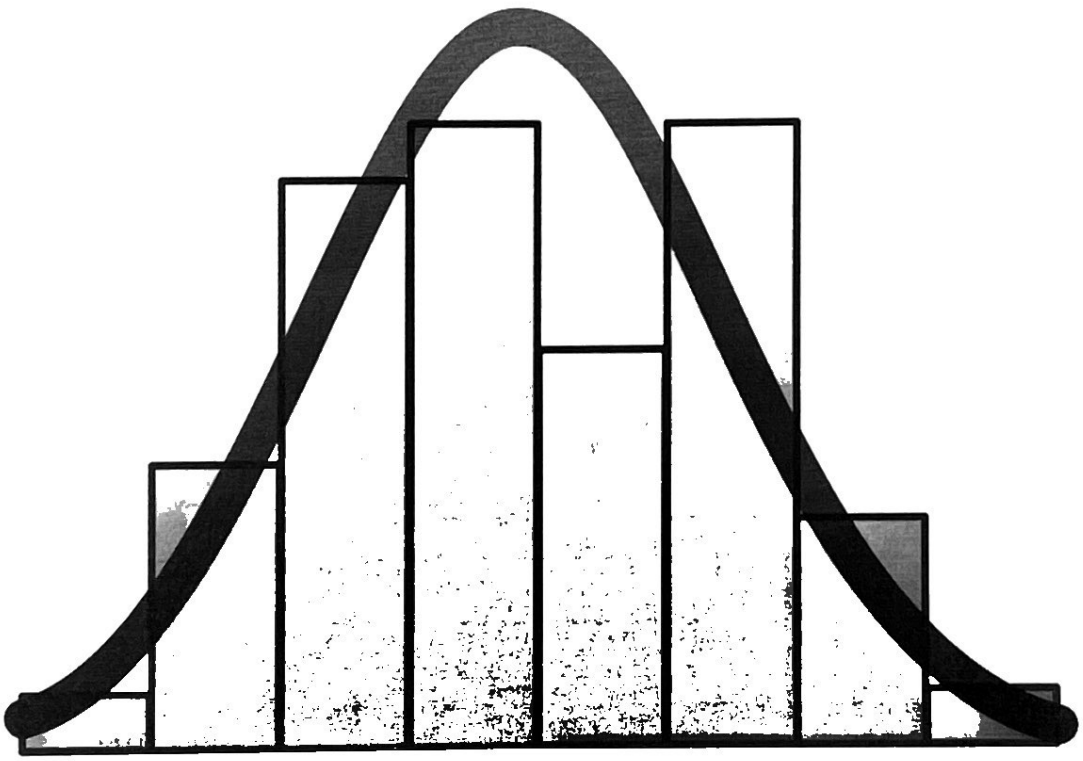
AP Statistics

Summer Packet



Dear Incoming AP Statistics Student,

Thank you for your interest in this course. During the year we will look at a variety of ways to interpret and display data, to analyze results and to determine the validity of statistical findings. Hopefully, the knowledge gained in this course will help you with many of your future endeavors.

Throughout the year you will need access to a TI-84 graphing calculator, both in class and at home. If you do not have access to a calculator over the summer, you can access one online or as an app for now; however, these will not be sufficient for the duration of the course.

In Algebra 1, you were introduced to some of the basic statistics functions on the calculator as well as data entry and sorting. In case you have forgotten how to use these features, the videos links listed on the next page are good resources that can be used to refresh your memory. The remaining pages of this packet will further remind you of some of the most common elements of statistics studied in previous courses and are designed to help you get started in AP Statistics successfully.

I look forward to meeting and working with you next year. Enjoy the summer!

Best regards,

Mrs. Gutheil

Statistics Video Links:

Basic statistics on the TI-84:

https://www.youtube.com/watch?v=sIGyOvTEZ1A

Box Plot and 5 Number Summary:

https://www.youtube.com/watch?v=VvCw5MRo1P4

Histogram:

https://www.youtube.com/watch?v=By0qU-YYBJA

AP Statistics Prerequisite Packet

Types of Data

Quantitative (or measurement) Data

These are data that take on numerical values that actually represent a measurement such as size, weight, how many, how long, score on a test, etc. For these data, it makes sense to find things like "average" or "range" (largest value – smallest value). For instance, it doesn't make sense to find the mean shirt color because shirt color is not an example of a quantitative variable. Some quantitative variables take on **discrete** values, such as shoe size (6, 6 ½, 7, ...) or the number of soup cans collected by a school. Other quantitative variables take on **continuous** values, such as your height (60 inches, 72.99999923 inches, 64.039 inches, etc,) or how much water it takes to fill up your bathtub (73.296 gallons or 185. 4 gallons or 99 gallons, etc.).

Categorical (or qualitative) Data

These are data that take on values that describe some characteristic of something, such as the color of shirts. These values are "categories" of a population, such as *M* or *F* for gender of people or "Don't Drive" or "Drive" for the method of transportation used by students to get to school. These are examples of **binary** variables. These variables only have two possible values. Some categorical variables have more than two values, such as hair color, brand of jeans, and so on.

Two Types of Variables

Exercises: Answer the following questions and then decide if the data is quantitative or categorical. (***Q*** or ***C***)

	ANSWER	**TYPE**
1. In what grade did you take your first algebra class (Math I, Coord. Alg., etc.)?	_____	_____
2. How many pairs of shoes do you own?	_____	_____
3. How old was your father when you were born?	_____	_____
4. How old was your mother when you were born?	_____	_____
5. Choose a random integer from 1 to 20.	_____	_____
6. How many siblings do you have? (**all**, whether you live with them or not)	_____	_____

7. How many cousins do you have? _____ _____

8. How tall are you (**in inches**)? _____ _____

9. How many AP classes will you be taking **THIS** year? _____ _____

10. What gender are you? _____ _____

11. Where did eat your last meal? _____ _____
(1 = home, 2 = restaurant, 3 = other)

12. How long have you lived in this area? _____ _____

13. How far away from school do you live? _____ _____

Numerical Descriptions of Quantitative Data

Measures of Center

Mean: The sum of all the data values divided by the number (n) of data values.

Example

Data: 4, 36, 10, 22, 9 $\quad \text{Mean} = \bar{x} = \frac{\sum x_i}{n} = \frac{4+36+10+22+9}{5} = \frac{81}{5} = 16.2$

Median: The middle element of an ordered set of data.

Examples

Data: 4, 36, 10, 22, 9 $\quad$ = 4 9 <u>10</u> 22 36 $\quad$ Median = 10

Data: 4, 36, 10, 22, 9, 43 $\quad$ = 4 9 10 | 22 36 43 $\quad$ Median = $\frac{10+22}{2} = 16$

Measures of Spread:

Range: Maximum value – Minimum value

Example

Data: 4, 36, 10, 22, 9 = 4 9 10 22 36

Range = Max. – Min. = 36 – 4 = **32**

AP Statistics Prerequisite Packet

Interquartile Range (IQR): The difference between the 75th percentile (Q3) and the 25th percentile (Q1). This is **Q3 – Q1.** Q1 is the median of the lower half of the data and Q3 is the median of the upper half. In neither case is the median of the data included in these calculations. The IQR contains the middle 50% of the data. Each quartile contains 25% of the data.

Examples

1. Data: 4, 36, 10, 22, 9 = 4 9 **10** 22 36

Q1 = 6.5 Q3 = 29

So, the IQR = 29 – 6.5 = 22.5

2. Data: 4 **9** 10 | 22 **36** 43

Q1 Q3

So, the IQR = 36 – 9 = 27

Exercises

Last year students collected data on the age of their moms and dads when they (the students) were born. The following are their results.

Dad:	41	27	23	31	30	33	26	32	43	25	34	27
	25	34	27	26	28	32	32	35	27	33	34	34
	34	35										
Mom:	39	26	23	30	28	33	23	32	38	23	35	24
	24	33	24	23	24	32	23	30	24	29	34	35
	26	31										

AP Statistics Prerequisite Packet

1. Find the mean and the median for the Dad data.

Mean: _____ Median: _____

Are they the same? If not, which is larger?

2. Find the mean and the median for the Mom data.

Mean: _____ Median: _____

Are they the same? If not, which is larger?

3. Now compare the two means you calculated. Which is larger? _____

Is this result what you expected? _____

Why/why not?

4. Calculate the range for each set of data. Dad _____ Mom _____

5. Are these ranges the same? _____ If not, what could account for the differences?

6. Find the Q1 and Q3 for the Dad data: Q1: _____ Q3: _____

7. Find the Q1 and Q3 for the Mom data: Q1: _____ Q3: _____

8. You have now calculated the "Five-Number Summary." This can also be used as a way to determine the spread of a set of data. The five-number summary consists of:

Minimum	**Q1**	**Median**	**Q3**	**Maximum**

Write the five number summary for the Dad data: _____

Write the five number summary for the Mom data: _____

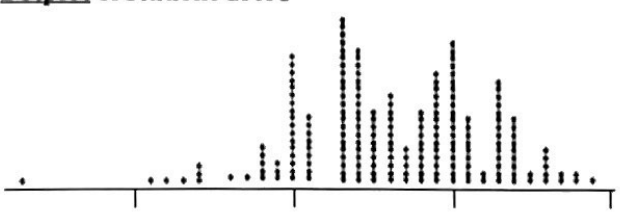
9. Now calculate the IQR for each of the two sets of data.

Dad: _____ Mom: _____

Graphical Displays of Univariate (one variable) Data

Quantitative Data: Dotplot
Boxplot (Box and Whiskers)
Stemplot (Stem and Leaf)
Histogram

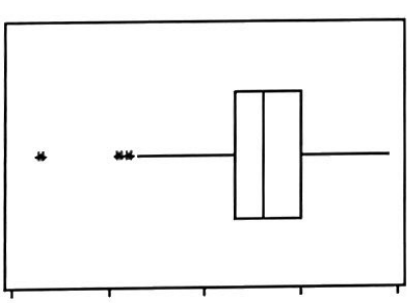
Dotplot of Student GPA's



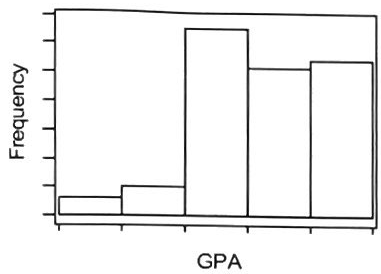
Stemplot of Student GPA's

1	23
1	444
1	67
1	88888999
2	000000000000000000111111111
2	3333333333333333333
2	444444444444444445555555555
2	66666666666677777
2	888888888899999999999999
3	000000000000000000111111111
3	2233333333333333
3	44444444455
3	6666677
3	889

Boxplot of Student GPA's



Histogram of Student GPA's



Categorical Data: Bar Graph
Circle Graph

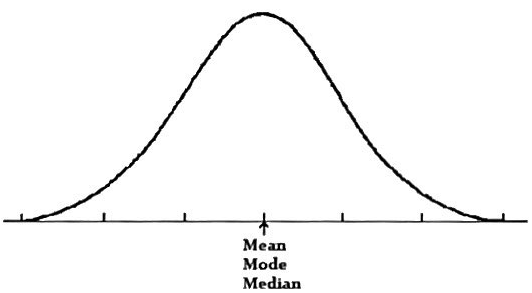
**I'm assuming that you already know how to make these two types of graphs.*

Assessing the Shape of a Graph

There are two basic shapes that we will examine: **Symmetric** and **Skewed.**

Symmetric: One can tell if a graph is symmetric if a vertical line in the "center" divides the graph into two fairly congruent shapes. (A graph does not have to be "bell-shaped" to be considered symmetric.)

Mean is *approximately equal* to the Median in a symmetric distribution



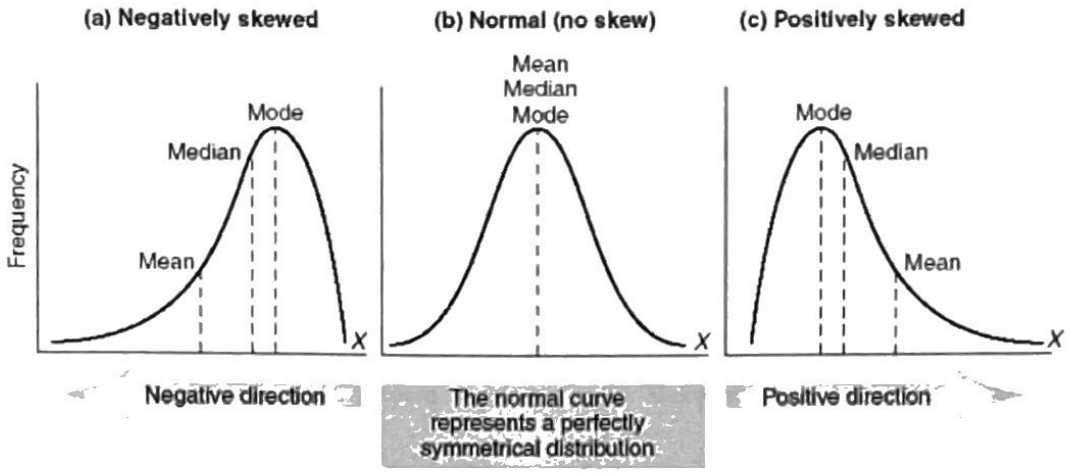
Skewed: One can tell that a graph is skewed if the graph has a big clump of data on either the left (skewed right) or on the right (skewed left) with a tendency to get flatter and flatter as the values of the data increase (skewed right) or decrease (skewed left). A common misconception is that the "skewness" occurs at the big clump – it does not!

Relationship between Mean and Median in a skewed distribution:

"Skewed Left, the mean is Less" or $mean < median$

"Skewed Right, the mean is Might" or $mean > median$

AP Statistics Prerequisite Packet



For the distributions above, graph (a) is more commonly referred to as a left-skewed distribution. Graph (c) is more commonly referred to as a right-skewed distribution.

Gathering Information from a Graphical Display

The first thing that should be done after gathering data is to examine it graphically and numerically to find out as much information about the various features of the data as possible. These will be important when choosing what kind of procedures will be appropriate to use to find out an answer to a question that is being investigated.

The features that are the most important are **S**hape, **O**utliers, **C**enter, **C**lusters and gaps, and **S**pread: **SOCCS**. Most of these can only be seen in a graph. However, sometimes the shape is indistinct – difficult to discern. So, in this instance (usually because of a very small set of data), it's appropriate to label the shape "indistinct."

Exercises

1. Construct a boxplot for each the following sets of data taken from consumer ratings of different brands of peanut butter in the September, 2013, issue of Consumer Reports. **Use the same number line for both graphs**. (You could do it this way: Draw a number line. Above this line construct the "Crunchy" boxplot. Then, above the "Crunchy" boxplot, construct the "creamy" boxplot.) Please place your boxplots below.

Crunchy:	62	53	75	42	47	40	34	62	52	50
	34	42	36	75	80	47	56	62		
Creamy:	56	44	62	36	39	50	53	45	65	40
	56	68	41	30	40	50	56	30	22	

(a) Find the range for: Creamy: _____ Crunchy: _____

(b) Find the median for: Creamy: _____ Crunchy: _____

(c) Looking at your boxplots and comparing the medians what type of peanut butter do consumers tend to prefer?

2. The following data is taken from the Statistical Abstract of the United States (112th Edition). These are the ages of drivers arrested for DUI from a random sample of size 50. Make a stemplot to show the distribution of this age data.

45	16	41	26	22	33	30	22	36	34
63	24	26	18	27	24	31	38	26	55
31	47	27	43	35	22	64	40	58	20
49	37	53	25	29	32	23	49	39	40
24	56	30	51	21	45	27	34	47	35

(a) What is the shape of this graph? _____

(b) Using your stemplot, find the median of this data. _____

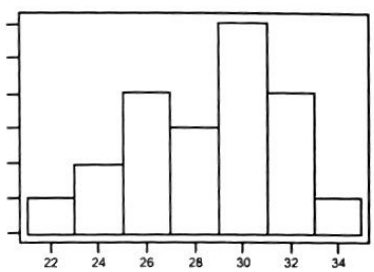
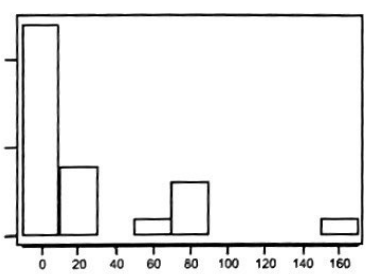
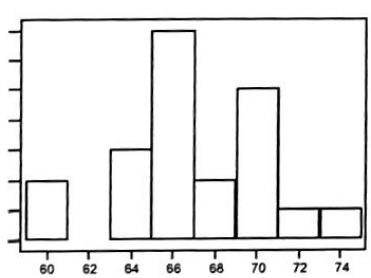
(c) Which data display is better – a boxplot or a stemplot? _____

Why? (Be specific.)

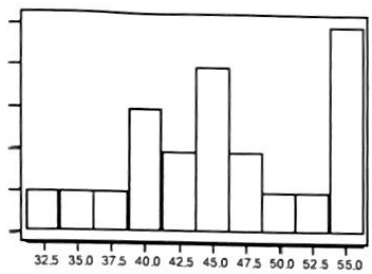
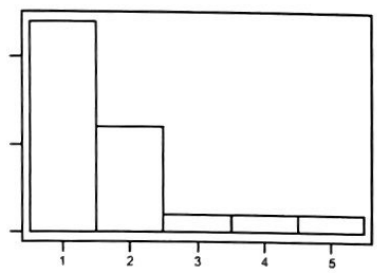
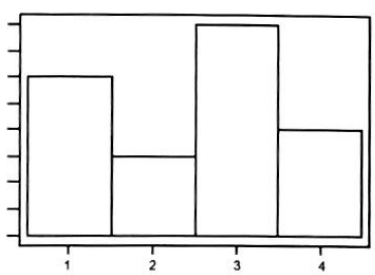
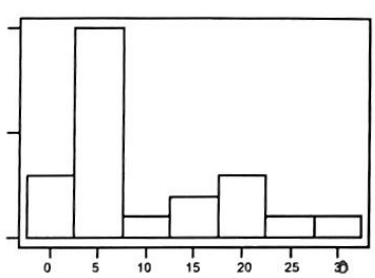
AP Statistics Prerequisite Packet

3. For the following graphs, find the center (**just do the median**), shape, and spread (find only the **range**). If there any other notable features evident in the graph (clusters, gaps, or outliers), then say where they are. Otherwise do not comment on clusters, gaps or outliers.

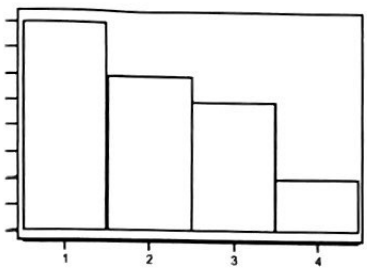
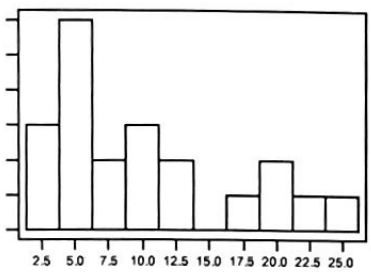
Note: To find the center of these graphs, use the frequencies found on the y-axis. Count how many are in each bar. Add these up and divide by two. This tells you where the median is located. Which bar is this value in? That's the median. For graph A, n = 21, so the middle value is 10.5. Starting with the first bar count 1 + 2 + 4 + 3 + 6 ... So the median is in the bar that contains the 10.5 value (bigger than 10, anyway). That's 30. So, the median is 30. To find a **VERY** rough estimate of the mean, take the frequency for each bar and multiply it by the value along the x-axis for that bar. Add these up for all the bars and then divide by 21. You get the mean = 28.571.

Graph A  	**Center?** **Shape?** **Spread?**
Graph B  	**Center?** **Shape?** **Spread?** **Clusters or gaps? Where?** **Outliers? Where?**
Graph C  	**Center?** **Shape?** **Spread?** **Clusters or gaps? Where?**

AP Statistics Prerequisite Packet

Graph D  	**Center?** **Shape?** **Spread?**
Graph E  	**Center?** **Shape?** **Spread?** **Clusters or gaps? Where?**
Graph F  	**Center?** **Shape?** **Spread?**
Graph G  	**Center?** **Shape?** **Spread?**

AP Statistics Prerequisite Packet

Graph H  	**Center?**
	Shape?
	Spread?
Graph I  	**Center?**
	Shape?
	Spread?
	Clusters or gaps? Where?

AP Statistics Prerequisite Packet Solutions

Types of Data

Quantitative (or measurement) Data

These are data that take on numerical values that actually represent a measurement such as size, weight, how many, how long, score on a test, etc. For these data, it makes sense to find things like "average" or "range" (largest value – smallest value). For instance, it doesn't make sense to find the mean shirt color because shirt color is not an example of a quantitative variable. Some quantitative variables take on **discrete** values, such as shoe size (6, 6 ½, 7, ...) or the number of soup cans collected by a school. Other quantitative variables take on **continuous** values, such as your height (60 inches, 72.99999923 inches, 64.039 inches, etc,) or how much water it takes to fill up your bathtub (73.296 gallons or 185. 4 gallons or 99 gallons, etc.).

Categorical (or qualitative) Data

These are data that take on values that describe some characteristic of something, such as the color of shirts. These values are "categories" of a population, such as *M* or *F* for gender of people or "Don't Drive" or "Drive" for the method of transportation used by students to get to school. These are examples of **binary** variables. These variables only have two possible values. Some categorical variables have more than two values, such as hair color, brand of jeans, and so on.

Two Types of Variables

Exercises: Answer the following questions and then decide if the data is quantitative or categorical. (*Q* or *C*)

	ANSWER*	TYPE
1. In what grade did you take your first algebra class (Math I, Coord. Alg., etc.)?	8	Q
2. How many pairs of shoes do you own?	20	Q
3. How old was your father when you were born?	31	Q
4. How old was your mother when you were born?	29	Q
5. Choose a random integer from 1 to 20.	3	Q
6. How many siblings do you have? (**all**, whether you live with them or not)	2	Q

* Answers will vary from person to person.

AP Statistics Prerequisite Packet

7. How many cousins do you have?	27	Q
8. How tall are you (**in inches**)?	64	Q
9. How many AP classes will you be taking **THIS** year?	3	Q
10. What gender are you?	F	C
11. Where did eat your last meal? (1 = home, 2 = restaurant, 3 = other)	1	C
12. How long have you lived in this area?	20 years	Q
13. How far away from school do you live?	8 miles	Q

Numerical Descriptions of Quantitative Data

Measures of Center

Mean: The sum of all the data values divided by the number (n) of data values.

Example

Data: 4, 36, 10, 22, 9 $\quad \text{Mean} = \bar{x} = \frac{\sum x_i}{n} = \frac{4+36+10+22+9}{5} = \frac{81}{5} = 16.2$

Median: The middle element of an ordered set of data.

Examples

Data: 4, 36, 10, 22, 9 $\quad$ = 4 9 **10** 22 36 $\quad$ Median = 10

Data: 4, 36, 10, 22, 9, 43 $\quad$ = 4 9 10 | 22 36 43 $\quad$ Median = $\frac{10+22}{2} = 16$

Measures of Spread:

Range: Maximum value – Minimum value

Example

Data: 4, 36, 10, 22, 9 = 4 9 10 22 36

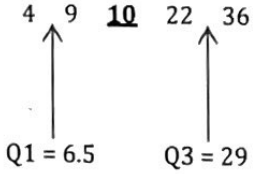
Range = Max. – Min. = 36 – 4 = **32**

AP Statistics Prerequisite Packet

Interquartile Range (IQR): The difference between the 75th percentile (Q3) and the 25th percentile (Q1). This is **Q3 – Q1.** Q1 is the median of the lower half of the data and Q3 is the median of the upper half. In neither case is the median of the data included in these calculations. The IQR contains the middle 50% of the data. Each quartile contains 25% of the data.

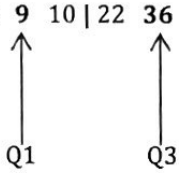
Examples

1. Data: 4, 36, 10, 22, 9 = 4 9 **10** 22 36



So, the IQR = 29 – 6.5 = 22.5

2. Data: 4 **9** 10 | 22 **36** 43



So, the IQR = 36 – 9 = 27

Exercises

Last year students collected data on the age of their moms and dads when they (the students) were born. The following are their results.

Dad:	41	27	23	31	30	33	26	32	43	25	34	27
	25	34	27	26	28	32	32	35	27	33	34	34
	34	35										
Mom:	39	26	23	30	28	33	23	32	38	23	35	24
	24	33	24	23	24	32	23	30	24	29	34	35
	26	31										

AP Statistics Prerequisite Packet

1. Find the mean and the median for the Dad data.

Mean: 31.077 Median: 32

Are they the same? If not, which is larger?
They are not the same. The median is larger.

2. Find the mean and the median for the Mom data.

Mean: 28.692 Median: 28.5

Are they the same? If not, which is larger?
They are not the same. The mean is larger.

3. Now compare the two means you calculated. Which is larger? Dad

Is this result what you expected? yes

Why/why not?
It is not uncommon for men to be older than women in a relationship.

43-23= 39-23=

4. Calculate the range for each set of data. Dad 20 Mom 16

5. Are these ranges the same? NO If not, what could account for the differences?
Women are more likely to have children at a younger age.

6. Find the Q1 and Q3 for the Dad data: Q1: 27 Q3: 34

7. Find the Q1 and Q3 for the Mom data: Q1: 24 Q3: 33

8. You have now calculated the "Five-Number Summary." This can also be used as a way to determine the spread of a set of data. The five-number summary consists of:

Minimum	Q1	Median	Q3	Maximum

Write the five number summary for the Dad data: 23, 27, 32, 34, 43

Write the five number summary for the Mom data: 23, 24, 28.5, 33, 39

9. Now calculate the IQR for each of the two sets of data.

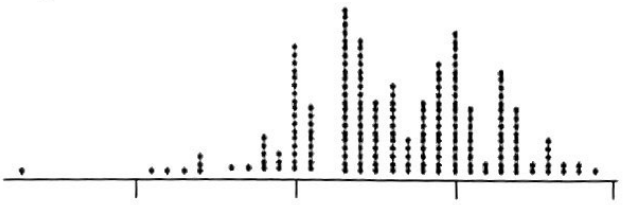
Dad: 7
34-27=

Mom: 9
33-24=

Graphical Displays of Univariate (one variable) Data

Quantitative Data: Dotplot
Boxplot (Box and Whiskers)
Stemplot (Stem and Leaf)
Histogram

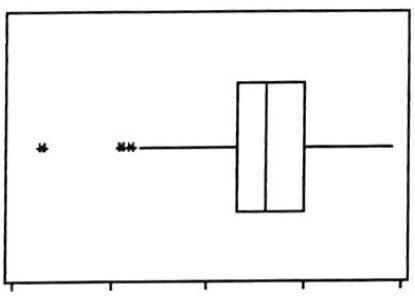
Dotplot of Student GPA's



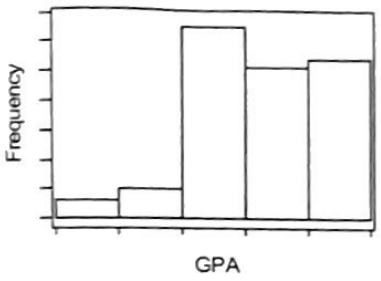
Stemplot of Student GPA's

1	23
1	444
1	67
1	88888999
2	00000000000000000111111111
2	3333333333333333333
2	44444444444444444455555555555
2	66666666666677777
2	88888888899999999999999
3	000000000000000000111111111
3	2233333333333333
3	44444444455
3	6666677
3	889

Boxplot of Student GPA's (*see TI-84 guide on how to make these*)



Histogram of Student GPA's



Categorical Data: Bar Graph
Circle Graph

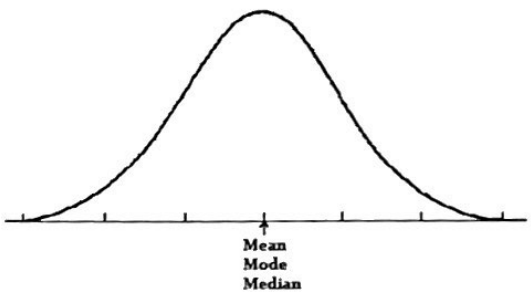
**I'm assuming that you already know how to make these two types of graphs.*

Assessing the Shape of a Graph

There are two basic shapes that we will examine: **Symmetric** and **Skewed.**

Symmetric: One can tell if a graph is symmetric if a vertical line in the "center" divides the graph into two fairly congruent shapes. (A graph does not have to be "bell-shaped" to be considered symmetric.)

Mean is *approximately equal* to the Median in a symmetric distribution



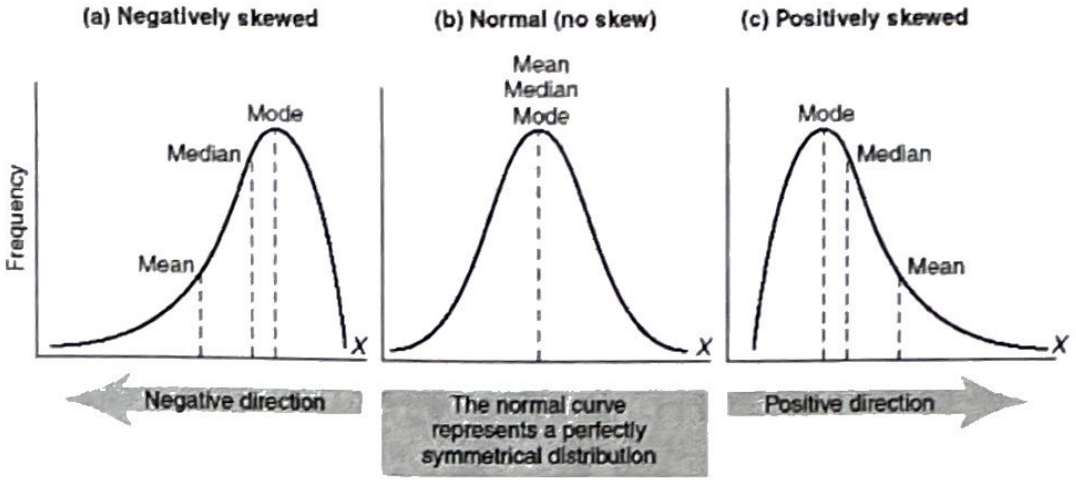
Skewed: One can tell that a graph is skewed if the graph has a big clump of data on either the left (skewed right) or on the right (skewed left) with a tendency to get flatter and flatter as the values of the data increase (skewed right) or decrease (skewed left). A common misconception is that the "skewness" occurs at the big clump – it does not!

Relationship between Mean and Median in a skewed distribution:

"Skewed Left, the mean is Less" or $mean < median$

"Skewed Right, the mean is Might" or $mean > median$

AP Statistics Prerequisite Packet



For the distributions above, graph (a) is more commonly referred to as a left-skewed distribution. Graph (c) is more commonly referred to as a right-skewed distribution.

Gathering Information from a Graphical Display

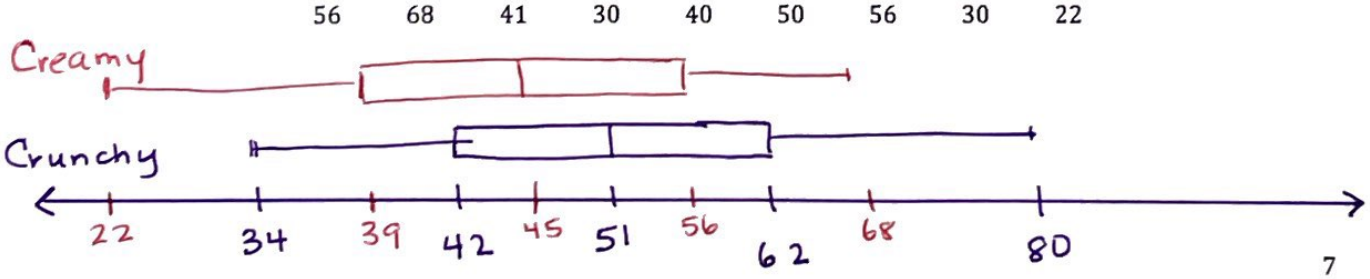
The first thing that should be done after gathering data is to examine it graphically and numerically to find out as much information about the various features of the data as possible. These will be important when choosing what kind of procedures will be appropriate to use to find out an answer to a question that is being investigated.

The features that are the most important are **S**hape, **O**utliers, **C**enter, **C**lusters and gaps, and **S**pread: **SOCCS**. Most of these can only be seen in a graph. However, sometimes the shape is indistinct – difficult to discern. So, in this instance (usually because of a very small set of data), it's appropriate to label the shape "indistinct."

Exercises

1. Construct a boxplot for each the following sets of data taken from consumer ratings of different brands of peanut butter in the September, 2013, issue of Consumer Reports. **Use the same number line for both graphs**. (You could do it this way: Draw a number line. Above this line construct the "Crunchy" boxplot. Then, above the "Crunchy" boxplot, construct the "creamy" boxplot.) Please place your boxplots below.

Crunchy:	62	53	75	42	47	40	34	62	52	50
	34	42	36	75	80	47	56	62		
Creamy:	56	44	62	36	39	50	53	45	65	40
	56	68	41	30	40	50	56	30	22	



AP Statistics Prerequisite Packet

(a) Find the range for: Creamy: 46 Crunchy: 46

(b) Find the median for: Creamy: 45 Crunchy: 51

(c) Looking at your boxplots and comparing the medians what type of peanut butter do consumers tend to prefer? Consumers prefer Crunchy peanut butter since the median is higher.

2. The following data is taken from the Statistical Abstract of the United States (112th Edition). These are the ages of drivers arrested for DUI from a random sample of size 50. Make a stemplot to show the distribution of this age data.

45	16	41	26	22	33	30	22	36	34
63	24	26	18	27	24	31	38	26	55
31	47	27	43	35	22	64	40	58	20
49	37	53	25	29	32	23	49	39	40
24	56	30	51	21	45	27	34	47	35

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

(a) What is the shape of this graph? Symmetric

(b) Using your stemplot, find the median of this data. 33.5

(c) Which data display is better – a boxplot or a stemplot? Stemplot

Why? (Be specific.)

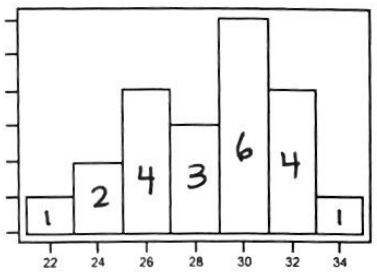
While a boxplot allows us to quickly identify the 5 number summary, a stem plot shows individual data points, gaps, clusters and peaks.

AP Statistics Prerequisite Packet

3. For the following graphs, find the center (**just do the median**), shape, and spread (find only the **range**). If there any other notable features evident in the graph (clusters, gaps, or outliers), then say where they are. Otherwise do not comment on clusters, gaps or outliers.

Note: To find the center of these graphs, use the frequencies found on the y-axis. Count how many are in each bar. Add these up and divide by two. This tells you where the median is located. Which bar is this value in? That's the median. For graph A, n = 21, so the middle value is 10.5. Starting with the first bar count 1 + 2 + 4 + 3 + 6 ... So the median is in the bar that contains the 10.5 value (bigger than 10, anyway). That's 30. So, the median is 30. To find a **VERY** rough estimate of the mean, take the frequency for each bar and multiply it by the value along the x-axis for that bar. Add these up for all the bars and then divide by 21. You get the mean = 28.571.

Graph A



$\frac{21}{2} = 10.5$

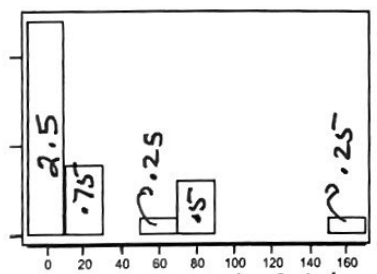
$1+2+4+3+6+4+1 = 21$

Center? 30

Shape? Symmetric

Spread? $34-22=12$

Graph B



$\frac{4.25}{2} = 2.125$

$2.5+.75+.25+.5+.25=4.25$

Center? 0

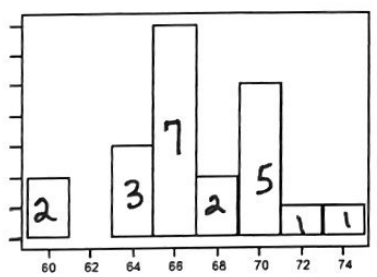
Shape? Skewed right

Spread? $160-0=160$

Clusters or gaps? Where? Yes. Cluster at 0. Gap at 40, 100, 120, 140

Outliers? Where? Yes 160.

Graph C



$\frac{21}{2} = 10.5$

$2+3+7+2+5+1+1=21$

Center? 66

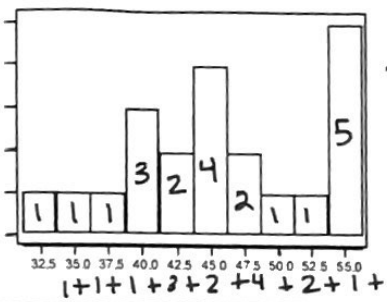
Shape? Symmetric

Spread? $74-60=14$

Clusters or gaps? Where? NO clusters. Gap at 62

AP Statistics Prerequisite Packet

Graph D



$\frac{21}{2} = 10.5$

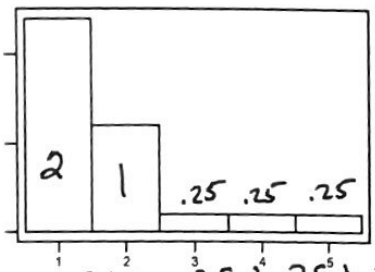
1+1+1+3+2+4+2+1+1+5 = 21

Center? 45.0

Shape? Skewed left

Spread? 55 − 32.5 = 22.5

Graph E



$\frac{3.75}{2} = 1.875$

2+1+.25+.25+.25 = 3.75

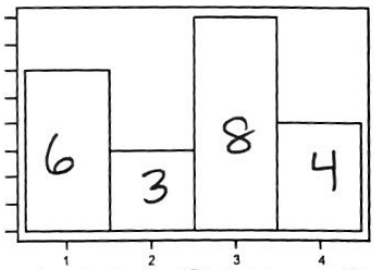
Center? 1

Shape? Skewed right

Spread? 5 − 1 = 4

Clusters or gaps? Where? Cluster at 1. No Gaps

Graph F



$\frac{21}{2} = 10.5$

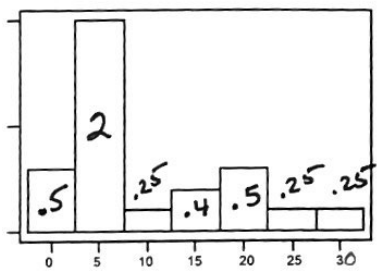
6+3+8+4 = 21

Center? 3

Shape? Indistinct

Spread? 4 − 1 = 3

Graph G



$\frac{4.15}{2} = 2.075$

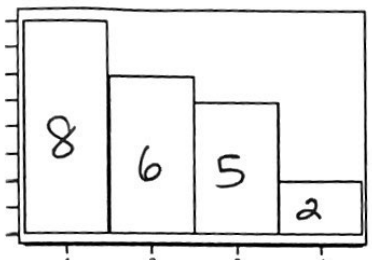
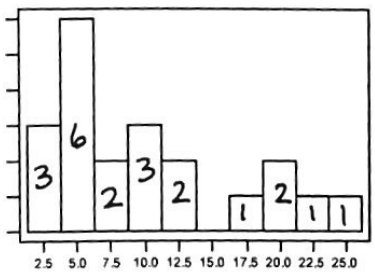
.5+2+.25+.4+.5+.25+.25 = 4.15

Center? 5

Shape? Skewed right

Spread? 30 − 0 = 30

AP Statistics Prerequisite Packet

Graph H  $\frac{21}{2} = 10.5$ $8+6+5+2=21$	**Center?** 2 **Shape?** skewed right **Spread?** $4-1=3$
Graph I  $\frac{21}{2} = 10.5$ $3+6+2+3+2+1+2+1+1=21$	**Center?** 7.5 **Shape?** Skewed right **Spread?** $25-2.5=22.5$ **Clusters or gaps? Where?** No clusters. Gap at 15.0