

A. P. STATISTICS

SUMMER PACKET

Welcome to *Advanced Placement Statistics*! Every student will be better prepared for any subsequent area of endeavor upon the successful completion of this course. That is the true beauty of Statistics.

Many students take A. P. Statistics because they have been convinced that it is an “easy” course. While that is true regarding one aspect of Statistics, it is completely false in other regards.

The math required to succeed in this course is relatively easy; however, the logical thought required to succeed can often be difficult. Good reading comprehension is also a necessary skill for success in Statistics, since practically every problem that will be encountered will be a “word problem”. Finally, once the statistical truth has hopefully been divined, the “answer” must be conveyed to the appropriate audience in a clear and concise fashion. In other words, while success in A. P. Statistics requires only decent math abilities, it also requires outstanding skills in logic, reading and writing.

Therefore, to aid in the development of the skills necessary for success, this packet must be printed and completed before the first day of school. Every effort must be made to complete the entire packet, which will be turned in for a grade. Afterwards, the packet will be reviewed. Part of the packet will include instructions for the use of a *Texas Instruments* TI-83 or TI-84; however, any of the problems can be completed with a basic calculator.

Once again, welcome to Statistics! We can all look forward to a fulfilling year together.

P. L. Migli (pmigli@dadeschools.net)

Types of Data

Quantitative (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 “mean” or “range” (largest value – smallest value). In contrast, it makes no sense to try to find the mean color of shirts because shirt color is not an example of a quantitative variable.

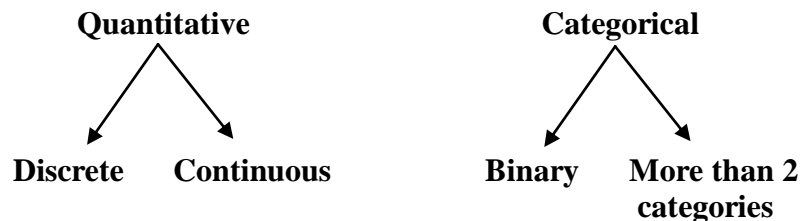
Some quantitative variables take on **discrete** values, such as shoe size (6, 6½, 7, 7½...) or the number of soup cans collected by a school. Other quantitative variables take on **continuous** values, such as height (60 in., 72.99999923 in., 64.039 in., etc.) or how much water it takes to fill up your bathtub (73.296 gal. or 185.4 gal. or 99 gal., etc.)

Categorical (label) Data

These are data that take on labels that describe a characteristic of something, such as the color of shirts. These labels are “categories” of a population, such as *male* or *female* for gender. Gender is an example of a **binary** variable, which only has two possible outcomes. Categorical variables that have more than two outcomes, such as hair color, brand of automobile, religion, et cetera, are **not binary**.

Note that numerical labels (such as numbers on racing cars) are categorical variables and *not* quantitative variables.

Two types of variables:



Exercises: Identify the following variables as *quantitative* or *categorical* (Q or C). If the variable is quantitative, identify it as *discrete* or *continuous* (D or C).

	Q* or C	* - D or C
1. Your favorite football player's jersey number.	_____	_____
2. The number of DVDs that you own.	_____	_____
3. Your father's age (in whole years) at your birth.	_____	_____
4. Your mother's actual age when you were born.	_____	_____
5. An integer between 1 and 20, inclusive.	_____	_____
6. The number of siblings that you have.	_____	_____
7. Your actual weight.	_____	_____
8. Your height (to the nearest inch).	_____	_____
9. The number of AP classes that you have taken.	_____	_____
10. Your boss' gender.	_____	_____
11. The location of your next meal. (1 = home, 2 = restaurant, 3 = other)	_____	_____
12. Your favorite season of the year.	_____	_____
13. The great-circle distance from home to school.	_____	_____
14. Whether a person has triskaidekaphobia or not.	_____	_____

Numerical Descriptions of Quantitative Data:

Measures of Center

1) **Mean:** The sum of quantitative data values divided by the number (n) of observations.

Example

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

Measures of Center (continued)

2) **Median:** The middle element of a numerically sorted set of data (or the mean of the middle two elements).

Examples

Data: 4, 36, 10, 22, 9 \longrightarrow 4 9 10 22 36 \longrightarrow Median = 10

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

3) **Mode:** The most frequently occurring observation (quantitative *or* categorical).

Measures of Spread

1) **Range:** Maximum value – Minimum value

Example

Data: 4, 36, 10, 22, 9 \longrightarrow 4 9 10 22 36

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

2) **Interquartile Range (IQR):** The difference between the 75th percentile (Q_3) and the 25th 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.

The IQR contains the middle 50% of the data. Each quartile contains 25% of the data.

Examples

1. Data: 4, 36, 10, 22, 9 \longrightarrow 4 \uparrow 9 10 22 \uparrow 36
 $Q_1 = 6.5$ $Q_3 = 29$

$$\text{IQR} = 29 - 6.5 = 22.5$$

2. Data: 4 9 10 | 22 36 43
 \uparrow \uparrow
 Q_1 Q_3

$$\text{IQR} = 36 - 9 = 27$$

Exercises:

Last year students collected data on the ages of their parents when the students were born. Here is the data:

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

Now type the data into your calculator using the list function: **STAT** → **ENTER** → type the Dads' ages into L₁. If you make a mistake, you can highlight the error and hit **DELETE**. If you forget an item, you can go to the line below where it is supposed to be and press **2nd DEL** to insert it. Do the same for the Mom data, but put into L₂.

NOTE: If the lists you are using already have numbers in them before you start, you can clear them this way: Arrow up (↑) to highlight L₁. Press **CLEAR**, then the down arrow (↓). *Do not* press **DEL**. If you do by mistake, then press **STAT** → **SET UP EDITOR**.

1. Find the mean and the median for the “Dad” data. To find the mean using your calculator, go to **2nd STAT** → **MATH** → **5** and then type in L₁ by typing **2nd → 1**. This will add all the values in the list. Then divide by 26 to get the mean.

To find the median, sort the data in the lists: **STAT** → **2** → **L₁** → **ENTER**. The median is the mean of the 13th and the 14th values.

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 about the same? _____ If not, 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_1 Median Q_3 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_____

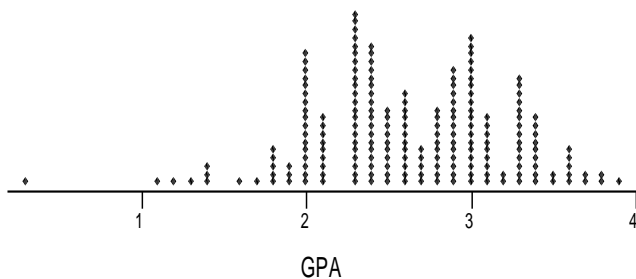
Mom_____

Graphical Displays of Univariate (one variable) Data

Quantitative Data:

- 1) Dotplot
- 2) Boxplot (Box and Whiskers)
- 3) Stemplot (Stem and Leaf)
- 4) Histogram

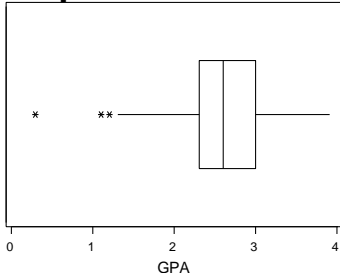
Dotplot of Student GPAs



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.

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 1st and 2nd dot, showing the “lower quartile”; and then draw a line from the 4th to the 5th dot to show the “upper quartile.” These are commonly called the “whiskers.”
4. Draw a rectangular box from the 2nd to the 4th dot and draw a line through the box on the middle dot – the median.

NOTE: In AP Statistics, a “modified boxplot” is often 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 (below Q_1 and above Q_3). These outliers are shown with dots, stars, or another small symbol.

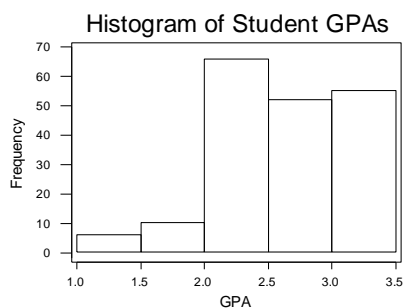
Stemplot of Student GPAs

1	23
1	444
1	67
1	88888999
2	000000000000000000111111111
2	333333333333333333333333333
2	444444444444444444444444555555555
2	66666666666677777
2	888888888999999999999999999
3	00000000000000000000111111111
3	22333333333333333
3	44444444455
3	6666677
3	889

Key: 2 | 5 Rep. a 2.5 GPA

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 differs 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 left. 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.



To make a histogram:

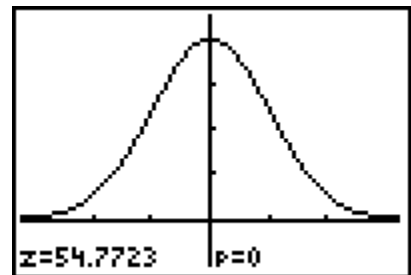
1. Put the data into ascending order.
2. Decide upon evenly spaced intervals into which to divide the set of data (such as 0, 10, 20, 30, etc.) and then count the number of values that fall within each interval. This number is called the “frequency.” If you divide each of these frequencies by the size of the data set, n , making percents, then you have what are called “relative frequencies.”
3. Draw and label a 1st 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:**
- 1) Bar Graph – Similar to a histogram, but bars do not touch and order is arbitrary.
 - 2) Pie Chart (circle graph) – Size of “slice” represents fraction of total.
-

Assessing the *Shape* of a Graph

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

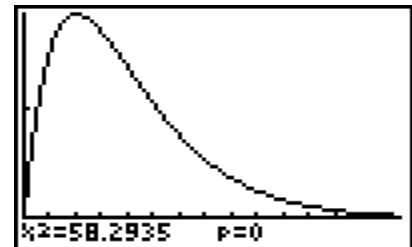
Symmetric: One can tell if a graph is symmetric if a vertical line in the “center” divides the graph into two fairly congruent shapes. (A graph does *not* have to be “bell-shaped” to be considered symmetric.)



Symmetric

Mean \approx 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 peak. The longer tail “points” in the direction of the “skewness”.



Skewed Right

Relationship between Mean and Median in a skewed distribution:

Skewed *Left*, the mean is *Less*.

Skewed *Right*, the mean is *More*.

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. This analysis is important for choosing appropriate procedures to answer questions about the data and the population from which it was drawn.

The features that are the most important are Shape, Center, Spread, Clusters/gaps, and Outliers: **SCSCO**. Most of these can only be seen in a graph. However, sometimes the shape is indistinct or difficult to discern. In those instances (usually because of a very small set of data), it is appropriate to label the shape as “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 "Creamy" boxplot. Then, above the "Creamy" boxplot, construct the "Crunchy" 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: Crunchy _____ Creamy _____
- b. Find the median for: Crunchy _____ Creamy _____
- c. Looking at your boxplots and comparing the medians, which type of peanut butter do consumers seem to prefer?

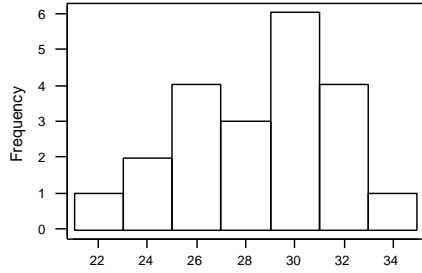
2. The following data is taken from the *Statistical Abstract of the United States* (112th Edition). These are the ages of a random sample of 50 drivers arrested for DUI. Make a stemplot to show the distribution of this age data.

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? _____
 - Explain why you chose your answer for part c.
-
3. For the following graphs, find the *shape*, *center* (just the **median**), and *spread* (just the **range**). If there are 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 $1 + 2 + 4 + 3 + 6 \dots$ 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.)

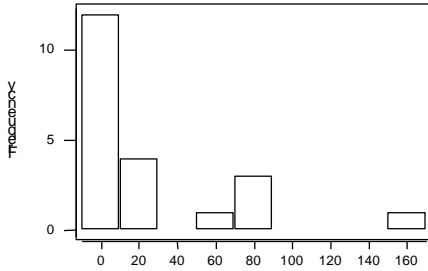
A



Shape _____
 Center _____
 Spread _____
 Clusters, Gaps? ____ Where?

Outliers? _____ Where?

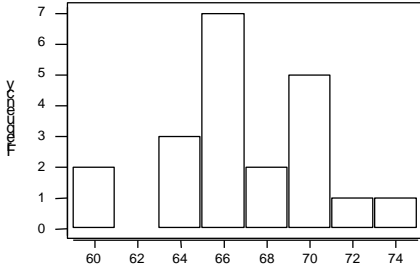
B



Shape _____
 Center _____
 Spread _____
 Clusters, Gaps? ____ Where?

Outliers? _____ Where?

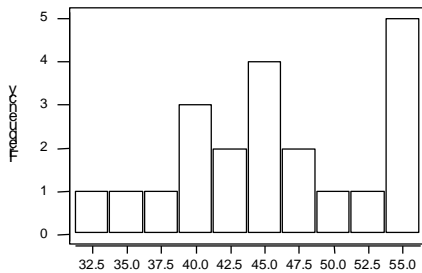
C



Shape _____
 Center _____
 Spread _____
 Clusters, Gaps? ____ Where?

Outliers? _____ Where?

D



Shape _____
 Center _____
 Spread _____
 Clusters, Gaps? ____ Where?

Outliers? _____ Where?

Complete the packet to the best of your abilities. Once again, welcome to *A.P. Statistics!*

P. L. Migli (pmigli@dadeschools.net)