Week One:
Graphs and Numerical Summaries

Week One Goals

- Learn vocabulary to describe measures of center and spread (Part I).
- Use statistical software to produce numerical summaries and graphical displays. Video demonstrations provided. Look for the yellow shaded text.
- Learn vocabulary to describe the shape of a distribution (Part II).
- Describe shape, center and spread using graphs and numerical summaries.
- Compare data sets.

Vocabulary You Should Know Part I

Measures of Central Tendency: Mean, Median (and Mode)

There are 3 measures of central tendency or measures of 'center': The mean, the median and the mode.

Mean

When we study a sample of numerical data, it is important to be able to report a value that is considered 'typical' for the sample. One measure of 'typical' is the measure of center called the 'mean'. The sample mean can be calculated for you using a graphing calculator, Minitab (Express) or any other statistical software. If you want to calculate the mean by hand, just add up all the data points and divide by the total sample size, n.

The mean is a measure of center that is affected by unusually small or large data points/outliers; the mean is influenced and changes when outliers are in a data set; we say that the mean is not resistant to outliers. If it is determined that there are outliers are in a data set, the mean is no longer a good measure of center and should not be used (we would use the median instead).

If you are calculating a population mean, you will run the same calculation and represent the population mean with the symbol μ (pronounced mu). The sample mean is expressed with the symbol $\bar{x}$ (pronounced x-bar).

Sample mean: $\bar{x}$

Population mean: μ

Interpret Mean

If you are asked to interpret the sample mean, you could use a template sentence such as:

"A typical <state the variable> for the <describe the sample> in this sample is <state mean>."

Median

When you line up your data from smallest value to largest value, you can find the value that sits in the middle. This is considered the 50th percentile and is known as the median or the second quartile (Q2). The median is a good measure of 'center' or ‘typical’ to report. The median is resistant to outliers. Resistant means that if your data set has outliers---a really small data point or a really large data point; a data that is unusual---the median will not be affected; it is still the value that sits in the middle. Contrast that with the mean; the mean is a statistic that is calculated using each and every data point and an unusual data point/outlier will change the mean value. Therefore, if your data set has unusual data points/outliers, the median is a preferred measure of center.

Mode

The mode is the data value that occurs most often in a data set. The highest peak of a histogram or dot plot represents the location of the mode of a data set.

In summary…

Mean, median and mode are all measures of describing the center of a data set. Median is a good measure of center when a data set has unusually small or large data points (or outliers). Mean and median are both good measures of center, and are likely close in value, when a data set has no unusual data points. Mode is an easy measure of center to report when you are viewing a histogram or dot plot.

Outlier or Unusual Data Point

An outlier is a data point that differs significantly from other observations. The terms outlier and unusual can be used interchangeably and may be used a bit casually. If, however, you are asked to calculate if a data point is an outlier, then a more formal calculation will be made. See box plot outlier fences below.

Measures of Spread: Standard Deviation, Interquartile Range and Range

There are 3 measures of spread: standard deviation, Interquartile Range (IQR) and range. When we study a set of numerical data, we will also want to report a measure of variability; in other words, a measure of how spread out the data is.

Standard Deviation

One measure of spread is the standard deviation. The sample standard deviation can be calculated for you using a graphing calculator, Minitab (Express) or any other statistical software. I found an online calculator that does a nice job of showing you the steps by hand.

Sample standard deviation: s

Population standard deviation: σ

Sample Standard Deviation

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

The sample standard deviation is expressed by a lower-case 's' and represents a typical distance that a data point is from the sample mean, for that sample. In the calculation of standard deviation by hand, you would:

1. calculate the mean.
2. subtract each data point from the sample mean.
3. square each of these differences.
4. add up all the squared differences.
5. divide by n-1 (degrees of freedom). At this point in the calculation, you have what is called **variance (s^2)** and you are in squared units.
6. To get back into your original units, take the square root of the variance. Now you have the **standard deviation**.

Interpret Standard Deviation

If you are asked to interpret the sample standard deviation, you could use a template sentence such as:

"Within the sample, the typical distance of individual <state the variable>from the sample mean is <sample standard deviation>."

Population Standard Deviation

Although we won't calculate the population standard deviation in this class (because we always work with a random sample of data), I want you to know that the population standard deviation is denoted by sigma, σ, and is calculated by dividing by n (instead of n - 1).

$$\sigma = \sqrt{\frac{\Sigma\,(x - \mu)^2}{\mathbf{N}}}$$

Video Explaining Measures of Center and Standard Deviation from Prof. Coffey

Here is a short video discussing the vocabulary on measures of center and standard deviation: https://youtu.be/Wh5FPejg7Sc

Here is the same video with no sound and INTERPRETING: https://youtu.be/MzI0Huo_hSQ

Quartiles

When you divide your data into 4 equal sections, you describe the data set in quartiles:
The first quartile is the 25th percentile. We will refer to it as Q1. Q1 is the value such that 25% of values from the sample are less than it and 75% are greater than it. The second quartile is the 50th percentile. We will refer to it as the median. The third quartile is the 75th percentile. We will refer to it as Q3.
Q3 is the value such that 75% of values from the sample are less than it and 25% are greater than it.

Interquartile Range

The distance between the third and first quartiles is known as the Interquartile Range (IQR) and represents the middle 50% of the data.

$$IQR = Q3 - Q1$$

The interquartile range (IQR) is a measure of spread that is resistant to unusual data points/outliers. The IQR can be visualized on a box plot –it is the size of the rectangle -- and is easily compared to another IQR. Remember, outliers are really large or really small data values. Since the IQR represents the middle 50% of the data, the outliers do not affect the calculation of the IQR.

Range

The difference between the lowest value and the highest value in a data set. When comparing two histograms, the range is easy to identify and compare. The range is not resistant to outliers and is, therefore, not a reliable measure of spread in a situation where you have unusually small or large data points.

In summary…

Standard deviation, IQR and range are all measures of describing how spread out a data set is. IQR is a good measure of spread when a data set has outliers/unusual values. Standard deviation and IQR are both good measures of spread when the data set does not have outliers. Range is easy to identify when you are viewing a histogram or dot plot but is rarely used otherwise.

The Five-Number Summary

The five-number summary is made up of the following five values:

- Minimum value (min)
- First quartile (Q1; 25th percentile)
- Median (Q2, second quartile; 50th percentile)
- Third quartile (Q3; 75th percentile)
- Maximum value (max)

The Modified Box Plot

A box plot is a graphical display that shows the five-number summary. It is also known as a box and whisker plot. The modified box plot shows if any outliers are present by indicating an outlier with an asterisk *.

The Five Number Summary

List the following values, in this order:

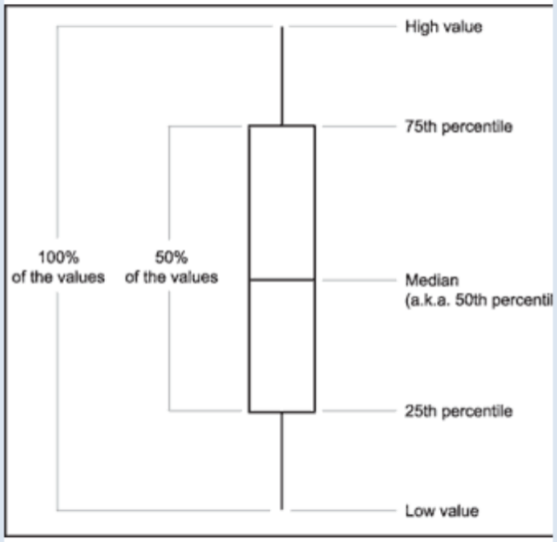
Min, Q1, Median, Q3, Max

The middle 50% of the values is Q3 – Q1 and called the Interquartile Range (IQR).

An outlier (an unusual observation) is a data value beyond the lower fence or upper fence.

Lower fence: Q1 – 1.5(IQR)

Upper fence: Q3 + 1.5(IQR)



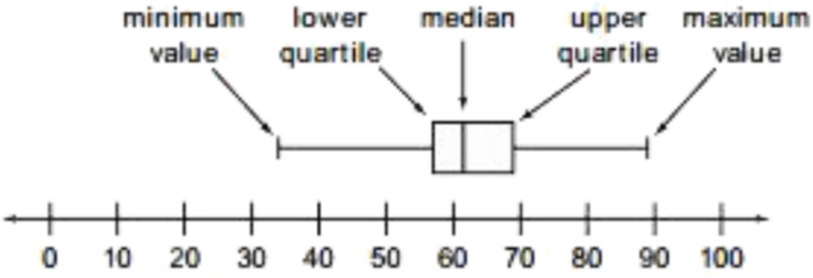
High value (max): The largest #.

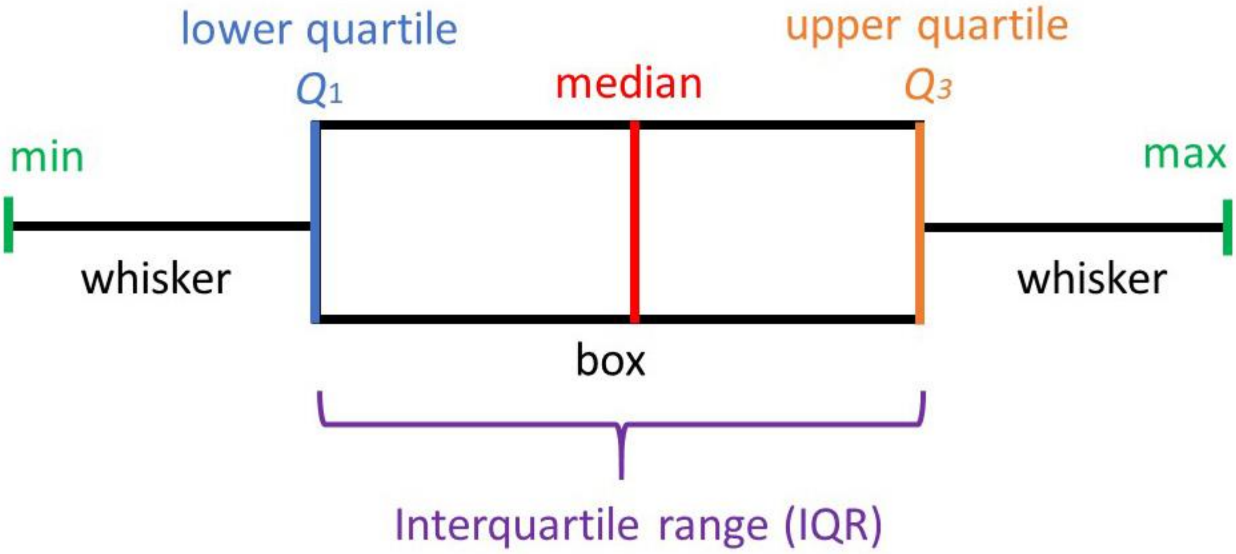
75th percentile (Q3): The data point such that 75% of data points are smaller than it and 25% are larger.

median (50th percentile, Q2): The data point such that 50% of data points are smaller than it and 50% are larger.

25th percentile (Q1): The data point such that 25% of data points are smaller than it and 75% are larger.

Low Value (Min): The smallest #.





Box plots can be drawn vertically or horizontally. Technology will assist us in building box plots; specifically, a modified box plot.

To construct a modified box plot by hand, a lower fence and upper fence are first calculated. These outlier fences represent imaginary lines such that any data point that falls beyond either of the fences would be considered an outlier.

Outlier Fences

The outlier fences can be calculated by following these steps:
Generate the 5-number summary for your sample.
1. Calculate the interquartile range (IQR) = Q3 - Q1.
2. Multiply the IQR by 1.5.
3. Lower fence = Q1 - 1.5 (IQR) That is, subtract 1.5(IQR) from Quartile 1. The result is the lower fence.
4. Upper fence = Q3 + 1.5 (IQR) That is, add 1.5(IQR) to Quartile 3. The result is the upper fence.

Data points smaller than the lower fence or larger than the upper fence would be considered outliers.

Note: The lower and upper fences are calculated for you in StatHelper descriptive statistics. When technology is used to build a box plot with outliers, the technology is using these fence calculations to determine if a data point should be represented with an asterisk; an asterisk indicates an outlier.

Video Explaining Box Plots from Prof. Coffey

Here is a short video discussing box plots: https://youtu.be/KqiFQKYAH_k

Here is the same video with no sound and INTERPRETING: https://youtu.be/wCbdQtaT_OY

Statistical Software and Descriptive Statistics

In this course, we will use technology to generate the graphical displays and numerical summaries (mean, median, standard deviation, IQR, etc.) If you have used statistical technology in the past, you should feel welcome to use whatever method you choose (Excel, TI graphing calculator, etc.)

NOTE: RIT has a site license for Minitab, Minitab Express and JMP Pro and you will use such software in STAT 146 Intro to Statistics II, assuming your program requires STAT 146. These programs work with Windows and Macintosh computers. Read the course syllabus for more information.

If you are interested in working with web-based technology, you should use StatHelper. I have secured free logins for everyone in class. E-mail to get more information. This is a website where you upload the data and ask for the graphs, summaries that you want. There is no software to download.

I will provide video demonstrations of StatHelper and Minitab 19. If you need assistance with a different technology, please reach out to me. When a video is embedded in the notes, you will find it quickly by looking for the shaded area with the links (see below).

Your Username and Password were e-mailed to you in a class email at the start of Week One

StatHelper by Responsum Analytics is being provided to our class at no charge. As long as the data is stored in an Excel spreadsheet, the descriptive statistics and explanations are provided using this web-based software.

StatHelper Video from Prof. Coffey

Here is a short video demonstrating how to generate descriptive statistics and graphical displays using StatHelper on the **Salaries** data: https://youtu.be/0Eh2eUpwwsE

Minitab and Minitab Express are available as a download from the RIT ITS website.

Minitab Video from Prof. Coffey

Here is a short video demonstrating how to generate descriptive statistics and graphical displays using Minitab on the **Salaries** data: https://youtu.be/sgZL7rw7DQU

Minitab Express Video from Prof. Coffey

Here is a short video demonstrating how to generate descriptive statistics and graphical displays using Minitab Express on the **Salaries** data: https://youtu.be/R_gCb7gHjls

Examples

All of the week one data can be found in Excel spread sheet: 1_Week 1 STAT 145 Data.xlsx.
The data with this example is under the sheet labeled **Salaries**.

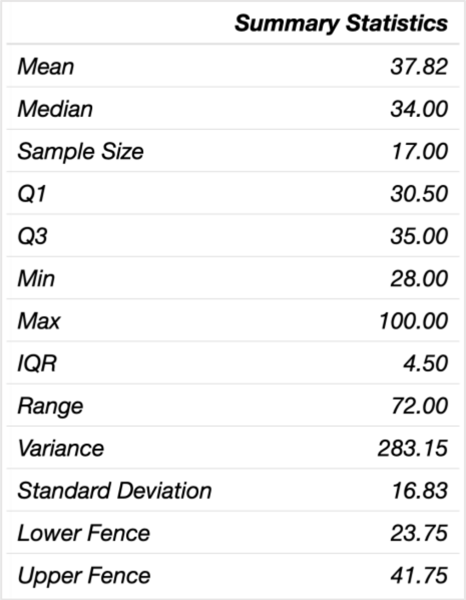
Salaries Example

A company is made up of 15 manufacturing employees, 1 employee in accounting/sales and the owner. Here is a list of company salaries (in thousands of dollars). Are there any salaries that are outliers? What is a typical salary for this company? What value is a good measure of spread?

28 28 28 29 32 33 34 34 34 34 35 35 35 35 39 50 100

StatHelper

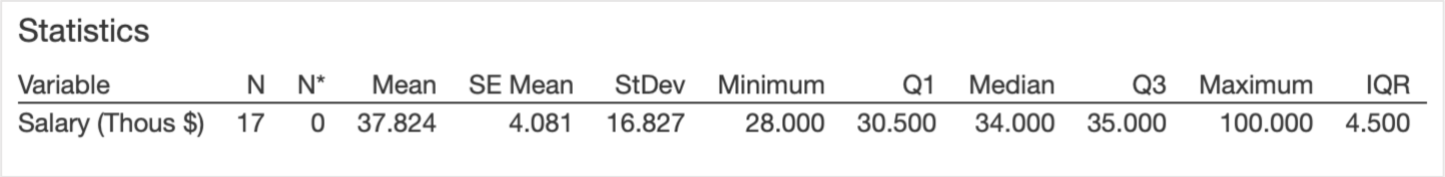
I used StatHelper to calculate the descriptive statistics on the salary data. The results from the Summary Statistics Tab are seen below:



	Summary Statistics
Mean	*37.82*
Median	*34.00*
Sample Size	*17.00*
Q1	*30.50*
Q3	*35.00*
Min	*28.00*
Max	*100.00*
IQR	*4.50*
Range	*72.00*
Variance	*283.15*
Standard Deviation	*16.83*
Lower Fence	*23.75*
Upper Fence	*41.75*

Minitab Express

I used Minitab Express to calculate the important descriptive statistics on the salary data. The results are below.



Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Salary (Thous $)	17	0	37.824	4.081	16.827	28.000	30.500	34.000	35.000	100.000	4.500

Outliers

The owner's salary is $100,000. This is considered an outlier because it is larger than the upper fence. The accounting/sales employee salary is also considered an outlier for the same reason.

Typical Salary

According to the output, the mean salary for this company is $37, 824. Does this seem to be a good representation of a typical salary for this company? No, it does not. Look again at the salaries lined up; you will notice that the median salary is 34 thousand.

28 28 28 29 32 33 34 34 **34** **34** 35 35 35 35 37.6 39 50 100

It is our job to recognize that we have two outliers in the data set and, therefore, to decide that the mean is not the best measure of what is typical for this data set. In this case, the best measure of center is the median. The median, $34,000 is a typical salary for this company.

Spread

Similarly, since the standard deviation is affected by outliers, we will choose not to report the standard deviation as a measure of spread. The Interquartile range (IQR) of $4.5 thousand is a reasonable measure of spread to study for these data. That is, the middle 50% of the data has a range of 4.5 thousand dollars.

All of the week one data can be found in Excel spread sheet: 1_Week 1 STAT 145 Data.xlsx.
The data with this example is under the sheet labeled **Expenditures**.

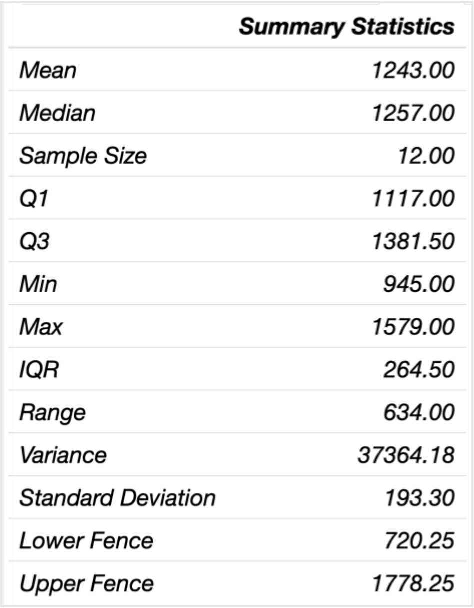
Expenditures Example

The National Retail Federation reported that first-year college students in the US spend more on back-to-school items than any other college group (USA Today 8/4/2006). Back-to-school expenditures ($) for a sample of twelve first-year students are provided below. Answer the following questions about this sample.

1479	983	1133	1286	1409	1259	1579	1113	945	1354	1255	1121

StatHelper

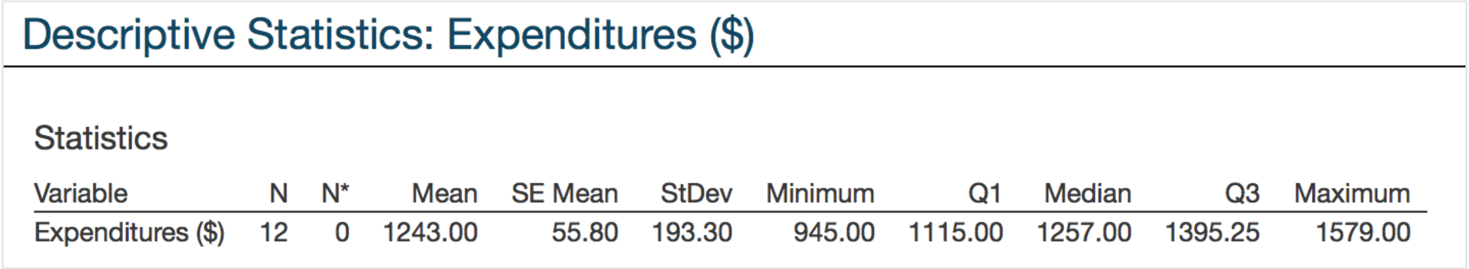
I used StatHelper to calculate the descriptive statistics on the salary data. The results from the Summary Statistics Tab are seen below:



	Summary Statistics
Mean	*1243.00*
Median	*1257.00*
Sample Size	*12.00*
Q1	*1117.00*
Q3	*1381.50*
Min	*945.00*
Max	*1579.00*
IQR	*264.50*
Range	*634.00*
Variance	*37364.18*
Standard Deviation	*193.30*
Lower Fence	*720.25*
Upper Fence	*1778.25*

Minitab

I used Minitab to calculate the important descriptive statistics on the salary data. The results are below.



Descriptive Statistics: Expenditures ($)

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Expenditures ($)	12	0	1243.00	55.80	193.30	945.00	1115.00	1257.00	1395.25	1579.00

1. Describe the objects in the population.

Answer: All first-year college students in the US.

2. Identify the variable being studied and its units.

Answer: The variable being studied is: Back-to-school costs (expenditures) and the units are in dollars ($).

3. State the sample size.

Answer: n = 12

4. Calculate the mean by using technology.

Answer: $\bar{x} = \$1243.00$

5. Interpret the mean of the sample by using the template sentence:

Answer: A typical back-to-school cost for the first-year students in this sample is $1,243.00.

6. Calculate the sample standard deviation by using technology.

Answer: s = $193.30

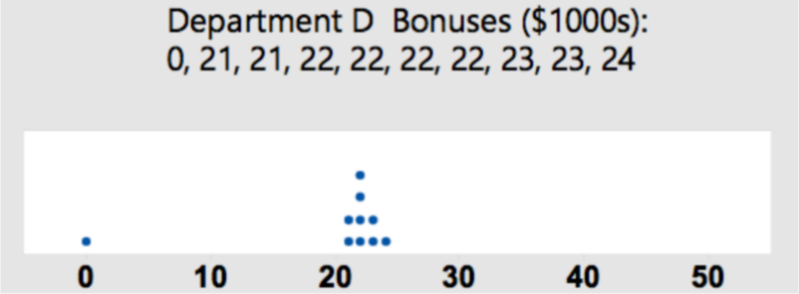
7. Interpret the sample standard deviation by using the template sentence:

Answer: The typical distance of individual back-to-school costs within the sample from the sample mean is $193.30.

All of the week one data can be found in Excel spread sheet: 1_Week 1 STAT 145 Data.xlsx.
The data with this example is under the sheet labeled **Bonus $**.

Bonus Example

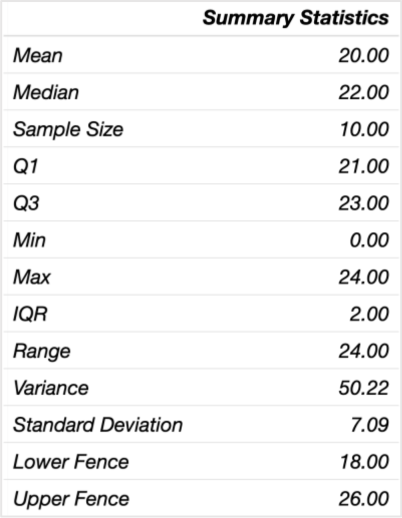
Sunshine Financial Corporation is handing out bonuses to its employees. The manager of each department is responsible for splitting up **$200,000** in bonus money among ten employees. The following dot plot shows the distribution of money for department D. The bonuses are listed in thousands of dollars at the top of the image.



Department D Bonuses ($1000s):
0, 21, 21, 22, 22, 22, 22, 23, 23, 24

StatHelper

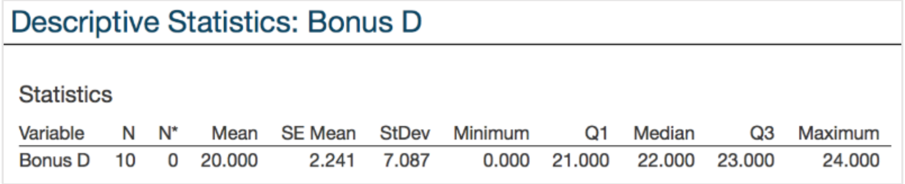
I used StatHelper to calculate the descriptive statistics on the salary data. The results from the Summary Statistics Tab are seen below:



	Summary Statistics
Mean	*20.00*
Median	*22.00*
Sample Size	*10.00*
Q1	*21.00*
Q3	*23.00*
Min	*0.00*
Max	*24.00*
IQR	*2.00*
Range	*24.00*
Variance	*50.22*
Standard Deviation	*7.09*
Lower Fence	*18.00*
Upper Fence	*26.00*

Minitab

I used Minitab to calculate the important descriptive statistics on the salary data. The results are below.



Descriptive Statistics: Bonus D

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Bonus D	10	0	20.000	2.241	7.087	0.000	21.000	22.000	23.000	24.000

1. Using technology, calculate the mean and median bonus amounts for Department D. Which is the preferred measure of CENTER for this sample? Explain.

ANSWER: The mean bonus amount is 20 (or $20,000). The median bonus amount is 22 (or $22,000). Since the data appears to have an unusually low bonus amount, the *median is the better measure of center*. Even if I did not see the dot plot, since the mean is 2 thousand dollars less than the median bonus amount, I would consider the distribution to be skewed left. Since it is skewed, the median is the better measure of center. It is resistant to outliers/unusually large or small values.

2. Calculate the standard deviation and interquartile range. Which is the preferred measure of SPREAD for this sample? Explain.

ANSWER: The standard deviation is 7.087. This means that a typical distance that an individual bonus amount is from the sample mean, for this sample is $7,087. The interquartile range (IQR) is 23 - 21 = 2 (or $2000). Since the data appears to have an unusually low bonus amount, the *IQR is the better measure of spread*. As stated earlier, the distribution is skewed left. Since it is skewed, the IQR is the better measure of spread. It is resistant to outliers/unusually large or small values.

All of the week one data can be found in Excel spread sheet: 1_Week 1 STAT 145 Data.xlsx.
The data with this example is under the sheet labeled **ACT Scores**.

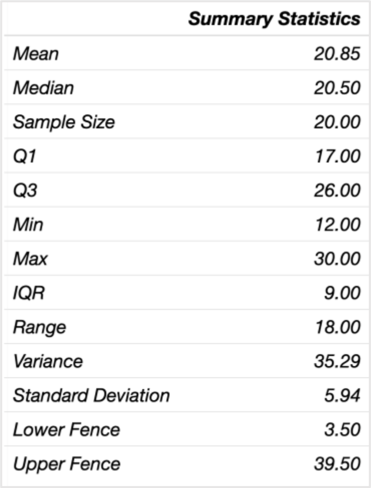
ACT Scores Example

A national random sample of 20 ACT scores from 2010 is listed here: 29, 26, 13, 23, 23, 25, 17, 22, 17, 19, 12, 26, 30, 30, 18, 14, 12, 26, 17, 18

Using technology, calculate the sample mean, sample standard deviation, median and IQR. Calculate the lower and upper fences and interpret these values.

StatHelper

I used StatHelper to calculate the descriptive statistics on the salary data. The results from the Summary Statistics Tab are seen below:



	Summary Statistics
Mean	*20.85*
Median	*20.50*
Sample Size	*20.00*
Q1	*17.00*
Q3	*26.00*
Min	*12.00*
Max	*30.00*
IQR	*9.00*
Range	*18.00*
Variance	*35.29*
Standard Deviation	*5.94*
Lower Fence	*3.50*
Upper Fence	*39.50*

ANSWER: The sample mean, $\bar{x} = 20.85$. The sample standard deviation, s = 5.94. The median = 20.5. The Interquartile Range (IQR) = 9. Values less than **3.5** and greater than **39.5** are considered outliers.

All of the week one data can be found in Excel spread sheet: 1_Week 1 STAT 145 Data.xlsx.
The data with this example is under the sheet labeled **Male Ages**.

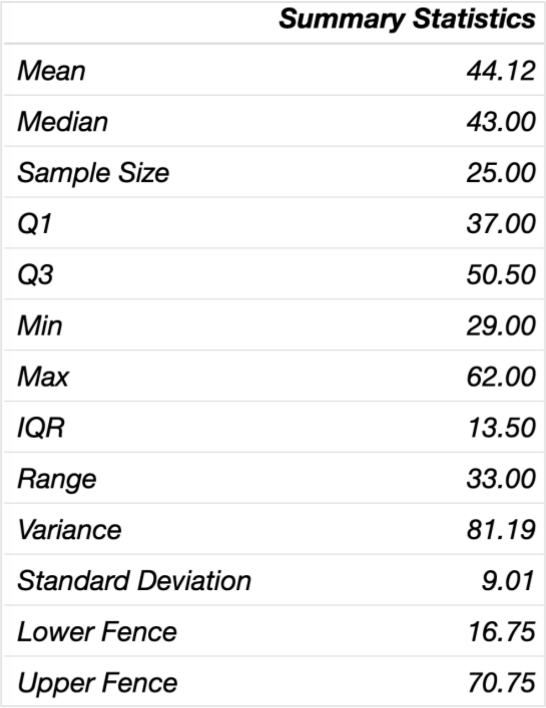
Male Ages Example

Jon is interested in knowing about the ages at which recent Oscar-winning male **actors** won their awards. He is unable to collect all the data he needs but gathers a random sample of 25 Oscar-winning male **actors** and their ages are given below. Using technology, find the five-number summary. Round to the nearest tenth.

60	48	50	45	38	37	43	29	47	36	40	46	60
45	32	38	37	52	54	42	32	51	43	62	36	

StatHelper

I used StatHelper to calculate the descriptive statistics on the salary data. The results from the Summary Statistics Tab are seen below:



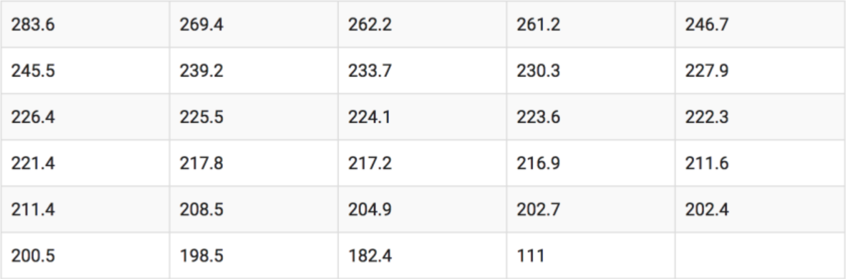
	Summary Statistics
Mean	*44.12*
Median	*43.00*
Sample Size	*25.00*
Q1	*37.00*
Q3	*50.50*
Min	*29.00*
Max	*62.00*
IQR	*13.50*
Range	*33.00*
Variance	*81.19*
Standard Deviation	*9.01*
Lower Fence	*16.75*
Upper Fence	*70.75*

ANSWER:
Minimum Value: 29.0
The First Quartile, Q1 (to the nearest tenth): 37.0
Median: 43.0
The Third Quartile, Q3 (to the nearest tenth): 50.5
Maximum Value: 62.0

All of the week one data can be found in Excel spread sheet: 1_Week 1 STAT 145 Data.xlsx.
The data with this example is under the sheet labeled **Ski Jump**.

Ski Jump Example

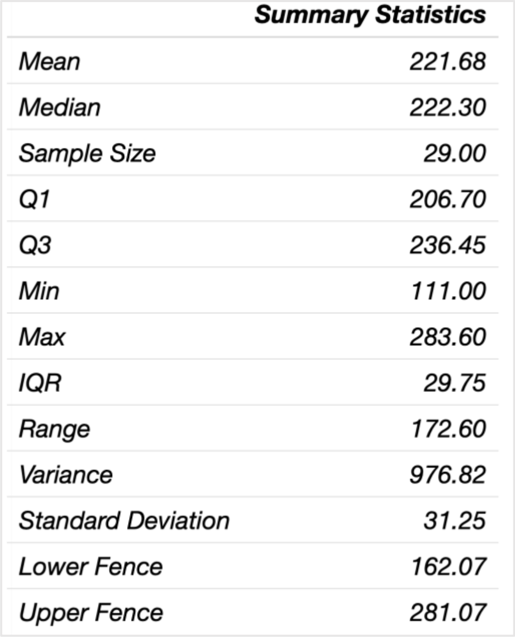
Below are the individual final results for the men's large hill ski jumping event at the Winter Olympics, measured in meters. Are any of the data points considered outliers (use the fences and a modified box plot to confirm)?



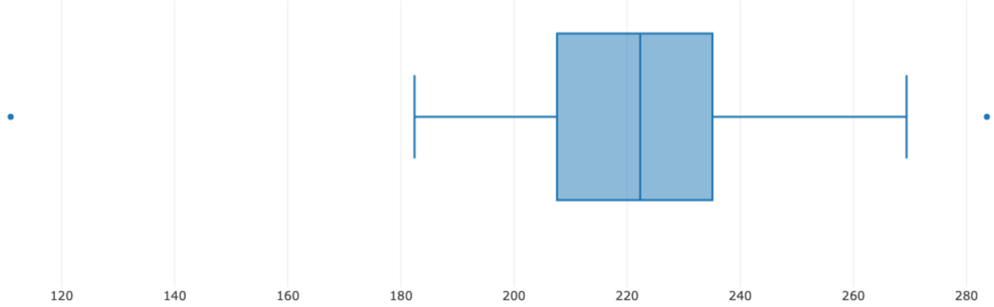
283.6	269.4	262.2	261.2	246.7
245.5	239.2	233.7	230.3	227.9
226.4	225.5	224.1	223.6	222.3
221.4	217.8	217.2	216.9	211.6
211.4	208.5	204.9	202.7	202.4
200.5	198.5	182.4	111	

StatHelper

I used StatHelper to calculate the descriptive statistics and box plot on the salary data. The results from the Summary Statistics Tab are seen below:



	Summary Statistics
Mean	*221.68*
Median	*222.30*
Sample Size	*29.00*
Q1	*206.70*
Q3	*236.45*
Min	*111.00*
Max	*283.60*
IQR	*29.75*
Range	*172.60*
Variance	*976.82*
Standard Deviation	*31.25*
Lower Fence	*162.07*
Upper Fence	*281.07*



ANSWER: The box plot shows outliers. The minimum value 111.0 is smaller than the lower fence and is an outlier. The maximum value 283.60 is larger than the upper fence and is an outlier.

Vocabulary You Should Know Part II

Describe the Shape of a Distribution

Dot plots, histograms and box plots are graphical displays that help us understand if the distribution of the data can be considered symmetric or skewed.

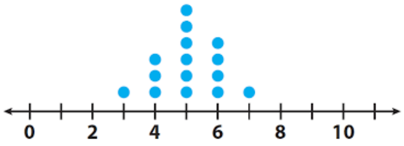
Dot Plot

Discrete, numerical data or categorical data can be visualized nicely in a dot plot. Look for the highest frequency of dots and you can find what might be considered 'typical' for that sample of data.

Describe the shape of the distribution of data in a dot plot.

Symmetric

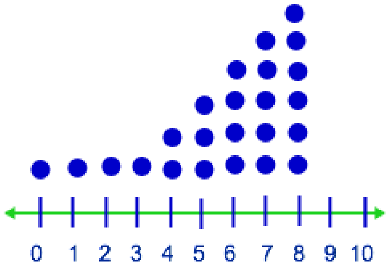
We say that a distribution is 'Somewhat Symmetric' if the mode (the value that occurs most often) is fairly central and the left side of the graph looks like a mirror image of the right side of the graph.



Somewhat Symmetric Distribution

Skewed Left or Skewed Negatively

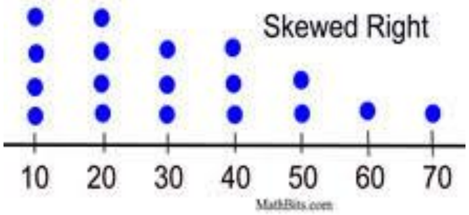
We say that a distribution is 'Skewed Left' if the mode (the value that occurs most often) is on the right side of the distribution and there are fewer responses for smaller values (i.e. the values on the left have fewer responses). We say 'Skewed left' when the graph has a long tail on the left.



Skewed Left Distribution

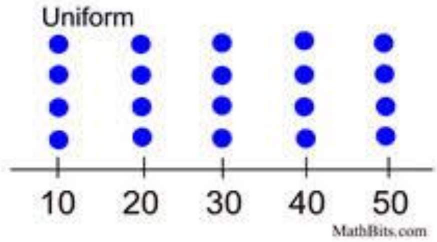
Skewed Right or Skewed Positively

We say that a distribution is 'Skewed Right' if the mode (the value that occurs most often) is on the left side of the distribution and there are fewer responses for larger values (i.e. the values on the right have fewer responses). We say 'Skewed right' when the graph has a long tail on the right.



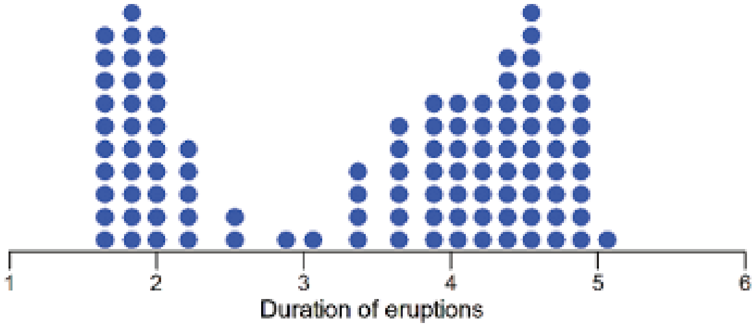
Uniform

A distribution would be described as 'uniform' if the frequency of the responses is somewhat the same across the graph



Bimodal

A distribution is said to have multiple modes when you see multiple peaks in the graph. In the case of having two distinct peaks, we would describe the shape of the graph as 'bimodal'.



Histogram

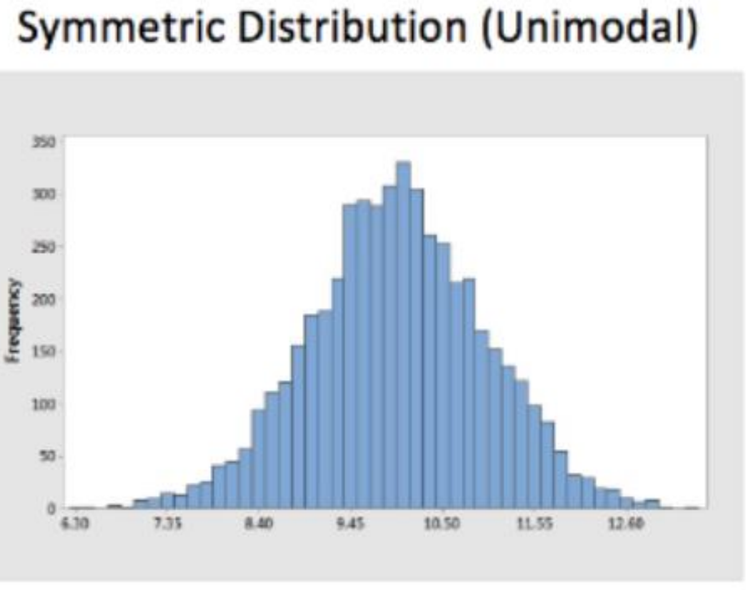
A histogram is a display of statistical information that uses rectangles to show the frequency of data items in successive numerical intervals of equal size. In the most common form of histogram, the numerical variable is organized into bins and plotted along the horizontal axis and the frequency is plotted along the vertical axis. The data appears as colored or shaded rectangles. Look for the highest frequency rectangle and you can find what might be considered 'typical' for that sample of data.

Symmetric Distribution

In a symmetric (or somewhat symmetric) distribution, we expect the mean, median and mode to be roughly the same value. We are not able to identify the mean on a histogram, but since the graph has its peak near the center and the left and right sides are close to being mirror images, we would be able to deduce that the mean and median would be near the mode (the mode is the highest frequency rectangle).

In a symmetric distribution, we expect that the $mean \approx median \approx mode.$

If asked to describe the center of the distribution, either the mean or median would be appropriate to report as typical. Both are good measures of center and we would expect them to be close in value.



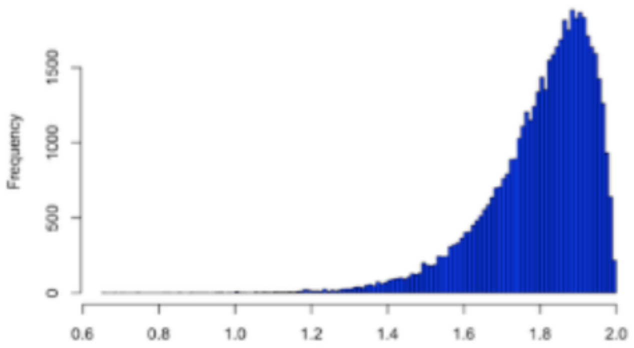
Skewed Left or Skewed Negatively

In a skewed left distribution (long tail on the left), we expect that the median is near the mode (the highest peak) and this value would be the best value to report to describe the center of this distribution. The graph indicates that there are some data points at the lower values (as seen by the tail on the left) and these values will pull the mean toward the smaller values. Although there are not many of these data points, they are considered unusual and are possibly outliers.

In a skewed left distribution, we expect that the *mean is less than the median*.

If asked to describe the center of the distribution, the median would be appropriate to report as typical since it is not affected by unusual values/outliers.

Skewed Left Distribution (unimodal)



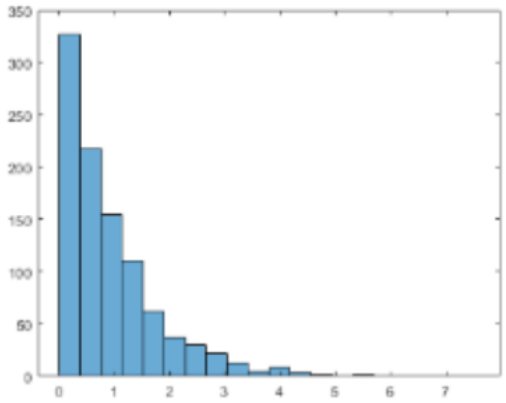
Skewed Right or Skewed Positively

In a skewed right distribution (long tail on the right), we expect that the median is near the mode (the highest peak) and this value would be the best vale to report to describe the center of the distribution. The graph indicates that there are some data points at the higher values (as seen by the tail on the right) and these values will pull the mean toward the higher values. Although there are not many of these data points, they are considered unusual and are possibly outliers.

In a skewed right distribution, we expect that the *mean is greater than the median*.

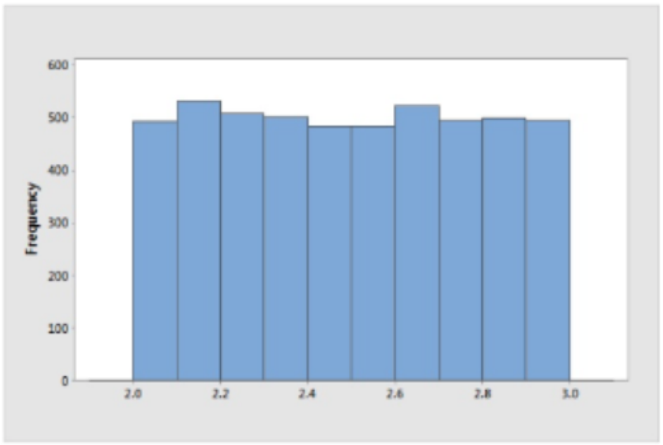
If asked to describe the center of the distribution, the median would be appropriate to report as typical since it is not affected by unusual values/outliers.

Skewed Right Distribution (Unimodal)

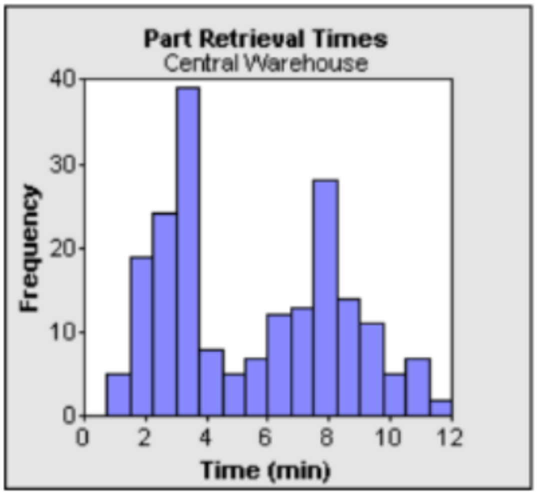


If we determine that a distribution is uniform or bimodal, then that is all that we have to say. We would rely on numerical summaries (mean, median, standard deviation, IQR) to report center and spread.

Uniform Distribution



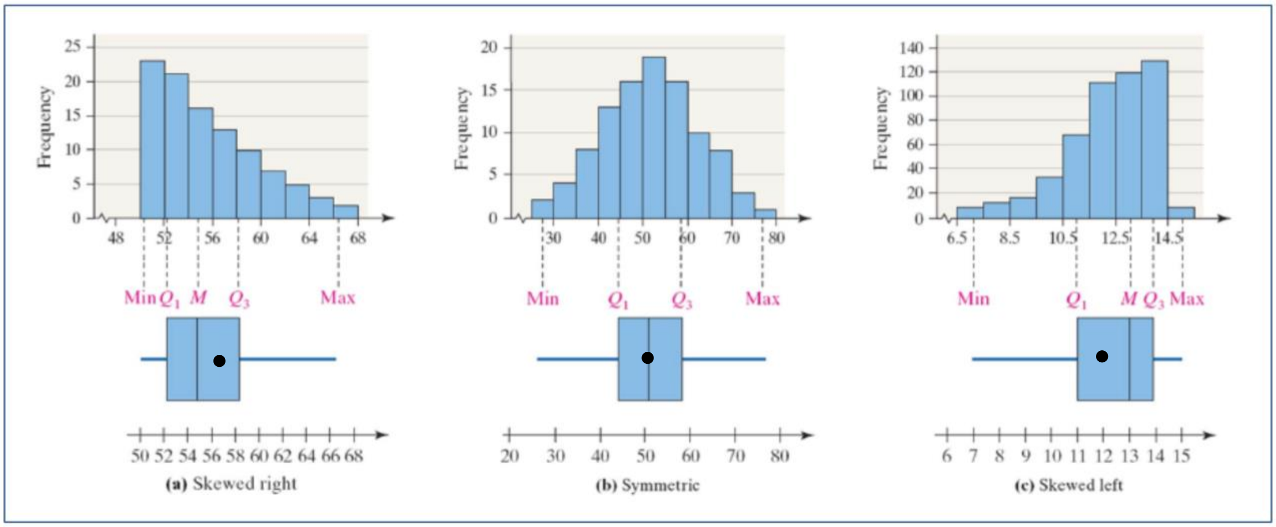
Bimodal Distribution



Box Plot

A box plot displays the Five-number summary. I would like to suggest that knowing the **mean** can be very helpful when asked to describe the shape of a box plot. In Minitab 19, the mean can be added to the box plot. Even if you cannot add a mean symbol to the graph, knowing the mean value and comparing it to the median value can be helpful:

- If the mean and median are very close in value, then we can assume the data distribution is somewhat symmetric.
- If the mean is less than the median, then we can assume the data distribution is skewed **LEFT**.
- If the mean is greater than the median, then we can assume the data distribution is skewed **RIGHT**.



(a) Skewed right

(b) Symmetric

(c) Skewed left

If the mean value is 57, then we would describe the shape of the distribution as Skewed Positively (Skewed Right) since the mean is greater than the median.

Mean > Median ⇒Skewed Positively

If the mean value is 50, then we would describe the shape of the distribution as Symmetric since the mean is approximately equal to the median.

Mean approx. = Median ⇒Symmetric

If the mean value is 12, then we would describe the distribution as Skewed Negatively (Skewed Left) since the mean is less than the median.

Mean < Median ⇒Skewed Negatively

Video Explaining How to Describe the Shape of a Distribution from Prof. Coffey

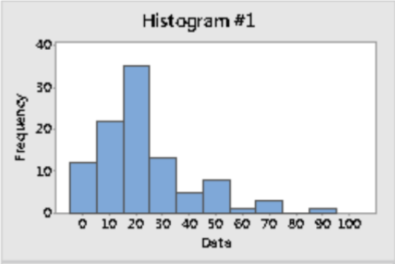
Here is a short video discussing the shape of a distribution: https://youtu.be/KqiFQKYAH_k

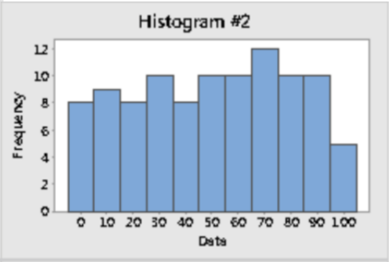
Matching Histograms to Box Plots

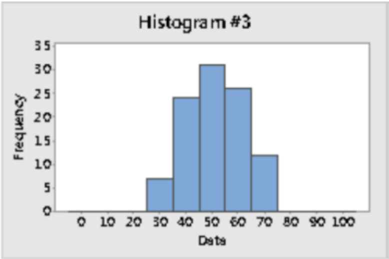
Given the following four histograms, match them to the corresponding box plot. The answers can be found below.

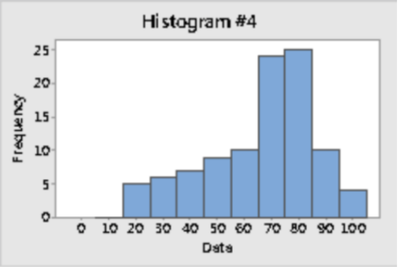
Video Explaining this Example

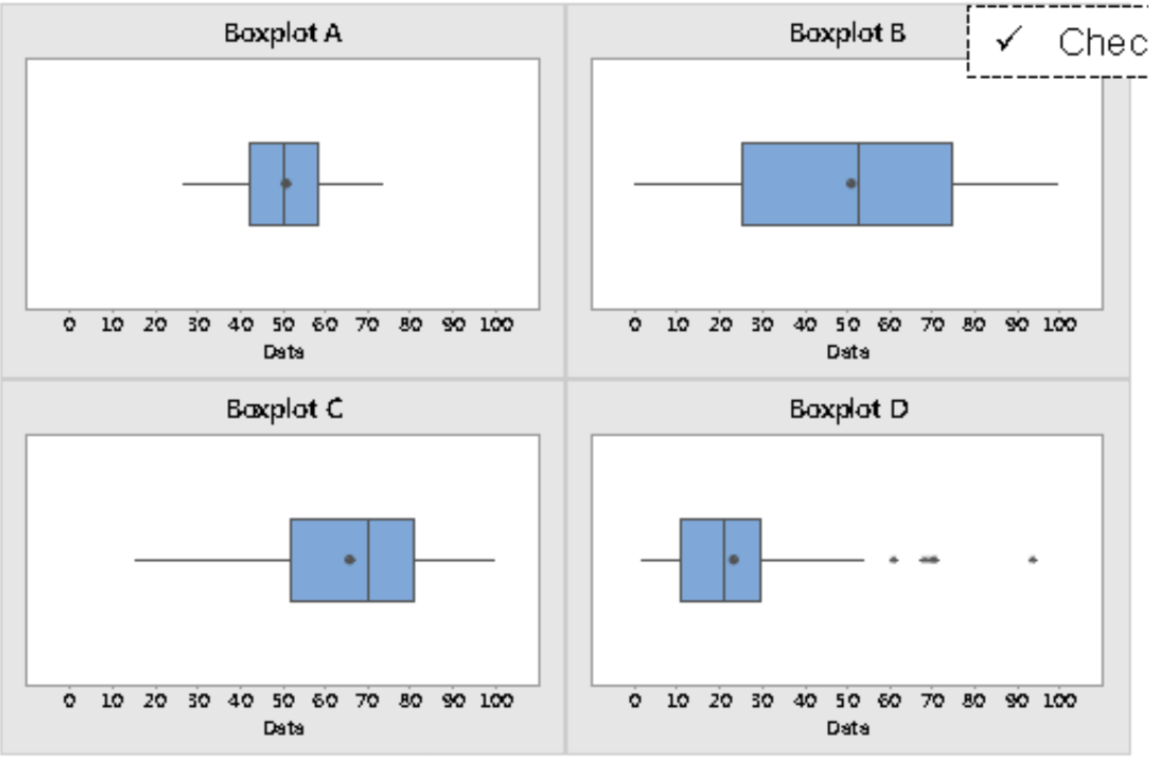
Here is a short video discussing this example—matching a histogram to the box plot: https://youtu.be/hPyBIQIYZ58











ANSWERS:

Histogram #1 matched with box plot D (histogram appears skewed right; matched it with box plot showing mean greater than median)

Histogram #2 matches with box plot B (histogram has a large spread; matched it with box plot showing IQR largest)

Histogram #3 matches with box plot A (histogram is symmetric; matched it with box plot where the mean is close in value to median)

Histogram #4 matches with box plot C (histogram appears skewed left, matched it with box plot showing mean less than median)

Get to Know the Data: Describe Shape, Center and Spread

When we are provided raw data or summarized data, we should understand what the data represents and take a moment to identify the population and sample, identify the variable being studied and know what type of variable it is.

When we are given raw data, we will use statistical software to generate numerical and graphical summaries so that we can get to know the data and describe it. When I say ‘describe’, I am suggesting that we describe the shape, center and spread for a set of data.

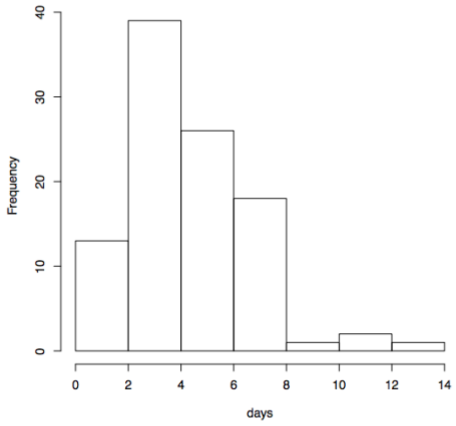
When we are given summarized data, then we will review the numerical summaries and any graphs provided to describe the shape, center and spread for the data.

In summary, when asked to know ‘Get to Know the Data’, be prepared to:

- Identify the objects in the population and in the sample.
- Identify the variable being studied (and indicate if it is numerical or categorical).
- If you have raw data, produce a graphical display and summary statistics that are appropriate for the variable being studied. If you have summarized data and graphs provided, move on to the next step.
- Describe the shape (if you can) and describe the center and spread of the distribution.

Histogram Example – Describe Shape and Center

The histogram below represents the lifespan (in days) of a random sample of a particular type of insect. Describe the shape and center of the distribution.



ANSWER: The lifespan histogram appears to be skewed right; A typical lifespan for this sample is approximately 2 – 4 days.

All of the week one data can be found in Excel spread sheet: 1_Week 1 STAT 145 Data.xlsx.
The data with this example is under the sheet labeled **Part Thickness;** use the variable **Day Shift**. Then use the variable **Night Shift**.

Part Thickness Example – Compare Shape, Center and Spread

For a summer job, you were working in the quality control department for a computer company that manufactures computer parts. The specific part that you are to evaluate the quality of is supposed to be 8 micrometers in thickness. You obtained samples of fifteen of these parts manufactured by the day shift and twelve parts manufactured by the night shift workers. Here are the findings:

Day Shift	Night Shift
7.9	2
8.0	4
8.2	12
8.3	14
7.8	6
8.0	10
7.7	5
8.1	4
8.1	11
8.2	13
7.9	9
8.2	8
8.3	
7.9	
8.0	

**NOTE: When asked to know 'Get to Know the Data', be prepared to:

- Identify the objects in the population and in the sample.
- Identify the variable being studied (and indicate if it is numerical or categorical).
- If you have raw data, produce a graphical display and summary statistics that are appropriate for the variable being studied. If you have summarized data and graphs provided, move on to the next step.
- Describe the shape (if you can) and describe the center and spread of the distribution.

A. Using technology, produce summary statistics to help **get to know the data****. Write up your findings.

B. Interpret the mean and standard deviation for the day shift sample.

C. Interpret the mean and standard deviation for the night shift sample.

D. Which shift, day or night, results in a mean thickness that was closer to the target thickness? Explain.

ANSWER:

A. Get to know the data:

- The population would be all the computer parts produced in the day shift and all the computer parts produced in the night shift.
- We are working with a sample of the data; the sample of the day shift computer parts consists of 15 parts; the sample of the night shift consists of 12 parts.
- We are studying the variable: Part Thickness and it is a numerical variable.
- **Shape**: Since the mean of the day shift is close to the median, I would consider the distribution for day shift part thickness to be somewhat symmetric. Since the mean is less than the median, I would consider the distribution for night shift part thickness to be skewed left.

  (SIDE NOTE: when comparing two data sets, do not report a value for center, compare the center values; do not report a value for spread, compare them and decide which has less spread)

  Center: The mean thickness values are approximately the same for both shifts.
  Spread: The part thickness values for the day shift have less variability than the night shift. The day shift produces parts that are more consistent.

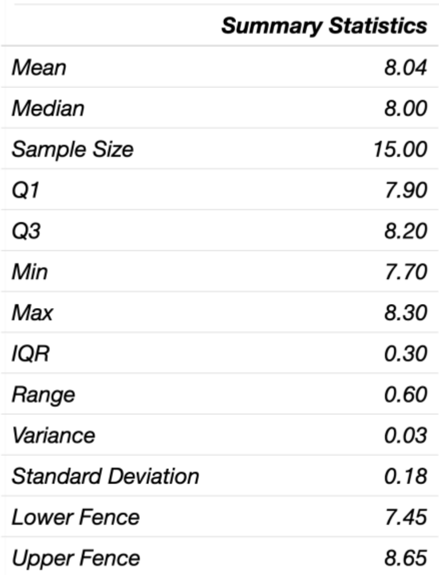
B. Interpret Mean: A typical thickness for the parts in the day shift sample is **8.04 micrometers**. Interpret Standard Deviation: The typical distance of the day shift individual part thickness within the sample from the sample mean is **0.18 micrometers**.

C. A typical thickness for the parts in the night shift sample is **8 micrometers**. Interpret Standard Deviation: The typical distance of the night shift individual part thickness within the sample from the sample mean is **3.77 micrometers**.

D. Which shift, day or night, results in a mean thickness that was closer to the target thickness? Explain.

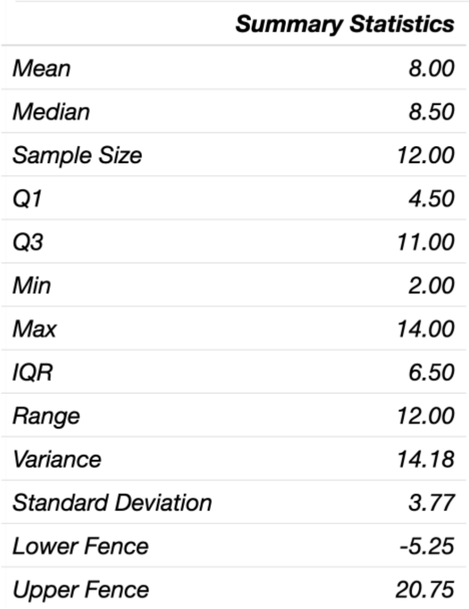
   The mean thickness for the day shift is closer to the target of 8 micrometers because the variability in the thicknesses is much smaller. Despite the mean thickness for the night shift being 8 micrometers, the standard deviation is larger. For this sample, the typical distance is 3.77 micrometers from the mean. The small standard deviation is an indication of more consistent results.

Day Shift



	Summary Statistics
Mean	*8.04*
Median	*8.00*
Sample Size	*15.00*
Q1	*7.90*
Q3	*8.20*
Min	*7.70*
Max	*8.30*
IQR	*0.30*
Range	*0.60*
Variance	*0.03*
Standard Deviation	*0.18*
Lower Fence	*7.45*
Upper Fence	*8.65*

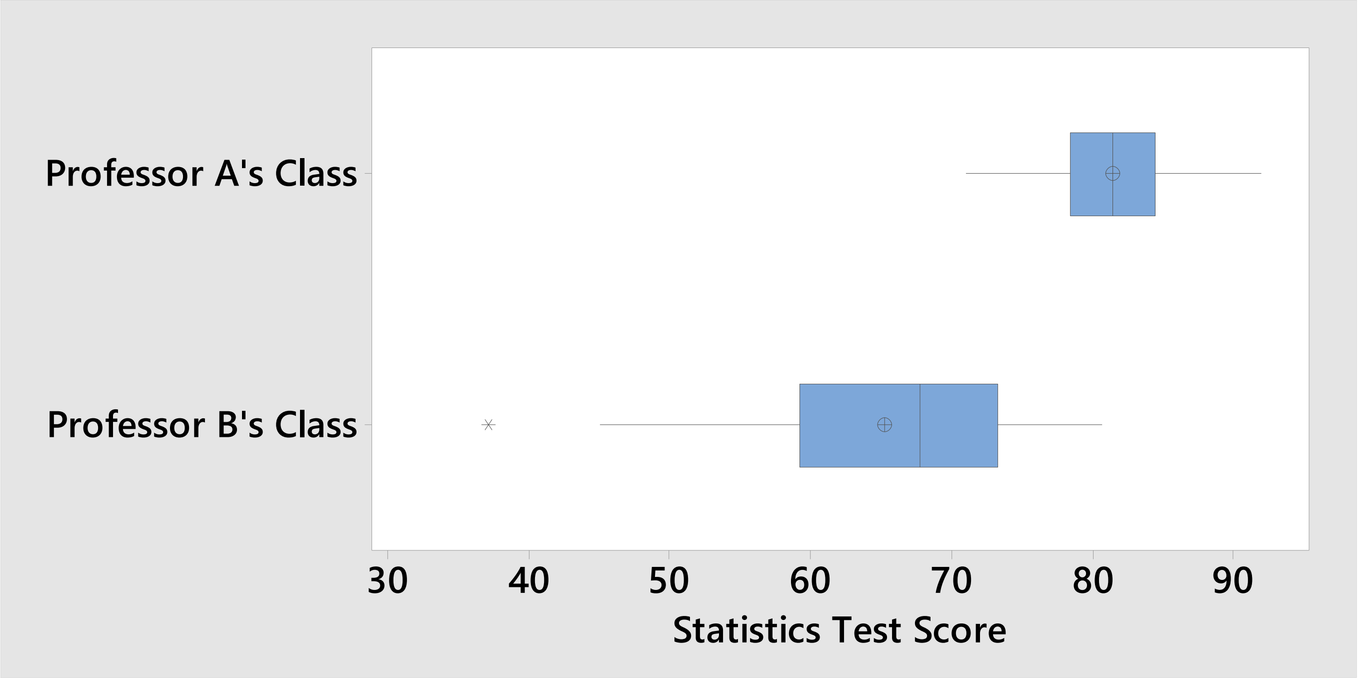
Night Shift



	Summary Statistics
Mean	*8.00*
Median	*8.50*
Sample Size	*12.00*
Q1	*4.50*
Q3	*11.00*
Min	*2.00*
Max	*14.00*
IQR	*6.50*
Range	*12.00*
Variance	*14.18*
Standard Deviation	*3.77*
Lower Fence	*-5.25*
Upper Fence	*20.75*

Comparative Box Plot Example

Professor A and Professor B have given their statistics classes the same exam.
See the exam results in the comparative box plots below.



1. Which class had the higher typical score?
2. Which class had the larger spread in exam scores?
3. Which class has a distribution of scores that would be considered somewhat symmetric?

ANSWER:

1. Prof. A's class has the higher typical score since the mean (dot) and median (vertical line in rectangle) are higher than with Prof. B.

2. Prof. B's class has the larger spread in exam scores as seen by the size of the interquartile range. When comparing box plots built on the same axes, you can compare the size of the IQR rectangle. The larger rectangle indicates the larger spread. Prof. A's class has less variability in the exam scores.

3. Prof. A's class has exam scores that have a somewhat symmetric distribution. We can see this by noticing the mean and median are close in value. Prof. B's class distribution would be considered skewed left since the mean is less than the median (the mean is smaller since there is a small outlier data point).