Chapter 1 – Exploring Data

Lesson Objectives:
In this chapter we will focus on creating appropriate **graphs** based upon both categorical and quantitative data sets. We will also learn how to **describe the main features** of a given data set (Shape, Center, Spread, and Outliers). The purpose of learning how to do this kind of exploratory data analysis is to gain an understanding about the topic that is being analyzed so that in latter chapters we can gather information from the data collected from samples in order to draw conclusions about the whole population.

Date	Topics	**Objectives:** Students will be able to…	Assignment
Jan 29	Chapter 1 Introduction	• Identify the individuals and variables in a set of data. • Classify variables as categorical or quantitative. Identify units of measurement for a quantitative variable.	**TPS: Read pg. 4 – 5** Complete page 2 of Chapter 1 Packet
	1.1 Bar Graphs and Pie Charts,	• Make a bar graph of the distribution of a categorical variable or, in general, to compare related quantities. • Recognize when a pie chart can and cannot be used. •	**TPS: Read pg. 8 – 10.** Watch Video #1 – Take notes. Complete pg. 3 - 4 of Chapter 1 Packet
	1.2 Dotplots, Describing Shape, Comparing Distributions, Stemplots	• Make a dotplot or stemplot to display small sets of data. • Describe the overall pattern (shape, center, spread) of a distribution and identify any major departures from the pattern (like outliers). • Identify the shape of a distribution from a dotplot, stemplot, or histogram as roughly symmetric or skewed. Identify the number of modes.	**TPS: Read pg. 11 - 16.** Watch Video #2 – Take notes. Complete page 5 -6 of Chapter 1 Packet
Jan 30	1.2 Histograms, Using Histograms Wisely	• Make a histogram with a reasonable choice of classes. • Identify the shape of a distribution from a dotplot, stemplot, or histogram as roughly symmetric or skewed. Identify the number of modes. • Interpret histograms.	**TPS: Read pg. 18 -22.** Watch Video #3 – Take notes. Complete pg. 7 - 9 of Chapter 1 Packet
	1.2 Measuring Center: Mean and Median, Comparing Mean and Median, Measuring Spread: IQR, Identifying Outliers	• Calculate and interpret measures of center (mean, median) • Calculate and interpret measures of spread (*IQR*) • Identify outliers using the 1.5 × *IQR* rule.	**TPS: Read pg. 37 - 44.** Watch Video #4 – Take notes. Complete pg. 10 -12 of Chapter 1 Packet
Feb 1	1.2 Five Number Summary and Boxplots, Measuring Spread: Standard Deviation, Choosing Measures of Center and Spread	• Make a boxplot. • Calculate and interpret measures of spread (standard deviation) • Select appropriate measures of center and spread • Use appropriate graphs and numerical summaries to compare distributions of quantitative variables.	**TPS: Read pg. 44 - 52.** Watch Video #5 – Take notes. Complete pg. 14 -17 of Chapter 1 Packet
Chapter 1 Packet Due Feb. 2, 2018. Start Chapter 2 Packet.			

Chapter 1 - Introduction

Answer the following questions as you read Chapter 1. READ FOR UNDERSTANDING!

1. What is statistics? ____

2. Define Individuals. ____

____

3. Define a variable. ____

____

4. Define a categorical variable. ______________________________________

____

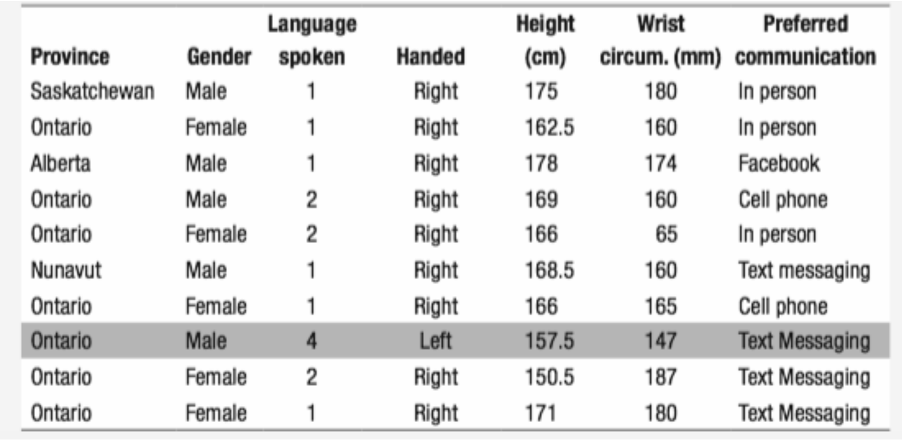
5. Define a quantitative variable. ______________________________________

____

6. Give three examples of categorical variables that are numerical (do NOT include zip codes!!!)

   a. ____________________ b. ____________________ c. ____________________

7. Look over Example “Census at School” and answer the following questions.
   CensusAtSchool is an international project that collects data about students using surveys. Students from Australia, Canada, New Zealand, South Africa, and the United Kingdom have taken part in the project since 2000. The “Random Data Selector” was used to choose 10 Canadian students who completed the survey in a recent year. The table below displays the data.



Province	Gender	Language spoken	Handed	Height (cm)	Wrist circum. (mm)	Preferred communication
Saskatchewan	Male	1	Right	175	180	In person
Ontario	Female	1	Right	162.5	160	In person
Alberta	Male	1	Right	178	174	Facebook
Ontario	Male	2	Right	169	160	Cell phone
Ontario	Female	2	Right	166	65	In person
Nunavut	Male	1	Right	168.5	160	Text messaging
Ontario	Female	1	Right	166	165	Cell phone
Ontario	Male	4	Left	157.5	147	Text Messaging
Ontario	Female	2	Right	150.5	187	Text Messaging
Ontario	Female	1	Right	171	180	Text Messaging

   a. Who are the individuals? ____

   b. How many variables are listed for each individual? __________

   c. What are the units for wrist circumference? ____________

   d. Which variables listed are categorical? ______________________________

   e. Which are quantitative? ____

8. Define the distribution of a variable. ____________________________________

____

Section 1.1 – Analyzing Categorical Data [Watch VIDEO #1]

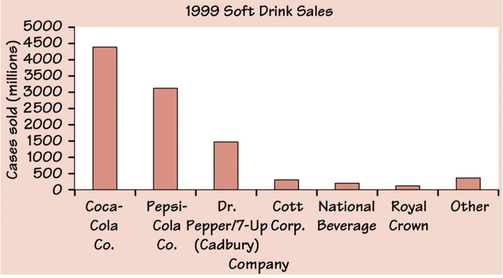
Graphing Categorical Variables: Pie Chart & Bar Graph

The following table displays the sales figures and market share (percent of total sales) achieved by several major soft drink companies in 1999. That year, a total of 9930 million cases of soft drink were sold.

Company	Cases sold (millions)	Market Share (percent)
Coca-Cola Co.	4377.5	44.1
Pepsi-Cola Co.	3119.5	31.4
Dr. Pepper / 7-Up	1455.1	14.7
Cott Corp.	310.0	3.1
National Beverage	205	2.1
Royal Crown	115.4	1.2
Other	347.5	3.4

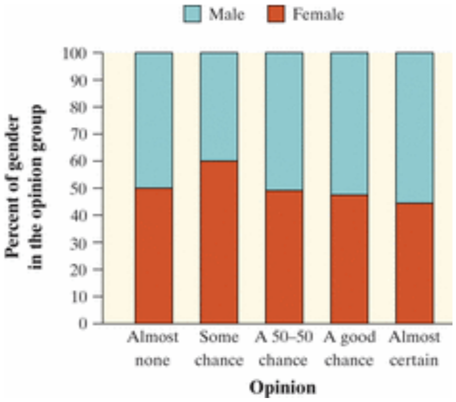
To Create a Bar Graph:

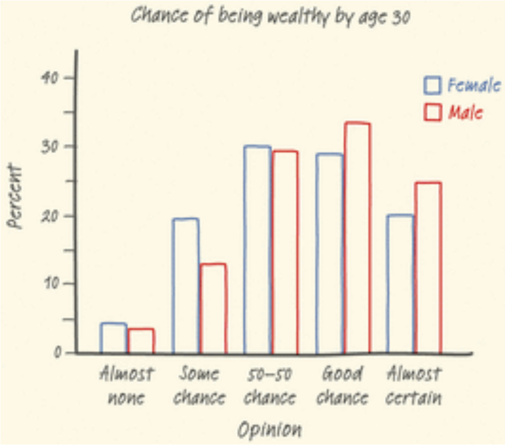
1. Label your axes and title your graph.
2. Scale your axes. Use the counts in each category to help you scale your vertical axis. Write the category names at equally spaced intervals beneath the horizontal axis.
3. Draw a vertical bar above each category name to the height that corresponds to the count in that category. *Leave spaces between the bars.*



<u>Variations</u> - Segmented and Side-by-Side Bar Graphs

- Segmented: used to compare two or more differences within the same variable (gender, grade level, etc.)
  - Each bar **ALWAYS** goes to 100%
  - Each category's bar is one-dimensional (length)
- Side-by-Side: like a segmented, but each category's bar can be two-dimensional (length and width)



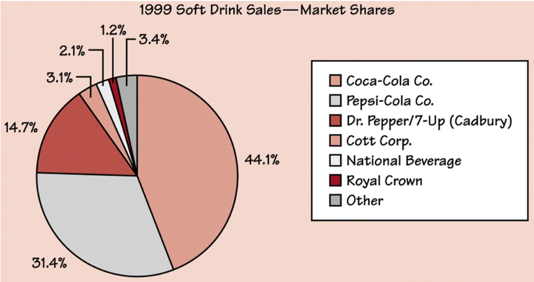


To Create a Pie Chart:

1. Calculate what percent of the whole each category is (if its not given to you!).

2. Multiply the percent by 360 to determine what portion of a circle the category should take up.

3. Draw a circle and draw the “slices” the appropriate size. Don’t forget to label!



Company	Cases sold (millions)	Market Share (%)	Portion of a Circle
Coca-Cola Co.	4377.5	44.1%	159°
Pepsi-Cola Co.	3119.5	31.4%	113°
Dr. Pepper / 7-Up	1455.1	14.7%	53°
Cott Corp.	310.0	3.1%	11°
National Beverage	205	2.1%	8°
Royal Crown	115.4	1.2%	4°
Other	347.5	3.4%	12°

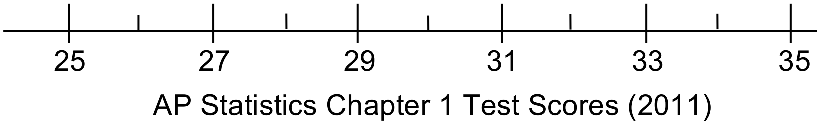
Section 1.2 – Displaying Quantitative Data with Graphs [Watch VIDEO #2]

Graphing Quantitative Variables: Dotplot

AP Statistics Chapter 1 Test Scores (2011)

25	27	27	27.5	30	30	30	30.5	30.5
31	31	31	31.5	32	32	32	32.5	33
33	33	34	34	34	34	34.5	35	

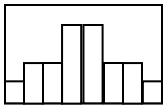
Construct a dotplot of the test scores below:



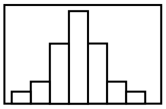
The whole purpose of making a graph is to gain a better understanding of the data set.
We will call this "*describing a distribution*". **When asked to describe a distribution use your S(O)CS!**

1. Describe the distribution's **Shape**. Some of the choices are:

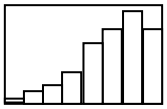
   a. Symmetric



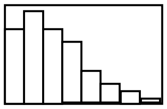
   b. Approximately Normal



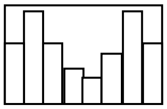
   c. Skewed to the Left



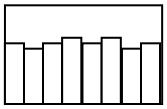
   d. Skewed to the Right



   e. Bimodal



   f. Uniform



2. Identify any **Outliers,** which are values that fall outside the overall pattern of the graph. Does the Chapter 1 data contain any outliers? ________ If so, which one(s)? __________________

3. Give the **Center**.
   a. Median
   b. Mean
   c. Mode*

4. State the **Spread**. Your options are:
   a. Range: ______________________
   b. IQR
   c. Standard Deviation

Graphing Quantitative Variables: Stemplot

Make a split and back-to-back stemplot of the following data.

AP Statistics Final Exam Scores (2011) By Gender

Female Scores: 64%, 69%, 74%, 81%, 85%, 86%, 90%, 91%, 91%, 94%, 95%, 96%, 97%, 98%

Male Scores: 51%, 65%, 72%, 73%, 79%, 81%, 88%, 89%, 91%, 101%, 101%, 105%

Regular Stemplot

5	1
6	4 5 9
7	2 3 4 9
8	1 1 5 6 8 9
9	0 1 1 1 4 5 6 7 8
10	1 1 5

Key:

Split Stemplot

5	
5	
6	
6	
7	
7	
8	
8	
9	
9	
10	
10	

Back-to-Back Stemplot

Males		Females
	5	
	6	
	7	
	8	
	9	
	10	

Stemplots with Decimal Data

Dr. Moore, who lives a few miles outside a college town, records the time he takes to drive to the college each morning. Here are the times (in minutes) for 42 consecutive weekdays, with the dates in order along the rows.

8.28	7.83	8.30	8.42	8.50	8.67	8.17	9.00	9.00	8.17	7.82
9.11	8.50	9.00	7.75	7.92	8.00	8.08	8.42	8.75	8.08	9.75
8.33	7.83	7.92	8.58	7.83	8.42	7.75	7.42	6.75	7.42	8.50
	8.67	10.17	8.75	8.58	8.67	9.17	9.08	8.83	8.67	

Make a stemplot of the data set and describe the distribution.

S:

O:

C:

S:

Key:

Graphing Quantitative Variables: Histogram [VIDEO #3]

A Histogram IS NOT a bar graph!

- It displays the distribution of a quantitative variable (Not a categorical variable)
- The bars will NOT have spaces between them (because all numerical x-axis values will be used).

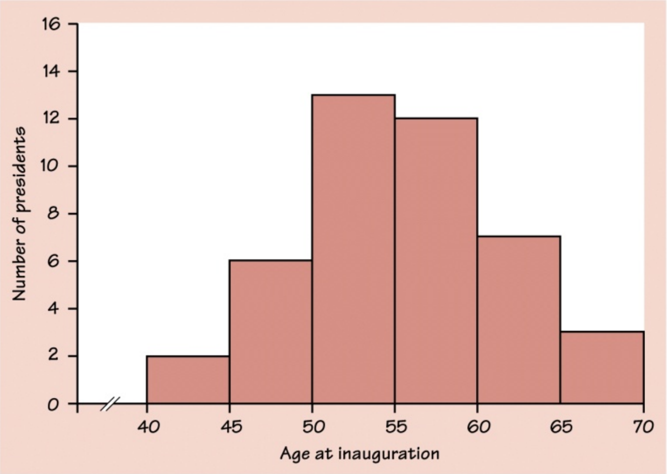
How old are presidents at their inaugurations? Was Barak Obama, at age 47, unusually young? The table below gives the ages of all U.S. presidents when they took office.

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F.D. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	61
Monroe	58	Garfield	49	Kennedy	43
J.Q. Adams	57	Arthur	51	L.B. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W.H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G. Bush	64
Taylor	64	Taft	51	Clinton	46
Fillmore	50	Wilson	56	G.W. Bush	54
Buchanan	65	Coolidge	51	Obama	47

To Create a Histogram:

1. Divide the range of the data into classes of equal width.
2. Count the number of observations in each class.
3. Label and scale your axes.
4. Create the **frequency histogram** by drawing a bar that represents the count in each class. A **relative frequency histogram** would use the percent of presidents that fall in each class.

Class	Count
40 – 44	2
45 – 49	7
50 – 54	13
55 – 59	12
60 – 64	7
65 – 69	3

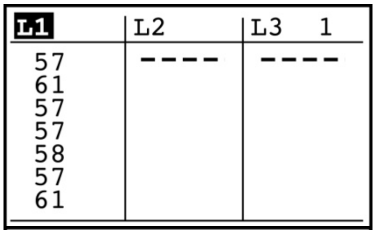


Describe the distribution of president ages at inauguration.

Graphing Quantitative Variables: Histograms on the Calculator

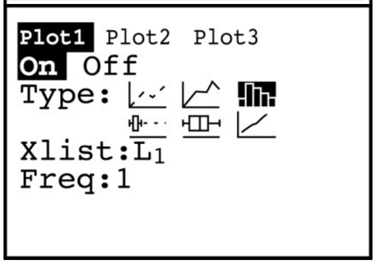
To Create a Histogram on the Calculator:

1. Press [STAT]
2. Select [EDIT]
3. Enter the data by hand



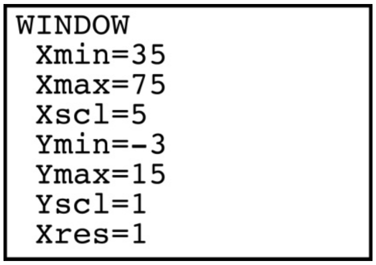
```
L1    |L2    |L3   1
 57   |----  |----
 61   |      |
 57   |      |
 57   |      |
 58   |      |
 57   |      |
 61   |      |
```

4. Press [2nd] [Y=] (Stat plot)
5. Press [ENTER] to go into Plot 1.
6. Adjust your settings as shown.



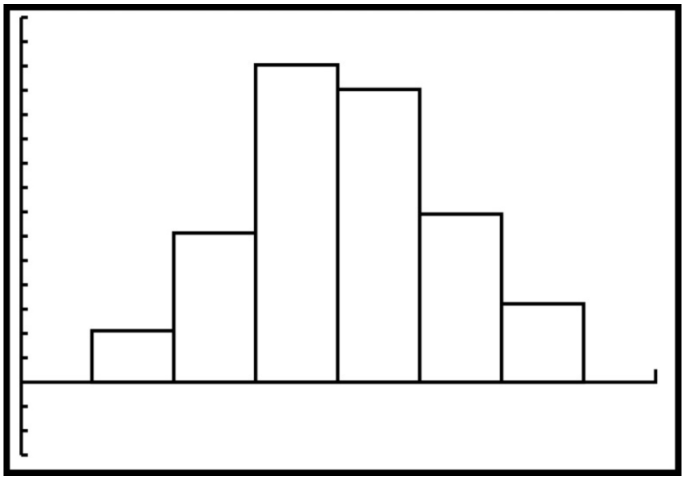
```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:1
```

7. Set the window to match the class intervals chosen by pressing [2nd] [WINDOW]. Enter the values as shown.
   PS – If you want the calculator to choose the window for you, then you can skip this step by pressing [ZOOM] [9: STAT] after step 6.



```
WINDOW
 Xmin=35
 Xmax=75
 Xscl=5
 Ymin=-3
 Ymax=15
 Yscl=1
 Xres=1
```

8. Press [GRAPH].



- This is called a ________________ histogram because it displays the count of presidents on the y-axis.

- A histogram that displays the percent of presidents in each age category is called a _________ _________ histogram.

1.14 CEO SALARIES In 1993, *Forbes* magazine reported the age and salary of the chief executive officer (CEO) of each of the top 59 small businesses.[8] Here are the salary data, rounded to the nearest thousand dollars:

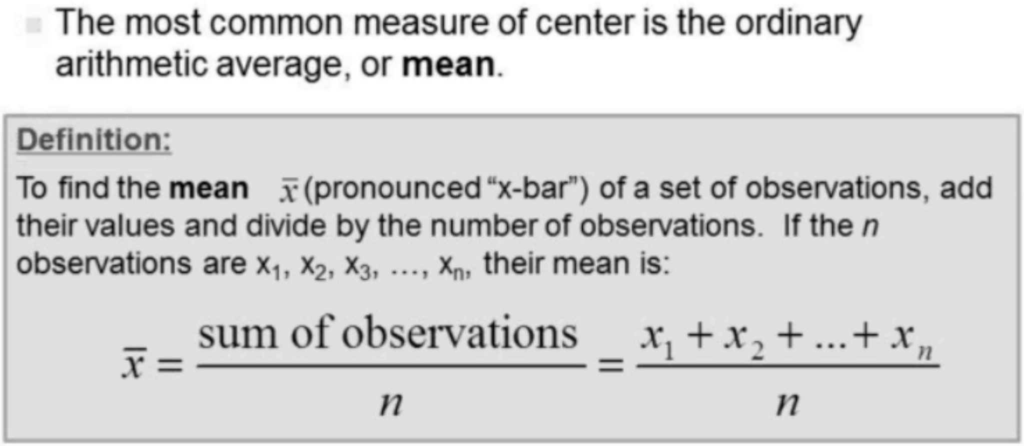
145	621	262	208	362	424	339	736	291	58	498	643	390	332
750	368	659	234	396	300	343	536	543	217	298	1103	406	254
862	204	206	250	21	298	350	800	726	370	536	291	808	543
149	350	242	198	213	296	317	482	155	802	200	282	573	388
250	396	572											

1. Construct a histogram for these data.

2. Describe shape, center, & spread of this distribution

3. Are there any outliers?

Section 1.3 – Describing Quantitative Data with Numbers [VIDEO #4] Measuring Center: Mean / Median



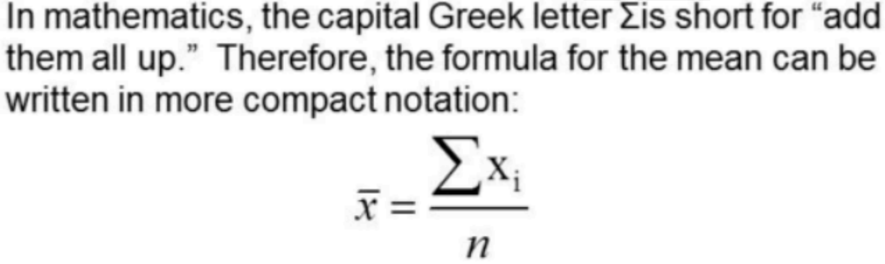
- The most common measure of center is the ordinary arithmetic average, or **mean**.

Definition:

To find the **mean** $\bar{x}$ (pronounced "x-bar") of a set of observations, add their values and divide by the number of observations. If the *n* observations are $x_1, x_2, x_3, \ldots, x_n$, their mean is:

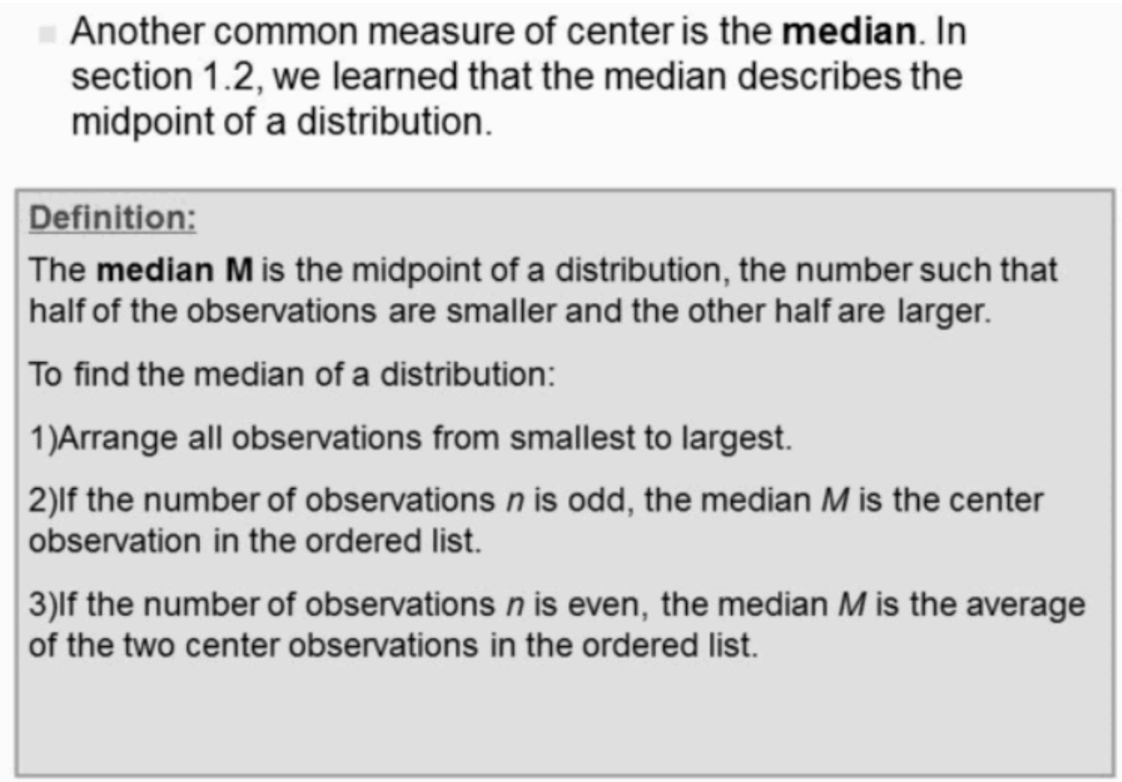
$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

sigma



In mathematics, the capital Greek letter Σis short for "add them all up." Therefore, the formula for the mean can be written in more compact notation:

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



- Another common measure of center is the **median**. In section 1.2, we learned that the median describes the midpoint of a distribution.

Definition:

The **median M** is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger.

To find the median of a distribution:

1)Arrange all observations from smallest to largest.

2)If the number of observations *n* is odd, the median *M* is the center observation in the ordered list.

3)If the number of observations *n* is even, the median *M* is the average of the two center observations in the ordered list.

Ex 1: You get a part time job working at a restaurant. When you were interviewing for the position, the manager told you that the average wage earned by employees at the restaurant is $20/hr. You took the job without hesitation! When you got your first paycheck you realized that you were being paid minimum wage. Outraged, you began to ask your coworkers what they make, and found out that they all make minimum wage. There are 4 employees other than your boss. How much does your boss have to make per hour to have told the truth about the average hourly wage? Assume minimum wage is $5.50/hr.

The mean is **nonresistant!**

Definition: ____

The mean is not the only way to describe the center of a distribution. Another natural idea is to use the "middle value". What is the median wage earned by employees at the restaurant? ___________

Suppose we exclude the boss' salary:

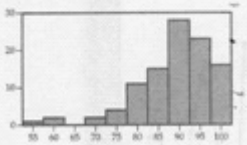
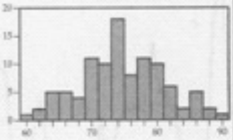
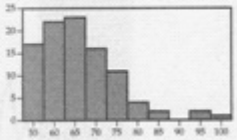
- How would that affect the mean?

- How would that affect the median?

Extending the Ideas

1. What does it mean to be resistant?

2. Is the mean or median resistant?

Graph	LEFT-SKEWED	SYMMETRIC	RIGHT-SKEWED
Histogram			

Generalizations:

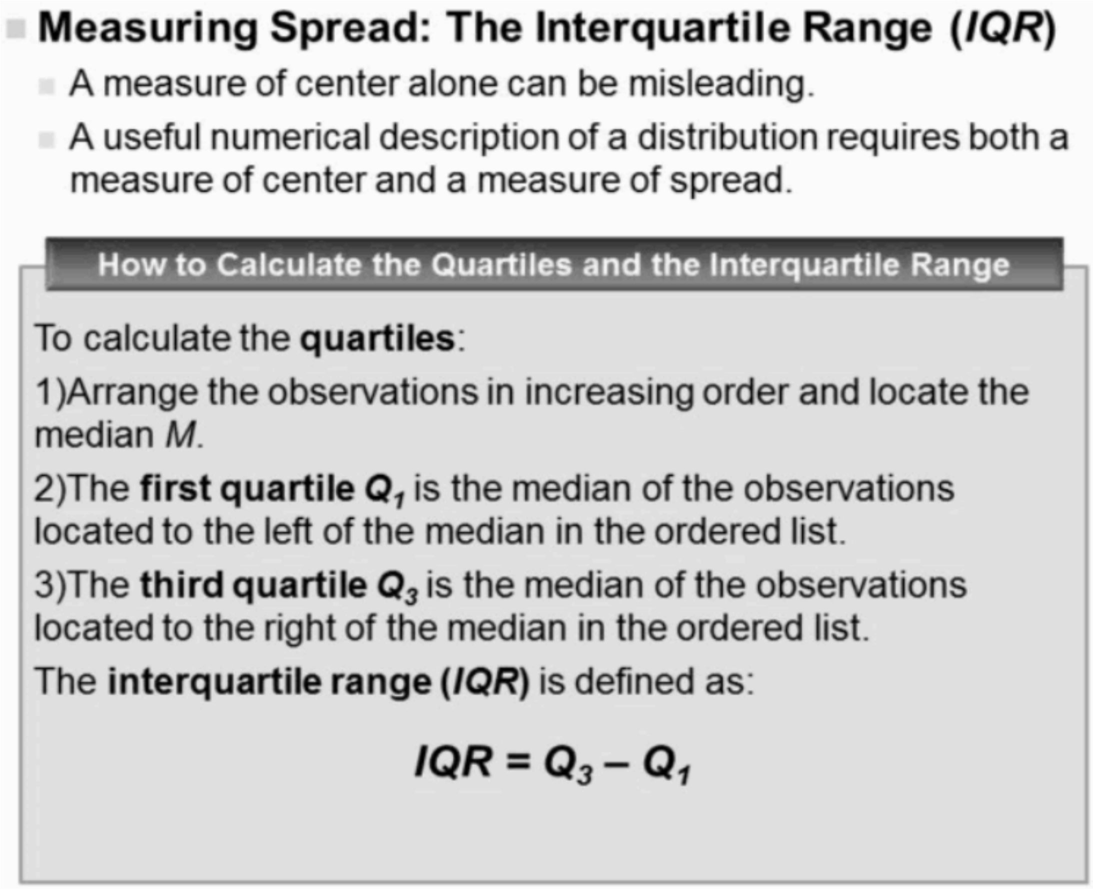
1. When the distribution is skewed to the left, the mean is _________________ the median.
2. When the distribution is symmetrical, the mean is _____________________ the median.
3. When the distribution is skewed to the right, the mean is ___________________ the median.

The mean is always pulled towards _______________ of the distribution!

If you are to describe a set of data by center, which do you choose? Mean or Median???

- ___________________________ works for symmetric (or approximately normal) distributions.
- _____________________ will give a more accurate picture of the distribution if it is skewed left or right.

****



Measuring Spread: The Interquartile Range (*IQR*)

- A measure of center alone can be misleading.
- A useful numerical description of a distribution requires both a measure of center and a measure of spread.

How to Calculate the Quartiles and the Interquartile Range

To calculate the **quartiles**:

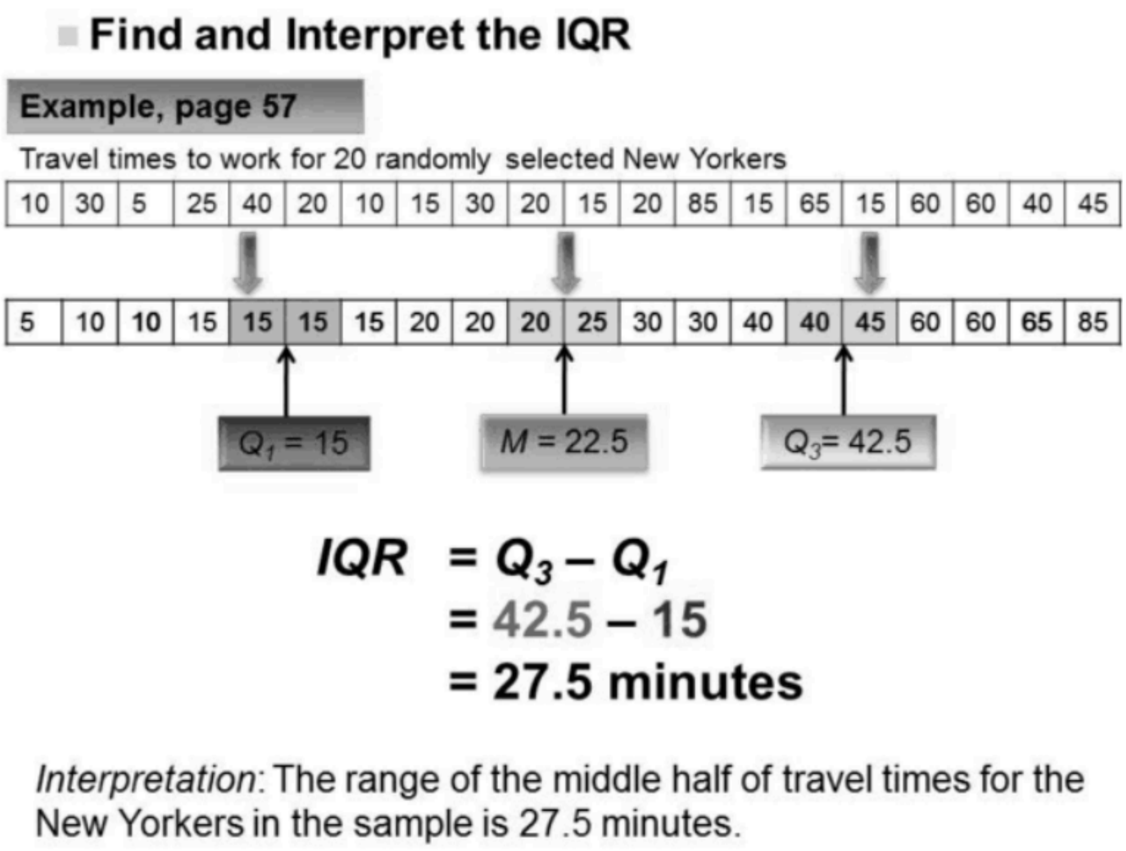
1)Arrange the observations in increasing order and locate the median *M*.

2)The **first quartile Q_1** is the median of the observations located to the left of the median in the ordered list.

3)The **third quartile Q_3** is the median of the observations located to the right of the median in the ordered list.

The **interquartile range (*IQR*)** is defined as:

$$IQR = Q_3 - Q_1$$



Find and Interpret the IQR

Example, page 57

Travel times to work for 20 randomly selected New Yorkers

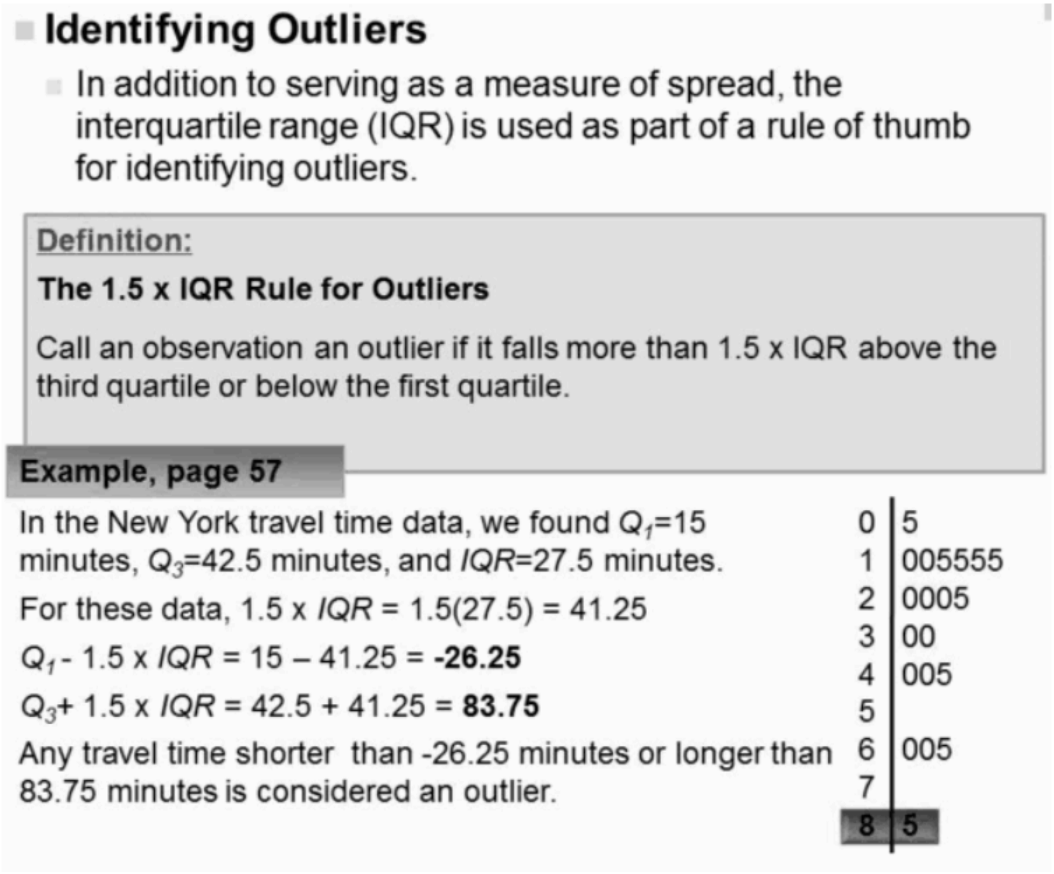
10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	15	15	15	20	20	20	25	30	30	40	40	45	60	60	65	85

$Q_1 = 15$ $M = 22.5$ $Q_3 = 42.5$

$$
\begin{aligned}
IQR &= Q_3 - Q_1 \\
&= 42.5 - 15 \\
&= 27.5 \text{ minutes}
\end{aligned}
$$

Interpretation: The range of the middle half of travel times for the New Yorkers in the sample is 27.5 minutes.



Identifying Outliers

- In addition to serving as a measure of spread, the interquartile range (IQR) is used as part of a rule of thumb for identifying outliers.

Definition:

The 1.5 x IQR Rule for Outliers

Call an observation an outlier if it falls more than 1.5 x IQR above the third quartile or below the first quartile.

Example, page 57

In the New York travel time data, we found Q_1=15 minutes, Q_3=42.5 minutes, and IQR=27.5 minutes.

For these data, 1.5 x IQR = 1.5(27.5) = 41.25

Q_1 - 1.5 x IQR = 15 – 41.25 = **-26.25**

Q_3+ 1.5 x IQR = 42.5 + 41.25 = **83.75**

Any travel time shorter than -26.25 minutes or longer than 83.75 minutes is considered an outlier.

0	5
1	005555
2	0005
3	00
4	005
5	
6	005
7	
8	5

Graphing Quantitative Variables: Boxplots [VIDEO #5]

There are five main features to any data set:

1. ________________________ 2. ________________________

3. ________________________ 4. ________________________

5. ________________________

These five figures make up the **five-number summary** and lead to a new graph, the **boxplot**.

The ***EASIEST*** way to get these values is with your ***CALCULATOR***!!!!

A **modified boxplot** shows outliers as dots or asterisks that are separate from the rest of the boxplot. The outlier rule will help us to determine whether a data set has outliers.

Use your calculator to determine the five-number summaries for each data set below:

Set 1

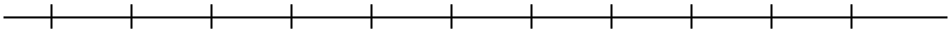
→ Min = ____, Q_1 = ____, M =____, Q_3 = ____ , Max = ____

Set 2

→ Min = ____, Q_1 = ____, M =____, Q_3 = ____ , Max = ____

Test both data sets for outliers.

Construct a modified boxplot:



Measuring Spread: Standard Deviation

Although the five-number summary is a common way of describing the distribution of a data set, the most common way to describe a distribution is by noting the mean and standard deviation.

- Description:____

  ____
- Symbol for the standard deviation of a sample: ______
- To calculate standard deviation, you must first calculate variance (symbol: _____)
- Formula:

Ex 1: In summer school, there was a math class with 5 students in it. On the chapter 1 test, the scores of the 5 students were as follows: 20%, 80%, 85%, 92%, 97%. Calculate the standard deviation of the test scores.

1. Calculate the mean: ___________.

2. Calculate how much each score deviates from the mean.

Observation	Obs – Mean	(Obs – Mean)^2
x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
20		
80		
85		
92		
97		

3. Divide the total (Obs – Mean)^2 by (n-1)
4. Take the square root.

FAQ's:

1. If we want to know the average deviation about the mean, why don't we just take the average of the second column $(x_i - \bar{x})$?

2. What are the properties of the standard deviation?

   a. s measures ___________________________ and should be used only when _________________ ____.

   b. s = 0 only when ______________________________. This happens only when all observations have __________________________________. Otherwise s > 0. As the observations become more spread out about their mean, s gets _____________.

   c. s, like the mean $\bar{x}$, is strongly influenced by _____________________________.

3. May I use my calculator to calculate s? __________________________________

Chapter 1 Review

1. What are your two **MAIN** options for the measure of the center of a distribution?

2. What are the three basic options for the measure of the spread of a distribution?

3. Draw a picture of a distribution that is:
   a. Skewed to the right
   b. Skewed to the left

4. Give an example of a type of data that tends to be symmetric. ________________________________

5. Give an example of a type of data that tends to be skewed right. ______________________________

6. Give an example of a type of data that tends to be skewed left. _______________________________

7. An outlier is any value that falls more than _____________ above _____ or below ______.

8. Write the outlier rule symbolically. ____

9. When describing a distribution you must address the distribution's __________________________ ,

   _______________________ , _____________________, and _________________________________.

10. Give an example of a categorical variable. ____

11. Give an example of a quantitative variable. ____

12. In a left skewed distribution, the mean is __________________ the median.

13. The total return on a stock is the change in its market price plus any dividends payments made. Total return is usually expressed as a percent of the beginning price. The figure shown is a histogram of the distribution of total returns for all 1528 stocks listed in the New York Stock Exchange in one year.

Use the graph to complete the following:

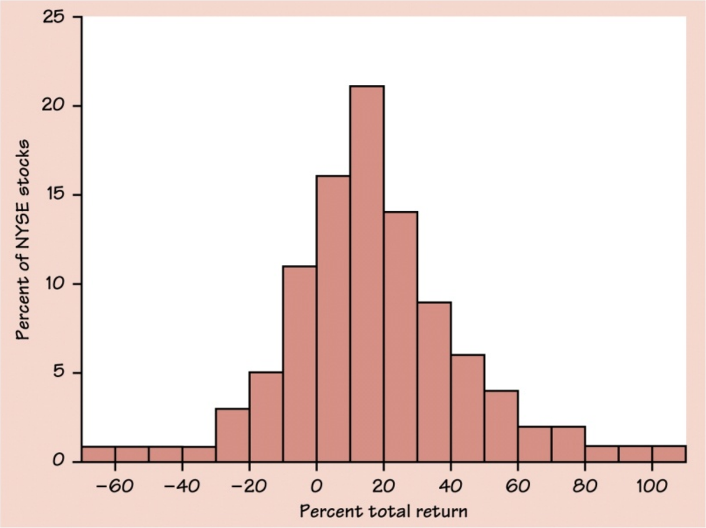
a. Shape: ______________________________

b. Center: ______________________________

c. Smallest: ______________________________

d. Largest: ______________________________

e. What % of stocks lost money? __________



14. The figure shown is a histogram of the number of days in the month of April on which the temperature fell below freezing at Greenwich, England. The data cover a period of 65 years. Use the graph to complete the following:

d. Shape: ______________________________

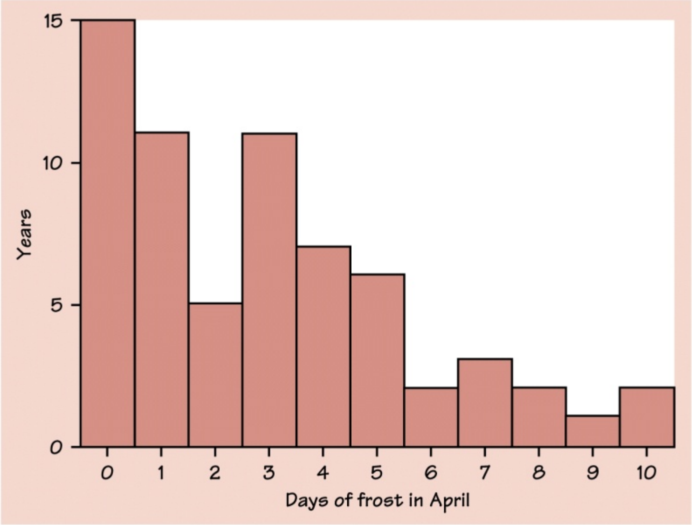
e. Center: ______________________________

f. Spread: ______________________________

g. Outliers? ______________________________

h. In what % of these 65 years did the temp

never fall below freezing in April? _______



After completing Chapter 1, you should know:

- ❑ How to construct and analyze a stemplot, split stem plot, back-to-back stemplot. Know which one to use depending on the situation.
- ❑ The difference between categorical and quantitative variables and be able to identify them.
- ❑ How to draw and analyze a histogram.
- ❑ How to describe a distribution with SOCS
- ❑ How to describe and analyze data in specific terms by calculating:
  - Mean
  - Median
  - Five Number Summary (Min, Q_1, Med, $Q_{3,}$ Max)
  - Range
  - IQR
  - Outliers
  - Standard Deviation
- ❑ How to draw and analyze a regular and modified box plot.
- ❑ The properties of the mean and median (resistant or not, and how they can be used to determine whether a distribution is skewed left, right, or is symmetrical).
- ❑ The properties of standard deviation, what standard deviation represents, and how to calculate it. (You may use the calculator to do this).