

Chapter 3 – Analyzing Normal Quantitative Data

Introduction: In chapters 1 and 2, 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.

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, unusual values. It is a very different approach than if the data was categorical.



Section 3A – Finding the Shape of a Quantitative Data Set 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 and unusual values. In statistics, there are many types of averages, many types of spreads. Shape helps us determine which averages and spreads 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 data set.

To create a dot plot in Statcato, you first need a quantitative data set. Open the Health Data. This data describes the medical statistics of forty women and forty men. It includes the ages, heights, weights, blood pressure, cholesterol, and body mass index (BMI).

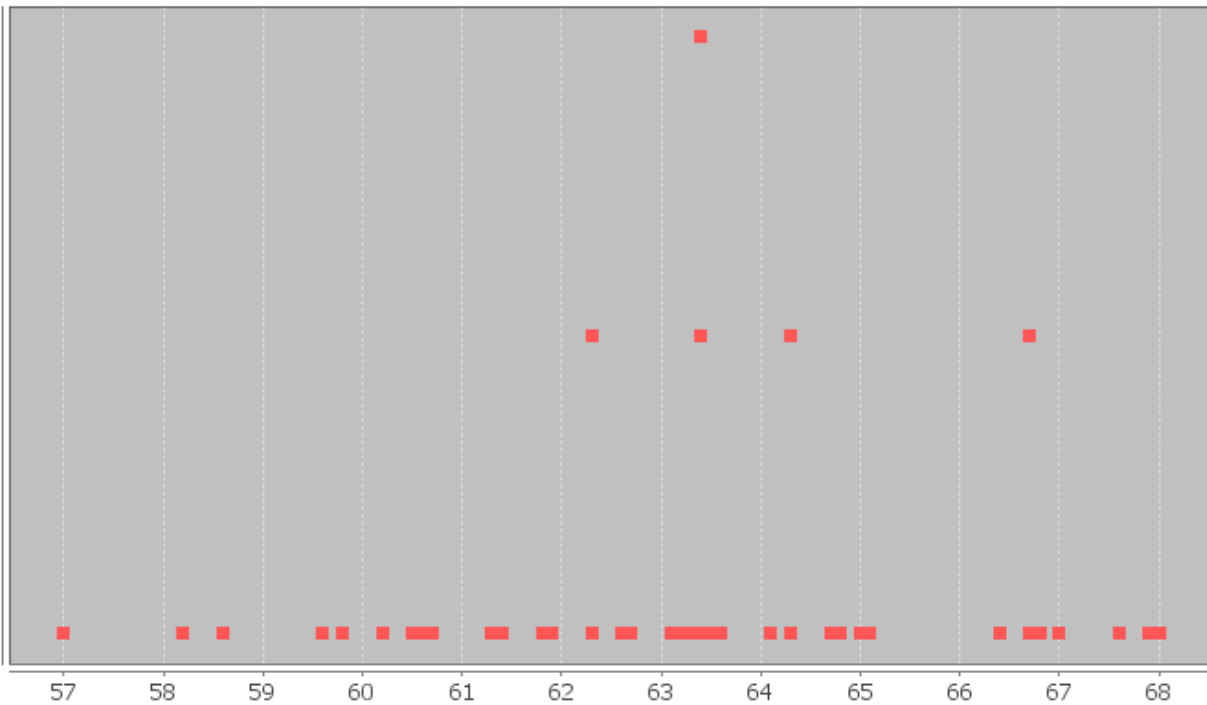
Copy and paste women's heights into a column of Statcato. The data set is only 40 values, so you will not need to add rows to Statcato. Notice the data is quantitative. It measures the height in inches of the women and it seems reasonable to look for an average height of these women.

To make a dot plot, go to the graph menu and click on dot plot. Then click on the column of data you want to use. Then push ok.

Making a dot plot in Statcato: *Graph => Dot plot => Pick a column => OK*

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

Dot Plot Women's Heights (Inches)



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.

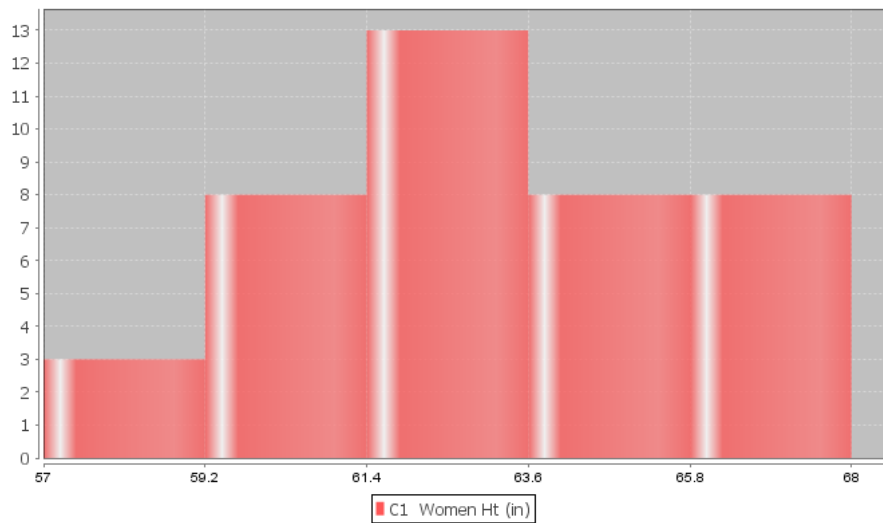
When determining shape, it is better to make a histogram. Think of a histogram as breaking the scale up into sections and counting how many dots are in each section. Then drawing a bar that represents the number of dots in that section (frequency).

To make a histogram in Statcato, go to the graph menu, and then click on histogram. Choose a column of data and how many bars (bins) you want. Then choose ok.

Making a histogram in Statcato: *Graph => Histogram => Pick a column => Chose number of bins => OK*

Note about bins: If you chose too many bars then the histogram starts to look very crazy and you will have a hard time seeing the shape. Remember the goal is to break the dots up into groups. For example, in this health data there are only 40 women. I would not want 40 bins since that would give me about one bar per dot. If it were a small data set like the health data, I would do about five bins. Remember, the more bins you have, the more difficult it is to see the shape.

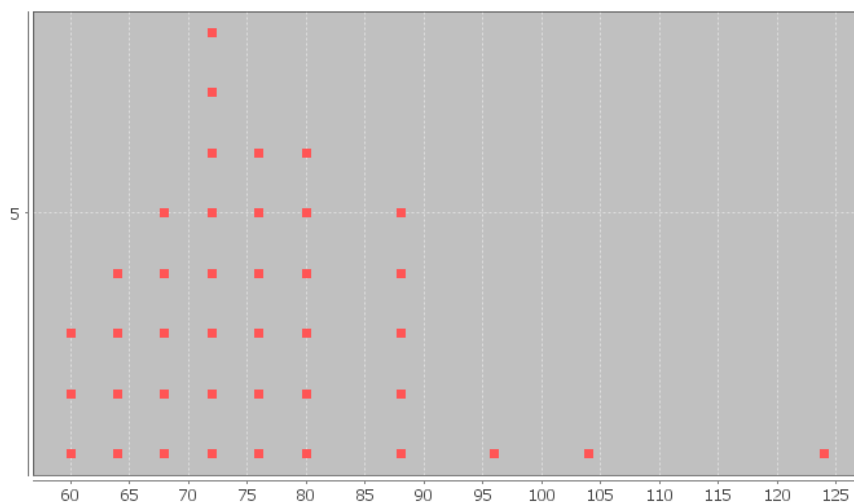
Histogram of Women's Height (Inches)



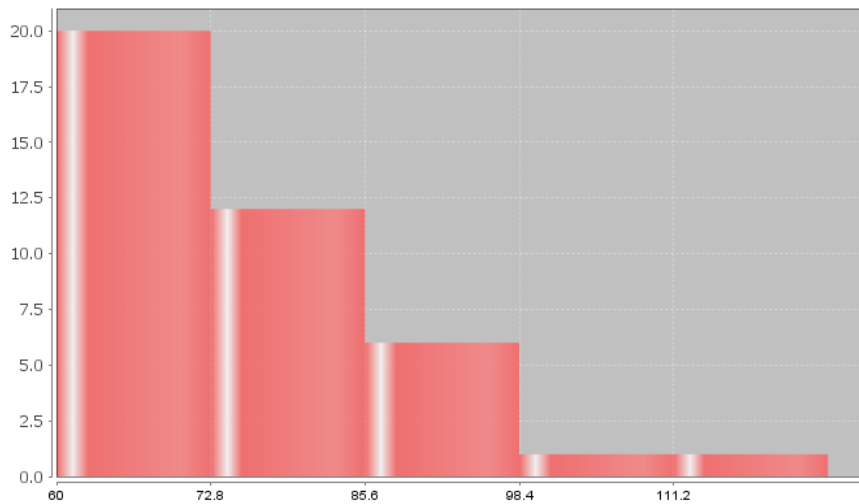
This is a very important shape in statistics. Notice the highest bar is close to the middle and the bars get smaller as we move away from the middle. This is often called “Bell Shaped” or “Normal Data”. Some like to describe this shape as unimodal (1 hill) and symmetric (left and right side look about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

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). If we copy and paste the women’s pulse data into Statcato, we can create the following dot plot and histogram. I used five bins again for the histogram.

Dot Plot of Women's Pulse Rates (Beats Per Minute)



Histogram of Women's Pulse Rates (Beats Per Minute)

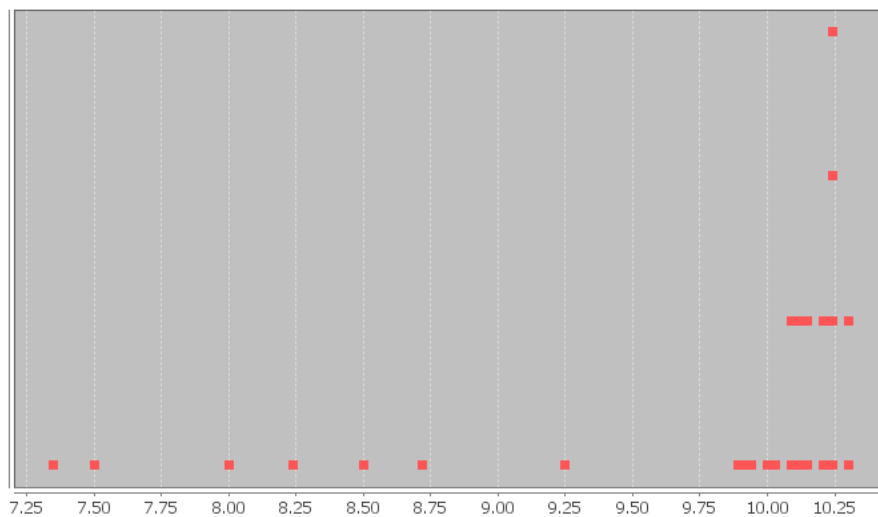


Notice this has a very different shape. There are more dots in the dot plot congregated on the far left. The highest bar in the histogram is on the far left and there are more bars to the right of the highest bar. There is a long tail to the right of the highest bar. This is called “Skewed Right”. Some 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 bell shaped.

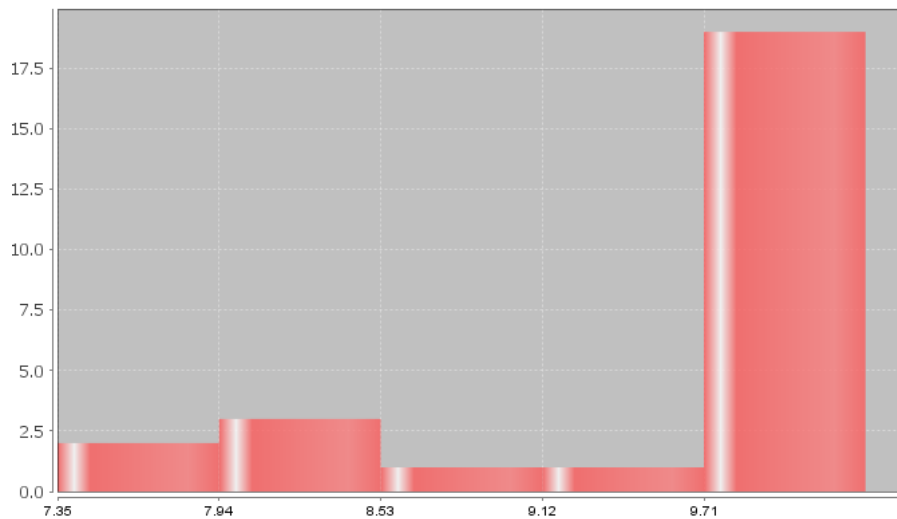
Let us look at another example.

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.

Dot Plot of Salary in Dollars per Hour



Histogram of Salary in \$ per hour



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.

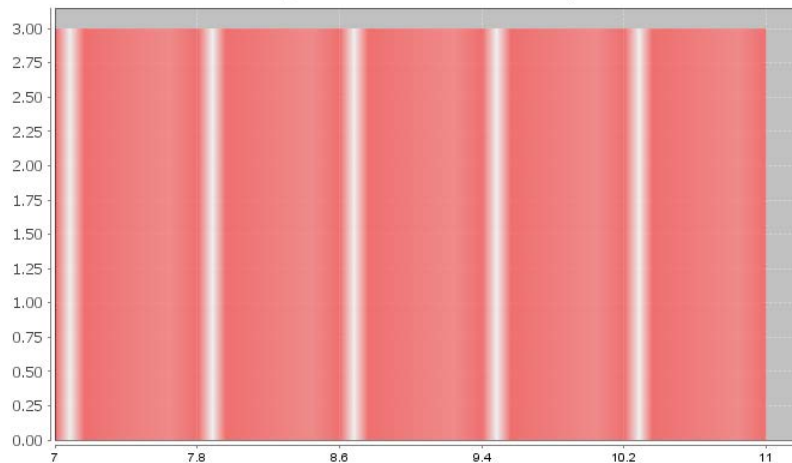
Note: Real data rarely has a perfect shape. Most data has a shape somewhere in between bell shaped and skewed, and you will need to make a decision. Look for a significant difference in the length of the tail to classify something as skewed. If my highest hill is toward the middle and I had 2 bars to 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.

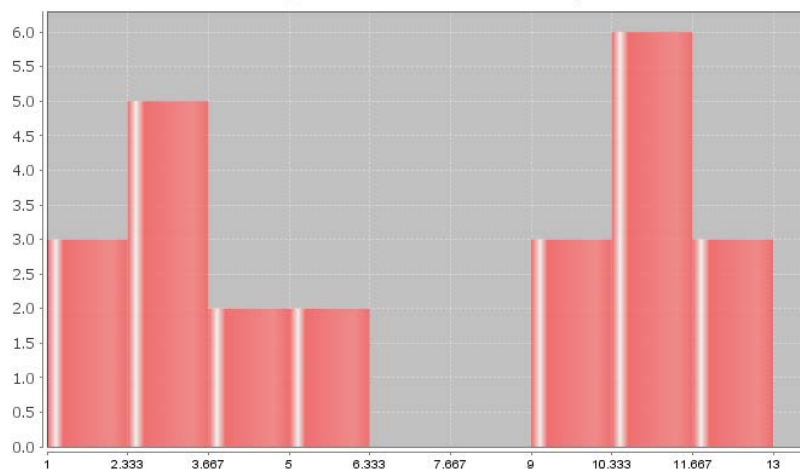
Here are a couple unusual shapes that sometimes appear.

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

Histogram with Uniform Shape



Histogram with Bimodal Shape



Problem Set Section 3A

Directions: Open the men and women's health data on my website www.matt-teachout.org Look under the "Int Alg for Stats" tab and then "data sets".

1. Use a Statistics software to create a dot plot and histogram of men's ages in years. Be sure to adjust the number of bins if necessary. Draw a rough sketch of the graphs on a sheet of paper or save the graphs in a word document. What is the shape of the data?

2. Use a Statistics software to create a dot plot and histogram of women's ages in years. Be sure to adjust the number of bins if necessary. Draw a rough sketch of the graphs on a sheet of paper or save the graphs in a word document. What is the shape of the data?

3. Use a Statistics software to create a dot plot and histogram of men's weight in pounds. Be sure to adjust the number of bins if necessary. Draw a rough sketch of the graphs on a sheet of paper or save the graphs in a word document. What is the shape of the data?

4. Use a Statistics software to create a dot plot and histogram of women's weight in pounds. Be sure to adjust the number of bins if necessary. Draw a rough sketch of the graphs on a sheet of paper or save the graphs in a word document. What is the shape of the data?

5. Use a Statistics software to create a dot plot and histogram of men's waist size in centimeters. Be sure to adjust the number of bins if necessary. Draw a rough sketch of the graphs on a sheet of paper or save the graphs in a word document. What is the shape of the data?

6. Use a Statistics software to create a dot plot and histogram of women's waist size in centimeters. Be sure to adjust the number of bins if necessary. Draw a rough sketch of the graphs on a sheet of paper or save the graphs in a word document. What is the shape of the data?

7. Use a Statistics software to create a dot plot and histogram of men's body mass index (BMI) in kilograms per meters squared. Be sure to adjust the number of bins if necessary. Draw a rough sketch of the graphs on a sheet of paper or save the graphs in a word document. What is the shape of the data?

8. Use a Statistics software to create a dot plot and histogram of women's body mass index (BMI) in kilograms per meters squared. Be sure to adjust the number of bins if necessary. Draw a rough sketch of the graphs on a sheet of paper or save the graphs in a word document. What is the shape of the data?

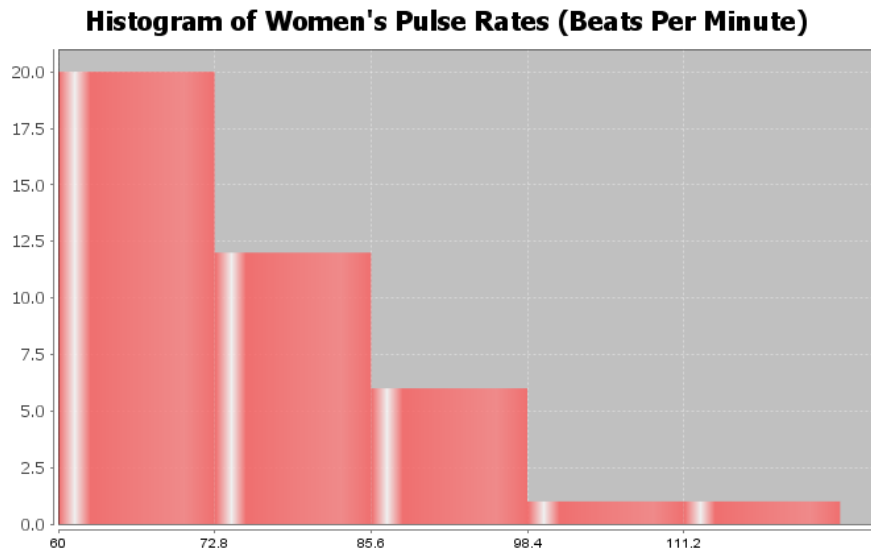
Section 3B – Shapes and Centers

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 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. Commonly used centers include the mean average, median average, mode or midrange. The key is to determine which center (average) is most accurate for the data.

Look at the following histogram describing the pulse rates of 40 women in beats per minute from the health data.



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.

Statistics software like Statcato can calculate these various measures of center or average very quickly. I used Statcato to calculate the following centers for the women's pulse data.

Mean Average = 76.3 beats per minute

Median Average = 74 beats per minute

Mode = 72 beats per minute

Midrange Average = 92 beats per minute

Let us compare these values to the histogram. Notice a few things. The mean and midrange are not very accurate measures of center since they are not close to the highest bar. The median and mode seem to be more accurate, since they are closer to the highest bar, i.e. closer to the center of the data.

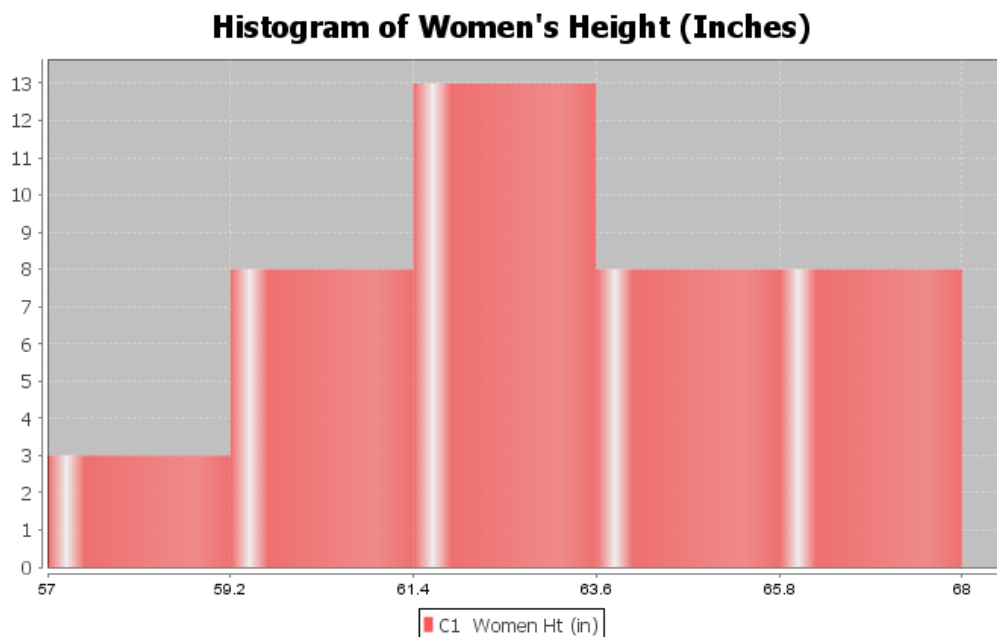
Here are a couple of things to keep in mind when finding an accurate average for a data set. The women's pulse data is skewed right. Mean averages get pulled in the direction of the skew

(long tail) tend to not be very accurate for skewed data sets. The midrange is a quick measure of center that can be calculated easily without a computer, but it is rarely accurate. The mode was accurate. However, the mode is the number that appears most often in the data set and not all data sets have a unique mode. 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 measure of center and we should use the median as the average for the data set.

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



Let us look again at the four measures of center.

Mean Average = 63.195 inches

Median Average = 63.35 inches

Mode = 63.4 inches

Midrange Average = 62.5 inches

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. All of the centers are close to the highest bar. It seems like all of these statistics are pretty accurate centers and any of them would be a decently accurate average for this data.

So which one should we use?

As we have said before, not all data sets have a mode. In addition, the midrange, though accurate in this data set, is not an accurate average most of the time. That leaves us with the median and mean. 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 average or 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 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 not use the mean average unless the data is bell shaped. If the data is not bell shaped then the mean is not accurate.

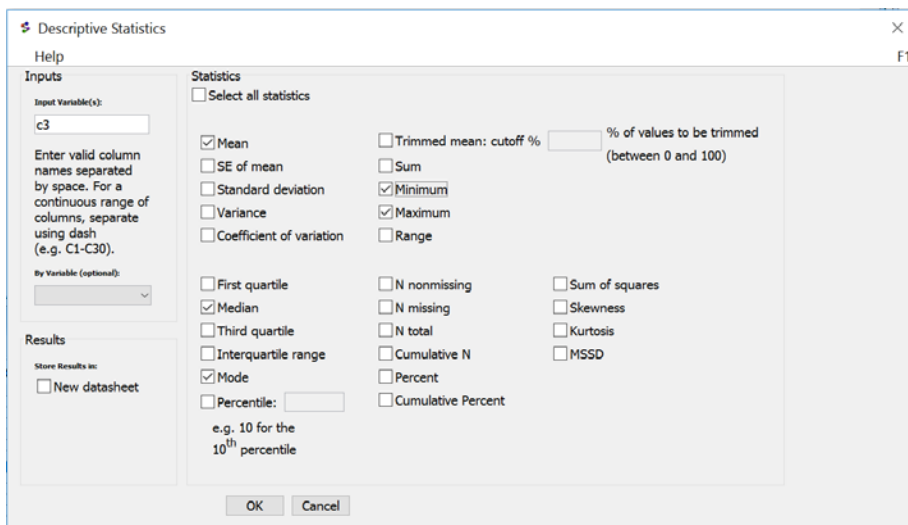
Calculating Centers with Technology

Remember that “statistics” are numbers that can 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.

Like most statistics software programs, Statcato can calculate statistics in fraction of a second. To calculate statistics with Statcato, go to the “statistics” menu, then “basic statistics”, then “descriptive statistics”. Tell the computer what column of data you would like to use. Type in “C1” if it is the first column, and “C2” if it is the second column. Then chose which statistics you would like to calculate. There is a huge list of possibilities. Most of these we have not discussed yet, but notice you can check “mean”, “median”, and “mode”. Midrange is a center not on the list, but if you check the boxes that say “minimum” and “maximum”, you can use these to calculate the midrange.

$$\text{Midrange} = \frac{(\text{Max} + \text{Min})}{2}$$

To calculate statistics with Statcato: *Statistics => Basic Statistics => Descriptive Statistics => Pick a column of data => Pick what statistics you want to calculate => OK*



Problem Set Section 3B

Directions: Open the men and women’s health data on my website www.matt-teachout.org Look under the “Int Alg for Stats” tab and then “data sets”.

1. Use a statistics software to create a histogram of men’s ages in years. Be sure to adjust the number of bins if necessary. You do not need to draw or save the graph. What is the shape of the data? Use the statistics software to calculate the four measures of center (mean, median, mode, and midrange). Based on the shape and rules discussed in this section should we use the mean average as our center or should we use the median average?

2. Use a Statistics software to create a histogram of women's ages in years. Be sure to adjust the number of bins if necessary. You do not need to draw or save the graph. What is the shape of the data? Use the statistics software to calculate the four measures of center (mean, median, mode, and midrange). Based on the shape and rules discussed in this section should we use the mean average as our center or should we use the median average?

3. Use a Statistics software to create a histogram of men's weight in pounds. Be sure to adjust the number of bins if necessary. You do not need to draw or save the graph. What is the shape of the data? Use the statistics software to calculate the four measures of center (mean, median, mode, and midrange). Based on the shape and rules discussed in this section should we use the mean average as our center or should we use the median average?

4. Use a Statistics software to create a histogram of women's weight in pounds. Be sure to adjust the number of bins if necessary. You do not need to draw or save the graph. What is the shape of the data? Use the statistics software to calculate the four measures of center (mean, median, mode, and midrange). Based on the shape and rules discussed in this section should we use the mean average as our center or should we use the median average?

5. Use a Statistics software to create a histogram of men's waist size in centimeters. Be sure to adjust the number of bins if necessary. You do not need to draw or save the graph. What is the shape of the data? Use the statistics software to calculate the four measures of center (mean, median, mode, and midrange). Based on the shape and rules discussed in this section should we use the mean average as our center or should we use the median average?

6. Use a Statistics software to create a histogram of women's waist size in centimeters. Be sure to adjust the number of bins if necessary. You do not need to draw or save the graph. What is the shape of the data? Use the statistics software to calculate the four measures of center (mean, median, mode, and midrange). Based on the shape and rules discussed in this section should we use the mean average as our center or should we use the median average?

7. Use a Statistics software to create a histogram of men's body mass index (BMI) in kilograms per meters squared. Be sure to adjust the number of bins if necessary. You do not need to draw or save the graph. What is the shape of the data? Use the statistics software to calculate the four measures of center (mean, median, mode, and midrange). Based on the shape and rules discussed in this section should we use the mean average as our center or should we use the median average?

8. Use a Statistics software to create a histogram of women's body mass index (BMI) in kilograms per meters squared. Be sure to adjust the number of bins if necessary. You do not need to draw or save the graph. What is the shape of the data? Use the statistics software to calculate the four measures of center (mean, median, mode, and midrange). Based on the shape and rules discussed in this section should we use the mean average as our center or should we use the median average?

Section 3C – 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 bell shaped data and data analysts prefer to use the mean average when data is bell shaped, 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 the data set.

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 analyst.

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

x : each individual number in the data set

Σ : summation symbol (tells us to add)

Σx : add up all the numbers in your data set

\bar{x} : mean average of a data set (sample mean average)

Formula for calculating the mean average

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

(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 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 frequency or sample size.)

Seven

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 = $4.871428571 \approx 4.87$ 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.

$$6.2 - 4.87 \approx 1.33$$

$$5.1 - 4.87 \approx 0.23$$

$$7.4 - 4.87 \approx 2.53$$

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

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

$$2.9 - 4.87 \approx -1.97$$

$$3.3 - 4.87 \approx -1.57$$

$$4.5 - 4.87 \approx -0.37$$

$$4.7 - 4.87 \approx -0.17$$

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

$$-1.97 + -1.57 + -0.37 + -0.17 \approx -4.08$$

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 from 20

16: four from 20

18: two from 20

19: one 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:

$$\text{Mean} = (14 + 16 + 18 + 19 + 33) / 5 = 100 / 5 = 20$$

More Examples

You can create tons of different data sets, if you understand this principle of the balancing point. For example, 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 and 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

Check this data set. Put these numbers into Statcato and calculate the mean and “N total”.

Problem Set Section 3C

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. Type the numbers into your statistics software and check your answers.

$$\text{Mean } (\bar{x}) = \frac{\sum x}{n} = \frac{\text{Sum of the Numbers}}{\text{Frequency (how many numbers in data set)}}$$

1. 2, 7, 7, 9, 8, 8, 4, 5, 1, 0, 3, 2, 11, 3, 1, 7, 2, 4

2. 17 , 21 , 23 , 24 , 25 , 27 , 28 , 29 , 31 , 32 , 33 , 36

3. 9.4 , 3.5 , 1.1 , 7.8 , 3.2 , 36.4 , 6.6

4. 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. 2.54 , 3.14 , 2.49 , 1.98 , 1.46 , 2.27 , 1.83 , 2.63 , 2.87 , 3.25 , 8.75

6. 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.

7. Add two numbers to your data set in #6, 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.

8. 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.

9. Add two numbers to your data set in #8, 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.

10. 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 10 without adding the numbers and without calculating the mean directly.

5, 6, 7, 8, 9, 11, 12, 13, 14, 15

Section 3D – Introduction to Spread, Standard Deviation, and Typical Values for Normal Data

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.

Small Spread => More Consistency and Greater Predictability

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

Large Spread => Not Consistent and 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”.

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 with Statcato?

To calculate statistics with Statcato: *Statistics => Basic Statistics => Descriptive Statistics => Pick a column of data => Pick what statistics you want to calculate => OK*

Example 1

Use Statcato to calculate the four measures of spread (Range, Interquartile range, Variance, and Standard Deviation) for the women’s height (inches) in the health data.

The screenshot shows the 'Descriptive Statistics' dialog box. The 'Inputs' section has 'c3' entered in the 'Input Variable(s):' field. The 'Statistics' section has the following options checked: 'Standard deviation', 'Variance', 'Range', 'Interquartile range', and 'Percentile'. The 'Results' section has 'New datasheet' selected. The 'OK' and 'Cancel' buttons are at the bottom.

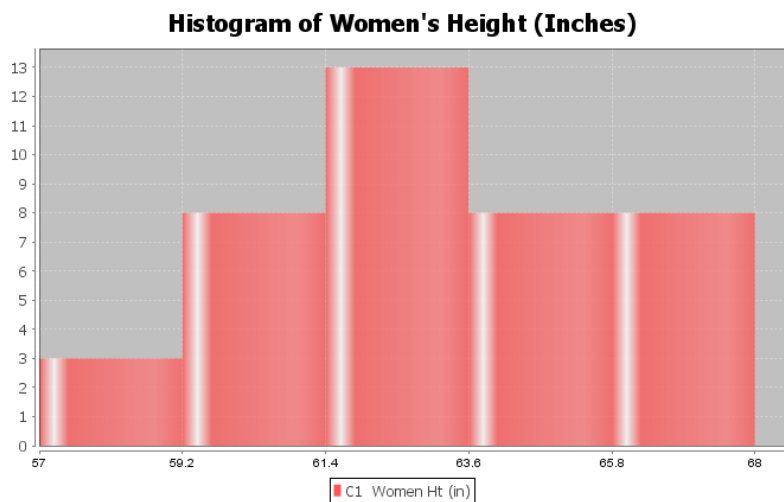
Descriptive Statistics

Variable	Standard Deviation	Variance
C3 Women Ht (in)	2.741	7.514

Variable	IQR
C3 Women Ht (in)	3.625

Variable	Range
C3 Women Ht (in)	11.0

Which of these measures of spread are most accurate? That depends on the shape. Recall that this data was bell shaped.



It turns out that the most accurate measure of spread for bell shaped data sets is the standard deviation. The range of a data set measures spread but does not measure typical values in the data. Interquartile range is the measure of spread that we use for skewed data sets. Variance is the square of the standard deviation and is useful in certain applications.

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

Example 1 Continued

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).

$$63.195 - 2.741 \leq \text{typical heights for these women} \leq 63.195 + 2.741$$

$$60.454 \leq \text{typical heights for these women} \leq 65.936$$

Typical women in this data set had a height between 60.45 inches (little over 5 feet) and 65.94 inches (little under 5 ½ feet).

To calculate the two values for the typical range in a bell shaped data set:

Add and subtract the mean and standard deviation. Be careful to subtract in the correct order.

$$\text{Mean} - \text{Standard Deviation} \leq \text{typical values} \leq \text{Mean} + \text{Standard Deviation}$$

Empirical Rule for 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. 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.

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.

$$\text{Mean Average} = (4.7 + 6.2 + 3.3 + 5.1 + 2.9 + 7.4 + 4.5) / 7 = 34.1 / 7 = 4.871428571 \approx 4.87$$

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 $(x - \bar{x})$. Remember we rounded the mean, so these are just approximate distances.

$$6.2 - 4.87 \approx 1.33$$

$$5.1 - 4.87 \approx 0.23$$

$$7.4 - 4.87 \approx 2.53$$

$$2.9 - 4.87 \approx -1.97$$

$$3.3 - 4.87 \approx -1.57$$

$$4.5 - 4.87 \approx -0.37$$

$$4.7 - 4.87 \approx -0.17$$

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

$$(1.33)^2 \approx 1.7689$$

$$(0.23)^2 \approx 0.0529$$

$$(2.53)^2 \approx 6.4009$$

$$(-1.97)^2 \approx 3.8809$$

$$(-1.57)^2 \approx 2.4649$$

$$(-0.37)^2 \approx 0.1369$$

$$(-0.17)^2 \approx 0.0289$$

Now we will add up all the squared distances and calculate the “Sum of Squares” $\sum(x - \bar{x})^2$.

This is a very important technique in statistics and occurs in many different applications.

$$\text{Sum of Squares} \approx 1.7689 + 0.0529 + 6.4009 + 3.8809 + 2.4649 + 0.1369 + 0.0289 \approx 14.6814$$

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

$$\sqrt{\frac{\sum(x - \bar{x})^2}{(n - 1)}}$$

Standard Deviation for weights of the bricks = square root (sum of squares / $(n-1)$)

$$\approx \text{square root}(14.6814 / (7-1)) \approx \text{square root}(14.6814 / 6) \approx \text{square root}(2.4469)$$

$$\approx 1.564 \text{ kilograms}$$

We calculated the standard deviation with Statcato and got approximately 1.567 kg. Statcato 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 data set is bell shaped, use the mean as the center and average. Use the standard deviation as the best measure of spread. Do not calculate these with formula and calculator. Use a computer program like Statcato.

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 calculate the typical range for a bell shaped data set by adding and subtracting the mean and standard deviation.

Key: The mean and standard deviation should only be used if the data set is bell shaped. They are not accurate if the data set is not bell shaped.

Problem Set Section 3D

Standard Deviation (s): Typical (or Average) Distance from the Mean

To calculate standard deviation: 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.

$$\bar{x} = \frac{\sum x}{n} \qquad s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Fill out the following tables and calculate the Mean and the Standard Deviation for each of the following three data sets.

1. 1, 2, 3, 11, 12, 13

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

Mean \bar{x} =???

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

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

Degrees of Freedom (n – 1) =???

Standard Deviation (s) =???

2. 2, 5, 6, 7, 9, 10, 11, 14

Values in data set (x)	Each Value – mean ($x - \bar{x}$)	Squares of Differences $(x - \bar{x})^2$
2		
5		
6		

7		
9		
10		
11		
14		

Mean \bar{x} =???

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

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

Degrees of Freedom (n - 1) =???

Standard Deviation (s) =???

(#3-#8 directions): Open “Bear” data from my website www.matt-teachout.org. (Look under “Int Alg for Stats” and then the “data sets” tab.) Use Statcato to create a histogram and then find the mean and four measures of spread (range, interquartile range (IQR), variance, and standard deviation).

3. Bear ages (months)

Shape = _____

Range = _____

IQR = _____

Variance = _____

Mean = _____

Standard Deviation = _____

How accurate are the mean and standard deviation for this data? _____

Write a sentence to explain the meaning of the standard deviation in this context.

4. Bear neck circumference (inches)

Shape = _____

Range = _____

IQR = _____

Variance = _____

Mean = _____

Standard Deviation = _____

How accurate are the mean and standard deviation for this data? _____

Write a sentence to explain the meaning of the standard deviation in this context.

5. Bear length (inches)

Shape = _____

Range = _____

IQR = _____

Variance = _____

Mean = _____

Standard Deviation = _____

How accurate are the mean and standard deviation for this data? _____

Write a sentence to explain the meaning of the standard deviation in this context.

6. Bear chest size (inches)

Shape = _____

Range = _____

IQR = _____

Variance = _____

Mean = _____

Standard Deviation = _____

How accurate are the mean and standard deviation for this data? _____

Write a sentence to explain the meaning of the standard deviation in this context.

7. Bear weight (pounds)

Shape = _____

Range = _____

IQR = _____

Variance = _____

Mean = _____

Standard Deviation = _____

How accurate are the mean and standard deviation for this data? _____

Write a sentence to explain the meaning of the standard deviation in this context.

8. Bear head length (inches)

Shape = _____

Range = _____

IQR = _____

Variance = _____

Mean = _____

Standard Deviation = _____

How accurate are the mean and standard deviation for this data? _____

Write a sentence to explain the meaning of the standard deviation in this context.

9. Find a data set with eight numbers that has a mean of 20 and a standard deviation less than four. Check your answer by calculating the mean and standard deviation with Statcato.

10. Find a data set with ten numbers that has a mean of 20 and a standard deviation greater than six. Check your answer by calculating the mean and standard deviation with Statcato.

Section 3E – 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 bell shaped (normal) 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? (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? What is the average? (If the data is 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 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 bell shaped then we should add and subtract the mean and standard deviation.

Mean – Standard Deviation ≤ Typical Values ≤ 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 Unusual Values in a Bell Shaped (Normal) Data Set

So how do you find unusual values in a bell shaped data set? It has long been considered that 1 standard deviation from the mean is considered typical. Any value more than two standard deviations from the mean is considered unusual. So any value in the data set more than two standard deviations above the mean is considered “unusually high” and any value more the two standard deviations below the mean is considered “unusually low”.

Unusual High Cutoff: $mean + (2 \times Standard\ Deviation)$

Unusual Low Cutoff: $mean - (2 \times Standard\ Deviation)$

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. If a value in the data set is less than or equal to the unusual low cutoff, then it is considered unusually low.

Use a dot plot: Once you find the unusual cutoffs, I like to use a dot plot to identify those values in the data set that are unusually high or unusually low. In Statcato, as in most statistics software programs, you can hold your curser over the dot and the computer will tell you its value.

Note: Not all data sets have unusual values. If that is the case, you can just say that there were no unusually high or unusually low values in the data set.

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.

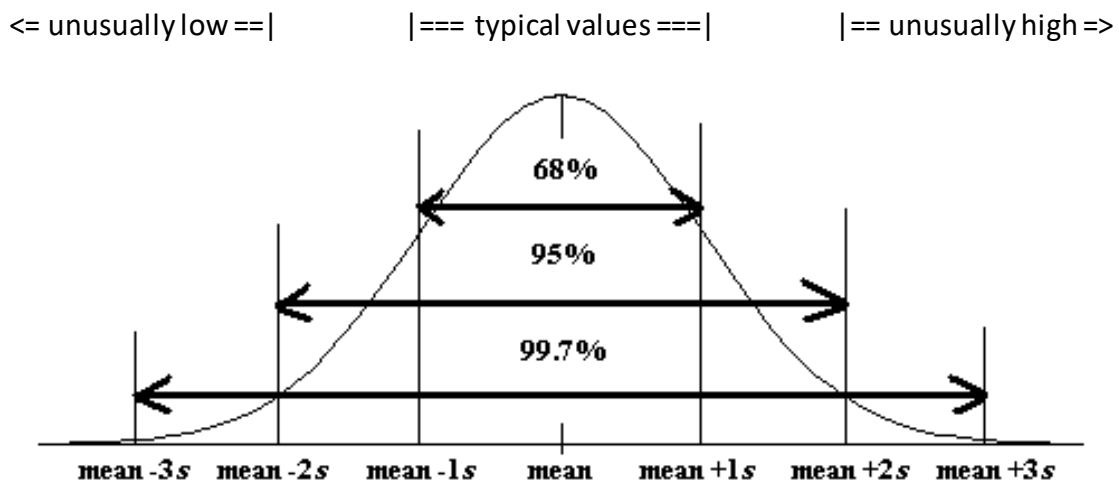
The Empirical Rule discussed in the last section can shed some light on this issue.

Empirical Rule for 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 in a 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 in a 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 bell shaped the data set is the more accurate the percentages. The Empirical Rule does not apply to skewed data sets. The following diagram describes the Empirical Rule for bell shaped data sets.



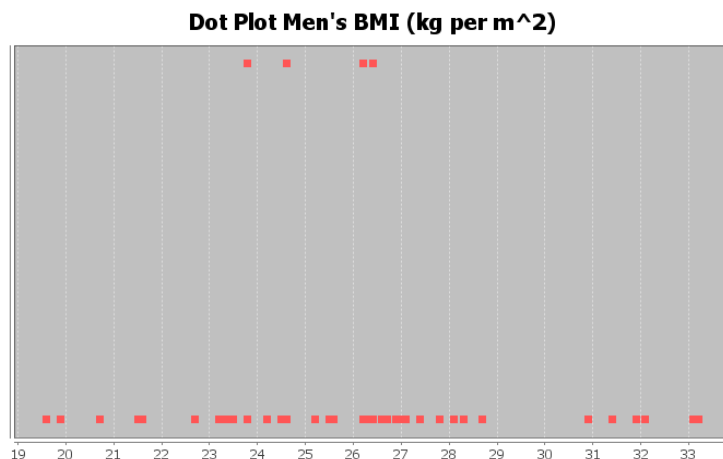
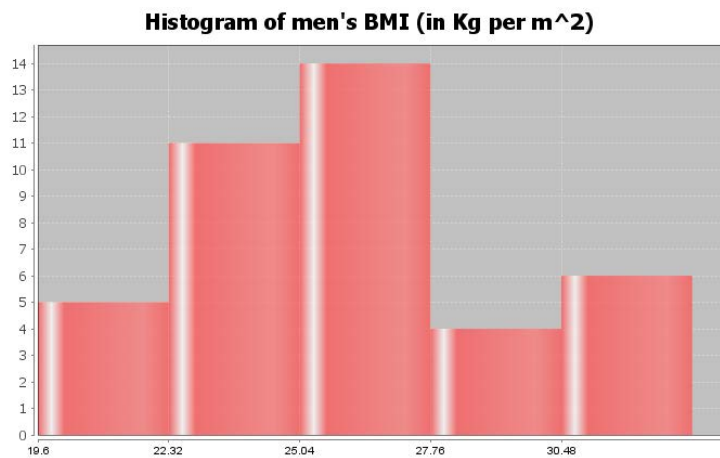
Example 1

Let us look at the body mass index (in kilograms per meters squared) for forty men in the health data. Let us see if we can summarize the important information about this data.

Quantitative Data Analysis Summary

- 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? What is the average? (If the data is 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 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 bell shaped then we should add and subtract the mean and standard deviation.
 $\text{Mean} - \text{Standard Deviation} \leq \text{Typical Values} \leq \text{Mean} + \text{Standard Deviation}$
- Find any unusual values in the data set. (Some call these unusual values “outliers”.)
 Unusual High Cutoff: $\text{mean} + (2 \times \text{Standard Deviation})$
 Unusual Low Cutoff: $\text{mean} - (2 \times \text{Standard Deviation})$
 Use a dot plot to identify unusual values in the data set.

Putting this data into Statcato, we get the following graphs and statistics.



Descriptive Statistics

Variable	Mean	Standard Deviation
C5 Men Body Mass Index BMI (kg per m ²)	25.998	3.431

Variable	Min	Max
C5 Men Body Mass Index BMI (kg per m ²)	19.6	33.2

Variable	N total
C5 Men Body Mass Index BMI (kg per m ²)	40

This data is measuring the body mass index (BMI) of 40 men. The units are kilograms per meters squared.

Notice that the histogram is not perfectly bell shaped. Real data rarely is. It is however closer to bell shaped than skewed. Notice the highest bar is still in the middle and there are two bars to the right and two bars to the left of the highest bar. We can classify this shape as close to bell shaped or “nearly normal”.

Since the data set is bell shaped, we should use the mean average as our measure of center and the standard deviation as our measure of spread. The original data set ended in the tenths place so we can round our statistics to the hundredths place (one more place to the right than the data has).

The best measure of center is the mean average of 25.998 kg/m². The average body mass index for these forty men is approximately 26.00 kg/m².

The best measure of spread or variability is the standard deviation of 3.431 kg/m². This tells us that typical men in the data have a body mass index 3.43 kg/m² from the mean. This tells us that typical men in the data set have a body mass index between 22.57 kg/m² and 29.43 kg/m².

$$26.00 - 3.43 \leq \text{Typical Men's BMI} \leq 26.00 + 3.43$$

$$22.57 \leq \text{Typical Men's BMI} \leq 29.43$$

Let us see if there are any unusually high or unusually low values in the data set. We will start with calculating the cutoffs. Then we will look at the dot plot.

$$\text{Unusually Low Cutoff: } 26.00 - (2 \times 3.43) \approx 26.00 - 6.86 \approx 19.14$$

$$\text{Unusually High Cutoff: } 26.00 + (2 \times 3.43) \approx 26.00 + 6.86 \approx 32.86$$

Hold your cursor over the dots in the dot plot in Statcato. We are looking for any dots that are 19.14 or lower as well as any dots that are 32.86 or higher.

The lowest BMI in the data set was 19.6 kg/m². This is not lower than 19.14, so it is not considered unusually low.

There were two unusually high values of 33.2 kg/m² and 33.1 kg/m². These were greater than the unusual high cutoff of 32.86 so were considered unusually high. The next highest value in the dot plot was 32.1 kg/m². This was not above 32.86 so it was not unusually high.

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 1 continued (Summary Report Paragraph)

Men's Body Mass Index Summary Report Paragraph: This data describes the body mass index (BMI) of forty men in kilograms per meters squared (kg/m²). The data was bell shaped. The best measure of center was the mean average of 26.00 kg/m². So the average BMI for these forty men was 26.00 kg/m². The best measure of spread was the standard deviation of 3.43 kg/m². So typical men in the data set had a BMI 3.43 kg/m² from the mean. This tells us that typical men had a BMI between 22.57 kg/m² and 29.43 kg/m². The lowest BMI in the data set was 19.6 kg/m². This was not considered unusual though. There were two unusually high BMI values in the data set. They were 33.2 kg/m² and 33.1 kg/m².

Problem Set Section 3E

1. Answer the following questions:

In a bell shaped data set, what measure of center should we use? _____

In a bell shaped data set, what measure of average should we use? _____

In a bell shaped data set, what measure of spread should we use? _____

In a bell shaped data set, how many standard deviations from the mean is considered typical? _____

In a bell shaped data set, approximately what percentage is typical? _____

In a bell shaped data set, how many standard deviations from the mean is considered unusual? _____

In a bell shaped data set, approximately what percentage of the data is unusually high? _____

In a bell shaped data set, approximately what percentage of the data is unusually low? _____

Explain how we can use a Dot Plot to find the unusually high values in the data set.

Explain how we can use a Dot Plot to find the unusually low values in the data set.

Directions: Now analyze the following data sets. Open “Bear” data and the “Health” data from my website www.matt-teachout.org. (Look under “Int Alg for Stats” and then the “data sets” tab.) Use Statcato to create a histogram to verify that each data set is bell shaped and that the mean and standard deviation are accurate. Then use Statcato to find the center (mean), average (mean), and spread (standard deviation). Use a calculator to calculate the typical range, and the unusual cutoff values. Use Statcato to create a dot plot. Use the dot plot to identify any numbers in the data set that are unusually high. Use the dot plot to identify any numbers in the data set that are unusually low.

2. Bear Head Length (inches)

What is the data measuring and what are the units?

How many numbers are in the data set?

Is the data set bell shaped? (Yes or No)

Minimum = _____

Maximum = _____

Mean = _____

Standard Deviation = _____

_____ ≤ typical values ≤ _____

Usually High Values ≥ _____

Unusually Low Values ≤ _____

List all the numbers in this data set that are unusually high.

List all the numbers in this data set that are unusually low.

Now write a summary report paragraph for this data set.

3. Bear neck circumference (inches)

What is the data measuring and what are the units?

How many numbers are in the data set?

Is the data set bell shaped? (Yes or No)

Minimum = _____

Maximum = _____

Mean = _____

Standard Deviation = _____

_____ ≤ typical values ≤ _____

Usually High Values ≥ _____

Unusually Low Values ≤ _____

List all the numbers in this data set that are unusually high.

List all the numbers in this data set that are unusually low.

Now write a summary report paragraph for this data set.

4. Bear Chest Size (inches)

What is the data measuring and what are the units?

How many numbers are in the data set?

Is the data set bell shaped? (Yes or No)

Minimum = _____

Maximum = _____

Mean = _____

Standard Deviation = _____

_____ ≤ typical values ≤ _____

Usually High Values ≥ _____

Unusually Low Values ≤ _____

List all the numbers in this data set that are unusually high.

List all the numbers in this data set that are unusually low.

Now write a summary report paragraph for this data set.

5. Women's Diastolic Blood Pressure

What is the data measuring and what are the units?

How many numbers are in the data set?

Is the data set bell shaped? (Yes or No)

Minimum = _____

Maximum = _____

Mean = _____

Standard Deviation = _____

_____ \leq typical values \leq _____

Usually High Values \geq _____

Unusually Low Values \leq _____

List all the numbers in this data set that are unusually high.

List all the numbers in this data set that are unusually low.

Now write a summary report paragraph for this data set.

6. Women's Wrist Circumference (Inches)

What is the data measuring and what are the units?

How many numbers are in the data set?

Is the data set bell shaped? (Yes or No)

Minimum = _____

Maximum = _____

Mean = _____

Standard Deviation = _____

_____ \leq typical values \leq _____

Usually High Values \geq _____

Unusually Low Values \leq _____

List all the numbers in this data set that are unusually high.

List all the numbers in this data set that are unusually low.

Now write a summary report paragraph for this data set.

7. Men's Height (Inches)

What is the data measuring and what are the units?

How many numbers are in the data set?

Is the data set bell shaped? (Yes or No)

Minimum = _____

Maximum = _____

Mean = _____

Standard Deviation = _____

_____ \leq typical values \leq _____

Usually High Values \geq _____

Unusually Low Values \leq _____

List all the numbers in this data set that are unusually high.

List all the numbers in this data set that are unusually low.

Now write a summary report paragraph for this data set.

8. Men's Weight (Pounds)

What is the data measuring and what are the units?

How many numbers are in the data set?

Is the data set bell shaped? (Yes or No)

Minimum = _____

Maximum = _____

Mean = _____

Standard Deviation = _____

_____ \leq typical values \leq _____

Usually High Values \geq _____

Unusually Low Values \leq _____

List all the numbers in this data set that are unusually high.

List all the numbers in this data set that are unusually low.

Now write a summary report paragraph for this data set.

Chapter 3 Review

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 \leq Typical Values \leq 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.

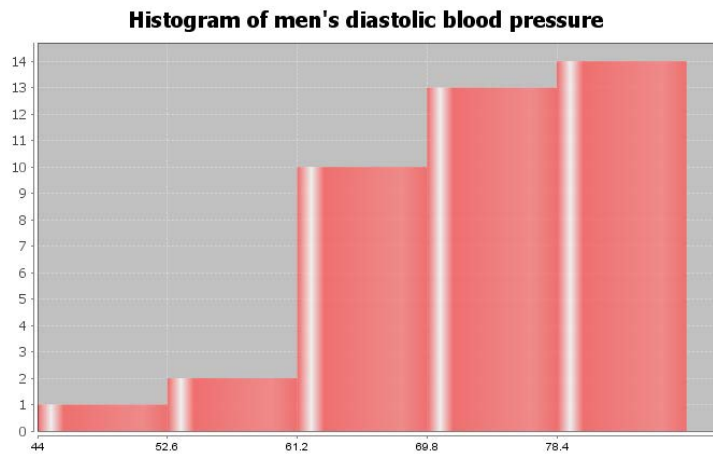
Problem Set Review Chapter 3

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 = _____

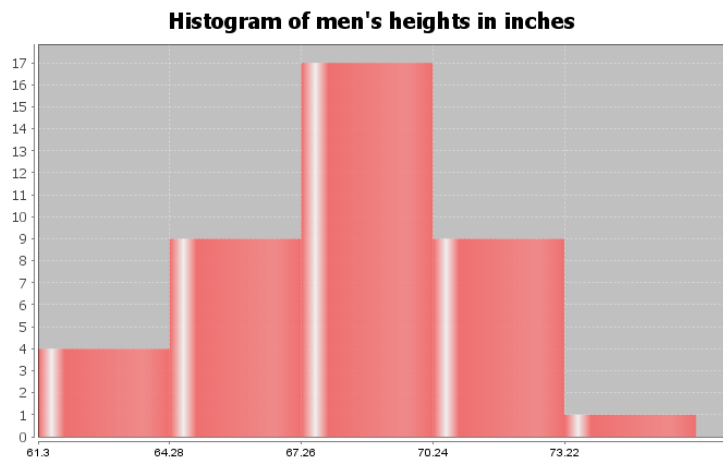
Mean or Median?



2. Men's Heights (inches)

Shape = _____

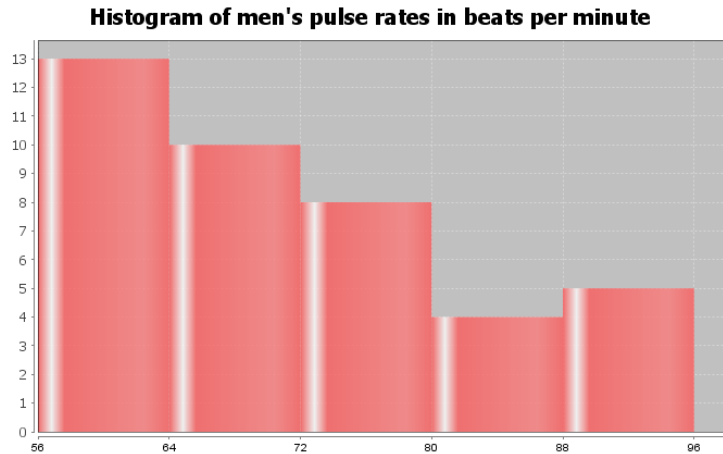
Mean or Median?



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).

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

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?

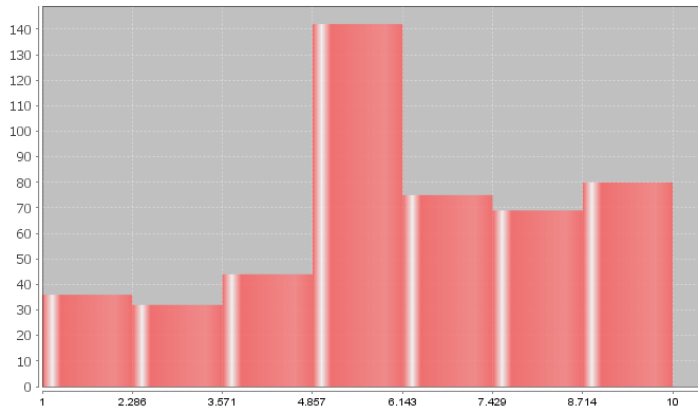
7. What percentage of the values in a bell shaped data set are considered typical?

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.

Histogram of Math 075 students Math Intimidation Scale



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 \leq typical math intimidation scores \leq Mean + Standard Deviation

_____ \leq typical math intimidation scores \leq _____

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

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

Unusual High Cutoff = _____

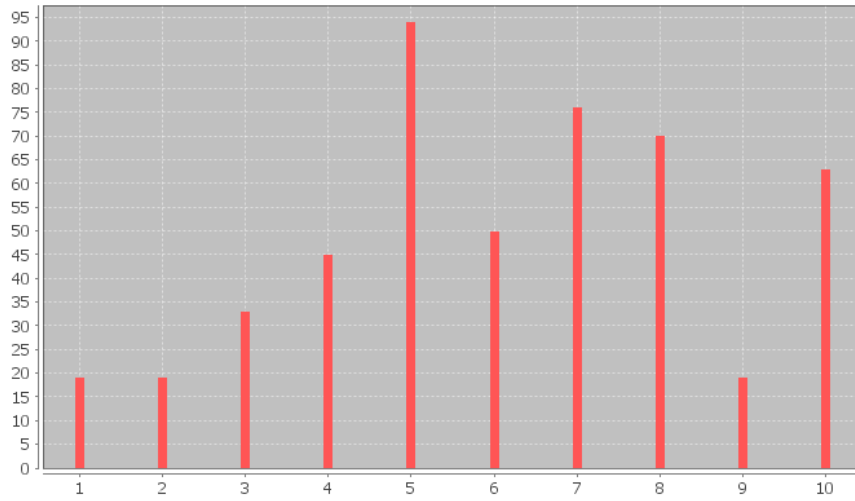
17. What is the cutoff for an unusually low math intimidation score?

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.

Dot Plot for Math Intimidation Score Data



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? _____

Project Chapter 3 - Quantitative Data Analysis Poster for Bell Shaped Data

Directions: *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 Bell Shaped Project 3 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 Bell Shaped Project 3 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?**
- **Graph from Software: Histogram and Dot Plot**
- **Sample Statistics from Software: Mean, Standard Deviation, Min, Max**
- **What is the data measuring?**
- **What are the units?**
- **How many numbers are in the data set : sample size (n)**
- **Shape**
- **Center (Mean): Write a sentence to explain the mean.**
- **Average (Mean)**
- **Spread (Standard Deviation): Write a sentence to explain the standard deviation.**
- **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 in the data set. (Find these on the dot plot.)**
- **Calculate Unusually low cutoff (Mean – 2 x Stand Dev)**
- **List all unusually high values in the data set. (Find these on the dot plot.)**
- **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.

