a dot plot and histogram of men's systolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

17. Use a StatKey to create a dot plot and histogram of men's diastolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

18. Use a StatKey to create a dot plot and histogram of men's cholesterol in milligrams per deciliter. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

19. Use a StatKey to create a dot plot and histogram of men's body mass index (BMI) in kilograms per meters squared. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

20. Use a StatKey to create a dot plot and histogram of men's wrist circumference in inches. Adjust the histogram to have three bars (3 buckets).

a) Draw a rough sketch of the dotplot on a sheet of paper.
b) Draw a rough sketch of the histogram (with 3 bars) on a sheet of paper.
c) What is the shape of the data?

Section 4B – Shapes and Centers

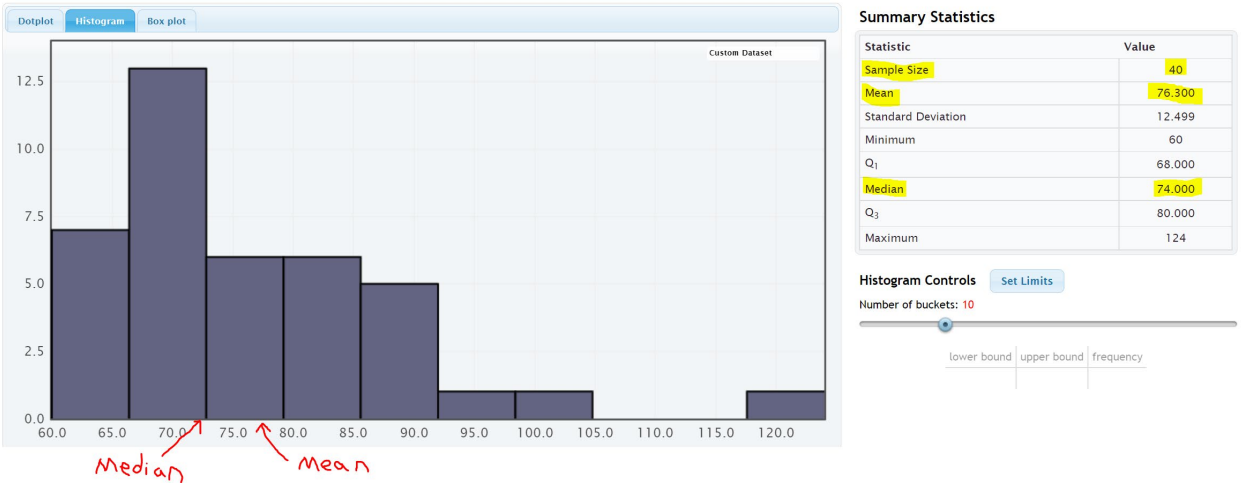
When analyzing quantitative (numerical measurement) data, we want to find the average. In statistics, we often refer to an average as a "Center". When a person asks about the center, they are really asking about the average.

Definition of Statistics: The word "statistics" refers to numbers that are calculated to describe sample data sets. For example, a mean average is one of many types of statistics. Therefore, the study of "statistics" is the study of numbers calculated from data sets that help describe the characteristics of that data and hopefully what that data tells us about the world around us. We are not there yet though.

In statistics, there are many types of centers or averages. The two most commonly used centers (averages) are the mean average and the median average. The key is to determine which center (average) is most accurate for the data. An accurate center should be close to your highest bar in your histogram.

Example 1

The following histogram and statistics were calculated with StatKey and are describing the pulse rates of 40 women in beats per minute. This data was found in the "health data". We see that this quantitative data is skewed right since the histogram has a long right tail.



We see that there was 40 women in the data since the "sample size" was 40. We also see that the mean average was 76.3 bpm and the median average was 74 bpm.

The center of a data set is where the most people or objects are located. The highest bar or bars represent the center of the data. An accurate center or average should be close to the highest bar in the data set and therefore be representative of the data values. An average that is not close to the highest bar is not a very good average.

Let us compare these values to the histogram. Notice a few things. The mean is not very accurate measures of center since they are not close to the highest bar. So the mean is not a very good average for this data. The median seems to be more accurate, since it is closer to the highest bar. So the median is closer to the center of the data.

Here are a couple of things to keep in mind when finding an accurate average for a data set. The women's pulse data is skewed right. Mean averages get pulled in the direction of the skew (long tail) and tend to not be very accurate for skewed data sets. The median average does not get pulled in the direction of the skew and remains close to the highest bar. All this leads to an important principle. When a data set is skewed, statisticians use the median average as best measure of center and the average of the data set.

Center Principle for Skewed Data

If a data set has a skewed shape, the median average is usually the most accurate average (measure of center) and we should use the median as the average for the data set.

Example 2

Let us look at another data set from the health data. Here is the StatKey histogram and sample statistics from the women's height data. The data set gives the heights in inches of 40 women.

Let us compare these values to the histogram. Notice a few things. First, look at the shape. This data set is bell shaped (normal) data. The highest bar is in the middle and the right and left tails are about the same distance from the center bar.

Notice that both the mean average and the median average are close to the highest bar. It seems like either of these statistics are pretty accurate averages (centers) since they are both close to the highest bar in the histogram. Either of them would be a decently accurate average for this data.

So which one should we use?

If a data set is bell shaped, statisticians prefer to use the mean. There are several reasons for this. One being that people are most familiar with the mean. It is after all the most common type of average. That is not the real reason why we should use the mean for bell shaped data though. The real reason has to do with the spread of the data set. Bell shaped data has a very specific spread that is measured most accurately with standard deviation. Standard deviation is the most accurate spread for normal (bell shaped) data. It measures the typical distance from the mean. Therefore, in a bell shaped data set we need to use the mean as our center or average so that we can use the standard deviation to accurately measure the spread.

Center Principle for Normal (Bell Shaped) Data

If a data set is bell shape (normal), then the mean average is usually accurate and we should use the mean as the average and center for the data set.

Key: Do <u>not</u> use the mean average unless the data is bell shaped (normal). If the data is <u>not</u> normal then the mean is <u>not</u> accurate.

Calculating Centers with Technology

Remember that "statistics" are numbers that describe characteristics of data sets. The calculations though are very difficult by hand or by calculator, especially with large data sets. Always use a statistics software to calculate statistics.

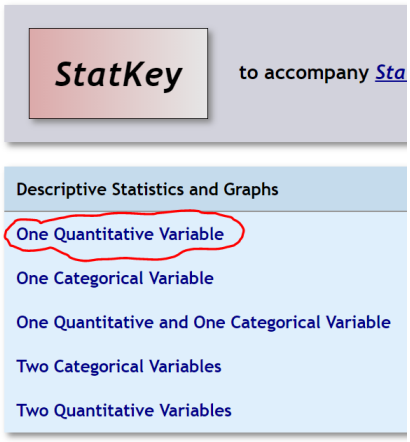
<u>Finding Quantitative Statistics with StatKey</u>

The heights of women used earlier in this section, may be found in the "Health Data" on Canvas or at www.matt-teachout.org. Open the data set and copy the column of data that says women's heights data. Notice the data is quantitative. The data is made up of numbers that measure the height in inches of the women. It also seems reasonable to look for an average height for these women.

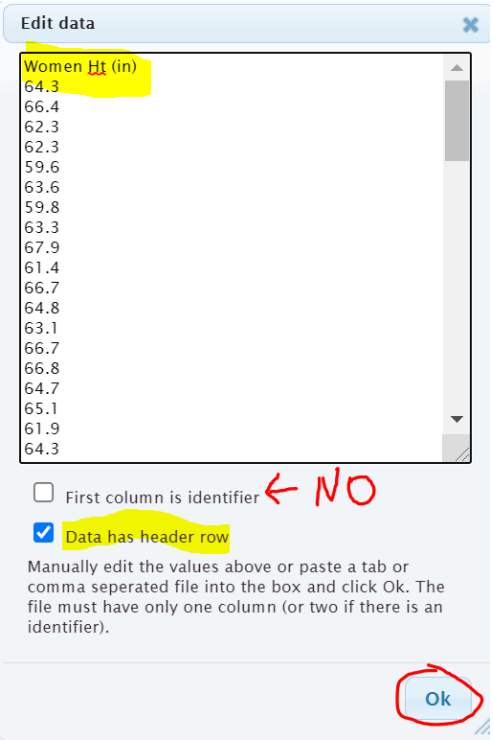
P	Q	R	S	T
Women Age (years)	Women Ht (in)	Women Wt (Lbs)	Women Waist (cm)	Women Pulse (Beats per min)
17	64.3	114.8	67.2	76
32	66.4	149.3	82.5	72
25	62.3	107.8	66.7	88
55	62.3	160.1	93	60
27	59.6	127.1	82.6	72
29	63.6	123.1	75.4	68
25	59.8	111.7	73.6	80
12	63.3	156.3	81.4	64
41	67.9	218.8	99.4	68
32	61.4	110.2	67.7	68
31	66.7	188.3	100.7	80
19	64.8	105.4	72.9	76

To copy the column of data, hold your cursor over the top of the column until it turns into a downward arrow "↓". Left click your mouse and the whole column will be highlighted. Then push "Contol C" on your keyboard to copy.

Go to www.lock5stat.com and click on the "StatKey" button. Under the "Descriptive Statistics and Graphs" menu, click on "One Quantitative Variable".



Click on the "Edit Data" button. Push "Control A" on your keyboard and then "delete" in order to get rid of any old data. Make sure your curser is at the top of the "edit data" field. Copy and paste the women's height data into StatKey. Do NOT check the box that says, "First column is an identifier". An identifier is a word next to every number. This data set does not have that. If your data has a title, then check the box that says, "Data has a header row". Now push "OK". Notice StatKey calculates many sample statistics and creates a dot plot and a histogram.



Once we push "OK" we see the quantitative statistics calculated on the right of the page.

Summary Statistics

Statistic	Value
Sample Size	40
Mean	63.195
Standard Deviation	2.741
Minimum	57
Q_1	61.350
Median	63.350
Q_3	64.900
Maximum	68

We will be discussing these statistics in greater detail later on. Notice the summary statistics for one quantitative data set include the following:

Mean Average: Average (or center) used for normal quantitative data.

Median Average: Average (or center) used for skewed or non-normal quantitative data.

Sample Size: Total number of people or objects that we collected data from. (For quantitative data, it also tells us how many numbers were in our quantitative data set.)

Minimum: Smallest number in the quantitative data set.

Maximum: Largest number in the quantitative data set.

Practice Problems Section 4B

Directions: Open the Health data in Canvas or at www.matt-teachout.org. You will be using the men and women's combined data (columns B - N). This data has columns of 80 values from 40 men and 40 women. Go to www.lock5stat.com, click on StatKey and then "One Quantitative Variable". Paste the column of data under "edit data" in StatKey. Click on "histogram" and change the histogram to have 3 bars (3 buckets). What was the shape of the data set? Under "summary statistics" in StatKey, make a note of the minimum, maximum, sample size (n), mean average, and median average. Based on the shape, of the data, should we use the mean or the median as our most accurate average?

1. Use a StatKey to create a histogram and summary statistics for the ages in years. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

2. Use a StatKey to create a histogram and summary statistics for the heights in inches. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

3. Use a StatKey to create a histogram and summary statistics for the weights in pounds. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

4. Use a StatKey to create a histogram and summary statistics for the waist sizes in centimeters. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

5. Use a StatKey to create a histogram and summary statistics for the pulse rates in beats per minute. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

6. Use a StatKey to create a histogram and summary statistics for the systolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

7. Use a StatKey to create a histogram and summary statistics for the diastolic blood pressure in millimeters of mercury. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

8. Use a StatKey to create a histogram and summary statistics for cholesterol in milligrams per deciliter. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

9. Use a StatKey to create a histogram and summary statistics for body mass index (BMI) in kilograms per square meters. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

10. Use a StatKey to create a histogram and summary statistics for leg length in inches. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

11. Use a StatKey to create a histogram and summary statistics for elbow circumference in inches. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

12. Use a StatKey to create a histogram and summary statistics for wrist circumference in inches. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

13. Use a StatKey to create a histogram and summary statistics for arm length in inches. Adjust the histogram to have three bars (3 buckets). Do not copy or draw the histogram. Simply use the histogram to verify the shape.

a) What is the shape of the data?
b) What is the sample size (n)? (This tells us how many total people were measured.)
c) What is the minimum (smallest value) for the data?
d) What is the maximum (largest value) for the data?
d) What is the mean average for this data?
e) What is the median average for this data?
f) Based on the shape of this data set, should we use the mean or median as our most accurate average (center)?

Section 4C – Understanding the Mean Average

If you walked up to someone and asked them how to calculate an average, most would tell you to add up the numbers and divide by how many numbers are in the data set. In other words, most people equate the word "average" with the mean average. It is by far the most common average used.

We learned in the last section that in statistics there are many types of averages and the mean average is only accurate when the data is bell shaped (normal). While many people have an idea of how the mean is calculated, very few understand the complexities behind the mean average.

Since we are in the chapter on analyzing normal (bell shaped) data and data analysts prefer to use the mean average when data is normal, we will focus on understanding the mean average in this section.

Definition of the Mean Average: The mean average is the center or average that balances the distances between all of the numbers in a quantitative data set. The mean is only accurate if the data set has a normal (bell) shape.

Note on Calculating Statistics

Many people focus on how statistics are calculated instead of the true meaning of the statistic and how to use and explain it properly. Remember, calculations in statistics are extremely time consuming, which is why we prefer to have a computer program do the calculations. What a computer cannot do is tell you what the meaning behind the statistic and when and how it should be used. In statistics, always focus on understanding and being able to explain ideas. That is the real job of a statistician, data scientist, or data analysist.

Calculating the Mean Average
Formulas for calculating statistics are very difficult. Focus on understanding the ideas behind the formula, not on using the formula to calculate. Remember, the formulas are already programed into statistics software programs. The software should be the one doing the calculation. You should be focused on explaining the statistic and what it tells us about the data.

Here are some variables (letters) you often see in statistics formulas for the mean.

n: total frequency or sample size (the number of values in your data set)

x: each individual number in the data set

$\sum$: summation symbol (tells us to add)

$\sum x$: add up all the numbers in your data set

$\overline{x}$: "x-bar". This symbol is used for the mean average of a data set (sample mean average)

Formula for calculating the mean average

$$\overline{x} = \frac{\sum x}{n}$$

(Add up all the number in your data set and divide by how many numbers are in your data set.)

Example 1
As we have said, no statistician calculates the mean with a formula and calculator. The data sets are usually way too large. Since we are just learning about how mean averages work, it would be nice to calculate a couple. If anything, so you have an idea of what the computer is doing.

The following data describes the weights (in kilograms) of various bricks at a building site. Calculate the mean average for the following data:

4.7 , 6.2 , 3.3 , 5.1 , 2.9 , 7.4 , 4.5

How many numbers are in the data set? (This is the total frequency or sample size.)

Seven (n = 7)

Mean Average = (4.7 + 6.2 + 3.3 + 5.1 + 2.9 + 7.4 + 4.5) / 7 = 34.1 / 7 = 4.871428571

Be sure to add the numbers first and then divide by the frequency.

Where should we round the answer?

Rounding Rule for Quantitative Data: *Round statistics calculated from quantitative data to one more decimal place to the right than is present in the original data.*

Notice the numbers in the data set ended in the tenths place (one place to the right of the decimal). This means that we should round our statistic to one more place value to the right. Therefore, we would round to two places to the right of the decimal (hundredths place).

Mean Average Weight of the Bricks = 4.871428571 ≈ 4.87 kilograms

Remember; focus on interpreting the meaning of this statistic.

What does a mean average of 4.87 kilograms tell us about the data?

A mean average of 4.87 kg tells us that the balancing point for the distances for all the numbers in the data set is 4.87 kg. What does this tell us?

Look at the numbers in the data set above the mean: 6.2, 5.1, and 7.4

Let us look at how far are each of these numbers from the mean? Remember we rounded the mean, so these are just approximate distances.

6.2 – 4.87 ≈ 1.33

5.1 – 4.87 ≈ 0.23

7.4 – 4.87 ≈ 2.53

Therefore, for numbers in the data set above the mean, we have a total approximate distance from the mean of 1.33 + 0.23 + 2.53 ≈ 4.09

Now look at the numbers in the data set below the mean: 2.9, 3.3, 4.5, and 4.7

Approximately how far are these numbers from the mean? If we subtract in the same order with the value minus the mean we will get negative differences. This issue of negative number differences is a reoccurring problem in statistics that is usually addressed by squaring the values

2.9 – 4.87 ≈ -1.97

3.3 – 4.87 ≈ -1.57

4.5 – 4.87 ≈ -0.37

4.7 – 4.87 ≈ -0.17

Therefore, the total of the differences for numbers below the mean is

-1.97 + -1.57 + -0.37 + -0.17 ≈ -4.08

Technically distances are not negative so the total distance is approximately +4.08

Notice that the total distance for numbers above the mean is almost the same as the total distance for numbers below the mean. This is why the mean is called the "balancing point". Why is it not perfectly equal? It would be if we used the unrounded version of the mean.

Understanding the Balancing Point
If you understand that mean is the balancing point, you will not only have a much better understanding of the mean, but you will also be able to estimate the mean in situations and be able to create data sets with a specific mean.

Example 2
Suppose I want to create a data set five values that has a mean average of 20.

I can pick any numbers I want as long as I balance the distances.

Suppose I use 14, 16, 18, and 19 for my first four numbers. Look at the distance from 20.

14 (six units from 20)

16 (four units from 20)

18 (two units from 20)

19 (one unit from 20)

All these numbers were below 20, so the total distance below so far is 6 + 4 + 2 + 1 = 13

If I want a total of five numbers in the data set, I will have to choose one number above 20 that has the same total distance. In this case 13 above 20 or 33.

Therefore, my created data set with five numbers and a mean of 20 is

14, 16, 18, 19, 33

Let us check it:

Mean = (14 + 16 + 18 + 19 + 33) / 5 = 100 / 5 = 20

More Examples

You can create tons of different data sets, if you understand this principle of the balancing point. Symmetric data sets are probably the easiest to create.

Suppose I want to create a data set with twelve numbers with a mean of 20.

An easy way to do this is to take six numbers above the mean (20) and six numbers below the mean (20). I will pick them so they have the same distances.

Below mean of 20: 14, 15, 16, 17, 18, 19

Above the mean of 20: 21, 22, 23, 24, 25, 26

Notice that 19 and 21 are both one from twenty, 18 and 22 are both two from twenty, and so on. The distances are balanced, so the mean of all of these numbers will be twenty.

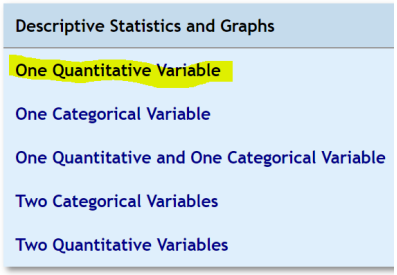
Data set with twelve numbers and a mean of twenty:

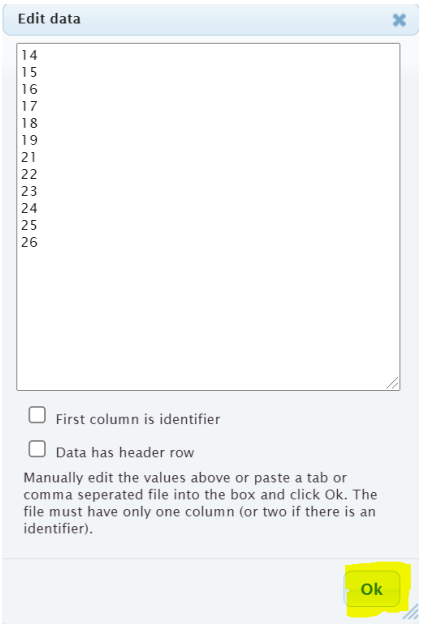
14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26

Let's check and see if this data set has a mean of 20.

Mean Average = (14 + 15 + 16 + 17 + 18 + 19+ 21 + 22 + 23 + 24 + 25 + 26) ÷ 12 = (240) ÷ 12 = 20. The mean is 20.

An easier way would be to go to "One Quantitative Variable" in StatKey at www.lock5stat.com. If we click the edit data button and type in the numbers. Do not check the box that says identity. Do not check the box that says "data has a header row". Just click "OK".





Summary Statistics

Statistic	Value
Sample Size	12
Mean	20.000

Practice Problems Section 4C

Directions for #1-7: Find the mean for the following data sets. You may use a calculator. When rounding is appropriate, round answers to one more decimal place than the numbers in the data set. Then write the definition of the mean average in context to explain the mean for each problem.

Mean $(\bar{x}) = \frac{\sum x}{n} = \frac{Sum\ of\ the\ numbers\ in\ the\ data\ set}{Sample\ size\ (how\ many\ numbers\ are\ in\ the\ data\ set)}$

1. Number of dogs at dog hotels: 2 , 7 , 7 , 9 , 8 , 8 , 4 , 5 , 1 , 0 , 3 , 2 , 11 , 3 , 1 , 7 , 2 , 4

2. Number of cars parked in various parking lots: 17 , 21 , 23 , 24 , 25 , 27 , 28 , 29 , 31 , 32 , 33 , 36

3. Temperature in degrees Celsius: 9.4 , 3.5 , 1.1 , 7.8 , 3.2 , 16.4 , 6.6

4. Grams of medicine: 1.6 , 5.2 , 3.3 , 9.4 , 1.7 , 1.9 , 2.8 , 12.5 , 8.6 , 1.8 , 2.6 , 2.4

5. Dollars spent for a hot dog: 2.54 , 3.14 , 2.49 , 1.98 , 1.46 , 2.27 , 1.83 , 2.63 , 2.87 , 3.25 , 8.75

6. Weight of building stones in kilograms: 1.362 , 5.714 , 3.199 , 2.285 , 4.477 , 9.251

7. Weight of a group of men in pounds: 146 , 157 , 181 , 193 , 226 , 158 , 176 , 187 , 216

8. Find a data set with six numbers that has a mean of 13 and without any repeating numbers. Check your answer by calculating the mean to make sure the data set works.

9. Add two numbers to your data set in #8, so that the mean remains 13. (You should now have eight numbers in your data set.) There should not be any repeating numbers. Check your answer by calculating the mean to make sure the data set works.

10. Find a data set with nine numbers that has a mean of 21.5 and without any repeating numbers. Check your answer by calculating the mean to make sure the data set works.

11. Add two numbers to your data set in #10, so that the mean remains 21.5. (You should now have eleven numbers in your data set.) There should not be any repeating numbers. Check your answer by calculating the mean to make sure the data set works.

12. Explain how the mean is the balancing point of the data in terms of distances. Look at the following data set. Use the distances to explain how the mean is really 11 without adding the numbers and without calculating the mean directly.

5, 6, 7, 8, 9, 13, 14, 15, 16, 17

Section 4D – Spread, Standard Deviation, and Typical Values for Normal Quantitative Data Sets

When analyzing a quantitative data set, we have seen so far that we want to look at the shape of the data set and we want to find the most accurate center (in which we get the average). There is another description of the data that is important to explore, and that is the "Spread" or "Variability" of a data set.

A measure of spread or variability in a data set tells us how spread out the data is. Why is this important? Let's look at an example.

Being a teacher, I like to look at quiz scores for my classes.

Class A: 90 , 92 , 99 , 100 , 97 , 96 , 98 , 94 , 91 , 90 , 89 , 100 , 93 , 93 , 88

This class has a very small spread. Virtually everyone in the class got an A or a high B. These kinds of scores make me very happy as a teacher. A data set with a small spread or small variability means it is more consistent and easier for us to predict future values. I predict quiz scores to be high for this class.

Class B: 26 , 97 , 35 , 84 , 55 , 72 , 61 , 44 , 88 , 69 , 77 , 38 , 51 , 99 , 86

This class has a very large spread with a lot of variability. The quiz scores are all over the place. This class is worrying me. Not only was there many low grades, but the class was very inconsistent. It will be very difficult to predict what quiz grades to expect from these students. I definitely need to review the material more with this class.

Notes on Spread (Variability)

- **Small Spread** (Small amount of Variability): Tells us the data values are close, more consistent and easier to predict.
- **Large Spread** (Large amount of Variability): Tells us that the data values are very spread out, less consistent, and more difficult to predict.

Measures of Spread
There are several statistics that measure spread or variability. The most common ones are the range, the interquartile range (IQR), the standard deviation (s), and the variance (s^2). Which are most accurate? Again it depends on the shape of the data set.

For normal (bell shaped) quantitative data sets, the standard deviation is the most accurate measure of spread.

Spread Principle for Normal Quantitative Data

When quantitative data is normal, use the standard deviation "s" as our most accurate measure of typical spread.

Note: The standard deviation is only accurate if the quantitative data is normal (bell shaped).

Definition of Standard Deviation: The standard deviation is how far typical values are from the mean in a normal (bell shaped) quantitative data set. The standard deviation can be thought of as an average distance from the mean or the typical distance from the mean, but is only accurate if the quantitative data is normal (bell shaped).

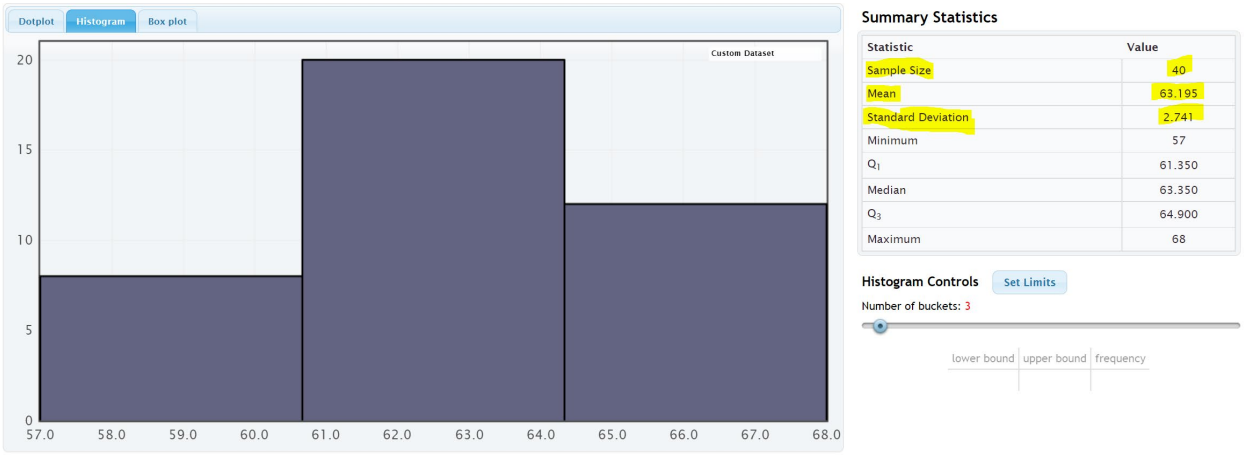
Calculations of spread are often even more difficult than measures of center, so it is even more important to use a statistics software program to calculate. For example, calculating standard deviation with a formula and calculator can take a long time, even for a small data set.

Remember how to calculate statistics for one quantitative variable with StatKey?

Go to www.lock5stat.com and click on StatKey. Click on "one quantitative variable" under the "descriptive statistics and graphs" menu. Under "edit data", copy and paste the column of quantitative data into StatKey. Do NOT check the box that says "identifier". If your data has a title, click on the box that says "data has a header row". If the data does not have a title, then do NOT check the box that says "data has a header row".

Example 1

We used StatKey to calculate the mean average and the standard deviation for the women's heights data. This column of data is located in the Health Data Set. Notice first that the data is normal (bell shaped). That tells us that we should use the Standard Deviation as the best measure of typical spread.



Remember to focus on interpretation, not on calculation: In the women's height data, the standard deviation is 2.741 inches. So typical heights for the women were 2.741 inches from the mean on average.

What does this tell us? The mean average for the women's height data was 63.195 inches. So typical women in the data set were within 2.741 inches from 63.195 inches. This gives us a "typical range" (two values that typical numbers in the data are in between).

63.195 – 2.741 ≤ typical heights for these women ≤ 63.195 + 2.741

60.454 ≤ typical heights for these women ≤ 65.936

Typical women in this data set had a height between 60.454 inches (little over 5 feet) and 65.936 inches (little under 5 ½ feet).

To calculate Typical Values for Normal Quantitative Data Sets: Add and subtract the mean and standard deviation. (Be careful to subtract in the correct order.)

Mean – Standard Deviation ≤ typical values ≤ Mean + Standard Deviation

Empirical Rule for Normal (Bell Shaped) Quantitative Data Sets

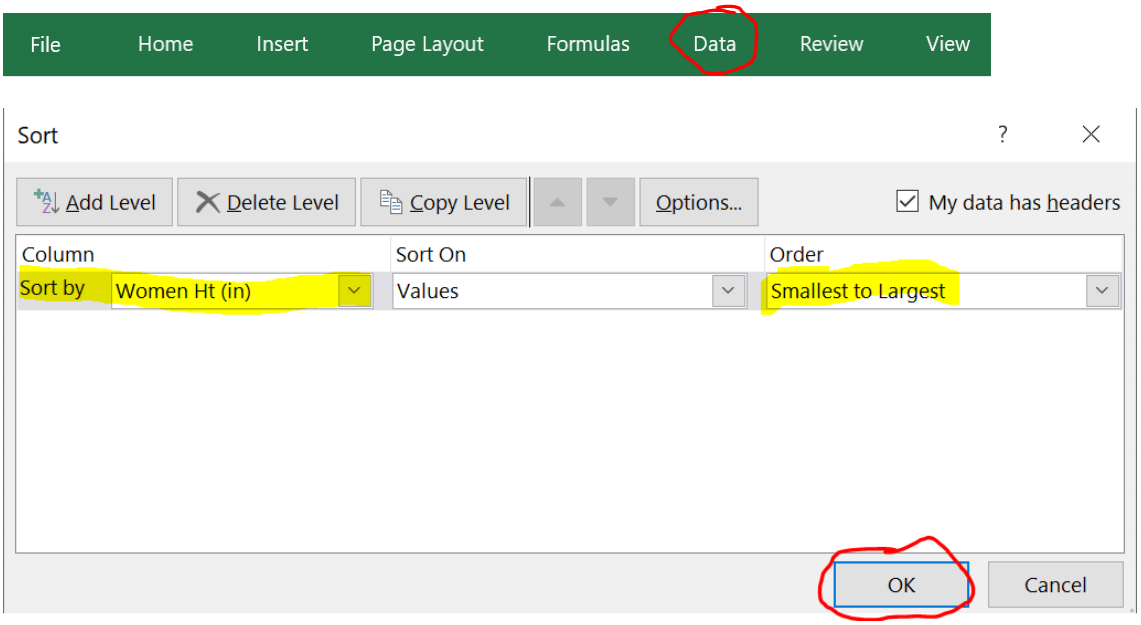
After looking at a lot of bell shaped data sets over the years, statisticians found that usually about 68% of the data values fall within one standard deviation of the mean. This means that in a bell shaped data set, approximately the middle 68% of the values are considered typical. Since this seemed to be the case for most bell shaped data sets, it is often referred to as the "Empirical Rule". The more bell shaped the data set is the more accurate the 68% is. The Empirical Rule does not apply to skewed data sets.

- **Typical Values in normal (bell shaped) data sets make up the middle 68% of the data values and are within ONE standard deviation from the mean.**

Example

In our last example, we saw that typical women in the health data had a height between 60.454 inches and 65.936 inches. If we look put the column of data in order from smallest to largest, we can identify which heights were considered "typical".

To put a data set in order in Excel, first highlight the entire column. Then click on the "data" tab, then click on "sort". You should see the data set you want to sort under "sort by" and under "order", you should see "Smallest to Largest". Now just push "OK".



Now that we have the data in order, we can identify the typical values in the data. The typical heights will be between 60.454 inches and 65.936 inches.

A	B
Women Ht (in)	
57	
58.2	
58.6	
59.6	
59.8	
60.2	
60.5	Typical
60.6	Typical
60.7	Typical
61.3	Typical
61.4	Typical
61.8	Typical
61.9	Typical
62.3	Typical
62.3	Typical
62.6	Typical
62.7	Typical
63.1	Typical
63.2	Typical
63.3	Typical
63.4	Typical
63.4	Typical
63.4	Typical
63.5	Typical
63.6	Typical
64.1	Typical
64.3	Typical
64.3	Typical
64.7	Typical
64.8	Typical
65	Typical
65.1	Typical
66.4	
66.7	
66.7	
66.8	
67	
67.6	
67.9	
68	

Notice that the middle 26 heights (65%) out of 40 total women were typical. This data was not perfectly normal, so we are not surprised it is not exactly 68%. Still it is close to what the empirical rule predicts.

Calculating Standard Deviation

As I said earlier, no one calculates standard deviation by hand. Always use a computer. I will show you the formula and calculation so that you can get a sense of what the computer is doing.

Let us look at the brick weight data from the previous section.

4.7 , 6.2 , 3.3 , 5.1 , 2.9 , 7.4 , 4.5

The standard deviation is the typical distance from the mean, so when calculating the standard deviation you need to know how many numbers are in the data set (seven) and you need to know the mean average.

Mean Average = (4.7 + 6.2 + 3.3 + 5.1 + 2.9 + 7.4 + 4.5) / 7 = 34.1 / 7 = 4.871428571 ≈ 4.87

I will be using the rounded value of the mean. Computers are always much more accurate since they carry many decimal places of accuracy.

Let us look at how far are each of these numbers from the mean? We will subtract the mean from each number in the data set $\left(x-\overline{x}\right)$. Remember we rounded the mean, so these are just approximate distances.

6.2 – 4.87 ≈ 1.33

5.1 – 4.87 ≈ 0.23

7.4 – 4.87 ≈ 2.53

2.9 – 4.87 ≈ -1.97

3.3 – 4.87 ≈ -1.57

4.5 – 4.87 ≈ -0.37

4.7 – 4.87 ≈ -0.17

Notice that some of the differences are negative and some are positive. In fact, if we were to add the distances now, they would add up to approximately zero. (Remember the mean is the balancing point.)

The negative numbers are a problem. To average the distances we need to get rid of the negatives. There are two ways to deal with negative numbers in mathematics, absolute value or squaring the numbers. Absolute value can have issues with calculus applications, so early statisticians preferred to square all the numbers and then eventually take a square root.

Squares of the distances

(1.33)^2 ≈ 1.7689

(0.23)^2 ≈ 0.0529

(2.53)^2 ≈ 6.4009

(-1.97)^2 ≈ 3.8809

(-1.57)^2 ≈ 2.4649

(-0.37)^2 ≈ 0.1369

(-0.17)^2 ≈ 0.0289

Now we will add up all the squared distances and calculate the "Sum of Squares" $\sum\left(x-\overline{x}\right)^2$. This is a very important technique in statistics and occurs in many different applications.

Sum of Squares ≈ 1.7689 + 0.0529 + 6.4009 + 3.8809 + 2.4649 + 0.1369 + 0.0289 ≈ 14.6814

We now want to take an average of the sum of squares. When dealing with spread, we will divide by one less than the sample size. This is often called "degrees of freedom" in statistics. Therefore, we will divide by n-1 instead of the frequency n. There are seven numbers in the data set, so we will divide by 7 – 1 or 6. Then we will take the square root of the answer.

Standard Deviation Formula

$$\sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}} = \sqrt{\frac{14.6814}{(7-1)}} = \sqrt{\frac{14.6814}{6}} = \sqrt{2.4469} \approx 1.564 \text{ kilograms}$$

We calculated the standard deviation with StatKey and got approximately 1.567 kg. StatKey is more accurate since it has less rounding error.

Degrees of Freedom

Why do we divide by n-1 when calculating the standard deviation? That is a good question. Think of it this way. Suppose you take a history class and your grade is based on six exams. The first five exams can have some variability. Maybe you got an 88 on the first exam, a 93 on the second exam, and so on. You want to get a 90 overall mean average to get an A in the history class. Therefore, once you know your first five exam scores, you can calculate what you need to get on the last exam to get an A in class. In other words, the last exam score is fixed in the sense that we can calculate it.

That is how degrees of freedom works. If we have a given mean average, n-1 of the numbers have variability from that mean, but the last number can be calculated. Therefore, if we have the heights of forty women and we know the mean average, then we should only measure the variability of 39 of those numbers.

What is important?

If a quantitative data set is normal (bell shaped), use the mean as the center or average. Use the standard deviation as the best measure of spread. Do not calculate these with formula and calculator. Use a computer program like StatKey.

Remember focus on interpretation not calculation. You should be able to explain the mean and standard deviation for a data set to someone. You should also be able to use the mean and standard deviation to calculate the typical values by adding and subtracting the mean and standard deviation.

Key: The mean and standard deviation should only be used if the data set is normal. They are not accurate if the data set is skewed or non-normal.

Practice Problems Section 4D

1. What is the definition of standard deviation?
2. For what shape is the standard deviation an accurate measure of typical spread?
3. For what shapes is the standard deviation not an accurate measure of typical spread?
4. How can we use the mean and standard deviation to identify typical values in a normally distributed (bell shaped) data set?
5. How many standard deviations from the mean is considered typical for normally distributed (bell-shaped) data?
6. What percentage of the data values are considered typical for normally distributed (bell shaped) data?

Directions #7-9: Fill out the following tables in order to calculate the Standard Deviation (s) for each of the following data sets. The mean average has already been calculated for you. To calculate standard deviation (s), subtract each number from the mean and square the differences. Then add up the squared differences (sum of squares). Then divide by the degrees of freedom (n-1). Last, take the square root of the answer.

(Mean) $\overline{x} = \dfrac{\sum x}{n}$ (Standard Deviation) $s = \sqrt{\dfrac{\sum (x-\overline{x})^2}{n-1}}$

7. 1, 2, 3, 11, 12, 13 Mean $(\bar{x})$ = 7

Values in data set (x)	Differences: Each Value – mean ($x-\overline{x}$)	Squares of Differences $(x-\overline{x})^2$
1		
2		
3		
11		
12		
13		

Sum of squares $\sum (x-\overline{x})^2$ =

Sample Size (How many numbers in the data set) (n) =

Degrees of Freedom (n – 1) =

Standard Deviation (s) =

8. 2, 5, 6, 7, 17, 18, 19, 22 Mean $(\bar{x})$ = 12

Values in data set (x)	Differences: Each Value – mean ($x-\overline{x}$)	Squares of Differences $\left(x-\overline{x}\right)^2$
2		
5		
6		
7		
17		
18		
19		
22		

Sum of squares $\sum\left(x-\overline{x}\right)^2$ =

Frequency (How many numbers in the data set) (n) =

Degrees of Freedom (n – 1) =

Standard Deviation (s) =

9. 5 , 8, 14 , 21 , 30 , 35 , 41 Mean $(\bar{x})$ = 22

Values in data set (x)	Differences: Each Value – mean ($x-\overline{x}$)	Squares of Differences $\left(x-\overline{x}\right)^2$
5		
8		
14		
21		
30		
35		
41		

Sum of squares $\sum\left(x-\overline{x}\right)^2$ =

Sample Size (How many numbers in the data set) (n) =

Degrees of Freedom (n – 1) =

Standard Deviation (s) =

(Directions #10-11): The following histogram and statistics were calculated using the "Bear" data and Statcato. Use the histogram and statistics provided to answer the following questions.

10. Bear neck circumference (inches)

a) What is the data measuring?
b) What are the units?
c) How many numbers are in the data set?
d) Is the data set normally distributed (bell shaped)? (Yes or No)
e) Is the mean an accurate average in this data? Why or why not?
f) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
g) Write a sentence to explain the standard deviation in this context.
h) Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

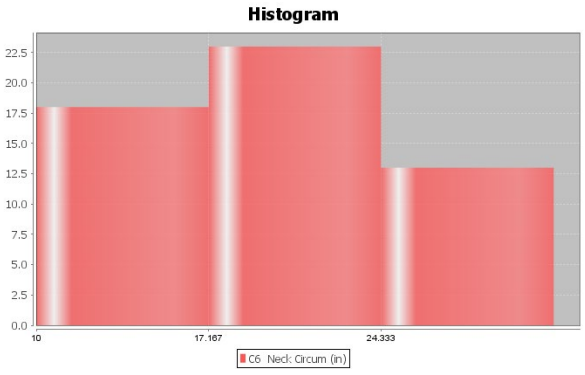
Mean – Standard Deviation ≤ Typical Values ≤ Mean + Standard Deviation

Descriptive Statistics

Variable	Mean	Standard Deviation
C6 Neck Circum (in)	20.556	5.641

Variable	Min	Max
C6 Neck Circum (in)	10.0	31.5

Variable	N total
C6 Neck Circum (in)	54



11. Bear Chest Size (inches)

a) What is the data measuring?
b) What are the units?
c) How many numbers are in the data set?
d) Is the data set normally distributed (bell shaped)? (Yes or No)
e) Is the mean an accurate average in this data? Why or why not?
f) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
g) Write a sentence to explain the standard deviation in this context.
h) Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

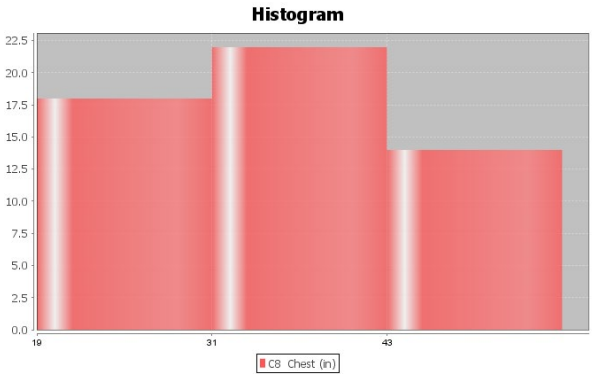
Mean – Standard Deviation ≤ Typical Values ≤ Mean + Standard Deviation

Descriptive Statistics

Variable	Mean	Standard Deviation
C8 Chest (in)	35.663	9.352

Variable	Min	Max
C8 Chest (in)	19.0	55.0

Variable	N total
C8 Chest (in)	54



(Directions #12-15): Open "Health" data in Canvas or at www.matt-teachout.org . Go to www.lock5stat.com and click on StatKey. Then click on "One Quantitative Variable" and "Edit Data". Copy and paste the indicated column of data into StatKey and push OK. Create a histogram with only 3 bars (3 buckets) and verify that the data looks normal. Look at the summary statistics to answer the following questions..

12. Women's Diastolic Blood Pressure (Millimeters of Mercury (mm of Hg))

    a) What is the data measuring?
    b) What are the units?
    c) How many numbers are in the data set?
    d) Is the data set normally distributed (bell shaped)? (Yes or No)
    e) Is the mean an accurate average in this data? Why or why not?
    f) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
    g) Write a sentence to explain the standard deviation in this context.
    h) Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

    Mean – Standard Deviation ≤ Typical Values ≤ Mean + Standard Deviation

13. Women's Wrist Circumference (Inches)

    a) What is the data measuring?
    b) What are the units?
    c) How many numbers are in the data set?
    d) Is the data set normally distributed (bell shaped)? (Yes or No)
    e) Is the mean an accurate average in this data? Why or why not?
    f) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
    g) Write a sentence to explain the standard deviation in this context.
    h) Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

    Mean – Standard Deviation ≤ Typical Values ≤ Mean + Standard Deviation

14. Men's Height (Inches)

    a) What is the data measuring?
    b) What are the units?
    c) How many numbers are in the data set?
    d) Is the data set normally distributed (bell shaped)? (Yes or No)
    e) Is the mean an accurate average in this data? Why or why not?
    f) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
    g) Write a sentence to explain the standard deviation in this context.
    h) Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

    Mean – Standard Deviation ≤ Typical Values ≤ Mean + Standard Deviation

15. Men's Weight (Pounds)

a) What is the data measuring?
b) What are the units?
c) How many numbers are in the data set?
d) Is the data set normally distributed (bell shaped)? (Yes or No)
e) Is the mean an accurate average in this data? Why or why not?
f) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
g) Write a sentence to explain the standard deviation in this context.
h) Use the mean, standard deviation and the following formula to calculate two numbers that typical values in this data fall in between.

Mean – Standard Deviation ≤ Typical Values ≤ Mean + Standard Deviation

---

Section 4E – Unusual Values in Normal Data, Using the Dot Plot, and Summarizing Quantitative Data

In this section, we will try to summarize how to analyze a normal (bell shaped) quantitative data set. When analyzing a quantitative data set there are a few key things to address.

Quantitative Data Analysis Summary

- What is the data measuring? What are the units?
- How many numbers are in the data set? (Sample Size "n")
- What is the shape of the data? Is the data bell shaped (normal), skewed right (long right tail), or skewed left (long left tail). (In this section, all of the data sets are normal.)
- What is the best measure of center? This will be the average. (If the data is normal (bell shaped), these should both be the mean average. Write a sentence to explain the mean average.
- What is the best measure of spread? (If the data set is normal (bell shaped), this should be the standard deviation. Write a sentence to explain the standard deviation.)
- Find two numbers that typical values fall in between. If the data is normal (bell shaped), then we should add and subtract the mean and standard deviation to calculate the two numbers.
  Mean – Standard Deviation ≤ Typical Values in Data ≤ Mean + Standard Deviation
- Find any unusual values in the data set. (Some call these unusual values "outliers".)
- I usually like to give the smallest and largest numbers in the data set, even if they are not unusual.

Finding Outliers (Unusual Values) in a Bell Shaped (Normal) Data Set

So how do you find unusual values in a normal (bell shaped) data set? Unusual values are often called "outliers" in statistics. It has long been considered that one standard deviation from the mean is considered typical. Any values that are two standard deviations or more from the mean is considered unusual. So any value in the data that is two standard deviations or more above the mean is considered "unusually high" or a "high outlier". Any value in the data that is two standard deviations or less below the mean is considered "unusually low" or a "low outlier".

Remember these rules only apply when data is normal (bell shaped). If the data is skewed right or skewed left, these rules do not apply.

High Outlier Cutoff (for normal quantitative data)**:** ***mean + (2 x Standard Deviation)***

Low Outlier Cutoff (for normal quantitative data)**:** ***mean – (2 x Standard Deviation)***

When calculating the cut off, be sure to multiply the standard deviation by 2 before doing the adding or subtracting. Also, be sure to add and subtract in the correct order.

The cutoff's themselves are not necessarily numbers in the data set. Think of them as fences. If a value in the data set is greater than or equal to the unusual high cutoff, then it is considered unusually high (high outlier). If a value in the data set is less than or equal to the unusual low cutoff, then it is considered unusually low (low outlier).

High Outliers (for normal quantitative data) **≥** ***mean + (2 x Standard Deviation)***

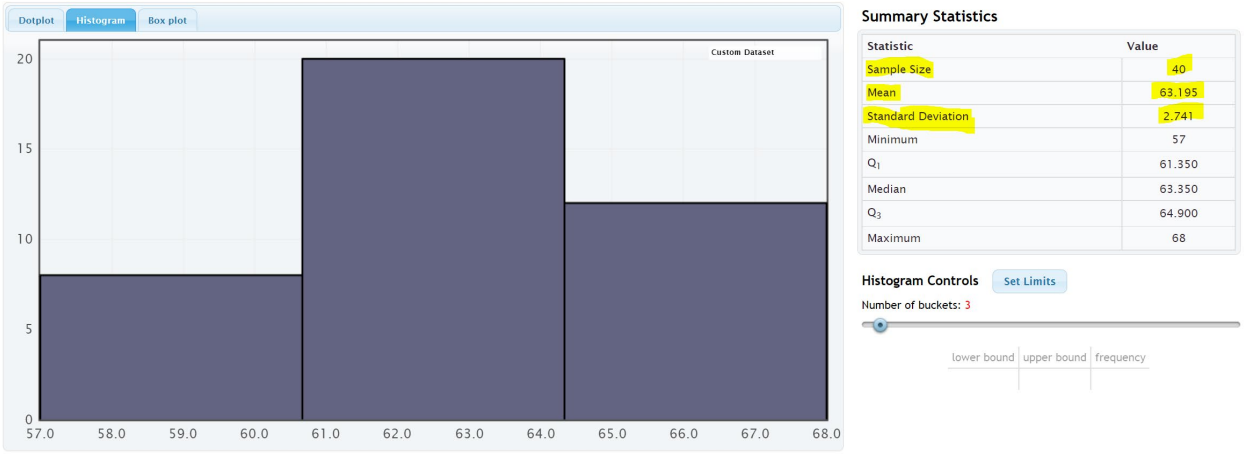
Low Outliers (for normal quantitative data) **≤** ***mean – (2 x Standard Deviation)***

Use the column of data

Once calculating the outlier cutoffs, I find it very easy to simply put the column of data in order from lowest to highest and then identify all of the values that are greater than or equal to the high cutoff and all the values that are less than or equal to the low cutoff.

Example

We used StatKey to calculate the mean average and the standard deviation for the women's heights data. This column of data is located in the "Health" data set. Notice first that the data is normal (bell shaped). That tells us that we should use the Mean as the center (average) and the Standard Deviation as the best measure of typical spread.



Let's see if there are any outliers (unusual values) in this data. Not all data sets have outliers.

High Outlier Cutoff = mean + (2 x standard deviation) = 63.195 + (2 x 2.741) = 63.195 + (5.482) = 68.677

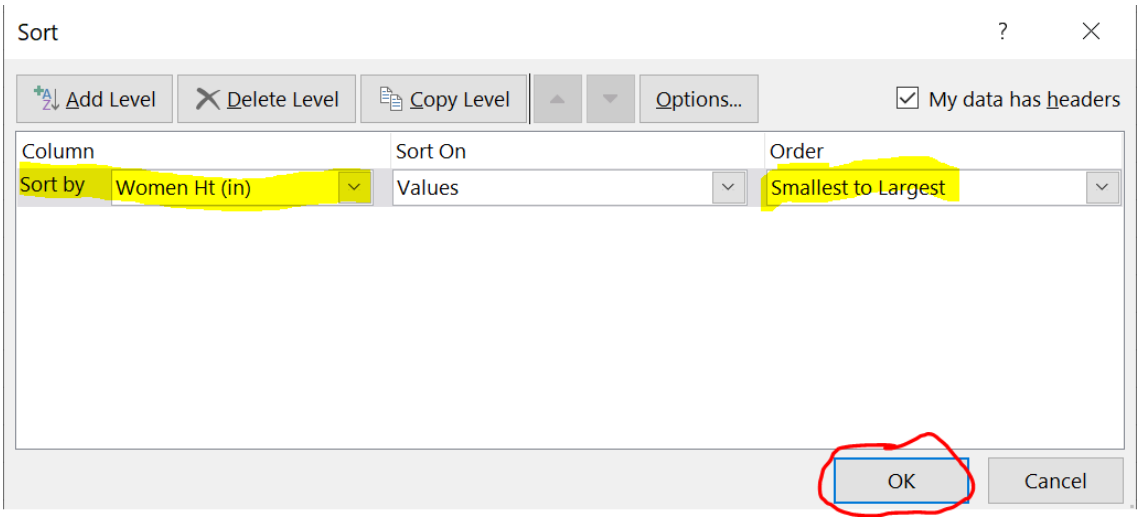
Low Outlier Cutoff = mean − (2 x standard deviation) = 63.195 − (2 x 2.741) = 63.195 − (5.482) = 57.713

So unusually tall women in this data will have a height of 68.677 inches or higher.

So unusually short women in this data will have a height of 57.713 inches or less.

Note: Remember, these cutoffs only apply to women in this data set and do not apply to all women.

To put a data set in order in Excel, first highlight the entire column. Then click on the "data" tab, then click on "sort". You should see the data set you want to sort under "sort by" and under "order", you should see "Smallest to Largest". Now just push "OK".



In the last section we saw that typical value were within one standard deviation from the mean. For the women's height data, all values between 60.454 inches and 65.936 inches were considered typical. We can now identify the unusual values (outliers) as well.

Women Ht (in)	
57	LOW OUTLER!!
	Low Outlier Cutoff (57.713)
58.2	
58.6	
59.6	
59.8	
60.2	
	Typical Values Cutoff (60.454)
60.5	Typcial
60.6	Typcial
60.7	Typcial
61.3	Typcial
61.4	Typcial
61.8	Typcial
61.9	Typcial
62.3	Typcial
62.3	Typcial
62.6	Typcial
62.7	Typcial
63.1	Typcial
63.2	Typcial
63.3	Typcial
63.4	Typcial
63.4	Typcial
63.4	Typcial
63.5	Typcial
63.6	Typcial
64.1	Typcial
64.3	Typcial
64.3	Typcial
64.7	Typcial
64.8	Typcial
65	Typcial
65.1	Typcial
	Typical Values Cutoff (65.936)
66.4	
66.7	
66.7	
66.8	
67	
67.6	
67.9	
68	
	High Outlier Cutoff (68.677)
	NO HIGH OUTLIERS!!

Notice that there is only one number (57) in the data set that is less than or equal to the low outlier cutoff of 57.713. So height is 57 inches is considered unusually low (low outlier) compared to the rest of the data.

Notice also that there are no numbers in the data set that are greater than or equal to the high outlier cutoff of 68.677. So there are no unusually high values in the data set (no high outliers). Notice the tallest woman in the data set was 68 inches tall, but her height is not considered unusual compared to the rest of the women.

Notice also that not all data values are either typical or unusual. There are values in between. These values are not typical and they are not unusual.

Using a dot plot: Once you find the unusual cutoffs, you can also use a dot plot to identify those values in the data set that are unusually high or unusually low. This works ok for smaller data sets, but for larger data sets, I prefer to use the actual column of data. In a data set of 10,000 values for example, you may have 500 outliers. It is hard to pick out 500 dots from a dot plot.

A common questions students ask me is if less than 1 standard deviation or less is considered "typical" and 2 standard deviation or more is considered "unusual", then what about all the values that are in between 1 and 2 standard deviations away from the mean? They are not typical and they are not unusual.

The Empirical Rule discussed in the last section can shed some light on this issue.

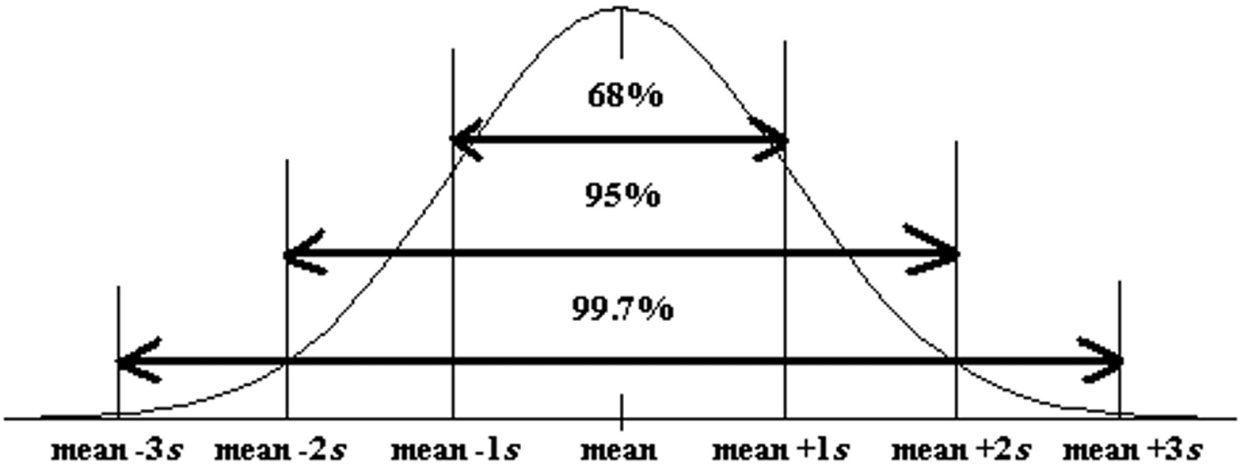
Empirical Rule for Normal (Bell Shaped) Data Sets

After looking at a lot of bell shaped data sets over the years, statisticians found that usually about 68% of the data values fall within one standard deviation of the mean. This means that in a bell shaped data set, approximately the middle 68% of the values are considered typical.

Unusually high values (high outliers) in a normal (bell shaped) data set are in the top 2.5% of the data and usually corresponds to about two standard deviations above the mean or higher. Unusually low values (low outliers) in a normal (bell shaped) data set are in the bottom 2.5% of the data and usually corresponds to about two standard deviations below the mean or less. The middle 95% of a bell shaped data set is not considered unusual.

It turns out that almost all of a bell shaped data set (99.7%) is within three standard deviations of the mean. Remember, the empirical rule percentages are rarely perfectly accurate. The more normal distributed the data set, the more accurate the Empirical Rule percentages. The Empirical Rule does not apply to skewed data sets or non-normal data sets. The following diagram describes the Empirical Rule for normal data. Notice the curve looks like a bell.

<= unusually low ==| |==== typical values ====| |== unusually high =>



Now let us see if we can summarize this chapter. This information is often summarized in a "Data Analysis Report" or a "Data Analysis Paragraph".

- **Normal (normally distributed) quantitative data is unimodal and symmetric and generally has a bell shape. The highest bar in the histogram is in the middle and the right and left tail are about the same length.**
- **For normally distributed data, we will use the mean as our average or center.**
- **For normally distributed data, we will use the standard deviation as our best measure of spread.**
- **For normally distributed data, typical values for normal datawill be within one standard deviation from the mean and will make up the middle 68% of the data.**
- **For normally distributed data, unusually high values in the data (high outliers) will be two or more standard deviations above the mean and make up the top 2.5% of the data.**
- **For normally distributed data, unusually low values in the data (low outliers) will be two or more standard deviations below the mean and make up the bottom 2.5% of the data.**
- **We should also include general information like what the data is measuring, the units, how many people or objects were measured (sample size), and the smallest and largest values in the data set.**

Quantitative Data Analysis Summary Report

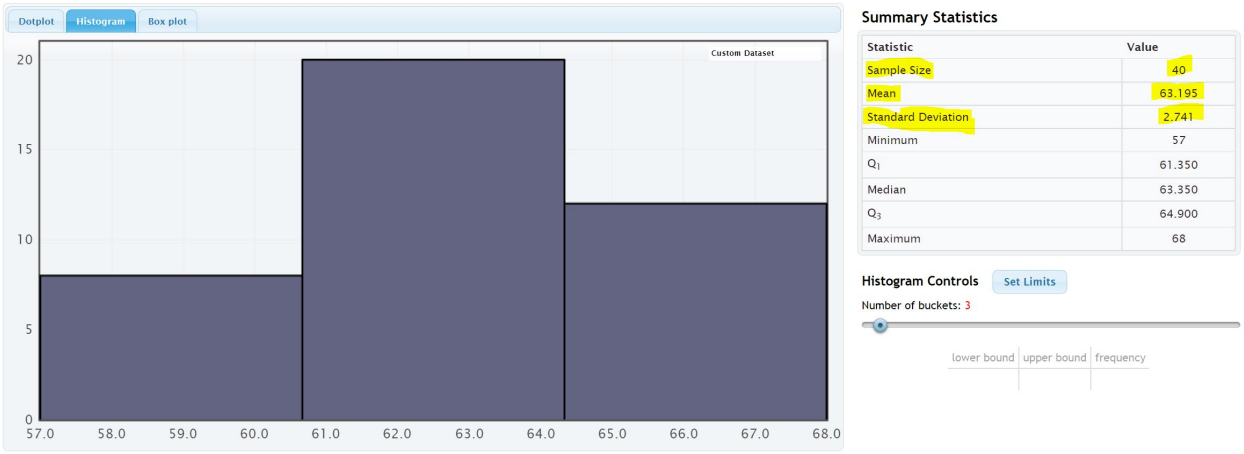
- What is the data measuring? What are the units?
- How many numbers are in the data set? (Frequency "N" or Sample Size)
- What is the shape of the data? (This will be bell shaped in this section.)
- What is the best measure of center? This is the average. (If the data is normal, the center or average should be the mean. Write a sentence to explain the mean average.)
- What is the best measure of spread? (If the data set is normal, this should be the standard deviation. Write a sentence to explain the standard deviation.)

- Find two numbers that typical values fall in between. If the data is normal then we should add and subtract the mean and standard deviation and look for data values in between.
  Mean – Standard Deviation ≤ Typical Data Values ≤ Mean + Standard Deviation
- Find any unusual values (high outliers) and unusually low values (low outliers) in the data set. Calculate the high outlier cutoff and the low outlier cutoff. Put the column of data in order from smallest to largest. Look for values in the column that are greater than or equal to the high cutoff or less than or equal to the low cutoff.
  Unusual High Cutoff: *mean + (2 x Standard Deviation)*
  Unusual Low Cutoff: *mean – (2 x Standard Deviation)*

Example

Analyze the women's heights data located in the "Health" data.

First we need to put the data into StatKey and create a histogram and find the summary statistics.



Notice that the women's height data has a relatively normal shape. The highest bar is in the middle and the left and right tails are about the same length.

We see that the sample size is 40. So the data describes the heights of 40 women.

The shortest woman was 57 inches and the tallest woman was 68 inches.

Since the data is normally distributed, the best center (average) is the mean of 63.195 inches. The average and the balancing point for this data is 63.195 inches.

Since the data is normally distributed, the best measure of spread is the standard deviation of 2.741 inches. So typical values in the data are within 2.741 inches from the mean.

Typical values in this data are between (63.195 – 2.741) and (63.195 + 2.741). So typical heights are between 60.454 inches and 65.936 inches.

Unusually high values (high outliers) ≥ 63.195 + (2 x 2.741)

Unusually high values (high outliers) ≥ 68.677 inches

Looking at the column of data with data values in order, we see that there are no high outliers in this data set.

Unusually low values (low outliers) ≤ 63.195 − (2 x 2.741)

Unusually low values (low outliers) ≤ 57.713 inches

Looking at the column of data with data values in order, we see that there is only one low outlier. The height of 57 inches is considered unusually low when compared to the other data values.

Writing a Summary Paragraph (Report)
Data analysts often summarize their findings in a paragraph. Think of this as a small report that explains the key features of the quantitative data set. You just need to write a sentence for each part of the summary.

Example Summary Report Paragraph

Women's Height Summary Report Paragraph: *This data describes the heights in inches of 40 women. The histogram showed a bell shape, so this data is considered normal or normally distibuted. The mean average height of the women was 63.195 inches (5 ft. 3 in.). So the center or balancing point for this data was 63.195 inches. The typical spread for this data was 2.741 inches (standard deviation). So typical values in the data were within 2.741 inches from the mean. This means that typical heights for these women were in between 60.454 inches (a little over 5 ft.) and 65.936 inches (about 5 ft. 6 in.). There were no unusually high values (no high outliers). The tallest woman in the data set was 68 inches (5 ft. 8 in), but this was not unusual. The shortest woman in the data set was 57 inches (4 ft. 9 in.). This height was the only outlier in the data. The height of 57 inches was considered to be unusually short (low outlier) comparted to the other women.*

Practice Problems Section 4E

1. What is the definition of the mean average?

2. What is the definition of standard deviation?

3. For what shape is the mean an accurate average (accurate measure of center)?

4. For what shapes is the mean NOT an accurate average (not accurate measure of center)?

5. For what shape is the standard deviation an accurate measure of spread?

6. For what shapes is the standard deviation NOT an accurate measure of spread?

7. High outliers (unusually high values) are how many standard deviations above the mean?

8. What percentage of values in a large normally distributed data set are usually considered high outliers?

9. How can we use the mean and standard deviation to identify unusually high values (high outliers) in a normally distributed data set?

10. Low outliers (unusually high values) are how many standard deviations below the mean?

11. What percentage of values in a large normally distributed data set are usually considered low outliers?

12. How can we use the mean and standard deviation to identify unusually low values (low outliers) in a normally distributed data set?

(Directions #13-14): Open "Bear" data in Canvas or at www.matt-teachout.org . Go to www.lock5stat.com and click on StatKey. Then click on "One Quantitative Variable" and "Edit Data". Copy and paste the indicated column of data into StatKey and push OK. Create a histogram with only 3 bars (3 buckets) and verify that the data looks normal. Use the histogram and the mean and standard deviation from the "summary statistics" printout to answer the following questions.

13. Bear neck circumference (inches)

a) Is the data set normally distributed (bell shaped)? (Yes or No)
b) Is the mean an accurate average in this data? Why or why not?
c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.
High Outlier Cutoff = Mean + (2 x Standard Deviation)
e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff for this data.
High Outlier Cutoff = Mean − (2 x Standard Deviation)
e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high ouliers. If there are none, simply say "no high outliers".
f) Use the column of data that is in order from smallest to largest. List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low ouliers. If there are none, simply say "no low outliers".

14. Bear Chest Size (inches)

a) Is the data set normally distributed (bell shaped)? (Yes or No)
b) Is the mean an accurate average in this data? Why or why not?
c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.
High Outlier Cutoff = Mean + (2 x Standard Deviation)
e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff for this data.
High Outlier Cutoff = Mean − (2 x Standard Deviation)
e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high ouliers. If there are none, simply say "no high outliers".
f) Use the column of data that is in order from smallest to largest. List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low ouliers. If there are none, simply say "no low outliers".

(Directions #15-18): Open "Health" data in Canvas or at www.matt-teachout.org . Go to www.lock5stat.com and click on StatKey. Then click on "One Quantitative Variable" and "Edit Data". Copy and paste the indicated column of data into StatKey and push OK. Create a histogram with only 3 bars (3 buckets) and verify that the data looks normal. Use the histogram and the mean and standard deviation from the "summary statistics" printout to answer the following questions.

15. Women's Diastolic Blood Pressure (Millimeters of Mercury (mm of Hg))

a) Is the data set normally distributed (bell shaped)? (Yes or No)
b) Is the mean an accurate average in this data? Why or why not?
c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.
High Outlier Cutoff = Mean + (2 x Standard Deviation)
e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff for this data.
High Outlier Cutoff = Mean − (2 x Standard Deviation)
e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high ouliers. If there are none, simply say "no high outliers".
f) Use the column of data that is in order from smallest to largest. List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low ouliers. If there are none, simply say "no low outliers".

16. Women's Wrist Circumference (Inches)

a) Is the data set normally distributed (bell shaped)? (Yes or No)
b) Is the mean an accurate average in this data? Why or why not?
c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.
High Outlier Cutoff = Mean + (2 x Standard Deviation)
e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff for this data.
High Outlier Cutoff = Mean − (2 x Standard Deviation)
e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high ouliers. If there are none, simply say "no high outliers".
f) Use the column of data that is in order from smallest to largest.
List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low ouliers. If there are none, simply say "no low outliers".

17. Men's Height (Inches)

a) Is the data set normally distributed (bell shaped)? (Yes or No)
b) Is the mean an accurate average in this data? Why or why not?
c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.
High Outlier Cutoff = Mean + (2 x Standard Deviation)
e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff for this data.
High Outlier Cutoff = Mean − (2 x Standard Deviation)
e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high ouliers. If there are none, simply say "no high outliers".
f) Use the column of data that is in order from smallest to largest.
List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low ouliers. If there are none, simply say "no low outliers".

18. Men's Weight (Pounds)

a) Is the data set normally distributed (bell shaped)? (Yes or No)
b) Is the mean an accurate average in this data? Why or why not?
c) Is the standard deviation an accurate measure of typical spread for this data? Why or why not?
d) Use the mean, standard deviation and the following formula to calculate the high outlier cutoff for this data.
High Outlier Cutoff = Mean + (2 x Standard Deviation)
e) Use the mean, standard deviation and the following formula to calculate the low outlier cutoff for this data.
High Outlier Cutoff = Mean − (2 x Standard Deviation)
e) Copy the column of data into a new spreadsheet. Put the data values in order from smallest to largest. List all the values in the data set that are greater than or equal to the high outlier cutoff. These are the high ouliers. If there are none, simply say "no high outliers".
f) Use the column of data that is in order from smallest to largest.
List all the values in the data set that are less than or equal to the low outlier cutoff. These are the low ouliers. If there are none, simply say "no low outliers".

Directions #19-20: Use the following information to write a quantitative data analysis paragraph. There should be a sentence written for each of the following.

- What is the quantitative data measuring?
- What are the units?
- How many numbers are in the quantitative data set?
- What is the shape of the quantitative data set?
- What is the average?
- What is the spread?
- What are the two values that typical numbers fall in between?
- What are the unusually high values (high outiers) in the quantitative data set?
- What are the unusually low values (low outiers) in the quantitative data set?

19. Quantitative Data: Restaurant Bill amounts in dollars.

Shape: Bell shaped (normal)

Sample Size (n) = 74

Mean = $84.31

Standard Deviation = $13.74

Calculate two numbers that typical values are in between: Mean ± Standard Deviation

Calculate the low outlier cutoff: Mean + (2 x standard deviation)

Calculate the high outlier cutoff: Mean – (2 x standard deviation)

High Outliers: $119.54 , $136.82

Low Outliers: $49.67 , $41.88

20. Quantitative Data: Weights of male lions in pounds

Shape: Bell shaped (normal)

Sample Size (n) = 40

Mean = 452.6 pounds

Standard Deviation = 59.1 pounds

Calculate two numbers that typical values are in between: Mean ± Standard Deviation

Calculate the low outlier cutoff: Mean + (2 x standard deviation)

Calculate the high outlier cutoff: Mean – (2 x standard deviation)

High Outliers: 596.5 pounds

Low Outliers: 324.7 pounds

Chapter 4 Review Sheet

Here is a list of important ideas in this chapter.

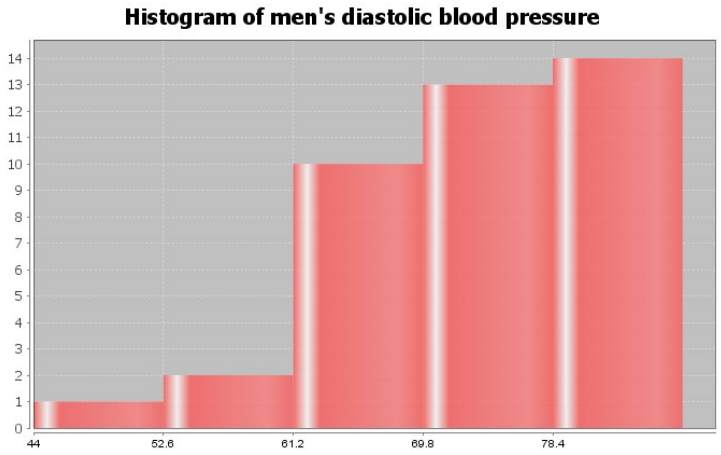
- Be able to distinguish between categorical data and quantitative (numerical measurement) data.
- Be able to create histograms and dot plots with technology and find the shape of a quantitative data set.
- Be able to find the mean, standard deviation, minimum, maximum, frequency (N) with technology.
- A center gives an average value for the data set is usually close to the highest bar or bars in the histogram.
- Statistics that measure center: Mean, Median, Mode and Midrange
- If a data set is bell shaped, we should use the mean average as our measure of center and our average for the data set. If a data set is not bell shaped, we should not use the mean.
- Mean Average definition: A statistic that measures the center or average of a bell shaped data set by balancing the distances.
- A measure of spread or variability tells us how spread out the data set is. The more spread out the data is, the less consistent the data is and the harder it is to predict. A small amount of spread tells us that the data is more consistent and easier to predict.
- Statistics that measure spread (variability): Standard Deviation, Variance, Range, Interquartile Range (IQR)
- If a data set is bell shaped, we should use the standard deviation as our measure of spread for the data set. If a data set is not bell shaped, then we should not use the standard deviation.
- Standard Deviation definition: A measure of spread that tells us how far typical values are from the mean in a bell shaped data set.
- Mean – Standard Deviation ≤ Typical Values ≤ Mean + Standard Deviation
- Unusually High Cutoff: Mean + (2 x Standard Deviation)
- Unusually Low Cutoff: Mean – (2 x Standard Deviation)
- Be able to use the unusual cutoffs and a dot plot to identify unusual values in the data set.
- Be able to write a summary report paragraph summarizing the key characteristics of a bell shaped quantitative data set.

Problems Chapter 4 Review Sheet

Give the shape of each of the following graphs from the men's health data. Then decide if the mean or the median is the most appropriate average for the data set.

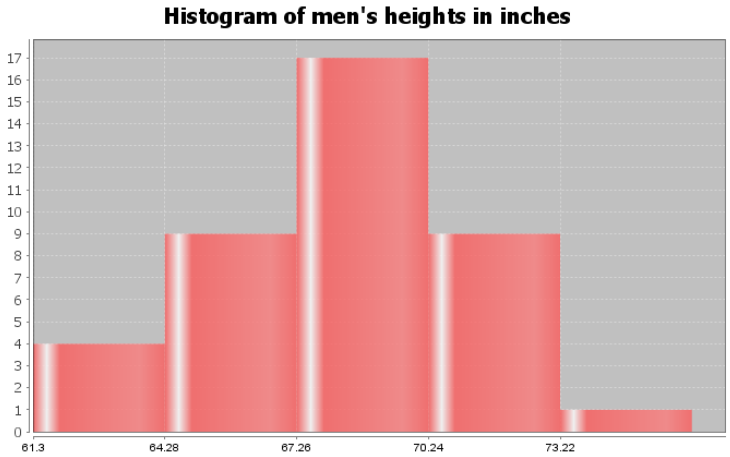
1. Men's Diastolic Blood Pressure

Shape = _____ Mean or Median? _____



2. Men's Heights (inches)

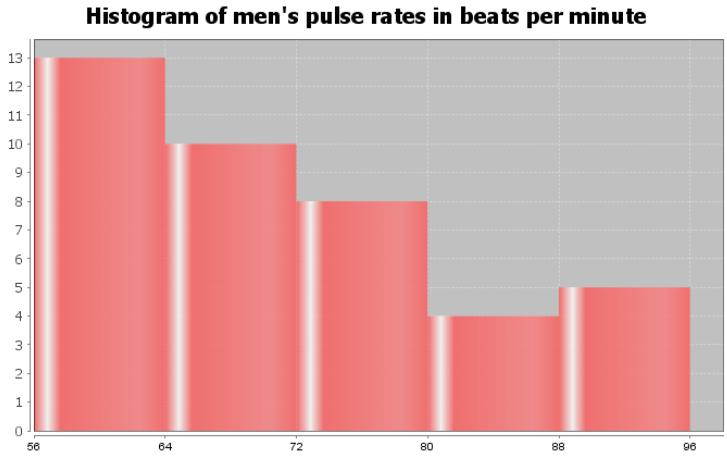
Shape = _____ Mean or Median? _____



3. Men's Pulse Rates (Beats per Minute)

Shape = _____

Mean or Median? _____



4. Calculate the Mean Average for the following data. Round your answer to the hundredths place (two numbers to right of decimal).

$$\overline{x} = \frac{\sum x}{n}$$

12.6	21.8	20.1	16.6
16.7	20.8	11.2	9.0
21.2	12.3	12.9	15.2
25.7			

Mean Average = _____

5. Standard Deviation is an important measure of spread or variability in statistics. Give the basic definition of Standard Deviation.

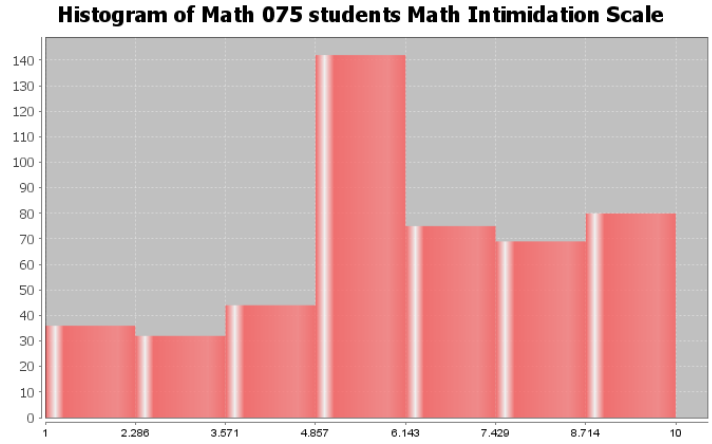
6. How can we tell if the mean and standard deviation are accurate?

7. What percentage of the values in a bell shaped data set are considered typical?

8. What percentage of the values in a bell shaped data set are considered unusually high?

9. What percentage of the values in a bell shaped data set are considered unusually low?

Math 075 Students in the Fall 2015 semester were asked on a scale of one to ten, how intimidated are you about math classes. Here is a histogram, dot plot, mean, standard deviation, frequency, minimum and maximum from Statcato.



Descriptive Statistics

Variable	Mean	Standard Deviation
C15 math intimidation	6.159	2.418

Variable	Min	Max
C15 math intimidation	1.0	10.0

Variable	N total
C15 math intimidation	478

10. What is the shape of the data set? _____

11. How many numbers are in the data set? _____

12. Are the mean and standard deviation accurate for this data? (Yes or No) _____

13. What is the average math intimidation score for the students? (Give a number.)

Average math intimidation score = _____

14. How far are typical values in the data set from the mean on average? (Give a number.)

Average distance from the mean = _____

15. Calculate two numbers that typical values fall in between and put your answer below.

Mean – Standard Deviation ≤ typical math intimidation scores ≤ Mean + Standard Deviation

_____ ≤ typical math intimidation scores ≤ _____

16. What is the cutoff for an unusually high math intimidation score?

Unusual High Cutoff = Mean + (2 x Standard Deviation)

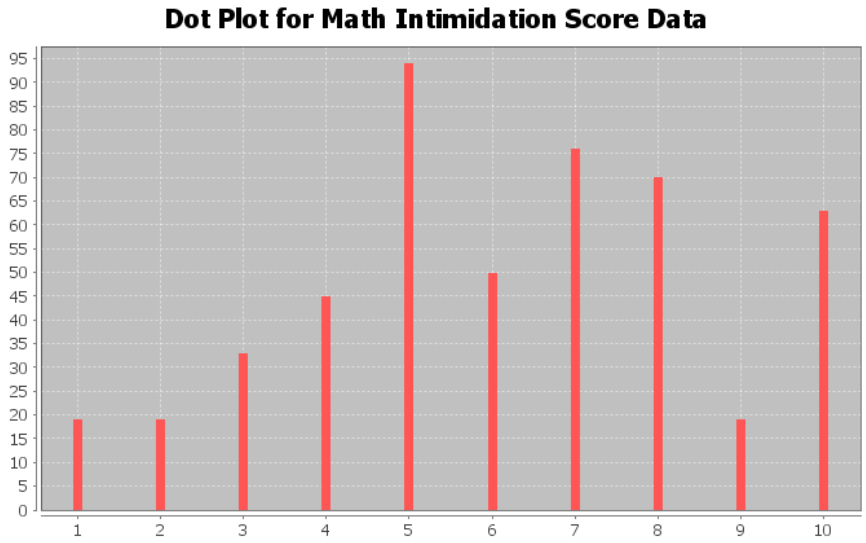
Unusual High Cutoff = _____

17. What is the cutoff for an unusually low math intimidation score?

Unusual Low Cutoff = Mean – (2 x Standard Deviation)

Unusual Low Cutoff = _____

Look at the following Dot Plot for the data and your answers to #16 and #17 to answer the following questions.



18. Are there any unusually high math intimidation scores in the data (yes or no)?

19. If you answered yes to #18, what are the unusually high scores? _____

20. Are there any unusually low math intimidation scores in the data (yes or no)?

21. If you answered yes to #20, what are the unusually low scores? _____

Chapter 4 Project
Normal Quantitative Data Analysis

Directions for Online Classes: *This will be an individual project. Each student will analyze one quantitative data set from the "Math 075 Chapter 4 Project Normal Data" and create a poster summarizing their findings, After submitting a picture of the poster to their instructor, students will then go to the "Chapter 4 Project Class Discussion" in Canvas and discuss their findings with other students in the class.*

Each student will pick one of the following data sets from the Math 075 Chapter 4 Project Normal Data to analyze: Male Body Temp Degrees Fahrenheit, Female Body Temp Degrees Fahrenheit, North Territory Australia Weekly Salary Dollars, Tasmania Australia Weekly Salary Dollars, Chicks Weight Gain (in grams) after 20 days on Normal Corn, January minimum temperature in degrees Fahrenheit of various U.S. Cities, Percent of Female Students at Universities around the world, Salamander Total Length (cm), Fat (grams) Fast Food Breakfast Items, Soil Surface temperature (degrees Celsius) in Comanche, Texas, NBA All-Star Player Heights.

The Individual Poster Should Have

- **First and Last Name of student**
- **Why is this data important or interesting to you?**
- **Go to www.lock5stat.com and open StatKey. Click on "one quantitative variable" under the "descriptive statistics and graphs" menu. Click on "edit data" and copy and paste your one column of quantitative data into Statkey. If you data has a title, click on "data has a header row". Do NOT click the box the says data has identifier. Press OK.**
- **Click on dot plot in StatKey and sketch the dot plot onto your poster.**
- **Click on histogram in StatKey, and pull the slider to "3 buckets". Your histogram should have 3 bars. Sketch the histogram onto your poster.**
- **Write down the Mean, Standard Deviation, Min, Max and Sample Size onto your poster. You will see them under the "Sample Statistics" menu in StatKey.**
- **What is the data measuring?**
- **What are the units?**
- **How many numbers are in the data set : sample size (n)**
- **What is the Shape? Look at your histogram. Should be normal (bell shaped).**
- **Write a sentence to explain the mean.**
- **What is the average? (Use the mean if normal data.)**
- **What is your spread for the data? (Use the standard deviation if normal data.)**
- **Write a sentence to explain the standard deviation.**
- **Find two numbers that typical values fall in between (Mean – Stand Dev , Mean + Stand Dev)**
- **Calculate Unusually high cutoff (Mean + 2 x Stand Dev)**
- **List all unusually high values (high outliers) in the data set. (Find these on the dot plot or excel spreadsheet.) If there are none, say "No high outliers".**
- **Calculate Unusually low cutoff (Mean – 2 x Stand Dev)**
- **List all unusually low values (low outliers) in the data set. (Find these on the dot plot or excel spreadsheet.). If there are none, say "No low outliers".**
- **Decorate Poster**

Now take a picture of your poster project and submit the picture to your instructor in Canvas.

After submitting the picture of the poster, go to the discussion menu in Canvas and complete the "Chapter 4 Project Discussion". You will be discussing your findings with other students in the class.

Directions for Face to Face Classes: *The class will be separated into groups. Each group is required to pick a "team name" for their group and analyze one quantitative data set from the "Math 075 Chapter 4 Project Normal Data", create a poster summarizing their findings, and present the poster to other students in the class.*

Each group will have a different topic and will pick one of the following data sets from the Math 075 Chapter 4 Project Normal Data to present it to their classmates: Male Body Temp Degrees Fahrenheit, Female Body Temp Degrees Fahrenheit, North Territory Australia Weekly Salary Dollars, Tasmania Australia Weekly Salary Dollars, Chicks Weight Gain (in grams) after 20 days on Normal Corn, January minimum temperature in degrees Fahrenheit of various U.S. Cities, Percent of Female Students at Universities around the world, Salamander Total Length (cm), Fat (grams) Fast Food Breakfast Items, Soil Surface temperature (degrees Celsius) in Comanche, Texas, NBA All-Star Player Heights.

The Poster Should Have

- **Group/Team Name**
- **First and Last Name of each team members on the poster**
- **Why is this data important or interesting to your group?**
- **Go to www.lock5stat.com and open StatKey. Click on "one quantitative variable" under the "descriptive statistics and graphs" menu. Click on "edit data" and copy and paste your one column of quantitative data into Statkey. If you data has a title, click on "data has a header row". Do NOT click the box the says data has identifier. Press OK.**
- **Click on dot plot in StatKey and sketch the dot plot onto your poster.**
- **Click on histogram in StatKey, and pull the slider to "3 buckets". Your histogram should have 3 bars. Sketch the histogram onto your poster.**
- **Write down the Mean, Standard Deviation, Min, Max and Sample Size onto your poster. You will see them under the "Sample Statistics" menu in StatKey.**
- **What is the data measuring?**
- **What are the units?**
- **How many numbers are in the data set : sample size (n)**
- **What is the Shape? Look at your histogram. Should be normal (bell shaped).**
- **Write a sentence to explain the mean.**
- **What is the average? (Use the mean if normal data.)**
- **What is your spread for the data? (Use the standard deviation if normal data.)**
- **Write a sentence to explain the standard deviation.**
- **Find two numbers that typical values fall in between (Mean – Stand Dev , Mean + Stand Dev)**
- **Calculate Unusually high cutoff (Mean + 2 x Stand Dev)**
- **List all unusually high values (high outliers) in the data set. (Find these on the dot plot or excel spreadsheet.) If there are none, say "No high outliers".**
- **Calculate Unusually low cutoff (Mean – 2 x Stand Dev)**
- **List all unusually low values (low outliers) in the data set. (Find these on the dot plot or excel spreadsheet.). If there are none, say "No low outliers".**
- **Decorate Poster**

Presentation

Make sure each person on the team understands the poster and can present your findings. Bring your poster to a designated presentation area in the classroom and hang or tape your poster to a wall. One person at a time will present the poster. We will then rotate so that each member of the team gets to present. Everyone else will listen to presentations and give feedback.