

# AP STATISTICS SUMMER FUNCS



# 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, 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)

	<b>ANSWER</b>	<b>TYPE</b>
1. In what grade did you take Algebra 1?	_____	_____
2. How many CDs 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? ( <b>all</b> , whether you live with them or not)	_____	_____
7. How many cousins do you have?	_____	_____
8. How tall are you ( <b>in inches</b> )?	_____	_____
9. How many AP classes will you be taking <b>THIS</b> 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**

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

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

**Examples**

$$\text{Data: } 4, 36, 10, 22, 9 = 4 \quad 9 \quad \underline{10} \quad 22 \quad 36 \longrightarrow \text{Median} = 10$$

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

---

## Measures of Spread:

**Range:** Maximum value - Minimum value

**Example**

$$\text{Data: } 4, 36, 10, 22, 9 = 4 \quad 9 \quad 10 \quad 22 \quad 36$$

$$\text{Range} = \text{Max.} - \text{Min.} = 36 - 4 = 32$$

**Interquartile Range (IQR):** The difference between the 75<sup>th</sup> percentile ( $Q_3$ ) and the 25<sup>th</sup> percentile ( $Q_1$ ). This is  $Q_3 - Q_1$ .  $Q_1$  is the median of the lower half of the data and  $Q_3$  is the median of the upper half. In neither case is the median of the data included in these calculations.



To find the median, sort the data in the lists: **STAT**® **2** ® **L**<sub>1</sub> The median is exactly in the middle between the 13<sup>th</sup> and the 14<sup>th</sup> value.

Mean\_\_\_\_\_ Median\_\_\_\_\_

Are they the same? \_\_\_\_\_

If not, which is larger? \_\_\_\_\_ Do you know why?

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

Mean\_\_\_\_\_ Median\_\_\_\_\_

Are they the same? \_\_\_\_\_

If not, which is larger? \_\_\_\_\_ Again, do you know why?

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 about the same? \_\_\_\_\_ If no, what are some reasons that might cause this difference?

6. Find  $Q_1$  and  $Q_3$  for the Dad data.  $Q_1$ \_\_\_\_\_  $Q_3$ \_\_\_\_\_

7. Find  $Q_1$  and  $Q_3$  for the Mom data.  $Q_1$ \_\_\_\_\_  $Q_3$ \_\_\_\_\_

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    Q<sub>1</sub>    Median    Q<sub>3</sub>    Maximum

Write the five number summary for the Dad data: \_\_\_\_\_

Write the five number summary for the Mom data: \_\_\_\_\_

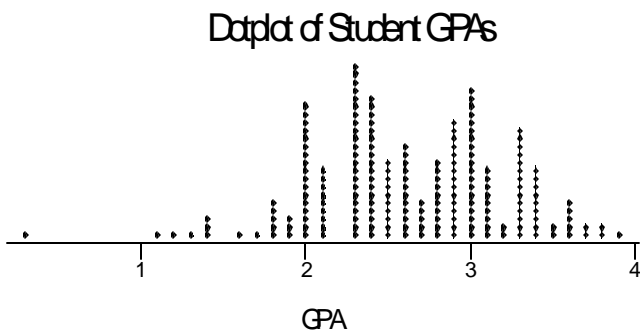
9. Now calculate the IQR for each of the two sets of data.

Dad \_\_\_\_\_

Mom \_\_\_\_\_

## Graphical Displays of Univariate (one variable) Data

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



**To make a Dotplot:**

1. Draw and label a number line so that all the values in your dataset will fit.
2. Graph each of the data values with a dot.  
Be sure to line the dots up vertically as well as horizontally so that you can really see the shape of the graph.

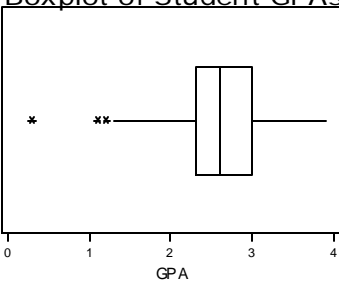
## Stemplot of Student GPAs

1		23
1		444
1		67
1		88888999
2		00000000000000000111111111
2		33333333333333333333333333
2		44444444444444444444444555555555
2		666666666666677777
2		8888888888999999999999999999
3		000000000000000000000111111111
3		223333333333333333
3		4444444445
3		6666677
3		889

### To make a Stemplot:

1. Put the data in ascending order.
2. Use only the last digit of the number as a leaf (see the numbers to the right of the line – each digit is the last digit of a larger number).
3. Use one, two, or more digits as the stem. (Sometimes, you can truncate data when there are too many digits in each data value – i.e. the number 20, 578 would become 20 | 5, where the “20” is in thousands. Note that this is **different** from rounding.)
4. Place the “stem” digit(s) to the left of the line and the leaf digit to the right of the line. Do this for each data value. You should then arrange the “leaves” in ascending order.
5. Sometimes, there are many numbers with the same “stem.” In this situation it might be useful to break the numbers with the same stem into either two distinct groups (each on a separate line; say, “leaves” from 0 – 4 on the first line and 5 – 9 on the second.) or into five distinct groups as is shown in the graph to the right. Here, the first line for each stem contains all the 0 – 1 leaves, the next line contains the 2 – 3 leaves and so on. This technique is called “splitting the stems.” It is useful in some cases in order to show the shape of the data more clearly.

Boxplot of Student GPAs

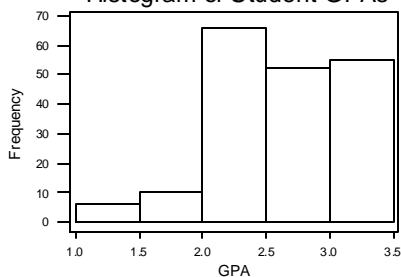


### To make a Boxplot:

1. **Draw and label a number line** that includes the minimum and the maximum values for the set of data.
2. Calculate the five-number summary and make a dot for each of these summary numbers above the number line.
3. Draw a line between the 1<sup>st</sup> and 2<sup>nd</sup> dot, showing the “lower quartile”; and then draw a line from the 4<sup>th</sup> to the 5<sup>th</sup> dot to show the “upper quartile.” These are commonly called the “whiskers.”
4. Draw a rectangular box from the 2<sup>nd</sup> to the 4<sup>th</sup> dot and draw a line through the box on the middle dot – the median.

**NOTE:** In AP Statistics, a “modified boxplot” is used. This shows any “outliers.” An outlier is a data point that does not fit the pattern of the rest of the data. When your calculator or computer software graphs a modified boxplot, an algorithm is used to determine what it takes to “not fit the pattern of the rest of the data.” This algorithm is:  $1.5 (IQR)$  away from the “box” part of the graph. (above and below the box). These outliers are shown with dots or stars, or any other small symbol.

Histogram of Student GPAs



### To make a histogram:

1. Put the data into ascending order.
2. Decide upon evenly spaced intervals into which to divide the set of data (such as 0, 10, 20, 30, etc.) and then count the number of values that fall within each interval. This number is called the “frequency.” If you divide each of these frequencies by the size of the data set,  $n$ , making percents, then you have what are called “relative frequencies.”
3. Draw and **label** a 1<sup>st</sup> quadrant graph using scales appropriate for the data. Be sure to include a title for the x- and for the y-axes.
4. Graph the frequencies that you calculated in step 2.

**Categorical Data:** Bar Graph  
Circle Graph

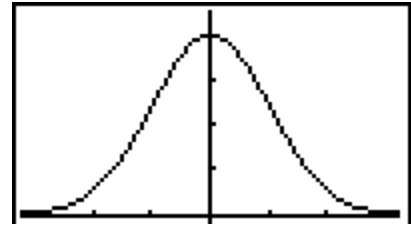
I'm assuming that you already know how to make these two types of graphs.

---

### Assessing the *Shape* of a Graph

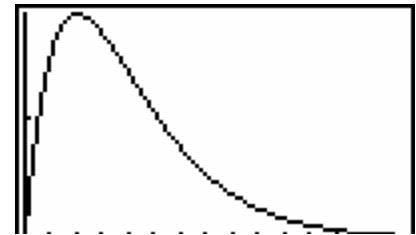
There are two basic shapes that we will examine: ***Symmetric*** and ***Skewed***.

***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.)



Symmetric

***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.



Skewed Right

### 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 Shape, Center, Spread, Clusters and gaps, Otliers: **SCSCO**. 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, 1990 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.)

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?

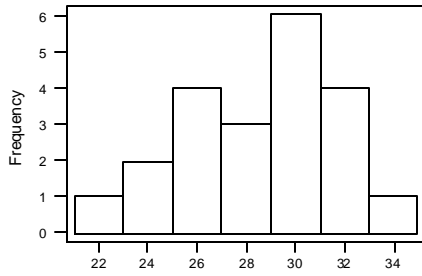
2. The following data is taken from the *Statistical Abstract of the United States* (112<sup>th</sup> Edition). These are the ages of drivers arrested for DUI from a random sample of size 50. Make a stemplot to show the distribution of this age data.

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

- What is the shape of this graph? \_\_\_\_\_
  - Using your stemplot, find the median of this data. \_\_\_\_\_
  - Which data display is better - a boxplot or a stemplot? \_\_\_\_\_  
Why?
3. For the following graphs, find the shape, center (just do the median), 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.

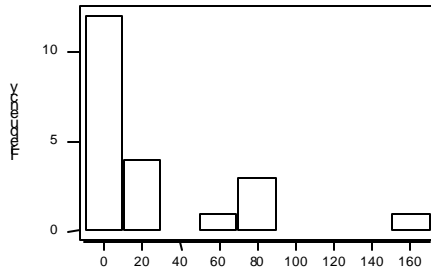
Starting with the first bar count 1 + 2 + 4 + 3 + 6... So the median is in the bar that contains the 10.5 value (bigger than 10 anyway). That's 30. So, the median is 30. To find a rough estimate of the mean, take the frequency for each bar and multiply it by the value along the x-axis for that bar. Add these up for all the bars and then divide by 21. You get the mean = 28.571.)

A



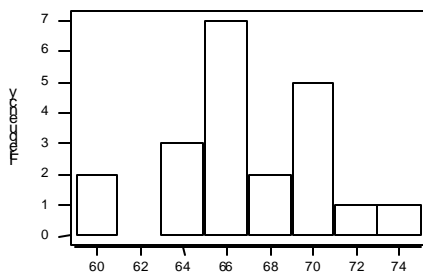
Shape \_\_\_\_\_  
 Center \_\_\_\_\_  
 Spread \_\_\_\_\_  
 Clusters, Gaps? \_\_\_\_\_ Where?  
 Outliers? \_\_\_\_\_ Where?

B



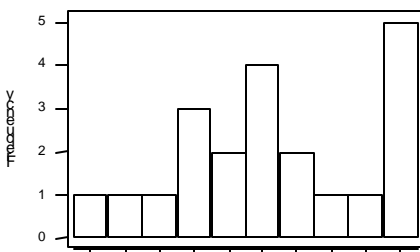
Shape \_\_\_\_\_  
 Center \_\_\_\_\_  
 Spread \_\_\_\_\_  
 Clusters, Gaps? \_\_\_\_\_ Where?  
 Outliers? \_\_\_\_\_ Where?

C



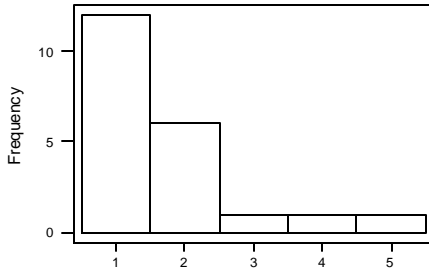
Shape \_\_\_\_\_  
 Center \_\_\_\_\_  
 Spread \_\_\_\_\_  
 Clusters, Gaps? \_\_\_\_\_ Where?  
 Outliers? \_\_\_\_\_ Where?

D



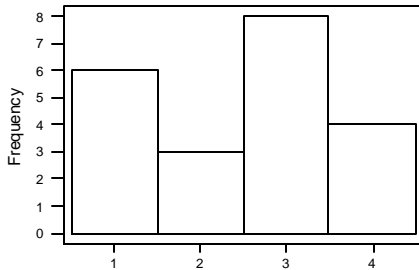
Shape \_\_\_\_\_  
 Center \_\_\_\_\_  
 Spread \_\_\_\_\_  
 Clusters, Gaps? \_\_\_\_\_ Where?  
 Outliers? \_\_\_\_\_ Where?

E



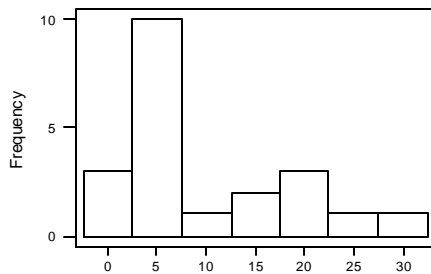
Shape \_\_\_\_\_  
Center \_\_\_\_\_  
Spread \_\_\_\_\_  
Clusters, Gaps? \_\_\_\_\_ Where?  
Outliers? \_\_\_\_\_ Where?

F



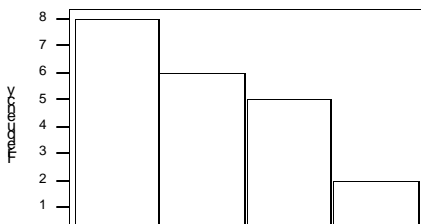
Shape \_\_\_\_\_  
Center \_\_\_\_\_  
Spread \_\_\_\_\_  
Clusters, Gaps? \_\_\_\_\_ Where?  
Outliers? \_\_\_\_\_ Where?

G



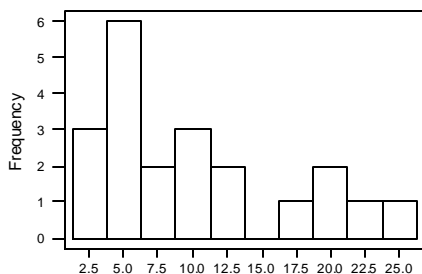
Shape \_\_\_\_\_  
Center \_\_\_\_\_  
Spread \_\_\_\_\_  
Clusters, Gaps? \_\_\_\_\_ Where?  
Outliers? \_\_\_\_\_ Where?

H



Shape \_\_\_\_\_  
Center \_\_\_\_\_  
Spread \_\_\_\_\_  
Clusters, Gaps? \_\_\_\_\_ Where?  
Outliers? \_\_\_\_\_ Where?

I



Shape \_\_\_\_\_  
Center \_\_\_\_\_  
Spread \_\_\_\_\_  
Clusters, Gaps? \_\_\_\_\_ Where?  
  
Outliers? \_\_\_\_\_ Where?

4. Use the following list of variables to identify which of the graphs in Question 4 **could** be a graphical display of the answers for a typical class of students. Write the letter of the correct graph in the blank provided. There are more variables than there are graphs, so don't worry if you have extras.

### Variables

- Grade when a student takes Algebra 1 \_\_\_\_\_
- Average # of CD's you own \_\_\_\_\_
- Age of your father when you were born \_\_\_\_\_
- Age of your mother when you were born \_\_\_\_\_
- Age of your stat teacher this next year \_\_\_\_\_(guess, even if you don't know me!!)
- # of siblings you have \_\_\_\_\_
- # of cousins you have \_\_\_\_\_
- Your height (in inches) \_\_\_\_\_
- # of AP classes you will be taking this next year \_\_\_\_\_
- How long you have lived in this area \_\_\_\_\_
- How far away from school you live (in miles) \_\_\_\_\_
- Amount of change in your pocket on the first day of school \_\_\_\_\_