Chapter 5: Analyzing Skewed Quantitative Data

Introduction: In our last chapter, we focused on analyzing bell shaped (normal) data, but many data sets are not bell shaped. How do we analyze quantitative data when it is not normal?

The main issue is that the mean and standard deviations are not accurate and should not be used in the analysis. Then what statistics should we use?

We will be introducing a new kind of graph that is specially designed for analyzing skewed data. It is called the "box and whisker plot" or "box plot" for short.

When data sets are not bell shaped, we will focus on the median, quartiles, interquartile range and boxplots to measure center and spread. Quartiles are more accurate because they are based on the order of the numbers instead of distances and so are not as effected by the skewed shape and extremely unusual values.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Section 5A – Review of Shapes and Centers with Histograms and Dot Plots

Let us start by reviewing shapes and centers.

Here are the directions for making dot plots and histograms in StatKey.

Making a dot plot in StatKey: Go to www.lock5stat.com. Click on "One Quantitative Variable" under the "Descriptive Statistics and Graphs" menu. Click on "Edit Data". Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says "Data has a header row". If the data does NOT has a title, do NOT check the box that says "Data has a header row". Do NOT check the box that says "First column is an identifier". (You would only check the "identifier" box if there is a word next to every number in the data.) Now push "OK". At the top left of the graph, click on the "dot plot" tab.

Making a histogram in StatKey: Go to www.lock5stat.com. Click on "One Quantitative Variable" under the "Descriptive Statistics and Graphs" menu. Click on "Edit Data". Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says "Data has a header row". If the data does NOT has a title, do NOT check the box that says "Data has a header row". Do NOT check the box that says "First column is an identifier". (You would only check the "identifier" box if there is a word next to every number in the data.) Now push "OK". At the top left of the graph, click on the "histogram" tab. Use the slider button on the right of the screen to adjust the number of bars (buckets) in your histogram. Three bars is usually the best for seeing the shape.

Center Principle for Quantitative Data

If a data set is normally distributed (bell shaped), the mean average is usually an accurate measure of center and we should use the mean as the average for the data set.

If a data set has a skewed or irregular 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.

Note: If a data set is not skewed, but just has an unusual shape like uniform, use the median also. Do not use the mean unless it is bell shaped. The mode is sometimes used as the center for bimodal or multimodal shaped data, since it can have multiple values and represent each hill in the data. That is why it is called bi-modal or multi-modal.

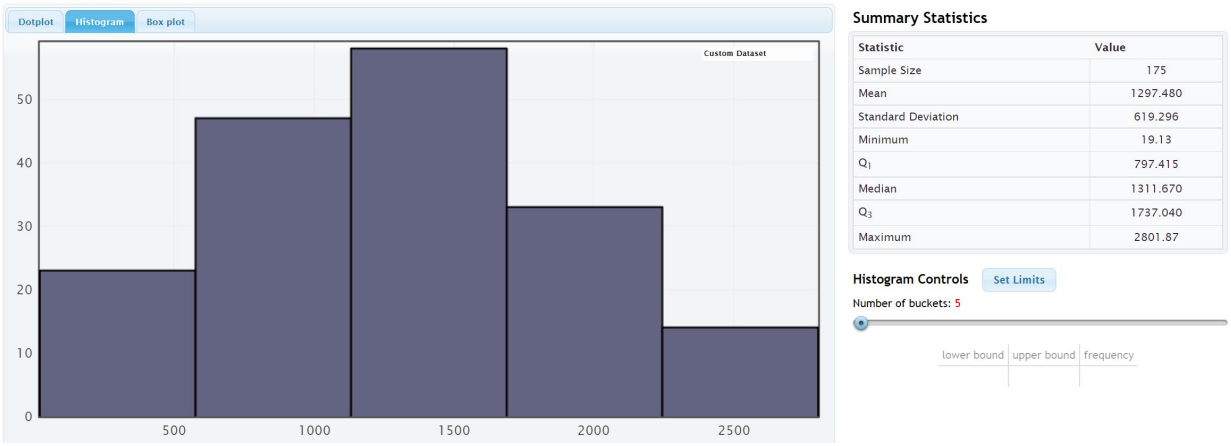
Example 1

We copied and pasted a random sample of 175 salaries from Australia into StatKey. Here is the dot plot and histogram created.

Summary Statistics

Statistic	Value
Sample Size	175
Mean	1297.480
Standard Deviation	619.296
Minimum	19.13
Q_1	797.415
Median	1311.670
Q_3	1737.040
Maximum	2801.87

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021



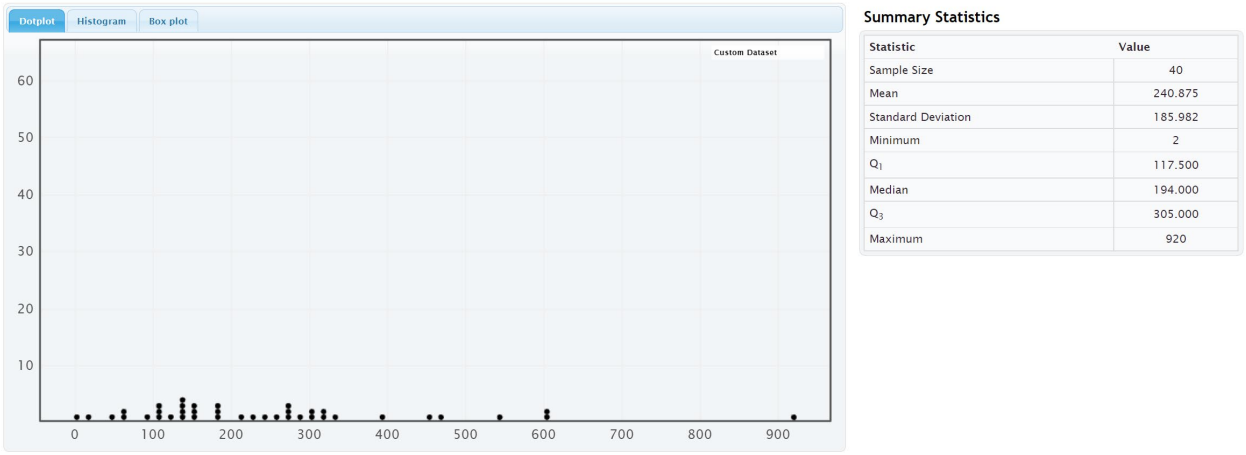
Notice that the salary data from Australia is normally distributed (bell shaped). The highest bar and the largest concentration of dots are in the center. The left and right tails are roughly the same length.

Notice that the mean and median are relatively close when data is normally distributed. However, we should use the mean as our center and average for this data.

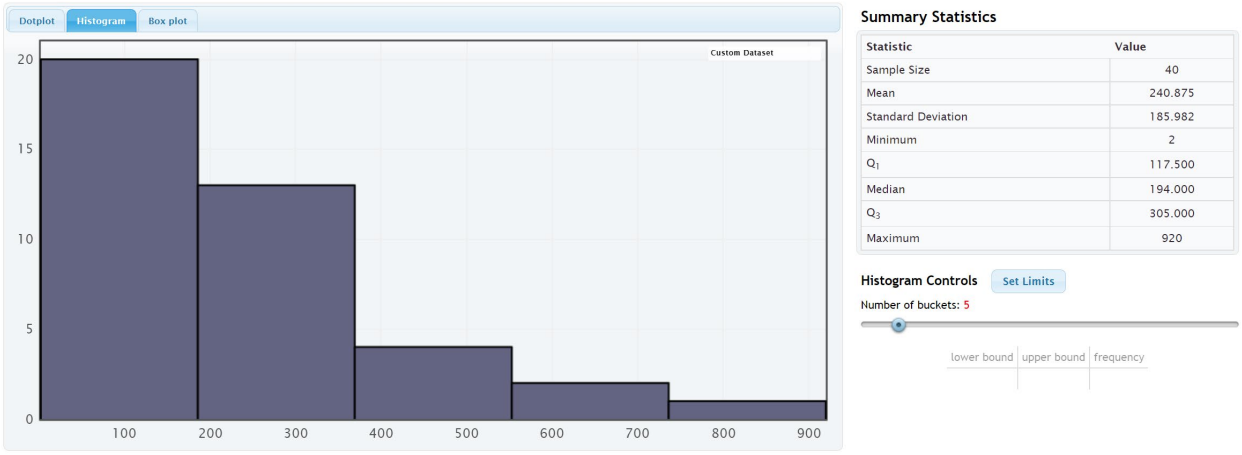
So the average salary for this data is $1297.48 (mean).

Example 2

We copied and pasted a random sample of 40 women's cholesterol in milligrams per deciliter into StatKey. Here is the dot plot and histogram created.



This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021



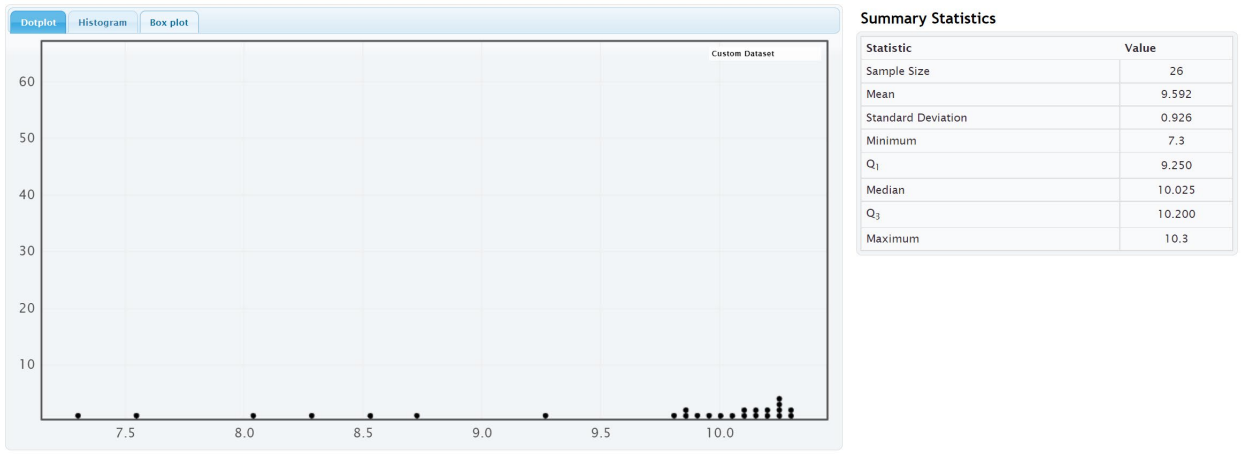
Notice that the shape of this data is not normal. The highest bars and the largest cluster of dots are on the far left and it has a long right tail. We call the shape of this data skewed right or positively skewed.

Notice that the mean average (240.875 mg/dL) is much larger than the median average (194 mg/dL). The mean has been pulled up in the direction of the long tail and is no longer accurate. The median average however is still close to the highest bar is a much more accurate average. Remember the rule for centers, when data is not normal, use the median as your average (center).

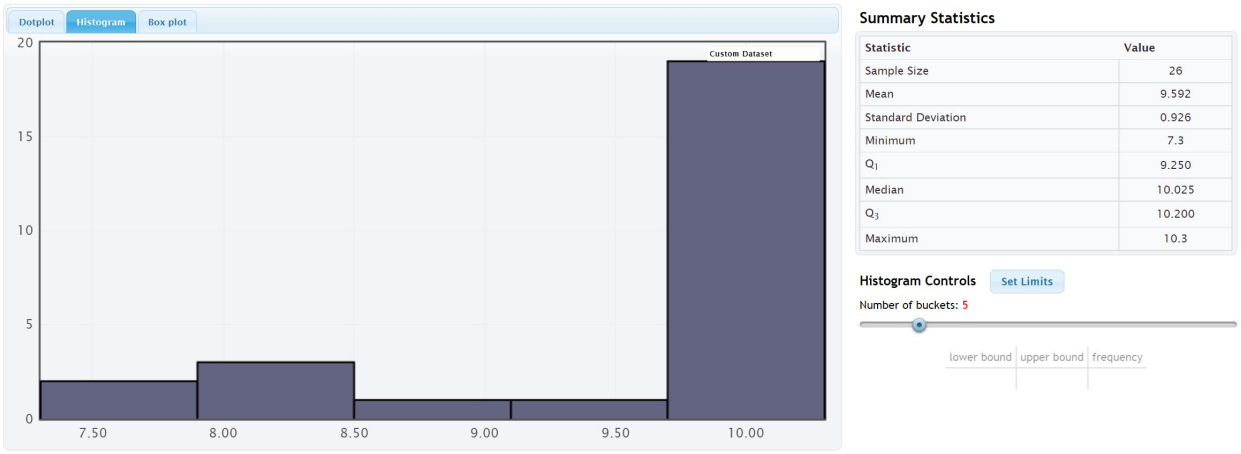
The average cholesterol for these 40 women is 194 mg/dL (median).

Example 3

We copied and pasted a sample of salaries in dollars per hour from a company into StatKey. Here is the dot plot and histogram created.



This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021



Notice that the shape of this data is not normal. The highest bars and the largest cluster of dots are on the far right and it has a long left tail. We call the shape of this data skewed left or negatively skewed.

Notice that the mean average (9.592 $/hour) is much smaller than the median average (10.025 $/hour). The mean has been pulled down in the direction of the long tail and is no longer accurate. The median average however is still close to the highest bar is a much more accurate average. Remember the rule for centers, when data is not normal, use the median as your average (center).

The average salary for the employees in this company is $10.025 (median).

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Problem Set Section 5A

Directions: Open the "Bear Data" and "Health Data" (columns AD-AP) in Canvas or at www.matt-teachout.org. Go to www.lock5stat.com. Click on "One Quantitative Variable" under the "Descriptive Statistics and Graphs" menu. Click on "Edit Data". Copy and paste in the column of quantitative data indicated in the problem. If the data has a title, check the box that says "Data has a header row". If the data does NOT has a title, do NOT check the box that says "Data has a header row". Do NOT check the box that says "First column is an identifier". (You would only check the "identifier" box if there is a word next to every number in the data.) Now push "OK". At the top left of the graph, click on the "dot plot" tab and the "histogram tab". When you click the histogram tab, pull the slider on the right to "3 buckets" so your histogram has 3 bars.

1. Bear Ages in Months
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
2. Bear Head Length in Inches
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
3. Bear Head Width in Inches
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
4. Bear Neck Circumference in Inches
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
5. Bear Length in Inches
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

6. Bear Chest Size in Inches
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
7. Bear Weight in Pounds
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
8. Men's Ages in Years
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
9. Men's Weight in Pounds
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
10. Men's Height in Inches
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
11. Men's Pulse Rate in Beats per Minute (BPM)
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

12. Men's Systolic Blood Pressure (mm of Hg)
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
13. Men's Diastolic Blood Pressure (mm of Hg)
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
14. Men's Cholesterol (mg per dL)
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?
15. Men's Body Mass Index (kg per m^2)
    a) Draw a rough sketch of the dot plot on a piece of paper or save the graph on a word document.
    b) Draw a rough sketch of the 3-bar histogram on a piece of paper or save the graph on a word document.
    c) What is the shape of the data set?
    d) Should we use the mean or median as the average (center) for the data?

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Section 5B – Understanding the Median Average

In the last section, we said that we should use the median as our center and average, when data is skewed or not normally distributed, but what is the median? Why is it more accurate than the mean for skewed data?

Let us see if we can get a better understanding of the median average.

Definition of the Median average: The median average is the center of the data when the values are put in order from smallest to largest. The median is also called the "50th percentile" since approximately 50% of the numbers in the data set will be greater than the median and 50% of the numbers in the data set will be less than the median. Think of the median as a marker that divides the data in half. That is why it is often called the true center of the data.

Skewed data tends to have extremely unusual values. These unusual values (outliers) are very far from the mean. That is why the mean and standard deviation (typical distance from the mean) are not accurate for skewed data. The median is based on how many numbers are in the data set (frequency) and the order of the numbers. If the highest value were 40 or 4000, it would not change the median.

Names for the Median Average: "Median", The 50th percentile ($\boldsymbol{P_{50}}$), or the Second Quartile ($\boldsymbol{Q_2}$).

How to Calculate the Median Average
As with all statistics, rely on technology to calculate. No statistician calculates the median by hand, especially for large data sets. All of them use statistics software or computer software programs. To get a better understanding of the median, we will look at a couple examples where we calculate the median with small data sets.

To calculate the median, put the data in order from smallest to largest. Computer programs like excel can sort the data for you if you do not want to put it in order. Once the data is in order, you will look for the center of the data.

Odd Number of Values: If you have an odd number of values in the data set, then your median will be the number in the exact middle of the data when it is in order. Suppose we have 17 numbers in order from smallest to largest in the data set. Then our median would be the ninth number in the data set. That would give us eight numbers below the median and eight numbers above the median. Remember the median separates the data into two equal groups.

Even Number of Values: If you have an even number of values in the data set, then your median will not be a value in the data set. The median will be half way in between the two numbers in the middle. Suppose you have 26 numbers in order from smallest to largest in the data set. Then the median will be half way between the 13th and 14th numbers in the data set. That way thirteen numbers will be below the median and thirteen numbers will be above the median. If you cannot think of what half way in between would be, you could use the following formula. Remember this formula only works if the data values are in order.

Median (even # of data values) = (first number in middle + second number in middle) / 2

Example 1
Find the median for the brick weight (in kilograms) data from last chapter.

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

The first thing to notice is that the data is not in order. It needs to be put in order before we can find the median.

Data in order:

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

Since there are seven numbers in the data set. The fourth number (4.7) will be the median.

Median Average = 4.7 kilograms

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Notice there are three numbers in the data set greater than the median (5.1, 6.2 and 7.4) and there are three numbers in the data set less than the median (2.9, 3.3 and 4.5).

Example 2
Let us look at a second example.

Here are the yearly salaries in thousands of dollars for employees from a small company.

36.5 , 51.2 , 47.9 , 44.1 , 37.2 , 39.6 , 41.8 , 45.4 , 43.2 , 253.5

(This last salary of 253.5 thousand dollars was the CEO of the company.)

Remember to put the numbers in order first.

Yearly Salary Data in order:

36.5 , 37.2 , 39.6 , 41.8 , 43.2 , 44.1 , 45.4 , 47.9 , 51.2 , 253.5

Since there are ten numbers (even), the median will not be a number in the data set. It will be half way between the two middle numbers that can divide the data in half. The two numbers in the middle are 43.2 and 44.1 thousand dollars.

Median Average = (43.2 + 44.1) / 2 = (87.3) / 2 = 43.65 thousand dollars

Notice again that there are five numbers above the median (44.1 , 45.4 , 47.9 , 51.2 , 253.5) and five numbers below the median (36.5 , 37.2 , 39.6 , 41.8 , 43.2). The data has been split in half.

This is a good example to explain why the median is a better average than the mean. The CEO is a large unusual value in the data set, making the data very skewed right. Let us compare the mean and median averages.

Mean Average =

(36.5 + 37.2 + 39.6 + 41.8 + 43.2 + 44.1 + 45.4 + 47.9 + 51.2 + 253.5) / 10

= 640.4 / 10 = 64.04 thousand dollars.

Median Average =

(43.2 + 44.1) / 2 = (87.3) / 2 = 43.65 thousand dollars

Notice no one in the company makes 64 thousand dollars. The mean is not a good average for this data. The median however is very accurate. Many people in the company make around 43 or 44 thousand dollars. Recently, companies have been using the median average as their “average salary” on their websites for this very reason.

Calculating the median average with technology
All statistics software programs can calculate the median. This is a much better way to find the median, especially if you have larger data sets.

Here are the steps to calculating graphs and quantitative statistics with StatKey.

Go to www.lock5stat.com. Click on “One Quantitative Variable” under the “Descriptive Statistics and Graphs” menu. Click on “Edit Data”. Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says “Data has a header row”. If the data does NOT has a title, do NOT check the box that says “Data has a header row”. Do NOT check the box that says “First column is an identifier”. (You would only check the “identifier” box if there is a word next to every number in the data set.) Push “OK”. You should see the median in the list of “Summary Statistics” on the top right of the page.

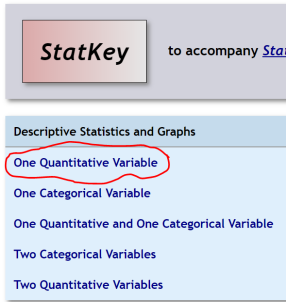
This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a “CC-BY” Creative Commons Attribution 4.0 International License – 3/17/2021

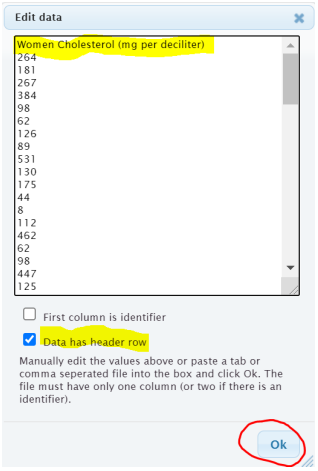
Example

In the last section, we looked at a random sample of 40 women's cholesterol in milligrams per deciliter. Let's put this data into StatKey, verify the shape and calculate the median. First we need to find the column of data in the "Health" data set.

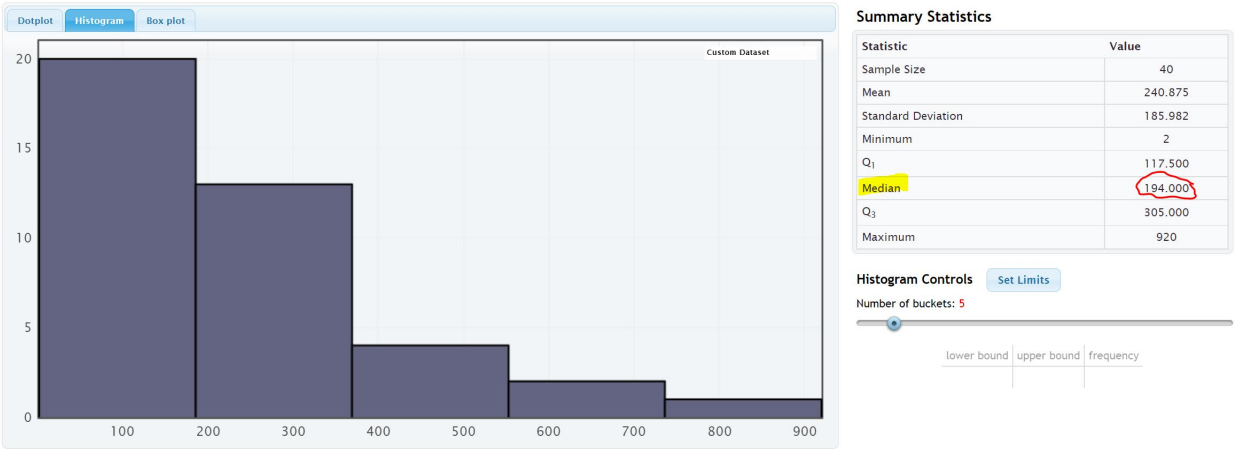
W
Women Cholesterol (mg per deciliter)
264
181
267
384
98
62
126
89
531
130
175
44
8
112
462
62
98

Now we will go to www.lock5stat.com and copy and paste the data into StatKey.





This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021



Notice that the shape of this data is skewed right.

Notice that the mean average (240.875 mg/dL) is much larger than the median average (194 mg/dL). The mean has been pulled up in the direction of the long tail and is no longer accurate. The median average however is still close to the highest bar is a much more accurate average. Remember the rule for centers, when data is not normal, use the median as your average (center).

The average cholesterol for these 40 women is 194 mg/dL (median).

Note: In a skewed left data set, the mean will also be pulled in the direction of the skew. This will make the mean average too small. You can often get a good idea of the shape of a data set by just looking at the mean and median.

Normal Data: The mean and median are very close.
Both the mean and median are accurate, but we use the mean.

Skewed Right Data: The mean is significantly larger than the median. Only the median is accurate.

Skewed Left Data: The mean is significantly smaller than the median. Only the median is accurate.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Problem Set Section 5B

1. Put each of the following data sets in order from smallest to largest. Then calculate the median average (50th percentile).

a) 5 , 7 , 8 , 8 , 9 , 11 , 14 , 16 , 17 , 19 , 21 , 25 , 26 , 29 , 31 , 33 , 36

b) 2.1 , 3.8 , 5.1 , 6.9 , 7.2 , 10.4 , 11.3 , 14.7 , 15.1 , 16.0

c) 31 , 34 , 41 , 52 , 68 , 71 , 79 , 83 , 88 , 90 , 103

d) 150 , 152 , 154 , 155 , 157 , 159 , 163 , 164 , 165

e) 7.5 , 2.3 , 4.6 , 1.9 , 2.8 , 9.4 , 8.3 , 6.1

f) 21 , 29 , 23 , 26 , 25 , 19 , 28 , 31 , 32 , 20 , 18

2. Use StatKey and the "Bear Data" to calculate median average for each of the following data sets.

Directions: Open the "Bear Data" in Canvas or at www.matt-teachout.org. Go to www.lock5stat.com. Click on "One Quantitative Variable" under the "Descriptive Statistics and Graphs" menu. Click on "Edit Data". Copy and paste in the column of quantitative data indicated in the problem. If the data has a title, check the box that says "Data has a header row". If the data does NOT has a title, do NOT check the box that says "Data has a header row". Do NOT check the box that says "First column is an identifier". (You would only check the "identifier" box if there is a word next to every number in the data.) Now push "OK". The median will be listed under "summary statistics" on the top right of StatKey.

a) Median average for Bear Ages = _____Months

b) Median average for Bear Head Length = _____ Inches

c) Median average for Bear Head Width = _____ Inches

d) Median average for Bear Neck Circumference = _____ Inches

e) Median average for Bear Length = _____ Inches

f) Median average for Bear Chest Size = _____ Inches

g) Median average for Bear Weight = _____ Pounds

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Section 5C – Spread and Typical Values for Skewed Data, Quartiles, Interquartile Range (IQR), and the Five Number Summary

We have now seen that when data is not normal, we should use the median average as our measure of center and average.

The median is actually a type of quartile. Quartile analysis is an important part of understanding skewed.

Definition of Quartiles: The quartiles are three numbers that break the data into four equal groups. Think of them as three fences that separate the data into quarters when the data is in order.

First Quartile (Q1): This statistic is also called the 25th percentile and is the number that approximately 25% of the data is less than and 75% of the data is greater than.

Second Quartile (Q2): This statistic is also called the Median or the 50th percentile and is the number that approximately 50% of the data is less than and 50% of the data is greater than.

Third Quartile (Q3): This statistic is also called the 75th percentile and is the number that approximately 75% of the data is less than and 25% of the data is greater than.

Remember, when data set is skewed or not normal, we should not use the standard deviation to measure spread. So what measure of spread should we use for skewed data? In normal (bell-shaped) data, typical values are closer to the center. The empirical rule implies that for bell shaped data, about 68% is typical. In skewed data, the data is more spread out with less values being typical. For skewed data, we look for the middle 50% of the data for typical values. This is called the interquartile range.

Interquartile Range (IQR): The interquartile range is how far typical values are from each other in a skewed data set. IQR is the length between the middle 50% of the data values and is calculated by subtracting the third quartile (Q3) minus the first quartile (Q1).

Interquartile range formula: IQR = Q3 – Q1

Center and Spread Rule for Skewed Right, Skewed Left, or Non-normal Data

Center (Average) = Median (2nd Quartile)

Spread = Interquartile Range (IQR)

Typical Values = Between the 1st Quartile (Q1) and 3rd Quartile (Q3)

The Five Number Summary

A common way to summarize skewed or non-normal quantitative data is by listing the following five statistics in this order. The five numbers are often referred to as the "five number summary".

Five Number Summary: Minimum Value, 1st Quartile (Q1), Median (Q2), 3rd Quartile (Q3), Maximum Value

Example 1

Different statistics programs sometimes give slightly different values for the quartiles. People sometimes overemphasize these differences. The key is to remember the idea, finding three fences that separate the data into four equal groups. Each quarter should have approximately the same number of values in that group.

Let us calculate the three quartiles and the interquartile range (IQR) for the following data set.

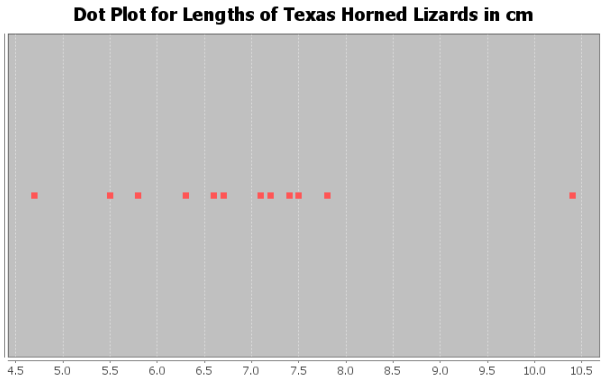
This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Texas Horned Lizard Lengths in Centimeters (cm)

6.7 , 10.4 , 7.8 , 4.7 , 5.5 , 5.8 , 7.2 , 7.5 , 7.1 , 6.3 , 6.6 , 7.4

When analyzing quantitative data, always determine the shape first.

Because of the one unusually large lizard of 10.4 cm, this data is probably skewed right. It is difficult to tell the shape of small data sets. Here is a dot plot of the data.



It does look like much of the data is bunched up on the left and the one unusual value on the right gives a longer tail to the right. So this data is skewed right. For skewed data, we should use the median for the average, IQR for the spread, and typical values in between the 1st and 3rd quartiles.

To calculate quartiles, we need to put the numbers in order first.

Texas Horned Lizard Length Data in order:

4.7 , 5.5 , 5.8 , 6.3 , 6.6 , 6.7 , 7.1 , 7.2 , 7.4 , 7.5 , 7.8 , 10.4

Now that the data is in order, think about quartering the data. Since there are 12 values in the data set, we should have 12 / 4 = 3 values in each quarter. Therefore, the quartiles should be placed between every three numbers.

4.7 , 5.5 , 5.8 | 6.3 , 6.6 , 6.7 | 7.1 , 7.2 , 7.4 | 7.5 , 7.8 , 10.4

Q1 Q2 Q3

Therefore, Q1 should be half way between 5.8 and 6.3

Q1 = (5.8 + 6.3) / 2 = 6.05

1st Quartile Sentence: About 25% of the horned lizards had a length less than 6.05 cm.

Q2 (median) should be half way between 6.7 and 7.1

Q2 (median) = (6.7 + 7.1) / 2 = 6.9

2nd Quartile (Median) Sentence: About 50% of the horned lizards had a length less than 6.9 cm. The second quartile is also the median average for skewed data. So the average length of the horned lizards is also 6.9 cm.

Q3 should be half way between 7.4 and 7.5

Q3 = (7.4 + 7.5) / 2 = 7.45

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

3rd Quartile Sentence: About 75% of the horned lizards had a length less than 7.45 cm.

This is how to think about quartiles. Notice we did not need a formula or fancy procedure to find the quartiles. We just needed to separate the data into four groups.

What about the interquartile range (IQR) for this data set?

IQR = Q3 – Q1 = 7.45 – 6.05 = 1.40 cm

Interquartile Range Sentence: Typical Horned Lizard Lengths (middle 50%) were within 1.40 cm from each other.

Horned Lizard Data Summary

Average lizard length = 6.9 inches

Spread = 1.40 cm

Typical Lizard Lengths = Between 6.05 cm and 7.45 cm.

Horned Lizard Length Five Number Summary: 4.7 cm, 6.05 cm, 6.9 cm, 7.45 cm, 10.4 cm

(Minimum Value, 1st Quartile (Q1), Median (Q2), 3rd Quartile (Q3), Maximum Value)

Notice the five number summary shows the smallest and largest values in the data set, as well as the average (6.9 cm) and the 1st and 3rd quartiles that typical values fall in between.

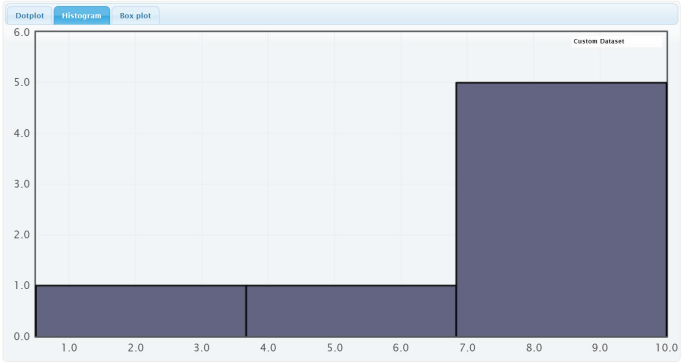
Example 2

There is some debate about how to calculate quartiles when the frequency is not divisible by four. This is especially true when there is an odd number of data values. Remember if there is an odd number of data values, the median (Q2) is an actual number in the data set. Since the median (Q2) is a value in the data set, it should be included in the top AND bottom halves of the data.

Let us look at the tips in dollars left at a bar. Here is the data in order.

$0.50 , $5.00 , $7.50 , $8.00 , $8.50 , $9.00 , $10.00

Let's start by looking at the shape of this data. Here is a histogram created with StatKey. Notice the data is skewed left. Hence we should use the median (Q2) as the average, the IQR as the spread, and typical values should be in between Q1 and Q3.



This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Quartile Calculations: To find the quartiles, always start by finding the median (second quartile). The data is already in order from smallest to largest. In the previous section, we learned that since the number of values is odd, Q2 would be the middle number of $8.00.

2nd Quartile (Median) Sentence: About 50% of the bar tips were less than $8.00. The second quartile is also the median average for skewed data. So the average bar tip was $8.00.

To find the first quartile (Q1), find the median of the bottom half of the data. Since the median is an actual value in the data, we will include it in our bottom half.

Bottom Half of Data: $0.50 , $5.00 , $7.50 , $8.00

Since we are including the median in the bottom half of the data, then there would be four numbers. The median of $0.50 , $5.00 , $7.50 and $8.00 would be just half way between $5.00 and $7.50 since this is an even number of values and there are two numbers in the middle.

Q1 = median of bottom half of the data $= \frac{(5.0+7.5)}{2} = \$6.25$

1st Quartile Sentence: About 25% of the bar tips were less than $6.25.

To find the third quartile (Q3), find the median of the top half of the data. Since the median is an actual value in the data, we will include it in the top half of the data.

Top Half of the Data: $8.00 , $8.50 , $9.00 , $10.00

If we include the median in the top half, then there would be four numbers. The median of $8.00 , $8.50 , $9.00 and $10.00 would be half way between $8.50 and $9.00 since this is an even number of values and there are two numbers in the middle.

Q3 = median of top half of the data $= \frac{(8.5+9.0)}{2} = \$8.75$

3rd Quartile Sentence: About 75% of the bar tabs were less than $8.75.

What about the interquartile range (IQR) for this data set?

IQR = Q3 – Q1 = $8.75 – $6.25 = $2.50

Interquartile Range Sentence: Typical bar tips (middle 50%) were within $2.50 from each other.

Bar Tips Data Summary

Average Bar Tip = $8.00

Spread = $2.50

Typical Bar Tips = Between $6.25 and $8.75.

Here is the summary statistics printout for the bar tip data from StatKey. Notice StatKey calculated the same median (Q2), the same 1st quartile (Q1) and the same 3rd Quartile (Q3). However, StatKey did not calculate the Interquartile Range (IQR). We would need to use the formula IQR = Q3 – Q1 to calculate it.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Summary Statistics

Statistic	Value
Sample Size	7
Mean	6.929
Standard Deviation	3.233
Minimum	0.5
Q_1	6.250
Median	8.000
Q_3	8.750
Maximum	10

Bar Tip Five Number Summary: $0.50 , $6.25 , $8.00 , $8.75 , $10.00

(Minimum Value, 1st Quartile (Q1), Median (Q2), 3rd Quartile (Q3), Maximum Value)

Notice again that the five number summary shows the smallest and largest values in the data set, as well as the average ($8.00) and the 1st and 3rd quartiles that typical values fall in between.

Take away

Some computer programs have slight differences in how the quartiles are calculated. One program might give the 1st quartile as 2.5 mm and another program might give 2.6 mm. This difference in how quartiles are calculated is not something to dwell on. In a data set with ten thousand numbers, the quartiles will be about the same no matter what program you are using. In small data sets like the previous example, there can be some discrepancy, but it is not something to worry about.

Again, the key is to explain the meaning of statistics like median, Q1, Q3 and IQR. Use technology to calculate the statistics. Take whatever value the program gives and use it. It matters very little if Q1 came out to be 78.4 degrees Fahrenheit or 78.3 degrees Fahrenheit. The important thing when analyzing skewed quantitative data, is to be able to explain that the average is the median (Q2), the spread is IQR, typical values are between Q1 and Q3, approximately 25% of the values in the data were less than Q1, and approximately 75% of data values were less than Q3 and the average is Q2 (median).

Calculating quartiles and IQR with technology
In large data sets, it is virtually impossible to calculate quartiles or any statistic for that matter with a calculator or by hand. We are now living in the age of "big data" where data sets often have hundreds of thousands of values or even millions of values. Surely, we cannot calculate the graphs and statistics we need from big data with a calculator. That is why it is so vital for data analysts to learn how to use statistics software like StatKey.

Here are the steps to calculating graphs and quantitative statistics with StatKey.

StatKey Directions: Go to www.lock5stat.com. Click on "One Quantitative Variable" under the "Descriptive Statistics and Graphs" menu. Click on "Edit Data". Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says "Data has a header row". If the data does NOT has a title, do NOT check the box that says "Data has a header row". Do NOT check the box that says "First column is an identifier". (You would only check the "identifier" box if there is a word next to every number in the data set.) Push "OK". You should see the median (average), 1st Quartile, and 3rd Quartile in the list of "Summary Statistics" on the top right of the page.

NOTE: You will need to calculate the interquartile range (IQR) by subtracting the 3rd quartile in StatKey minus the 1st quartile in StatKey. *(Interquartile Range Formula: IQR = Q3 – Q1)*

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Problem Set Section 5C

Directions: Put each of the following data sets in order from smallest to largest. Calculate the median (Q2), the first quartile (Q1), the third quartile (Q3) and the Interquartile Range (*IQR = Q3 – Q1).*

Give the five number summary for the data set *(Minimum, Q1, Median, Q3, Maximum)*.

1. { 5 , 7 , 8 , 8 , 9 , 11 , 14 , 16 , 17 , 19 , 21 , 25 , 26 , 29 , 31 , 33 , 36 }

   a) Calculate the Median (Q2).
   b) Calculate the 1st Quartile (Q1).
   c) Calculate the 3rd Quartile (Q3).
   d) Calculate the Interquartile Range IQR = Q3 – Q1
   e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

2. { 2.1 , 3.8 , 5.1 , 6.9 , 7.2 , 10.4 , 11.3 , 14.7 , 15.1 , 16.0 }

   a) Calculate the Median (Q2).
   b) Calculate the 1st Quartile (Q1).
   c) Calculate the 3rd Quartile (Q3).
   d) Calculate the Interquartile Range IQR = Q3 – Q1
   e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

3. { 31 , 34 , 41 , 52 , 68 , 71 , 79 , 83 , 88 , 90 , 103 }

   a) Calculate the Median (Q2).
   b) Calculate the 1st Quartile (Q1).
   c) Calculate the 3rd Quartile (Q3).
   d) Calculate the Interquartile Range IQR = Q3 – Q1
   e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

4. { 150 , 152 , 154 , 155 , 157 , 159 , 163 , 164 , 165 }

   a) Calculate the Median (Q2).
   b) Calculate the 1st Quartile (Q1).
   c) Calculate the 3rd Quartile (Q3).
   d) Calculate the Interquartile Range IQR = Q3 – Q1
   e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

5. { 7.5 , 2.3 , 4.6 , 1.9 , 2.8 , 9.4 , 8.3 , 6.1 }

   a) Calculate the Median (Q2).
   b) Calculate the 1st Quartile (Q1).
   c) Calculate the 3rd Quartile (Q3).
   d) Calculate the Interquartile Range IQR = Q3 – Q1
   e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

6. { 21 , 29 , 23 , 26 , 25 , 19 , 28 , 31 , 32 , 20 , 18 }

a) Calculate the Median (Q2).
b) Calculate the 1st Quartile (Q1).
c) Calculate the 3rd Quartile (Q3).
d) Calculate the Interquartile Range IQR = Q3 – Q1
e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

Direcitons: Use StatKey and the bear data to calculate minimum, maximum, median, Q1, Q3, and IQR for the following data sets. Then give the "five number summary" for each data set. Go to www.lock5stat.com. Click on "One Quantitative Variable" under the "Descriptive Statistics and Graphs" menu. Click on "Edit Data". Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says "Data has a header row". If the data does NOT has a title, do NOT check the box that says "Data has a header row". Do NOT check the box that says "First column is an identifier". (You would only check the "identifier" box if there is a word next to every number in the data set.) Push "OK". You should see the median (average), 1st Quartile, and 3rd Quartile in the list of "Summary Statistics" on the top right of the page.

NOTE: You will need to calculate the interquartile range (IQR) by subtracting the 3rd quartile in StatKey minus the 1st quartile in StatKey. *(Interquartile Range Formula: IQR = Q3 – Q1)*

7. Bear Ages in Months

a) Calculate the Median (Q2).
b) Calculate the 1st Quartile (Q1).
c) Calculate the 3rd Quartile (Q3).
d) Calculate the Interquartile Range IQR = Q3 – Q1
e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

8. Bear Head Length in Inches

a) Calculate the Median (Q2).
b) Calculate the 1st Quartile (Q1).
c) Calculate the 3rd Quartile (Q3).
d) Calculate the Interquartile Range IQR = Q3 – Q1
e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

9. Bear Head Width in Inches

a) Calculate the Median (Q2).
b) Calculate the 1st Quartile (Q1).
c) Calculate the 3rd Quartile (Q3).
d) Calculate the Interquartile Range IQR = Q3 – Q1
e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

10. Bear Neck Circumference in Inches

a) Calculate the Median (Q2).
b) Calculate the 1st Quartile (Q1).
c) Calculate the 3rd Quartile (Q3).
d) Calculate the Interquartile Range IQR = Q3 – Q1
e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

11. Bear Length in Inches

    a) Calculate the Median (Q2).
    b) Calculate the 1st Quartile (Q1).
    c) Calculate the 3rd Quartile (Q3).
    d) Calculate the Interquartile Range IQR = Q3 – Q1
    e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

12. Bear Chest Size in Inches

    a) Calculate the Median (Q2).
    b) Calculate the 1st Quartile (Q1).
    c) Calculate the 3rd Quartile (Q3).
    d) Calculate the Interquartile Range IQR = Q3 – Q1
    e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

13. Bear Weight in Pounds

    a) Calculate the Median (Q2).
    b) Calculate the 1st Quartile (Q1).
    c) Calculate the 3rd Quartile (Q3).
    d) Calculate the Interquartile Range IQR = Q3 – Q1
    e) Give the five number summary in order. (Minimum, Q1, Median, Q3, Maximum)

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Section 5D – Box Plots and Finding Unusual Values for Skewed Data

So far, we have seen that when a data set is skewed right, skewed left, or not normal, we should use the median as our center and average and the interquartile range (IQR) for the spread. We also learned that typical values will make up the middle 50% of data values and fall between the 1st and 3rd quartiles. What about finding unusual values (outliers) for skewed data sets?

Unusual Values
For bell shaped data sets, unusual values (outliers) are more than two standard deviations from the mean, but skewed data involves more extreme values and is more spread out. It therefore has a different rule for finding unusual values (outliers). Here are the unusual cutoff values for skewed or non-normal data.

Unusually High Cutoff for Skewed Data: Q3 + (1.5 x IQR)

Unusually Low Cutoff for Skewed Data: Q1 – (1.5 x IQR)

The good news is that the typical and unusual values for skewed data are summarized nicely with a box plot. The box plot is a fabulous graph to look at when your data is skewed right, skewed left or not normal.

I like to call the unusual cutoff values the "Box and a Half Rule", since 1.5 x IQR represents the length of a box and a half. So any values in the skewed data that is a box and half from the box are considered unusual (outlier).

Introduction to Box Plots
Let us look at how box plots work. Remember to use technology when you create a box plot. No statistician, data analyst, or data scientist creates graphs by hand, especially with big data sets.

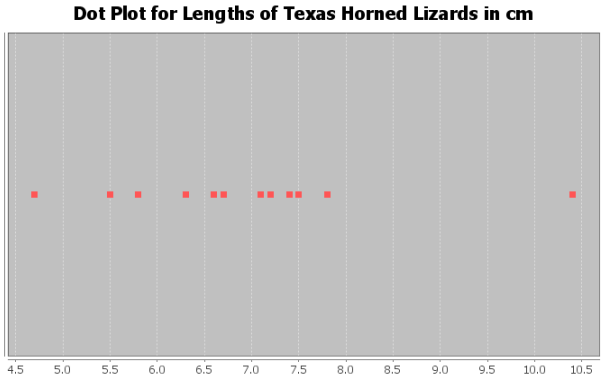
Let us look at an example where we do make the box plot by hand, just so we can understand the process.

Example 1
Let us look at the Texas Horned Lizard data and create a box plot for the data.

Texas Horned Lizard Length Data in order:

4.7 , 5.5 , 5.8 , 6.3 , 6.6 , 6.7 , 7.1 , 7.2 , 7.4 , 7.5 , 7.8 , 10.4

A dot plot of the data indicated a skewed right shape. Therefore, this works nicely for a box plot.



In the last section, we calculated the three quartiles and the interquartile range for this data.

4.7 , 5.5 , 5.8 | 6.3 , 6.6 , 6.7 | 7.1 , 7.2 , 7.4 | 7.5 , 7.8 , 10.4

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Q1 Q2 Q3

Q1 should be half way between 5.8 and 6.3

Q1 = (5.8 + 6.3) / 2 = 6.05

Q2 (median average) should be half way between 6.7 and 7.1

Q2 (median) = (6.7 + 7.1) / 2 = 6.9

Q3 should be half way between 7.4 and 7.5

Q3 = (7.4 + 7.5) / 2 = 7.45

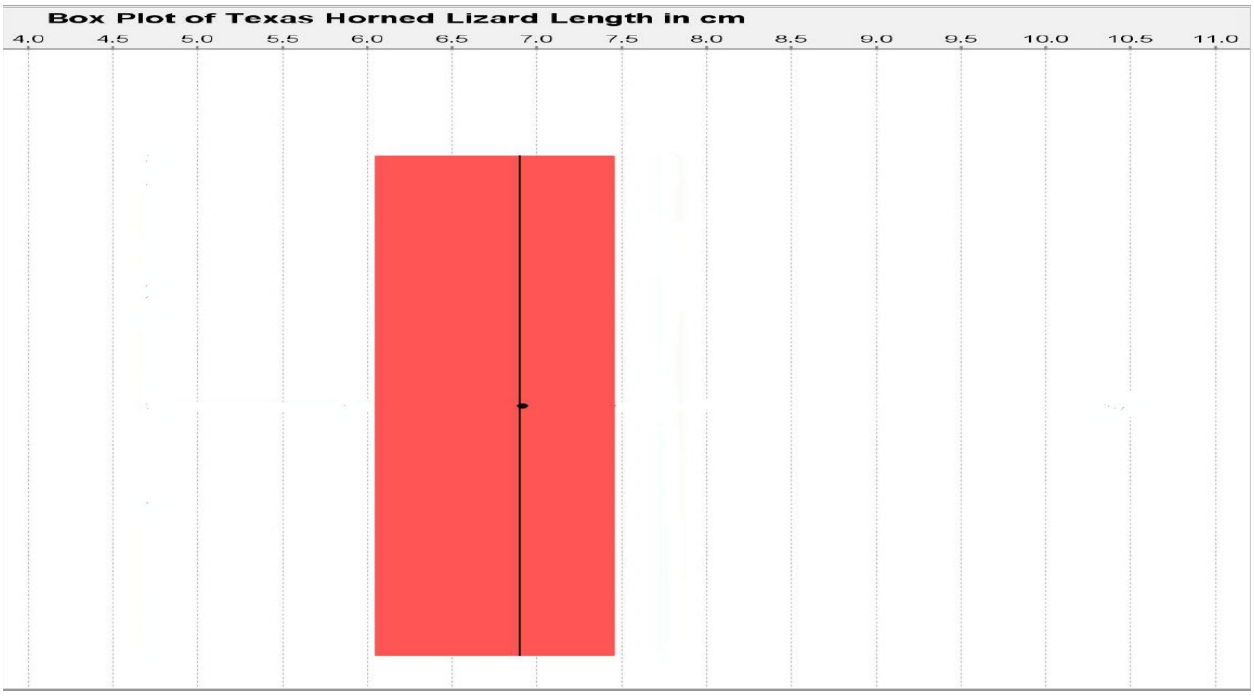
IQR = Q3 – Q1 = 7.45 – 6.05 = 1.40 cm

Typical Values: Since this is skewed data, typical values will fall in between Q1 and Q3. So typical horned lizards in this data set have a length between 6.05 cm and 7.45 cm.

Note: Q1 and Q3 do not accurately represent typical values in bell shaped data. You would need to use the standard deviation and the mean in that case.

Making the box plot

Start by drawing an even number line that goes from the smallest and largest values in the data set. Then draw a box from Q1 to Q3. Draw a line in the box at the median average (Q2).



Now we need to calculate the unusual cutoff fences to determine if there are any unusual values (outliers) in the data set. To calculate the outliers, we will need to find the distance of a box and a half (1.5 x IQR). In this data 1.5 x IQR

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

= 1.5 x 1.40 = 2.1. So any data values that are 2.1 or higher from Q3 are high outliers (unusually high values). Also any data values that are 2.1 or lower from Q1 are low outliers (unusually low values).

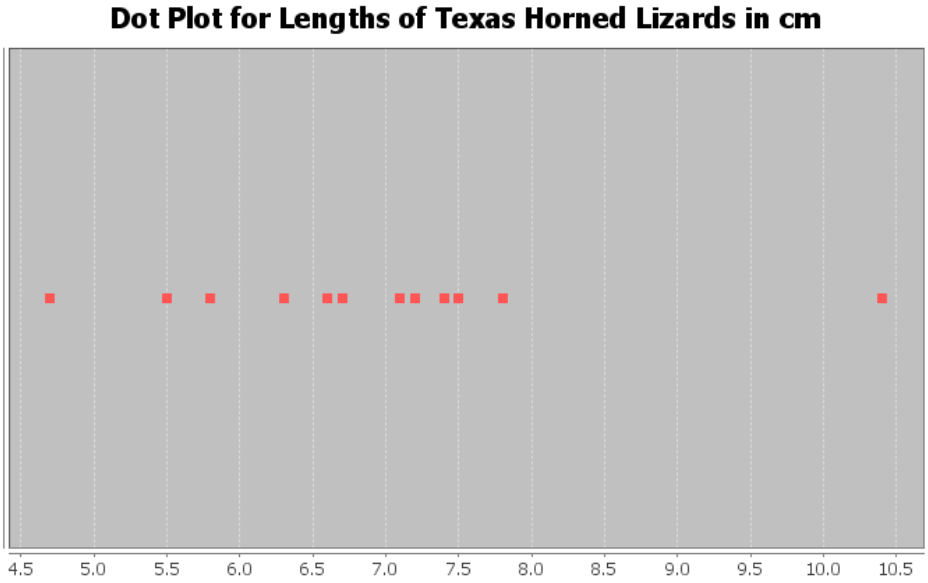
Unusually High Cutoff for Skewed Data:

Q3 + (1.5 x IQR) ≈ 7.45 + (1.5 x 1.40) ≈ 7.45 + 2.1 ≈ 9.55

Unusually Low Cutoff for Skewed Data:

Q1 – (1.5 x IQR) ≈ 6.05 – (1.5 x 1.40) ≈ 6.05 – 2.1 ≈ 3.95

Let us look at the dot plot again and see if there are any numbers that are 3.95 or lower. We can also look to see if there are any numbers that are 9.55 or higher.



Notice there are no values in the data set that are 3.95 cm or below. That means there are no unusually low values in the data set.

There is one value in the data set that is 9.55 cm or higher. It is the maximum value of the data set 10.4 cm. Therefore, 10.4 cm is an unusually high value (high outlier) in the data set. We need to designate that value as an outlier (unusual). Some computer programs draw their outliers with a circle, some draw it with a triangle, and some draw it with a star. I will draw it with a triangle.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

The box plot is also called the "box and whisker plot". Now we need to determine where to draw the whiskers. The whiskers are drawn to the highest and lowest numbers in the data set that are not outliers (not unusual). Be careful. The whiskers are not drawn to the unusual cutoff fences. They must be drawn to numbers that are actually in the data set and are not outliers.

There was no unusually low value in the data set. Therefore, the low whisker on the left should be drawn to the smallest number in the data set, which is 4.7 cm.

There was an unusually high value (outlier) at 10.4 cm. That means we cannot draw the whisker to that value. We must choose a new maximum value in the data set that is not an outlier. Looking at the dot plot, we see that the next biggest number in the data set was 7.8 cm. That is 9.55 cm or below so it is not unusual. We will draw the high whisker (on the right) to 7.8 cm since that is the largest number in the data set that is not an outlier (not unusual).

Texas Horned Lizard Length Data in order:

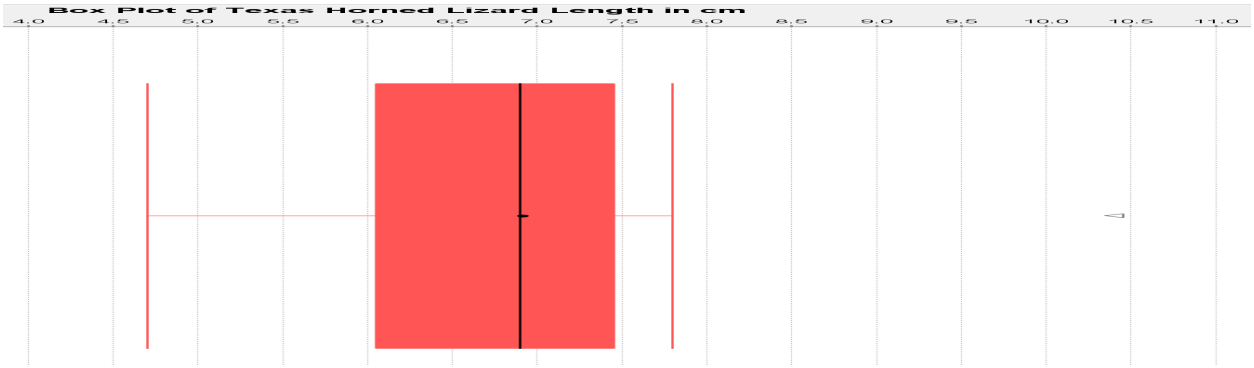
4.7 , 5.5 , 5.8 , 6.3 , 6.6 , 6.7 , 7.1 , 7.2 , 7.4 , 7.5 , 7.8 , 10.4

(No Low Outliers) Low Whisker — Typical Values — High Whisker — High Outlier

Note: Notice that there are values in the data that are neither typical nor an outlier. Some students make the mistake of thinking that if a data value is not typical, it is unusual. That is NOT true. There are many values in data sets that are not typical and not outliers.

Putting this information into our Box Plot, gives us the following graph.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021



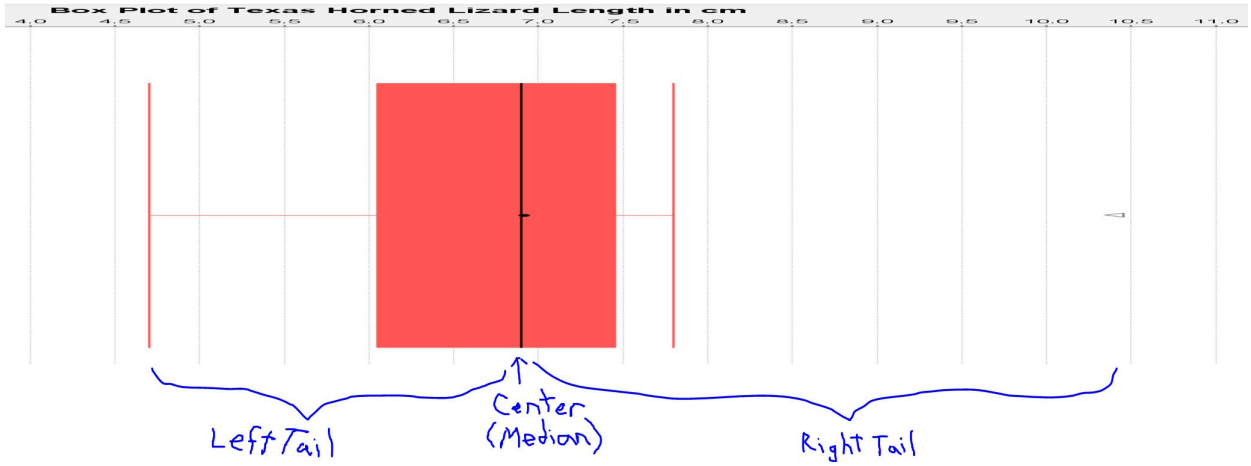
The whiskers probably get their name because they kind of look like cat whiskers. This is our complete box and whisker plot or box plot for short.

We can learn a lot by looking a box plot like this.

- The line in the box shows the median average.
- The edges of the box show the 1st and 3rd quartiles. Typical values will be in between them.
- The whiskers show the values that are neither typical, nor an outlier. The end of the left whisker shows the smallest number in the data set that is not an outlier. The end of the right whisker shows the largest number in the data set that was not an outlier.
- Stars, circles or triangles outside the whiskers are outliers (unusually low and high values).

Determining shape from a box plot

It is usually best to look at a histogram or dot plot when determining shape. Remember, a box plot is really a graph of the quartiles and outliers. However, since the median is the center and the line inside the box, we can at least compare the tails. In the box plot, we can look at the distance from the median to the largest value in the data and the distance from the median to the smallest value in the data. Remember this data did not have any low outliers so the smallest value is found at the end of the low whisker. Since the right tail is longer than the left tail, this is likely to be a skewed right data set.



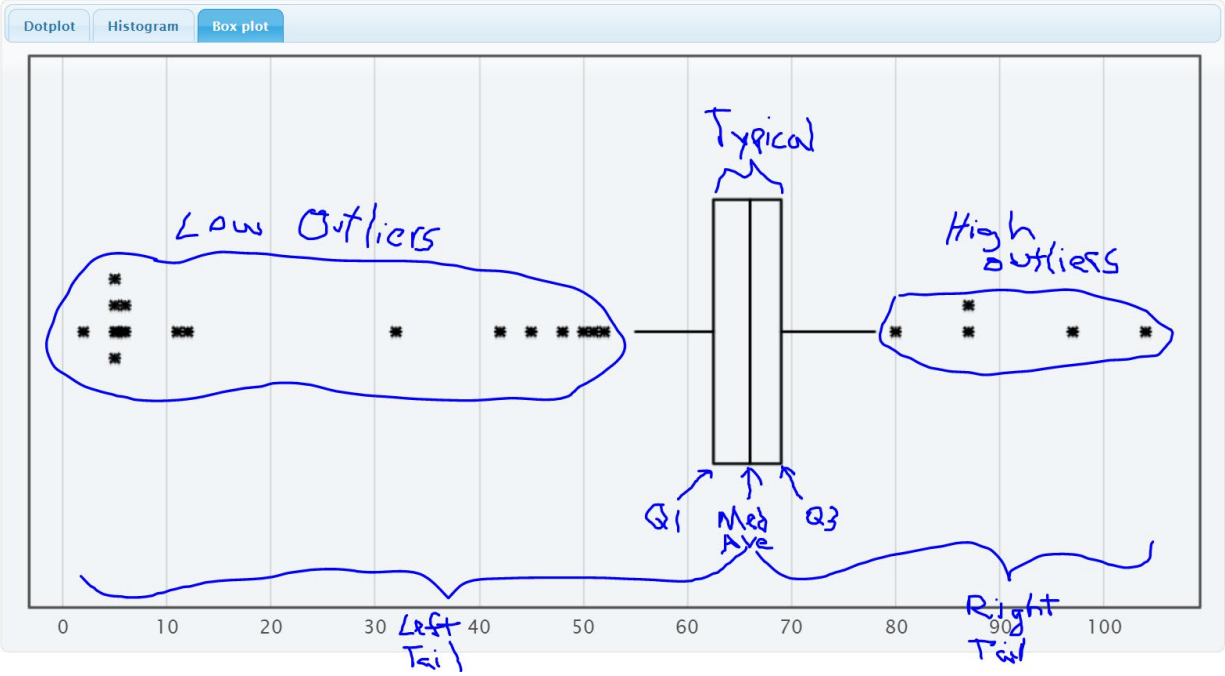
Creating Box Plots with Technology
Let us look at how to create box plots with StatKey.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Making a box plot in StatKey: Go to www.lock5stat.com. Click on "One Quantitative Variable" under the "Descriptive Statistics and Graphs" menu. Click on "Edit Data". Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says "Data has a header row". If the data does NOT has a title, do NOT check the box that says "Data has a header row". Do NOT check the box that says "First column is an identifier". (You would only check the "identifier" box if there is a word next to every number in the data.) Now push "OK". At the top left of the graph, click on the "box plot" tab.

Example

In the spring 2016 semester, we asked math 140 statistics students how tall they were in inches. We created a box plot using StatKey. It is good to refer the summary statistics to see the actual value of the median and quartiles. Also if you hold your cursor over the stars (outliers), StatKey will tell you the value of the outlier.



Summary Statistics

Statistic	Value
Sample Size	357
Mean	64.390
Standard Deviation	12.040
Minimum	2
Q_1	62.500
Median	66.000
Q_3	69.000
Maximum	104

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

We can see a lot of information from this graph.

- The left tail is longer than the right tail so the data is likely to be skewed left.
- Since it is skewed, we should use the median and quartiles instead of the mean and standard deviation.
- The median average height of the stat students was 66 inches $\left(5\frac{1}{2}\ feet\right)$.
- Typical heights fell between Q1 (62.5 inches) and Q3 (69 inches).
- The width of the box is the spread (IQR = 69 – 62.5 = 6.5 inches). So typical statistics students have a height within 6.5 inches of each other.
- The low outlier cutoff is Q1 – (1.5 x IQR) = 62.5 – (1.5x6.5) = 62.5 – 9.75 = 52.75 inches. So any height below 52.75 inches would be considered a low outlier (unusually low). There are 18 low outliers ranging from 2 inches to 52 inches. Putting the data in order, we can identify the outliers in the excel spreadsheet below.Some students that said they were only a few inches tall. Maybe they thought the question was in feet.
- The high outlier cutoff is Q3 + (1.5 x IQR) = 69 + (1.5x6.5) = 69 + 9.75 = 78.75 inches. So any height above 78.75 inches would be considered a high outlier (unusually high). There are five high outliers ranging from 80 inches to 104 inches. Putting the data in order, we can identify the outliers in the excel spreadsheet below.
  Some of these might also be a miscalculation. 80 inches (6 ft 8 in) is possible but 104 inches is 8 ft 8 inches. There is probably no stat student that tall.
- The whiskers indicate that there are some students whose heights are neither typical not unusual.

Excel spreadsheet showing low outliers

2	Low Outlier	
5	Low Outlier	
5	Low Outlier	
5	Low Outlier	
5	Low Outlier	
5.2	Low Outlier	
5.7	Low Outlier	
6	Low Outlier	
6	Low Outlier	
11	Low Outlier	
12	Low Outlier	
32	Low Outlier	
42	Low Outlier	
45	Low Outlier	
48	Low Outlier	
50	Low Outlier	
51	Low Outlier	
52	Low Outlier	
	Low Outlier Cutoff = 52.75	
55		
56		
56		

Excel Spreadsheet showing high outliers

76		
77		
77		
77.5		
78		
	High Outlier Cutoff = 78.75	
80	High Outlier	
87	High Outlier	
87	High Outlier	
97	High Outlier	
104	High Outlier	

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Summary Paragraph

In our previous chapter, we saw that we can write a summary paragraph with all the information about a quantitative data set. For normal data, our average, spread, typical values and outliers were calculated with the mean average and standard deviation. For data that is skewed left, skewed right, or not normal, the mean and standard deviation are not accurate. So we will use the median average, quartiles and interquartile range.

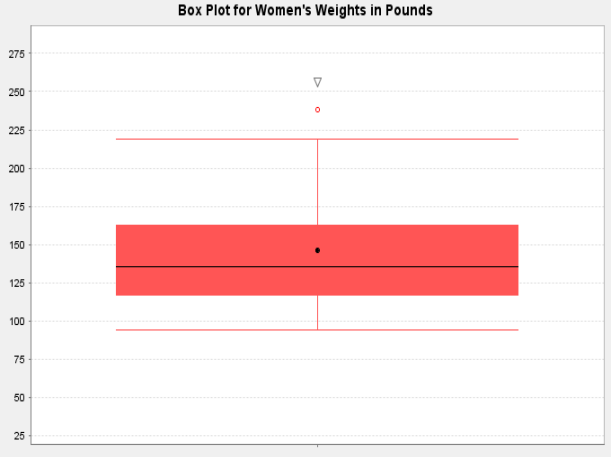
Here is the summary paragraph for the height of COC stat students in spring 2016. Notice since the data is skewed left we used the median average and IQR.

Summary Paragraph:

The data measured the heights in inches of 357 math 140 stat students in the spring 2016 semester. The shape of this quantitative data is skewed left. The median average height of the stat students was 66 inches $\left(5\frac{1}{2}\ feet\right)$*. The spread of the data was 6.5 inches (IQR). So typical statistics students have a height within 6.5 inches of each other. Typical heights fell between Q1 (62.5 inches) and Q3 (69 inches). There are 18 low outliers ranging from 2 inches to 52 inches. Some students that said they were only a few inches tall. These data values are obvious mistakes. There are five high outliers ranging from 80 inches to 104 inches. Some of these might also be a miscalculation. 80 inches (6 ft 8 in) is possible but 104 inches is 8 ft 8 inches. There is probably no stat student that tall.*

Vertical Box Plots

Box plots can also be drawn vertically. Here is an example. Now high outliers are at the top and low outliers are at the bottom.

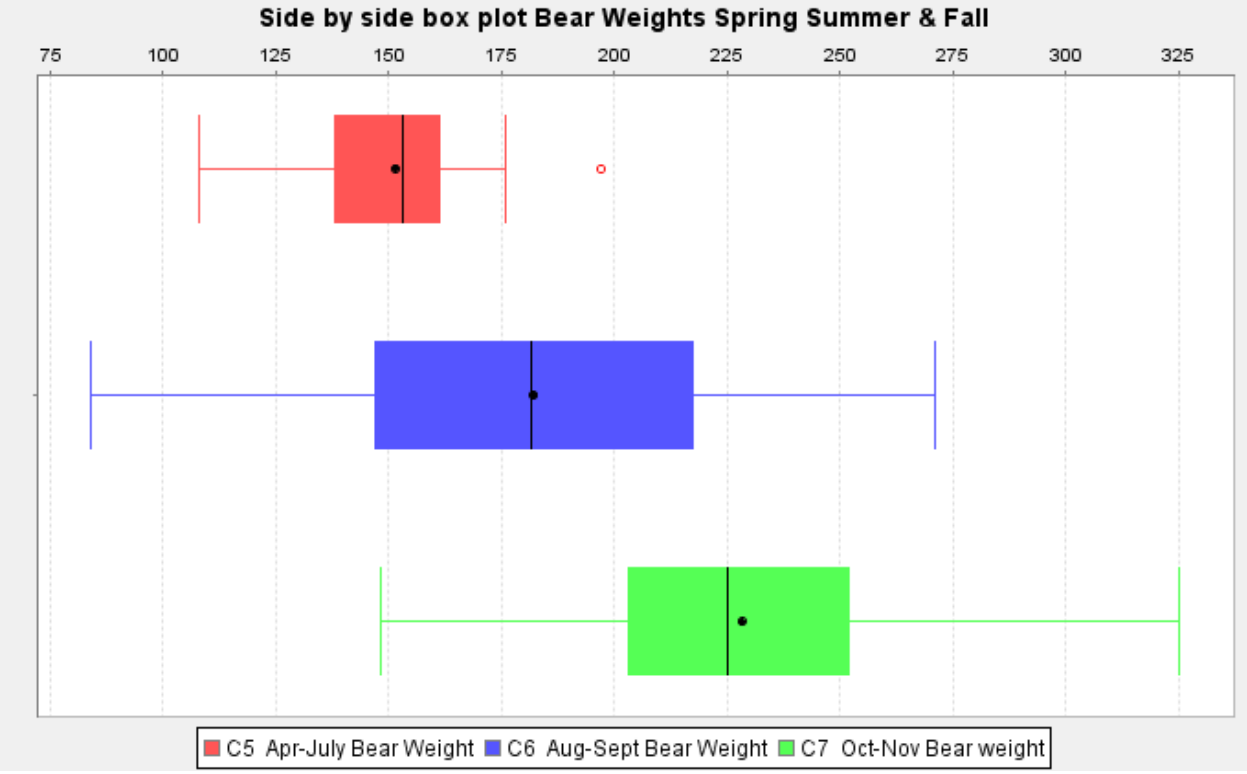


Interpreting Box Plots

Remember to use technology to create graphs and find statistics. The important part is being able to interpret what the graph and statistics are telling us.

Box plots are often used to compare quantitative data from different groups. We call this kind of graph a "side by side" box plots. These graphs can be drawn horizontally or vertically. The following example was found from the North American black bear data. The bear weights were separated into three groups depending on what time of the year the measurements were taken (Spring, Summer or Fall). The box on top is describing bears measured in spring (April – July), the box in the middle is describing bears measured in summer (August – September) and the box on the bottom is describing bears measured in fall (October – November).

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021



Graphs like this give us a lot of information.

Which group of bears had the highest average weight and what was the highest average weight?

Notice the lines inside the box are the medians, which are very accurate measures of center. The group whose median line is farthest to the right is the fall bears (October-November). The bears measured in fall had the highest average weight. The line in the box looks like it falls around 225 pounds. So the average weight of the North American black bears measured in fall was 225 pounds. The average weight of the bears measured in spring was about 155 pounds and the average weight of bears measured in summer was about 180 pounds.

Which group of bears had the most typical spread (variability) in their weights? *Typical spread (IQR) is the length of your box. So which group had the longest box? We can see it is the middle group of bears measured in summer (August – September). The bears measured in summer had the most variability in their weights. A typical bear measured in summer could have a weight from 145 pounds to 215 pounds.*

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Problem Set Section 5D

(#1-3) Directions: The median, 1st quartile and 3rd quartile are given for the following data sets. Calculate the IQR and outlier fences. Then identify the outliers, answer the questions, and draw a box plot for the data on a piece of paper. The data sets are already in order.

1. { 4 , 5 , 19 , 20 , 21 , 22 , 23 , 24 , 26 , 27 , 28 , 29 , 30 , 32 , 33 , 51 }

Median = 25
Q1 = 20.5
Q3 = 29.5

a) Calculate the IQR = Q3 – Q1

b) Calculate the high outlier fence Q3 + (1.5xIQR)

c) Calculate the low outlier fence Q1 – (1.5xIQR)

d) List all of the high outliers. (Data values larger than the high outlier fence.)

e) List all of the low outliers. (Data values smaller than the low outlier fence.)

f) What is the largest value in the data set that is NOT an outlier.
(This is where the right whisker will go.)

g) What is the smallest value in the data set that is NOT an outlier.
(This is where the left whisker will go.)

h) Draw the box plot including whiskers, outlier fences and outliers.

2. { 23 , 31 , 32 , 33 , 34 , 35 , 36 , 37 , 55 }

Median = 34
Q1 = 32
Q3 = 36

a) Calculate the IQR = Q3 – Q1

b) Calculate the high outlier fence Q3 + (1.5xIQR)

c) Calculate the low outlier fence Q1 – (1.5xIQR)

d) List all of the high outliers. (Data values larger than the high outlier fence.)

e) List all of the low outliers. (Data values smaller than the low outlier fence.)

f) What is the largest value in the data set that is NOT an outlier.
(This is where the right whisker will go.)

g) What is the smallest value in the data set that is NOT an outlier.
(This is where the left whisker will go.)

h) Draw the box plot including whiskers, outlier fences and outliers.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

3. { 8.4 , 9.6 , 10.8 , 10.9 , 11.0 , 11.2 , 11.3 , 11.4 , 11.6 , 11.7 , 12.9 , 13.1 }

Median = 11.25
Q1 = 10.85
Q3 = 11.65

a) Calculate the IQR = Q3 – Q1

b) Calculate the high outlier fence Q3 + (1.5xIQR)

c) Calculate the low outlier fence Q1 – (1.5xIQR)

d) List all of the high outliers. (Data values larger than the high outlier fence.)

e) List all of the low outliers. (Data values smaller than the low outlier fence.)

f) What is the largest value in the data set that is NOT an outlier.
(This is where the right whisker will go.)

g) What is the smallest value in the data set that is NOT an outlier.
(This is where the left whisker will go.)

h) Draw the box plot including whiskers, outlier fences and outliers.

(#4-7) Directions: Use StatKey and the Math 075 Survey Data Fall 2015 to create Box Plots for the following data sets. Draw a rough sketch of the box plot on a piece of paper or save the box plot on a word document.

Go to www.lock5stat.com. Click on "One Quantitative Variable" under the "Descriptive Statistics and Graphs" menu. Click on "Edit Data". Copy and paste in the column of quantitative data you want to analyze. If the data has a title, check the box that says "Data has a header row". If the data does NOT has a title, do NOT check the box that says "Data has a header row". Do NOT check the box that says "First column is an identifier". (You would only check the "identifier" box if there is a word next to every number in the data.) Now push "OK". At the top left of the graph, click on the "box plot" tab.

4. Age of Math 075 students in Fall 2015 (Column "C").

5. Work hours per week for Math 075 students in Fall 2015 (Column "J").

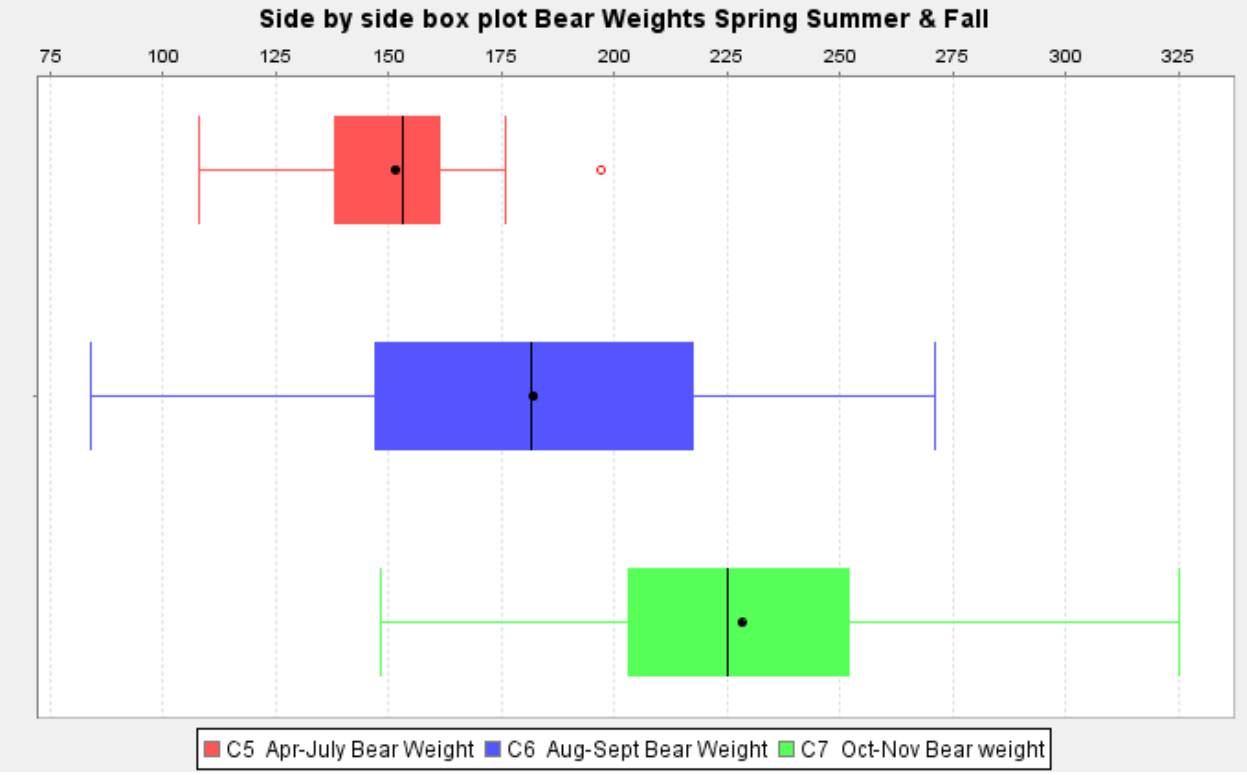
6. Exercise hours per week for Math 075 students in Fall 2015. (Column "L")

7. Commute time to campus in minutes for Math 075 students in Fall 2015. (Column "N")

Directions: Let us look again at the side-by-side box plot describing the weights of bears measured at different times of the year. The box on top is describing bears measured in spring (April – July), the box in the middle is describing

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY"
Creative Commons Attribution 4.0 International License – 3/17/2021

bears measured in summer (August – September) and the box on the bottom is describing bears measured in fall (October – November).



8. Which group of bears had the lowest median average weight? Estimate the lowest median average weight from the graph for the three groups?

9. Which group of bears had the lowest typical spread (IQR) in their weights? *(These are the bears whose weights were most consistent.)* Estimate the smallest IQR from the graph? Estimate the two values that typical values fall in between (Q1 and Q3) for the smallest spread group.

10. Were there any outliers (unusual bear weights) in any of the three groups (yes or no)? If so, what group had an unusual value (outlier)? Was it a high outlier or a low outlier? Estimate the bear weight or weights from the graph that were considered unusual.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Section 5E – Measures of Center, Spread and Position

Statistics: Numbers calculated from sample data in order to understand the characteristics of the data.

Though the mean, median, standard deviation and IQR are used most often to analyze quantitative data, there are many different types of statistics that can also be used to dig deeper into the data. We will not be covering these statistics in depth, but it is good to at least have an idea of what they measure.

Memorize the following definitions so that you can explain these statistics if needed. You should also know if the statistic is a measure of center, spread or position.

- Measures of center *(mean, median, mode and midrange)* are types of averages.
- Measures of spread *(standard deviation, variance, range, and interquartile range)* measure variability or how much the data is spread out.
- Measures of position *(min, max, 1st quartile (Q1) and 3rd quartile (Q3))* are statistics that we often use to identify where a data value falls compared to these positions.

Measures of Center
Mean Average: The balancing point in terms of distances. The measure of center or average used when a quantitative data set is bell shaped (normal). The mean average is calculated by adding all of the data values and then dividing that sum by how many numbers are in the data.

Median Average: The center of the data in terms of order. Also called the second quartile (Q2) or the 50th percentile. Approximately 50% of the data will be less than the median and 50% will be above the median. This is the measure of center or average used when a data set is skewed (not bell shaped).

Mode: The number that occurs most often in a data set. Data sets may have no mode, one mode, or multiple modes. It is also sometimes used in bimodal or multimodal data.

Midrange: A quick measure of center that is usually not very accurate, but can be calculated quickly without a computer. The midrange lies half way between the smallest and largest values in the data.
Midrange = (Max + Min) ÷2

Measures of Spread
Standard Deviation: How far typical values are from the mean in a normal (bell shaped) data set. It is the most accurate measure of spread for normal quantitative data. If you add and subtract the mean and standard deviation, you get two numbers that typical values in a bell shaped data set fall in between. It can also be used to find unusual values in bell shaped data. The standard deviation should not be used unless the data is bell shaped.

Variance: The standard deviation squared. A measure of spread used in ANOVA testing. Only accurate when the data is normal (bell shaped).

Range: The distance between the max and the min. All the data values are within this amount of each other. Range is a quick measure of spread that is not very accurate. It is based on unusual values and does not measure typical spread in the data set. It can be calculated quickly without a computer. The range is calculated by subtracting the maximum value in the data minus the minimum value in the data. (Range = Max – Min)

Interquartile range (IQR): How far typical values are from each other in a skewed data set. Measures the length of the middle 50% of the data. It is the most accurate measure of spread for skewed data sets. Interquartile range should not be used when data is bell shaped. Interquartile range is calculated by subtracting the 3rd quartile minus the 1st quartile. (IQR = Q3 – Q1)

Measures of Position
Minimum: The smallest number in the data set.

Maximum: The largest number in the data set.

First Quartile (Q1): The number that approximately 25% of the data is less than and 75% of the data is greater than. Used for finding typical values for skewed data sets.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Third Quartile (Q3): The number that approximately 75% of the data is less than and 25% of the data is greater than. Used for finding typical values for skewed data sets.

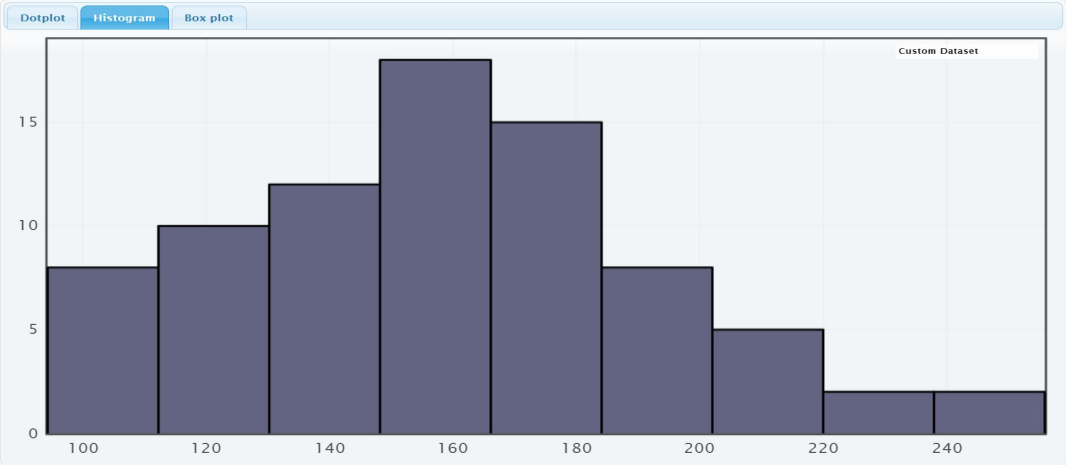
Total Frequency or Sample Size (n)
The total frequency or sample size of a data set (n) is not a measure of center, spread or position, but is important bit of information. It tells us how many numbers are in the data set.

Example

Here is a StatKey printout of the weights in pounds of 80 randomly selected adults. Explain each of the statistics listed above in context.

Summary Statistics

Statistic	Value
Sample Size	80
Mean	159.385
Standard Deviation	34.877
Minimum	94.3
Q_1	135.000
Median	161.000
Q_3	179.600
Maximum	255.9



This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Here is a similar printout from another computer program called Statcato. Notice it is similar to StatKey but has more statistics listed.

Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
Weight (Lbs)	159.385	34.877	1216.397

Variable	Q1	Median	Q3	IQR	Mode	N for mode
Weight (Lbs)	135.0	161.0	179.75	44.75	161.9, 135.0, 156.3	2

Variable	Min	Max
Weight (Lbs)	94.3	255.9

Variable	N total
Weight (Lbs)	80

Mean Average Weights: The measure of center or average weight for these adults that balances the distances is 159.385 pounds. This average is only accurate if the data is normal (bell shaped).

Median Average Weights: The measure of center or average when the data values are put in order is 161 pounds. So approximately 50% of the adults weighted less than 161 pounds and approximately 50% of the adults weighed more than 161 pounds.

Mode: StatKey does not calculate the mode. We do see it on the Statcato printout though. The weights that appear most often in a data set are 135.0 pounds, 156.3 pounds and 161.9 pounds. Looking at "N for mode" we see that these three weights all appeared twice in the data set. This data set would be considered multimodal since it has three modes. Notice that two of the modes were close to the median (center) of the data.

Midrange: The midrange is not listed in either printout, but can be easily calculated.
Midrange = (Max + Min) ÷2 = (255.9 + 94.3) ÷2 = (350.2) ÷2 = 175.1 pounds. The midrange average weight was 175.1 pounds. Notice it is not close to the actual center (median) of the data, so not a very accurate average.

Measures of Spread
Standard Deviation: Typical values are within 34.877 pounds of each other. The standard deviation is only accurate if the data is bell shaped.

Variance: StatKey does not list the variance, but it can be easily calculated by squaring the standard deviation. Variance = 34.877 x 34.877 ≈ 1216.405. The variance or standard deviation squared is approximately 1216.4 square pounds. The variance is only accurate when the data is normal (bell shaped). *(Notice the Statcato printout lists the variance as 1216.397. Statcato is more accurate in this case because it keeps more decimal places. StatKey rounds the standard deviation to the thousandths place, so when we use it to calculate, we have a little bit of a rounding error.)*

Range: StatKey does not list the range but it can be easily calculated. Range = Max – Min = 255.9 − 94.3 = 161.6 pounds. The distance between the max and the min for the weight data is 161.6 pounds. All the data values are within 161.6 pounds of each other.

Interquartile range (IQR): StatKey does not list the interquartile range (IQR) but it can be easily calculated.
IQR = Q3 – Q1 = 179.6 − 135 = 44.6.6 pounds. Typical data values are within 44.6 pounds of each other. This would only be accurate if the data is not normal. *(Notice that Statcato lists the IQR as 44.75 pounds. Computer programs often have slight differences in the quartile calculations. They are close though, so either would be ok to use.)*

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Measures of Position
Minimum: The lightest adult in the data set was 94.3 pounds.

Maximum: The heaviest adult in the data set was 255.9 pounds.

First Quartile (Q1): Aproximately 25% of the adults in the data weighed less than 135 pounds.

Third Quartile (Q3): Aproximately 75% of the adults in the data weighed less than 179.6 pounds (StatKey).
(Notice that Statcato lists Q3 as 179.75 pounds. Computer programs often have slight differences in the quartile calculations. They are close though, so either would be ok to use.)

Total Frequency or Sample Size (n): Weight data was collected from a total of 80 adults. *(Notice StatKey lists this statistic as "sample size" and Statcato lists it as "N total".*

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Problem Set Section 5E

1. For each of the following statistics, classify it as a measure of center, spread or position.

a) Q1
b) Mean
c) Variance
d) Midrange
e) Standard Deviation
f) Minimum
g) Q3
h) Mode
i) IQR
j) Median
k) Range
l) Maximum

2. The following statistics were created from some weekly salary data in dollars from people living in Victoria, Australia. Write a sentence or two explaining the meaning of each of these statistics in context.

Descriptive Statistics

Variable	Mean	Standard Deviation	Variance
Victoria Salary	1149.050	516.553	266826.719

Variable	Q1	Median	Q3	IQR	Mode	N for mode
Victoria Salary	703.45	1015.74	1496.11	792.660	1011	2

Variable	Min	Max	Range
Victoria Salary	371.57	2396.28	2024.710

Variable	N total
Victoria Salary	35

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Chapter 5 Review Sheet

Here is a list of important ideas in this chapter.

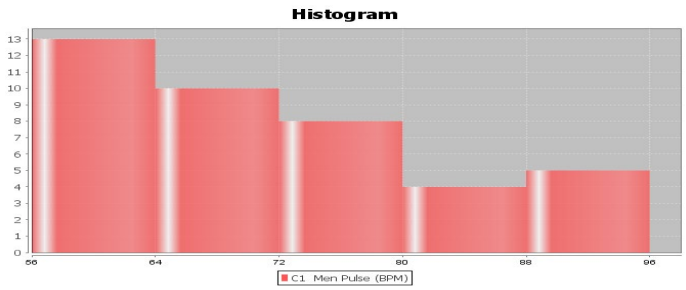
- Be able to distinguish between categorical data and quantitative (numerical measurement) data.
- Be able to create histograms and dot plots with technology and find the shape of a quantitative data set.
- Be able to find the five number summary (minimum, Q1, median, Q3, Maximum) with a calculator and with technology. Also find the interquartile range (IQR) and the total frequency (N).
- Write sentences to explain Q1, Median, Q3 and IQR.
- The interquartile range (IQR) tells us the maximum distance that typical values are from each other in a skewed data set. It measures the spread for the middle 50% and is the most accurate spread for skewed data sets.
- The first quartile Q1 is a divider that about 25% of the data values are less than and about 75% of the data values are greater than.
- The third quartile Q3 is a divider that about 75% of the data values are less than and about 25% of the data values are greater than.
- A center gives an average value for the data set is usually close to the highest bar or bars in the histogram.
- If a data set is skewed, we should use the median average as our measure of center and our average for the data set.
- 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.
- If a data set is skewed, we should use the interquartile range (IQR) as our measure of spread for the data set. If a data set is bell shaped, then we should not use the IQR.
- For Skewed Data: Q1 ≤ Typical Values ≤ Q3
- Unusually High Cutoff for Skewed Data: Q3 + (1.5 x IQR) *(Automatically calculated in a box plot)*
- Unusually Low Cutoff for Skewed Data: Q1 – (1.5 x IQR) *(Automatically calculated in a box plot)*
- Be able to read and use a box plot to understand quartiles and percentages and identify unusual values in the data set.
- Be able to write a summary report paragraph summarizing the key characteristics of a skewed quantitative data set.
- Be able to classify various statistics as a measure of center, spread or position.

Problems Chapter 5 Review Sheet

Directions: Give the shape of each of the following graphs from the men's health data. Then decide what the best measure of center and spread would be. (Mean/standard deviation or median/IQR?)

1. Men's Pulse Rate in Beats per Minute (BPM)

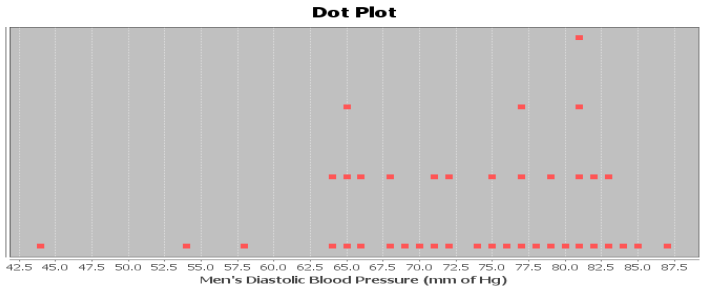
Shape = _____ Mean/Stand Dev **OR** Median/IQR? _____



This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

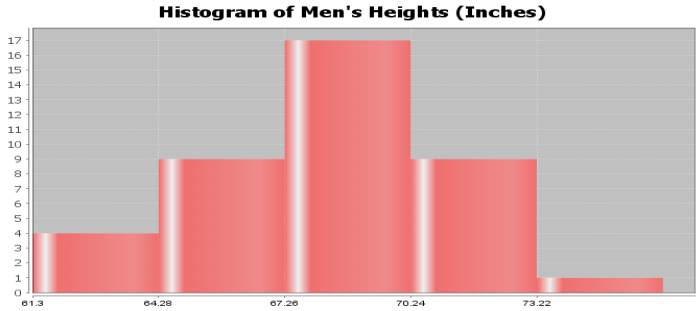
2. Men's Diastolic Blood Pressure in Millimeters of Mercury (mm of Hg)

Shape = _____ Mean/Stand Dev **OR** Median/IQR? _____



3. Men's Heights (inches)

Shape = _____ Mean/Stand Dev **OR** Median/IQR? _____



4. Calculate the Median, Q1, Q3 and IQR for the following data. The 16 numbers are already in order. Show work and put your answers in the spaces below.

17 , 19 , 20 , 26 , 28 , 31 , 35 , 37 , 41 , 43 , 44 , 48 , 51 , 53 , 55 , 62

Median Average = _____

Q1 = _____

Q3 = _____

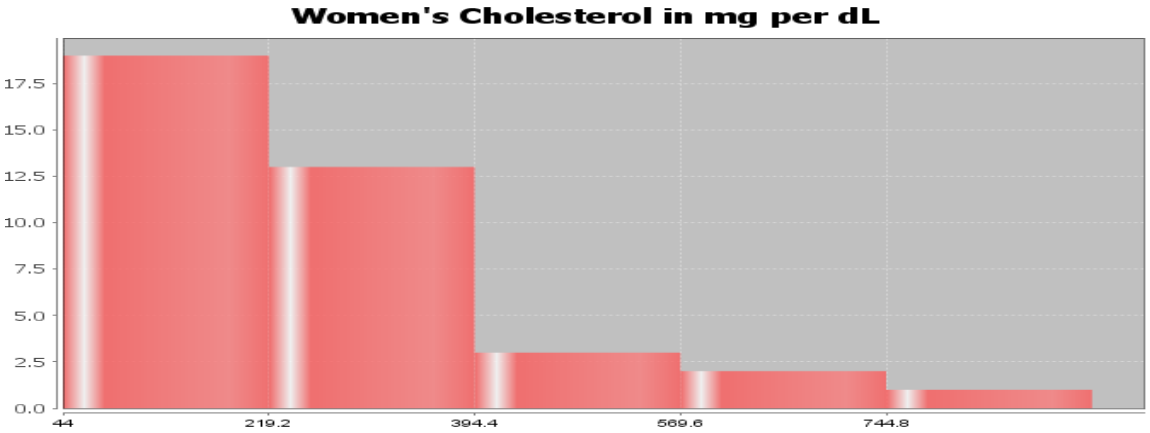
IQR = Q3-Q1 = _____

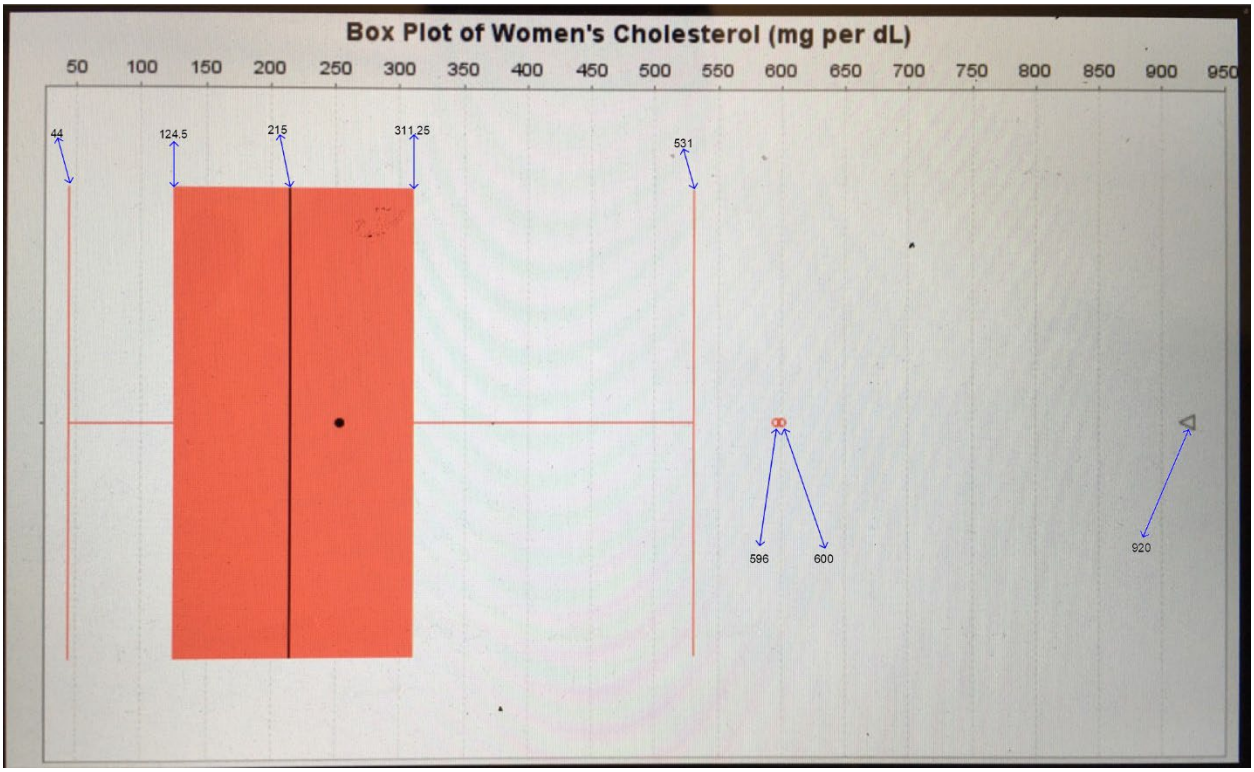
5. Interquartile Range (IQR) is an important measure of spread or variability in statistics. Give the basic definition of IQR.

6. How can we tell if we should use the median and IQR as our center and spread?

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

Look at the following Histogram, Box Plot and summary statistics of the women's cholesterol data and answer the following questions.





	Q1	Median	Q3	IQR	Min	Max	N total	
Women Cholesterol (mg per dL)	124.5	215.0	311.25	186.75	44.0	920.0	38	

7. What is this data measuring? _____

8. What are the units for the data set? _____

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

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

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

11. Are the median and IQR accurate for this data? (Yes or No) _____

12. What is the average cholesterol for these women? (Give a number) (No calculation needed)

Average Cholesterol = _____

13. How far are typical values in the data set from each other? (Give a number) (No calculation needed)

Average distance typical values are from each other = _____

14. Find two numbers that typical values fall in between and put your answer below. (No calculation needed)

_____ ≤ typical cholesterol for these women ≤ _____

15. Are there any unusually low values (low outliers) in the data set (yes or no)? _____

16. Are there any unusually high values (high outliers) in the data set (yes or no)? _____

17. List all unusual values (outliers) in the data set.
Give the actual numbers, not a cutoff point. _____

(For #18-22, refer to the boxplot.)

18. What percent of these women had a cholesterol below 311.25? _____

19. What percent of these women had a cholesterol below 124.5? _____

20. What percent of these women had a cholesterol higher than 215? _____

21. What was the largest value in the data set that was not an

outlier (not unusual)? _____

22. True or False? There were more numbers in the data set greater than 311.25 than there were numbers in the data set less than 124.5.

Chapter 5 Project
Skewed Data Analysis

Online Class Directions: *This will be an individual project. Each student will analyze one quantitative data set from the math 075-survey data fall 2015, create a poster summarizing their findings, and present the poster to other students in the class.*

Each student will pick one of the following data sets from the math 075 survey data fall 2015 to analyze: Hours work per week, Hours sleep per night, Hours of exercise per week, Number of Minutes to get to school, College GPA, Number of Units completed at COC, Average cell phone bill per month, Dollars spent on a meal when eat out, Number of times eat at restaurant or fast food per week, Number of U.S. states visited, Number of minutes spent on social media.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

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 histogram in StatKey, and pull the slider to "3 buckets". Your histogram should have 3 bars. Sketch the histogram onto your poster.**
- **Click on "Box Plot" in StatKey and sketch the box plot onto your poster.**
- **Write down the Median, 1st Quartile, 3rd Quartile, SMin, Max and Sample Size onto your poster. You will see them under the "Sample Statistics" menu in StatKey.**
- **Calculate the Interquartile Range IQR = Q3 – Q1**
- **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.**
- **Write a sentence to explain the median.**
- **What is the average? (Use the median if data is skewed.)**
- **What is your spread for the data? (Use the Interquartile Range if data is skewed.)**
- **Write a sentence to explain the interquartile range.**
- **Find two numbers that typical values fall in between (Q1 and Q3)**
- **Calculate Unusually high cutoff: Q3 + (1.5xIQR)**
- **List all unusually high values (high outliers) in the data set. Find these on the box plot. If there are none, say "No high outliers".**
- **Calculate Unusually low cutoff: Q1 − (1.5xIQR)**
- **List all unusually high values (high outliers) in the data set. Find these on the box plot. If there are none, say "No high outliers".**
- **Estimate the largest value in the data set that is not an outlier. Look at the right whisker on the box plot. Does not have to be exact.**
- **Estimate the smallest value in the data set that is not an outlier. Look at the left whisker on the box plot. Does not have to be exact.**
- **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 5 Project Discussion". You will be discussing your findings with other students in the class.

Face to face Class Directions: *The class will be separated into groups. Each group is required to pick a "team name" for their group and analyze one skewed quantitative data set from the math 075-survey data fall 2015, 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 survey data fall 2015 to present it to their classmates: Hours work per week, Hours sleep per night, Hours of exercise per week, Number of Minutes to get to school, College GPA, Number of Units completed at COC, Average cell phone bill per month, Dollars spent on a meal when eat out, Number of times eat at restaurant or fast food per week, Number of U.S. states visited, Number of minutes spent on social media.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021

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 histogram in StatKey, and pull the slider to "3 buckets". Your histogram should have 3 bars. Sketch the histogram onto your poster.**
- **Click on "Box Plot" in StatKey and sketch the box plot onto your poster.**
- **Write down the Median, 1st Quartile, 3rd Quartile, SMin, Max and Sample Size onto your poster. You will see them under the "Sample Statistics" menu in StatKey.**
- **Calculate the Interquartile Range IQR = Q3 – Q1**
- **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.**
- **Write a sentence to explain the median.**
- **What is the average? (Use the median if data is skewed.)**
- **What is your spread for the data? (Use the Interquartile Range if data is skewed.)**
- **Write a sentence to explain the interquartile range.**
- **Find two numbers that typical values fall in between (Q1 and Q3)**
- **Calculate Unusually high cutoff: Q3 + (1.5xIQR)**
- **List all unusually high values (high outliers) in the data set. Find these on the box plot. If there are none, say "No high outliers".**
- **Calculate Unusually low cutoff: Q1 − (1.5xIQR)**
- **List all unusually high values (high outliers) in the data set. Find these on the box plot. If there are none, say "No high outliers".**
- **Estimate the largest value in the data set that is not an outlier. Look at the right whisker on the box plot. Does not have to be exact.**
- **Estimate the smallest value in the data set that is not an outlier. Look at the left whisker on the box plot. Does not have to be exact.**
- **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.

This material is from *Introduction to Data Analysis*, 2nd edition by Matt Teachout, College of the Canyons, Santa Clarita, CA, USA, and is licensed under a "CC-BY" Creative Commons Attribution 4.0 International License – 3/17/2021