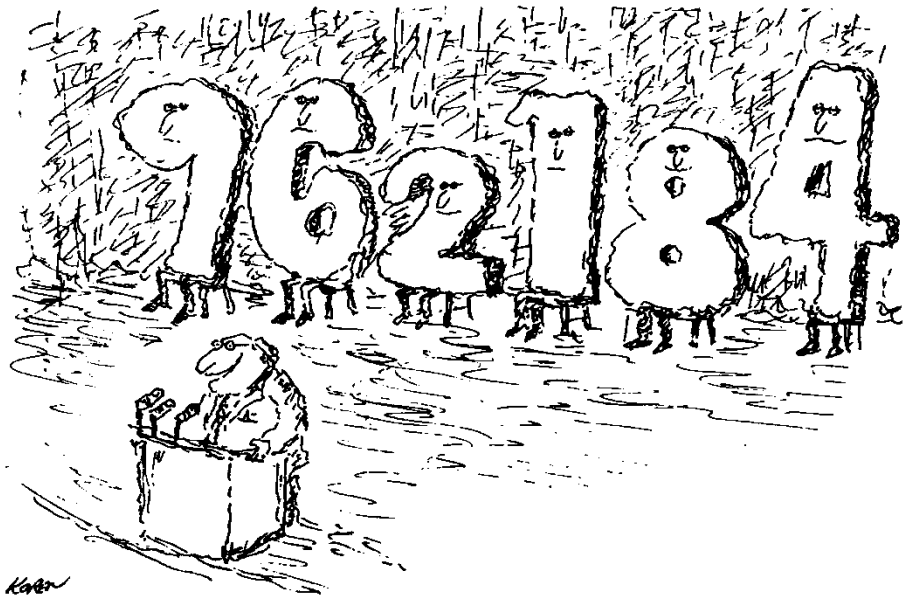
Statistics Review



"Tonight, we're going to let the statistics speak for themselves."

TJHSST

Summer Semester

Review & Self-Study for
Research Stat 1 (Summer School)

Discussion 1: Measures of Central Tendency

Measures of central tendency are numbers that are most representative of the data contained in the set. There are three measures of central tendency: mean, median, and mode. The **mean** is the arithmetic average of a set of data. The mean can be identified by the symbol $\overline{x}$ when we are talking about a sample or the symbol μ (the Greek letter *mu)* when we are talking about an entire population. The **median** is the middle number of the set of data once it has been ordered from least to greatest. If there are an odd number of pieces of data, the median is the middle number; if there is an even number of pieces of data, the median is the average of the two middle numbers. The **mode** is the data point that occurs most often. There may be more than one mode; there may be no modes.

There are advantages and disadvantages for each of these three measures of central tendency. It is important to choose the most representative measure, so you must understand the pros and cons of choosing each:

Mode	
Advantages	**Disadvantages**
Gives most frequently occurring measure	May not be central
Easy to find once data is put in order	May not be unique
	May change significantly with addition of new scores
	May not exist
Median	
Advantages	**Disadvantages**
Gives middle score	May not be part of data set
Not easily influenced by extremes	Data must be arrayed to identify it
Mean	
Advantages	**Disadvantages**
Most people are familiar with it	Can be greatly influenced by extremes
Easy to define algebraically	May not be part of data set
Gives information about total of scores	
Used in other statistical calculations	

Whichever measure of central tendency you decide to use, it is important to look at the reason you are choosing it. If you are a buyer for a clothes department, finding the mean or median size dress that is sold will not be helpful. In this instance, using the mode (the dress size most often sold) will be most appropriate. If you must identify the upper and lower half of a group of qualifying scores for a race, the median is the most appropriate. The numbers are used as standards of comparison and simple indicators of a population. Choose them in appropriate ways.

The graphing calculator will be used for the following example. You will use this data again in Discussion 4. Save it!

Example: The following are the weights in pounds of children in a fourth grade class:

64	71	57	67	74	65	59	62	60	72	84	60	68
72	91	55	69	71	69	75	59	60	70	76	62	

We will use the graphing calculator to enter data (1-variable data) into the calculator (the instructions are for the TI-84), and then use the calculator as an aid in determining the mean, median and mode.

To Enter Data and Calculate with 1-variable data:

(1) **Press: [STAT] <ENTER>**

This displays the **<EDIT>** menu that allows for data entry. The screen should read: **L1 L2 L3**

Note: To CLEAR data to make room for new entries, there are several options,

(a) Press: **[STAT] [4] [2nd] [1] <ENTER>**

The screen displays **DONE** when the data has been cleared and you are ready to begin again.

(b) Press: **[STAT] <ENTER> < ▲ > [CLEAR] <ENTER>**

(2) Enter Data:

Under L1, type in the first value, press **<ENTER>.** Enter the second value and continue until all the data is entered.

(3) To Calculate Mean, Median, Mode:

(a) To find the MEAN:

Press: **[STAT]< ► > .**The screen should now highlight the CALC menu.

Press: **<ENTER>**. This selects the **1-Var Stats** option.
Press: **[2nd] [1]** for L1 **<ENTER>**
You will obtain a display of data where $\overline{x}$ = MEAN.

(b) To find the MEDIAN, move the cursor down until you see **MED =.**
The MEDIAN can also be found by arranging the data in ascending order. To do so, press **[STAT]**
Choose **SORT A(** and press **<ENTER>**
Press: **[2nd] [1] <ENTER>**
The data is now arranged in ascending order in L1.
To view the sorted data, press: **[STAT] <ENTER>**
Move the cursor to the middle term to read the median (in this case, the thirteenth term).
If there is an even number of terms, the median is the average of the two middle terms.

(c) To find the MODE, scan the ordered data and find which data point occurs most often.

Using the previous results from above, answer the following questions:

1. What is the mean weight of these fourth graders?_____

   What is the median weight?_____ What is the mode?_____

2. Which do you think is the most representative number for these weights, the mean, median or mode? Explain.

Exercises for Discussion 1

1. If you wanted to estimate the total amount spent on junk food for a week by your class, would you prefer to know the daily mean, median or mode amount spent by the class on one day? Explain.

2. If you wanted to know if you read more or fewer books per month than most people in the class, would you prefer to know the mean, median or mode? Explain.

3. The Reston Town Center skating rink is ordering new skates. Which would be more useful to know, the mode, mean or median skate size? Explain.

4. You want to know which Virginia county has a large portion of people with low incomes. Which is most helpful to know for each county: the mean, mode or median income? Explain.

5. (Taken from Statistics and Information Organization: Math Resource Program by University of Oregon) A manufacturing company boasts that they pay an average salary of $30,000 to their employees. Study the chart below and answer the following questions:

Type of Job	Salary	Number Employed
President	$183,000	1
Vice-President	$90,000	2
Plant-Manager	$50,000	3
Foreman	$30,000	12
Skilled Operator	$22,000	21
Unskilled Operator	$18,000	36

a) Is the company telling the truth? To help you decide, find each of the following:

mean salary_____ median salary_____ mode salary_____

Note: In this problem, all but the first salary must be entered into the list a multiple number of times, i.e., 90,000 must be entered twice, 50,000 three times, 30,000 twelve times, etc. We can use **L2** in order to enter the frequency.

Using the calculator to find the 1-Variable Statistics:

Enter the data:

Press: **[STAT] <ENTER>**

In **L1**, enter the salaries.

In **L2**, enter the number employed.

To Calculate using the data

Press: **[STAT] < ► >** (to CALC) **<ENTER> to display** 1-Var Statistics.

Press: **[2nd] 1 , [2nd] 2** to select L1, L2. Press: **<ENTER>**

1-Variable Statistics will be displayed. Scroll down to see the five number summary.

b) Which do you think is a more representative number for these salaries, the mean, median or mode? Explain.

Discussion #2: Measures of Spread or Variability

Data points may cluster about the mean or be spread out. A measure of variability or dispersion is a single number that represents the spread or amount of dispersion in a set of data. The four most common measures of variability are range, interquartile range, variance, and standard deviation. Mean absolute deviation (MAD) is another measure of variability which is used to introduce and understand standard deviation.

The **range** of a set of numbers is the difference between the largest and smallest numbers in the set. For example, the range of salaries in Discussion 2, Exercise 5 is 183,000 – 18,000 = $165,000.

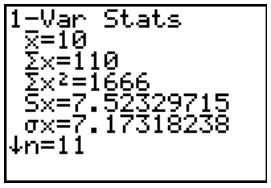
To find the **interquartile range**, array the numbers in increasing order. Find the median. This divides the data into two halves. If you have an odd number of values, leave the median out of the upper and lower half. Now find the median of each half of the data. The median of the lower half is the **first quartile** and the median of the upper half is the **third quartile**. The difference between the first and third quartiles is the **interquartile range** (IQR).

A deviation measures the distance between the value of a single observation and the mean of the data set and is calculated using the expression $(x - \mu)$. The deviation can be positive, negative, or zero. The preferred method is to use the squares of the deviations to calculate the variance which is then used to calculate the standard deviation. **Variance** is the **sum** of the squared deviations, divided by the number of values in the population; the symbol for population variance is σ^2. The **standard deviation** of a population is the square root of its variance so the symbol used is σ. The formula for standard deviation is $\sigma = \sqrt{\sigma^2}$.

Example 1: In today's world, calculators and computers make the computation of variance and standard deviation simple. Using the population {0, 3, 4, 4, 6, 9, 12, 13, 15, 21, 23} proceed through the following steps to familiarize yourself with the use of the calculator:

Enter the data as before, then:

Press: **[STAT]** < ▸ > (to CALC). **<ENTER>** **1-Var Stats L1**



This means that for this population of $n = 11$ data points the standard deviation is approximately 7.2 and the variance is approximately 51.5.

When we have data for an entire population, we use σ and μ. When we only have a sample to work with, we use the symbol s (or s_x) for the sample standard deviation and $\overline{x}$ for the mean. For example, finding the average GPA of all of the students in the school and standard deviation of the GPA would be very time consuming. Instead, take a sample of 100 students and estimate the GPA and standard deviation based on these students. There will be an error due to the fact that you are only working with a portion of the population. s has a correction factor built in to try to minimize this error. As you work with data sets and use the calculator to find standard deviation, it is important to consider whether you are dealing with all of the data (population) or a sample, and to choose the correct standard deviation in each case.

Exercise for Discussion 2

In math class, we follow AP rounding rules. Round the answers to three decimal places unless told otherwise. In science, decimals are reported to the nearest significant figure which is determined by the accuracy of the measurements.

1. Here are three sets of test scores.

Class A:	77, 77, 77, 82, 85, 85, 86, 88, 90, 91, 92, 92, 93, 93
Class B:	75, 75, 76, 76, 77, 85, 87, 88, 92, 94, 94, 94, 98, 98
Class C:	56, 60, 76, 77, 85, 85, 87, 88, 91, 93, 94, 94, 96, 100

a) For each class, find the mean, median, range, IQR, variance and standard deviation.

b) Discuss the similarities and differences among the three classes.

Discussion #3: Stem-and-Leaf Plots, Dotplots, and Box –and-Whisker Plots

There are many ways of organizing data. One is the ordered array and the second is the stem-and-leaf diagram. An **ordered array** is a set of data arranged in ascending order. Your calculator can perform this task.

A **stem-and-leaf plot** is a set a visual display of a data set that separates each observation into two pieces: "a stem" and "a leaf." When the data consist primarily of two digit numbers, the natural choice is to make the tens digit the stem and the ones digit the leaf. It is used to rank order data and provide an indication of the shape of the distribution. The procedure for drawing stem-and-leaf plots is as follows:

(1) Identify the stems (leading digit(s)).
(2) Place leaves with corresponding stems.
(3) Order stem-and-leaf plot.

Another useful plot is a **dotplot**. To create a dotplot, draw a horizontal or vertical axis and scale it according to the values you have in your data set. Make a dot, *x*, or other mark to indicate each data point. The following graph is a Fathom dotplot of the number of seats in a sample of 36 commercial airplanes.

A **box-and-whisker plot** is a means for illustrating measures of central tendency and the range of data in an easy-to-read format. The procedure for drawing box-and-whisker plots is as follows:

1. Put the numbers in order.
2. Find the minimum, first quartile, median, third quartile, and maximum. These measures are called the five number summary.
3. Draw a number line with a scale.
4. Put a short vertical line above the number line where the minimum and maximum occur.
5. Put a short vertical segment where the first and third quartiles and median occur.
6. Draw a horizontal segment from the minimum to the first quartile. Draw a box between the two quartiles. Draw a horizontal segment from the third quartile to the maximum.

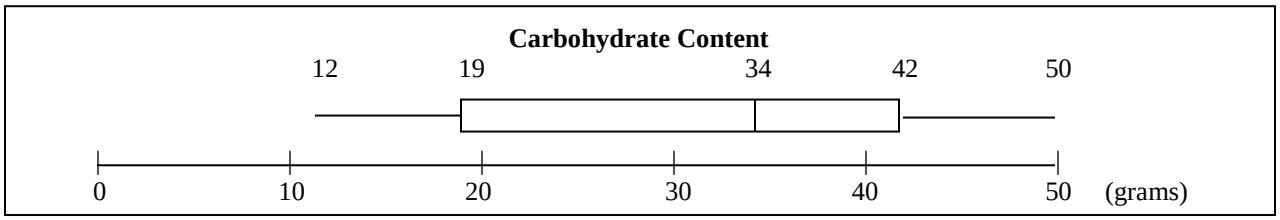
Example: Consider the stem-and-leaf plot for the carbohydrate content of fast foods (the full set of data is given in Exercise 1 below):

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

Key: 1|2 = 12 grams

The minimum is 12 g; the first quartile is 19 g; median is 34 g; the third quartile is 42 g; and the maximum is 50 g.

The following box-and-whisker plot illustrates this data:



Box-and-whisker plots can also be found on the TI-84. Enter the data as described before.

Press: **[2nd] [y=] <ENTER>** to display STAT PLOT Menu

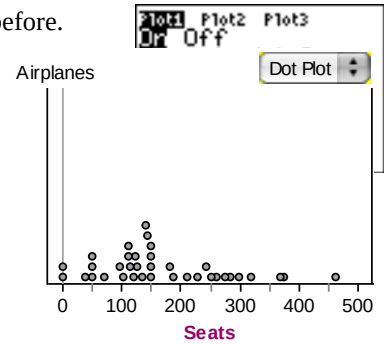
Move the cursor over **[On]** and press **<ENTER>**

Move the cursor down to the **Type** line.

Select box-and-whisker plot icon as shown at the right and Press: **<ENTER>**

Move the cursor down to **Xlist**.

Press: **<2nd> <STAT>**, highlight the list you want and press **<ENTER>.**



Press**:** **[ZOOM]** **<9>.** The calculator automatically chooses an appropriate scale.

Press **[TRACE]** and the cursor moves back and forth between the minimum (minX), first quartile (Q1), median (Med), third quartile (Q3), and maximum (maxX).

The information found in the example can be used to determine if there are any outliers in the data. An **outlier** is a data point that is very different from the other points and can consequently cause the mean of the data to be overly influenced in one direction. Recall the set of salaries in exercise 5 of discussion 2. The top 4 salaries are all considered outliers because they are far larger than the majority of salaries. The formula for determining an outlier follows:

1. Find the interquartile range or IQR (the difference between the third and first quartiles).
2. Multiply the IQR by 1.5.
3. Add this number to the third quartile; subtract this number from the first quartile. If a data point falls above or below the resulting values, it is an outlier.

Example: Consider the carbohydrate content information shown in the stem-and-leaf plot above. The interquartile range is (42 - 19) = 23. 1.5(23) = 34.5. 42 + 34.5 = 76.5. There are no data points above 76.5 so there are no high outliers. 16 - 34.5 = -18.5. There are no data points below -18.5, so there are no low outliers.

There is also a modified box-and-whisker plot that shows the outliers. Rather than drawing the whiskers to the extremes, we draw whiskers to the smallest or largest value in the data set that is not an outlier. The outliers are plotted as a point or asterisk. For this problem, the whiskers will still extend to 11 since that is the minimum value and there are no low outliers. But the right whisker would only extend to 30.5. The last value, the outlier, 31, is plotted as a point.

The TI-84 draws this modified box-and-whisker plot.

Select the first type of box-and-whisker plot in the Stat Plot Menu.

Replot the data to graph the modified box-and-whisker plot.

Exercises for Discussion 3

1. The following data is taken from Fast Food Facts (1994):

Item	Calories	Fat(g)	Carbohydrates(g)	Sodium(mg)
Burger King Whopper	570	31	46	870
McDonald's Big Mac	500	26	42	890
Wendy's Single	440	23	36	850
Subway 6" Roast Beef	345	12	42	1140
Hardee's Roast Beef	380	18	29	1230
Arby's Roast Beef	383	18	35	936
Hardee's Fisherman's Filet	480	21	50	1210
McDonald's Filet-O-Fish	370	18	38	730
Burger King Ocean Catch	450	28	33	760
Kentucky Fried Chik'n	284	18	15	865
McDonald's Chicken McNuggets	270	15	17	580
Wendy's Chicken Nuggets	280	20	12	600
Hardee's Rise ' N Shine	320	18	34	740
Egg McMuffin	280	11	28	710
Burger King Bacon Croissant	353	23	19	780

a) Make a stem-and-leaf plot of the fat content.

b) Use this plot to help find the mean, median and mode for fat content. Confirm your answers by using your calculator

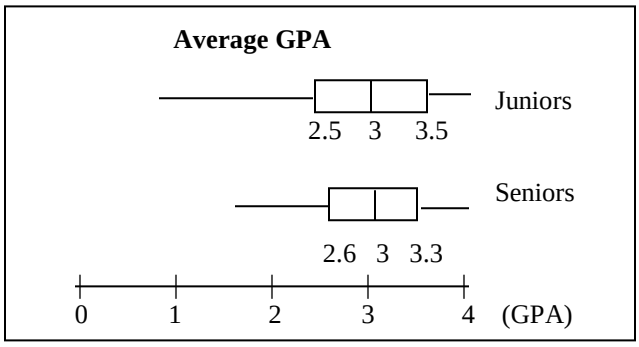
c) Make a stem-and-leaf plot of the carbohydrate content.

d) Use this plot to help find the mean, median and mode of the carbohydrate content.

e) Are there any outliers for the fat content? Justify your answer using the method above.

f) Draw a modified box-and-whisker plot for the fat content of fast foods, indicating any outliers that you find.

2. Refer to the plots below.



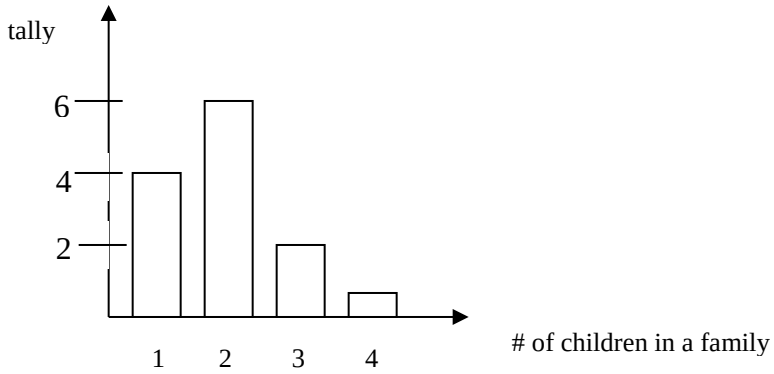
a) Which class had the higher median?

b) What was the interquartile range for each class?

c) Estimate each of the classes' best and worst grade point averages. Are there any outliers? Explain.

Discussion 4: Quantitative Data, Frequency Tables, and Histograms

Quantitative data is classified as either discrete or continuous, depending on the type of numerical values. **Discrete data** are separate measurements that most frequently occur as counts. For example: the number of softballs sold at Modell's on Tuesday, the number of students in your 6th period class, the number of children in a family. These quantitative values cannot overlap – you can have a family with 2 children or 3 children, but not with 2.5 children. The graph below illustrates the way that discrete quantitative data is pictured. Note that a **bar graph** is used just as it was for qualitative data because there is no overlap between the categories.



Continuous data may take on any value in an interval and can be subdivided into smaller and smaller increments depending on the precision of the measurement device. These include variables such as height, weight, time, distance, etc. This diagram of continuous data is called a **histogram,** not a bar graph.

When you work with continuous data, class intervals do not occur naturally. In order to determine a convenient interval, you should first find the range. The **range** is the difference between the smallest and largest values in the set. After examining the range and the number of pieces of data you have, you must determine how many intervals would be convenient. We generally follow Sturges' Rule to determine the number of intervals:

Number of Values in a Set	Appropriate Number of Intervals
10 to 100	4 to 8
100 to 1000	8 to 11
1000 to 10000	11 to 14

Follow the procedure outlined below in setting up the frequency table:

1. Establish the class limits--the smallest and largest values that would be placed in a given class. Decide on the lowest limit (make it convenient) and work from there. This limit is often included in the given class.
2. Tally your data.
3. Find the frequency--the number of data points in a class. The symbol *f* is often used to represent frequency.
4. Find the relative frequency--the fractional part of the data points in a class. If *n* data points are tallied, the relative frequency is *f/n*.

When plotting a histogram for continuous data, remember to label your axes and follow the procedure below when drawing the histogram on your TI-84 calculator (use the fourth grade weights data from the previous exercises to complete the following example):

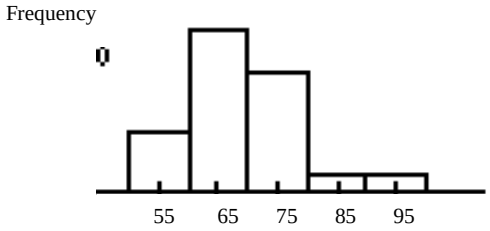
1. Clear any data in the calculator.
2. Press **<WINDOW>.**
3. Decide on the **Xmin** and **Xmax** that best suits the data (perhaps 50 and 100).
4. Decide on how wide you wish each bar to be. Your frequency table should aid in this decision. Enter this in **Xscl** (x-scale).
5. Enter 0 for **Ymin**. Decide on the **Ymax** and **Yscl** (perhaps 15 and 1, respectively).
6. Clear any previous graphs stored in your calculator **([Y=] [CLEAR])**
7. Press **[2nd] [Y=].** (This gives you **STAT PLOT**.) **<ENTER>.**
8. Turn on Plot 1. Press < ▼ > to Type and select histogram. Press <▼> to Xlist and select the location of your data (L1, L2, or named list, etc.)
9. Press [**GRAPH]**.

Generally you will be asked to reproduce this histogram on your own paper in order to get a frequency polygon.

Be sure to:

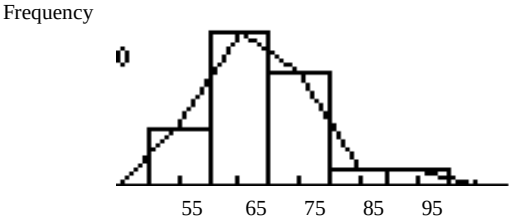
1. Label the axes, label the beginning and ending points of each bar, and label the scale on the y-axis.
2. Display the entire vertical axis (don't truncate).
3. Leave a space to the left and right of the histogram equal to the width of a bar. This is necessary to construct a frequency polygon.

Example: Use the data set of fourth grade weights from Discussion 1 and use the procedure outlined above to find a histogram. Add labels and scales and compare your results to the histogram shown here.



Weights of 4th Graders in lb.

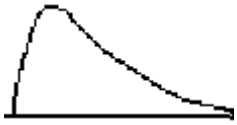
A frequency polygon is easy to plot using your histogram. Find the midpoint of the top of each bar of your histogram and of the initial and terminal empty intervals. Connect these points. The frequency polygon is shown below.



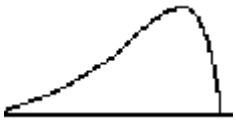
Weights of 4tn Graders in lb.

There are four commonly occurring shapes of smoothed frequency polygons:

Bell-Shaped or Symmetric



Right Skewed



Left Skewed

Bimodal

Exercises for Discussion 5

1. For the following examples, determine whether the data is continuous or discrete:
   a) Population in Fairfax County, Virginia

   b) Weight of newspapers collected for recycling on a single day at TJ

   c) Score on a math test

   d) GPA

2. For the following set of continuous data, determine an appropriate number of classes and set class limits. Then, set up a frequency table to organize the data.

A researcher suspects that the magnetite from a local region is below average. He collects 25 samples of magnetite from a local region and displays the data in a frequency table. Determine an appropriate number of classes, set class limits, find the frequency and relative frequency. Then, set up a frequency table to organize the data.

3.0	3.1	3.7	4.3	5.7	2.4	4.0
5.6	2.6	3.9	3.4	4.4	3.7	3.7
4.6	3.9	5.0	3.6	2.7	4.6	5.1
3.8	5.1	4.3	6.2			

Using the data from the frequency table, design a histogram and then a frequency polygon of the magnetite data. What shape does the frequency polygon appear to have? Explain.

Discussion 5: The Scatter Plot

Generally, more than three data points are collected in surveys and experiments. The data is generally collected as ordered pairs. A **scatter plot** is a graphic display of data points in a two-dimensional plane (xy plane). Each data point on a scatter plot represents two pieces of data for a single unit of observation. The most common plots are time plots--the time is always put along the horizontal axis.

The scatterplot is the first tool we have in determining the type of relationship that might exist between two variables. In this unit, we will be talking specifically about linear relationships. Two variables display an association (or relationship) if knowing the value of one variable is useful (to some degree) in predicting the value of the other variable.

Three aspects of the **association** between quantitative variables are direction, strength, and form. **Direction** refers to whether greater values of one variable end to occur with greater values of the other variable (positive association) or with smaller values of the other variable (negative association). The **strength** of the association indicates how closely the observations follow the relationship between the variables. In other words, the strength of the association reflects how accurately you could predict the value of one variable based on the value of the other variable. The **form** of the association can be linear, or it can follow some more complicated pattern (non-linear). (Rossman-Chance p. 571)

The graphing calculator may be used to find a scatter plot.

In order to find the scatter plot on the calculator, use the same procedure to enter the data as with univariate data:

First, clear old data from memory. Then, press **[STAT]** **<ENTER>**

Enter the first set of data under **L1** and the second set of data under **L2**.

Be sure that old graphs are cleared. Now, set the range by pressing **[WINDOW]**. Examine the data to determine appropriate entries.

To draw the scatter plot, press **[2nd] [y=] <ENTER>.**

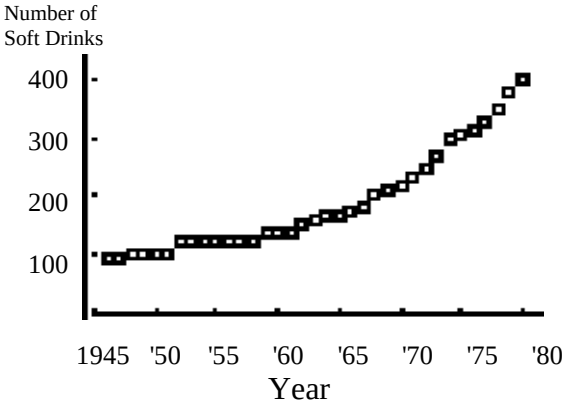
To turn on Plot 1, move the cursor to **<ON>** and press **<ENTER>.** Select scatter plot and press **<ENTER>**. Let Xlist be L1 and Ylist be L2. Select the type of mark you would like on your graph. Press **[GRAPH].**

Sketch the scatterplot on your paper. Label axes and scales clearly.

Exercises for Discussion 5: Put answers on separate sheet of paper.

1. Soft Drinks. The following is a plot over time showing how many 12 ounce soft drinks the average person in the U. S. drank each year from 1945 to 1980.

This problem was taken from the Exploring Data packet by James M. Landwehr and Ann E. Watkins prepared for the American Statistical Association and National Council of Teachers of Mathematics Joint Committee on the Curriculum in Statistics and Probability.



a) About how many soft drinks did the average person drink in 1950? in 1970?

b) About how many six-packs of soft drinks did the average person drink in 1980?

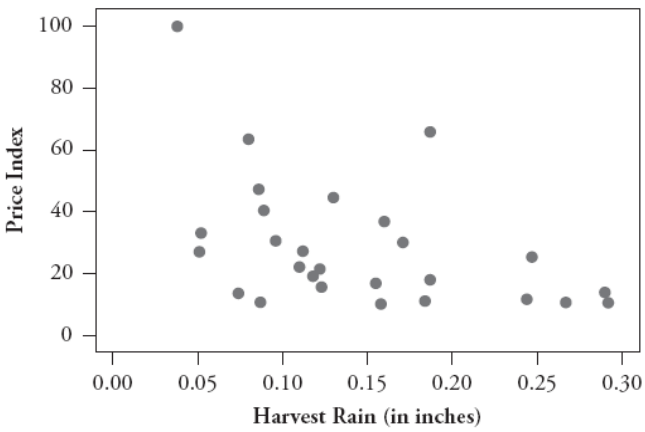
c) About how many soft drinks did the average person drink per week in 1950? in 1980?

d) If the trend in the plot continued, about how many 12 ounce soft drinks did the average person drink in the year 2000?

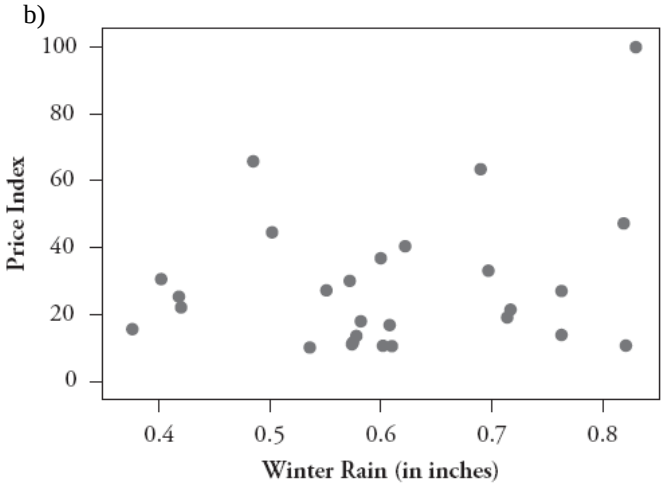
e) In what year did soft drink consumption start to "take off"? Can you think of any possible reason for this phenomenon?

2. Describe the direction, strength, and form of the association shown in each scatterplot.

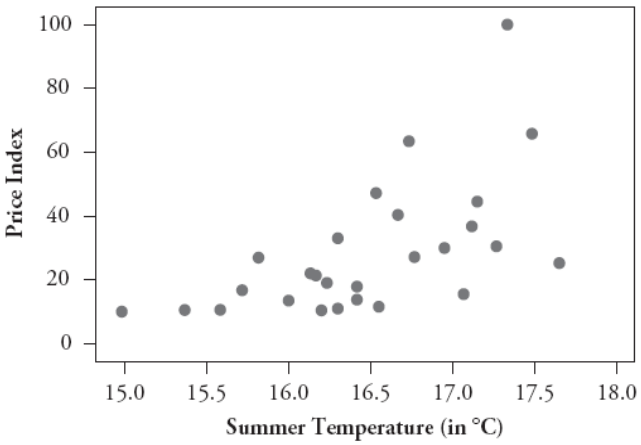
a)



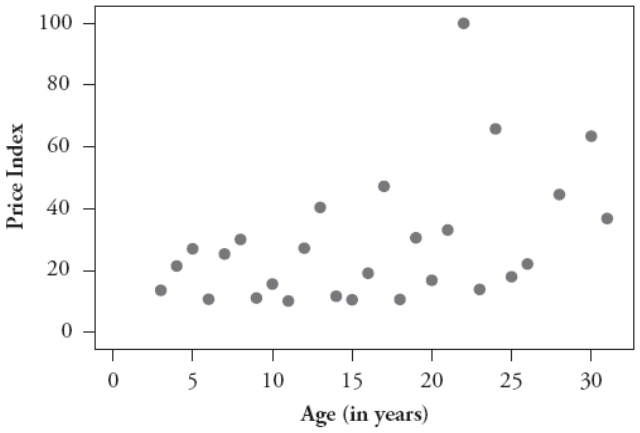
b)



c)



d)



Discussion 6: Linear Equations

Once bivariate data has been graphed on a scatter plot, the statistician wishes to determine (1) if there is some type of relationship between the two pieces of information from each observation, (2) how strong this relationship is and (3) an equation expressing this relationship so that predictions can be made. Lines or curves can result from the plots. We will begin by reviewing linear relationships and equations, and later we will discuss the process for fitting a line or curve to the data.

Exercises for Discussion 6:

Linear Equations Review: Recall the three forms of linear equations:
Slope Intercept: $y = mx + b$ **Standard:** $Ax + By = C$ **Point-Slope:** $y - y_1 = m(x - x_1)$

1. Given the slope and *y*-intercept: Write the equations for the lines below in point-slope and slope-intercept forms.
   a. $m = \frac{1}{2}$, $b = -3$ b. $m = -2$, $b = 5$ c. $m = 0$, $b = 6$ d. $m = 3$, $(0, -10)$

2. Given the slope and a point: Write the equations for the lines below in point-slope and slope-intercept forms.
   a. $P(-2, 1)$; $m = -3$ b. $P(-3, -3)$; $m = 4$ c. $P(-2, 4)$; $m = \frac{2}{3}$

3. Given two points: Write the equations for the lines below in point-slope and slope-intercept forms.
   a. $(-3, -1)$ and $(2, 1)$ b. $(-4, 3)$ and $(8, 0)$ c. $(-\frac{1}{2}, 2)$ and $(6, 4)$ d. $(-5, 4)$ and $(-5, -2)$

4. Parallel and perpendicular lines: Find the equations of the lines given each of the following:
   a. The equation of the line that is parallel to the line $y = -\frac{1}{4}x + 2$ through the point $(3, -2)$.

   b. The equation of the line that is perpendicular to the line $y = -3x + 6$ through the point $(-3, 4)$.

5. Find the slope and a point on the line given the equations:

a. $y + 6 = \frac{4}{3}(x - 2)$ b. $y - 1 = -4(x - 6)$ c. $y + 6 = -\frac{5}{4}(x + 2)$

6. Write each of the following equations into standard form.

a. $-5x + 11 = \frac{1}{2}y$ b. $y = \frac{2}{3}x + 4$ c. $y - 6 = -2(x + 3)$

7. Write the standard form of the linear equation for the line through the given the following points.

a. $(5, 2)$; $m = -\frac{5}{3}$ b. $(5, 2)$; $(-3, -2)$ c. $(-5, 2)$; $m = 0$

Discussion 7: Linear Equations as Mathematical Models

Discussion 6 reviewed how to write the equations of lines. What is more important is to use linear equations in order to make predictions about real world situations. A function used in this way is called a mathematical model. When you are given a situation in which two real-world variables are related by a linear equation, you must be able to

1. Sketch a graph
2. Find the linear equation
3. Use the equation to predict values of either variable
4. Interpret the real-world meaning of the slope and intercepts

Example: (from http://fordcalculuspages.wikispaces.com/file/view/3-5+notes.pdf) You pull out the plug from your bathtub. After 40 seconds, there are 13 gallons of water, there are 13 gallons of water left in the tub. One minute after you pull out the plug, there are 10 gallons left. Assume that the number of gallons varies linearly with time since the plug was pulled.

a. Write the particular equation expressing number of gallons left in the tub in terms of the number of seconds since you pulled the plug.

b. How many gallons would be left after i. 20 seconds? ii. 50 seconds?

c. Find the gallons-intercept. What does this number represent in the real world?

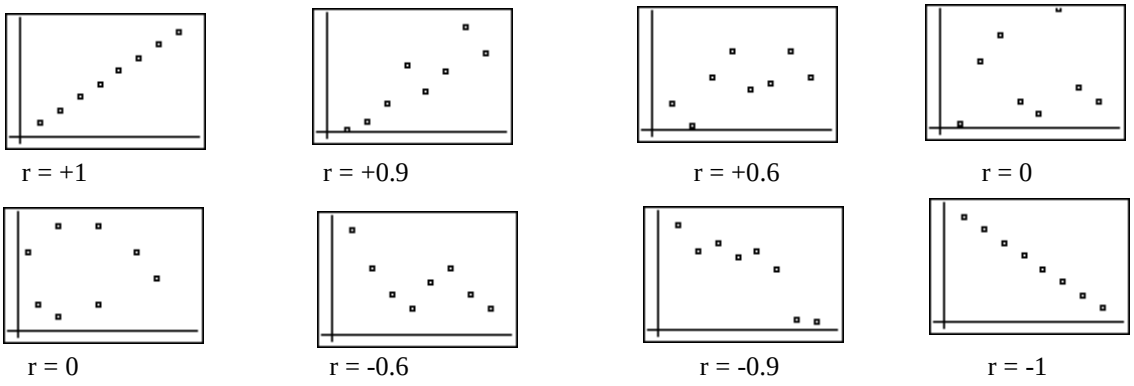
d. Find the time-intercept. What does this number represent in the real world?

e. Plot the graphs of this linear function using a suitable domain.

f. What are the units of the slope? What does this number represent in the real world?

Discussion 8: Correlation

The **correlation coefficient**, represented by the symbol *r*, measures the strength and direction of the **linear association** between two numerical variables, on a scale from -1 to +1. The closer the value of *r* is to +1 or to -1, the stronger the linear correlation. A correlation of 0 means that there is no *linear* relationship between the two variables. Below are eight diagrams illustrating various correlations. The larger the absolute value of *r*, the stronger the linear relationship. If most of the data points of a scatterplot lie close to the regression line, there is a high correlation.



r = +1 r = +0.9 r = +0.6 r = 0

r = 0 r = -0.6 r = -0.9 r = -1

The formula for finding the correlation coefficient, called ***r***, of a set of data is given below.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The calculator is programmed to find this value for you. In order to have the calculator display *r* when you calculate regression lines, press **<CATALOG>** and scroll down to **DiagnosticOn**. When the cursor is pointing to **DiagnosticOn**, press **<ENTER>, <ENTER>.** You should see the word DONE on the home screen.

The correlation coefficient is a third piece of information that can be used to determine whether there is truly a linear relationship between two variables. You still look at the scatterplot first, and then consider the form of the residual plot.

The calculator diagnostics returns another indicator of "best fit", the **coefficient of determination** or r^2. Note this is the only statistical diagnostic returned for non-linear models or regressions. The coefficient of determination is interpreted as the proportion of variability in the dependent variable that is explained by the independent variable. Note: In a linear model only, the coefficient of determination is the square of the correlation (*r*) between predicted *y* scores and actual *y* scores; thus, it ranges from 0 to 1, with numbers closer to 1 indicating a better fit regression and numbers closer to 0 indicating the opposite.

Exercises for Discussion 8:

1. Honda Resale. The following data indicates the resale value of 13 Honda Accords and the age of the car at the time of resale.

x: Age	*y*: Price
5	11995
5	9900
6	9000
6	8000
7	6900
7	6000
7	8700
7	5650
7	7300
7	7000
8	6999
9	5500
10	3500

   a) Construct the scatter plot for the data. Sketch it on your paper. Indicate the scale and labels.

   b) Before finding the correlation coefficient, predict what you think *r* will be and explain why.

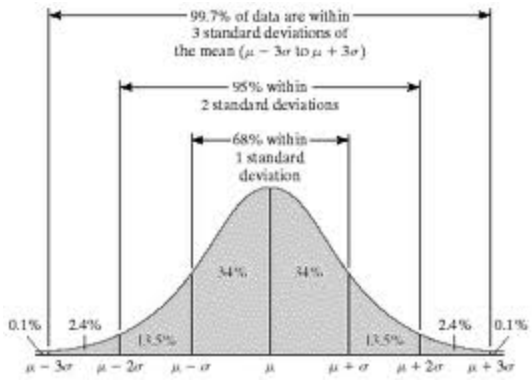
   c) Describe the relationship between the two variables.

   d) What is the linear regression equation?

   e) Find the correlation coefficient, *r*. Is there a high linear correlation between these two variables? Explain.

Discussion 9: The Normal Distribution

In nature, many processes and phenomena follow a particular shape called the **normal distribution**. This is the most commonly observed distribution. It has particular characteristics described by the **empirical rule**. The empirical rule states: If the population of measurements is symmetrical and bell-shaped, then approximately 68% of all the measurements in the set fall within the interval from $\mu - \sigma$ to $\mu + \sigma$; approximately 95% of all the measurements fall within the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$; and, approximately 99.7% of all measurements fall within the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$. Study the diagram at the right:



Notice that outliers occur beyond three standard deviations from the mean and represent less than 0.3% of the population.

http://wps.prenhall.com/esm_sullivan_statistics_1/10/2604/666710.cw/index.html%20parentloc

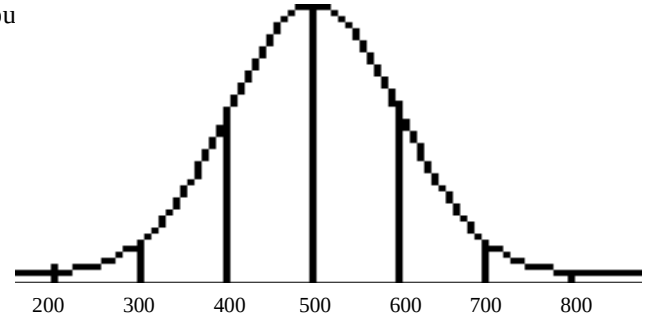
Exercises for Discussion 9: Consider the normal distribution while answering the following problems:

1. The SAT's are designed so that the distribution of scores wou
   a) What is the mean score?

   b) What is the standard deviation of the scores?

   c) What percentage of scores were between
      I. 500 and 600
      II. 500 and 700
      III. 500 and 800

Distribution of SAT Scores



   d) Some colleges do not admit applicants whose scores are less than 600. According to this distribution, what percentage of students would be expected to have scores of 600 or more?

   e) The most competitive colleges require scores over 700. What percentage of students would be considered by these colleges?

For each of the following problems apply the empirical rule to answer the questions. **Draw a curve for each problem.**

2. A ketchup company has fixed the weight of a bottle at 16 oz., with a standard deviation of 0.5 oz. The curve depicting the weights is bell-shaped. Approximately what percentage of bottles will be:
   a) greater than 15 oz.?

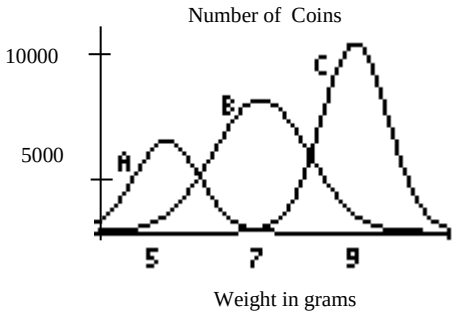
   b) greater than 17 oz.?

   c) less than 14 oz.?

   d) less than 13 oz.?

   e) between 15 and 17 oz.?

3. The curves in the figure at the right show the variations in the weights of quarters that were minted and put into circulation at the same time.* One curve shows the weight distribution when the coins were new and the other two show the distributions when they had been in circulation for five years and for ten years.



   a) Which curve do you think shows the weights of the newly minted quarters, which curve the coins after five years, and which curve after ten years?

   b) What happens to the average weight of the coins as time passes?

   c) What happens to the standard deviation of the weight of the coins as time passes?

*Adapted from a graph in the article "Scientific Numismatics" by D. D. Kosambi, *Scientific American*, February 1966.

Discussion 10: Standard Scores or *z* Scores

The standard score or *z* score is the number of standard deviations that a given value is above or below the mean, and it is found using the following formula: $z = \frac{x-\mu}{\sigma}$. It is used to compare samples that do not have the same mean and standard deviation. By standardizing the scores, the data can be compared. For example, which is better: a score of 65 on Test A or a score of 29 on Test B? The class statistics for the two tests are as follows:

Test A	Test B
$\mu = 50$	$\mu = 20$
$\sigma = 10$	$\sigma = 5$

For the score of 65 on Test A we get a z score 1.5, but for the score of 29 on Test B we get a z score of 1.8. That is, a score of 65 on Test A is 1.5 standard deviations above the mean, while a score of 29 on Test B is 1.8 standard deviations above the mean. This implies that the 29 on Test B is the better score. While 29 is below 65, it has a better *relative* position when considered in the context of the other test results.

Exercises for Discussion 10:

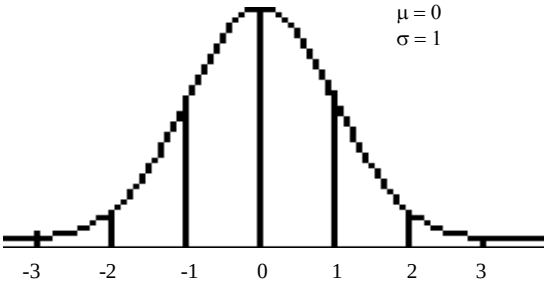
1. Transfer students to a new high school are sometimes given a standardized test with a mean of 100 and a standard deviation of 20. To three decimal places, convert the raw scores of the following students to z scores:
   Alice--105 Bob--72 Carol--142 David--133 Elliott--95

2. John weighs 220 pounds; his dog Fido weighs 90 pounds. If human males weigh an average of 160 pounds with a standard deviation of 20 pounds, and all dogs of Fido's breed have an average weight of 80 pounds with a standard deviation of 5 pounds, how do John and Fido compare, relative to their populations, with respect to weight?

Discussion 11: The Standard Normal or *z* Distribution

A **standard normal distribution** curve (called a *z*-distribution) is a symmetrical bell curve with the following characteristics:

1. The mean is 0.
2. The variance is 1.
3. The standard deviation is 1.
4. Approximately 68% of the data falls between z = ±1.
5. Approximately 95% of the data falls between z = ±2.
6. Approximately 99.7% of the data falls between z = ±3.
7. Total area under the curve is 1.0



We cannot talk about area under the curve at a specific point, but we can talk about area of a specific region. For example, the area left of the mean is 0.50. We can interpret that as $z < 0$ has an area of 0.50. The probability that $z < 0 = P(Z < 0) = 0.50$.
You can use normal tables or your calculator to find the probability that a randomly selected data point falls within an interval that is not defined by a whole 1, 2, or 3 standard deviations. Press **<2nd> <VARS> 2:normalcdf.** There are four parameters for normalcdf(lower bound, upper bound, mean, standard deviation). So, in order to find $P(Z < -2.13)$, enter **<2nd> < VARS> 2:normalcdf(-10^10, -2.13, 0, 1) (ENTER)**.

You should get approximately 0.017. This means there is a 1.7% probability that the *z*-score is less than -2.13.
Sometimes you want to find out what value of your variable will place you in a certain percentile. For example, if I wanted the top 25% of the population, I would be looking for a value of *z* such that $P(Z > z) > 0.25$ To reverse the process and find *z*, use the following **<2nd> < VARS> 3:invNorm**

There are three parameters for invNorm(area, mean, standard deviation). However, the area in the invNorm syntax must be the area to the *left* of the value of interest. So, in order to find $P(Z > z) > 0.25$, enter **<2nd> <VARS><3>invNorm(0.75, 0, 1) (ENTER)**
You should get approximately 0.674. This means there is a 25% probability that the *z*-score is greater than 0.6744.

Exercises for Discussion 11:

1. For a standard normal distribution, find z if
   a) $P(Z < z) = 0.0668$ b) $P(Z > z) = 0.9861$

2. Use your calculator to find the following probabilities, assuming a normal distribution. Include a sketch illustrating each area under the normal curve.
   a) $P(X < 3.5)$ when $\mu = 5$, $\sigma = 1$.

   b) $P(X > 130)$ when $\mu = 110$, $\sigma = 25$

   c) $P(14.2 < X < 15.0)$ when $\mu = 14$, $\sigma = 0.5$.

3. One part of a test administered to adults is an exercise in manual dexterity. The average time for the test is 165 seconds, with a standard deviation of 21 seconds. Assume the relative frequency distribution of the times needed to complete the test is approximately normal. What proportion of the adult population can complete the test in
   a) more than 190 seconds b) between 140 and 160 seconds

4. If the mean of a normal distribution is 10 feet and the standard deviation is 2 feet, for what values of y is it true that $P(Y < y) = 0.025$?

5. If the mean of a normal distribution is 20.5 inches and the standard deviation is 0.2 inches, find x such that $P(X > x) = 0.7995$.

6. If X is a continuous random variable that can be modeled as normal with a mean of 100 cm and a standard deviation of 10 cm, find x so that
   a) $P(X > x) = 0.0049$

   b) $P(X < x) = 0.9332$

Research 1 Practice Test

This is a one hour test.

There are many other questions that can be asked about the material for which you are responsible. Do NOT assume that just because you can do the material on this test that you should not study ALL of the material in the packet. This practice test has been provided so that you can see the type of question you might be asked and so that you understand the time constraints under which you will work.

Load the following data sets in your calculator before beginning the timed test.

Sleeping Habits

8	7.5	9	7.5	9	6	5	9	7.5	7	8	7	6.5
8.5	8	6.5	8.5	6	7	7.5	7	6	8.5	7	8	7
7.5	7	6	7	8	7.5	6	7					

Monthly Rents

Harrisburg (n = 10): 500, 549, 569, 575, 585, 600, 630, 680, 705, 790

Philadelphia (n = 15): 475, 525, 540, 575, 600, 600, 645, 700, 725, 755, 885, 930, 965, 1180, 1300

Multiple Choice

_____ 1. A small company that prints custom t-shirts has 6 employees, one of whom is the owner and manager. Suppose the owner makes $120,000 per year and the other employees make between $40,000 and $50,000 per year. One day, the owner decides to give himself a $30,000 raise. Which of the following describes how the company's mean and median salaries would change?

a. The mean and median would both increase by $5000.
b. The mean would increase by $5000 and the median would not change.
c. The mean would increase by $6000 and the median would not change.
d. The median would increase by $6000 and the men would not change.
e. The mean would increase by $6000, but we cannot determine the change in the median without more information.

_____2. A medical researcher collects health data on many women in each of several countries. One of the variables measured for each woman in the study is her weight in pounds. The following list gives the five-number summary for the weights of adult women in one of the countries.

Statislvania: 92, 110, 120, 160, 240

About what percent of Statislvania women weigh between 110 and 240 pounds?

a. 50%
b. 65%
c. 75%
d. 85%
e. 95%

_____3. The area under the standard normal curve corresponding to $-0.3 < Z < 1.6$ is

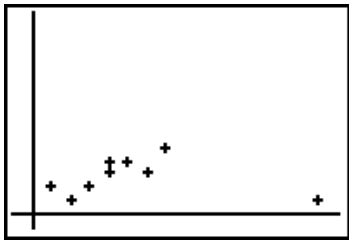
a. 0.3273
b. 0.3821
c. 0.4713
d. 0.5631
e. 0.9542

_____4. An agricultural economist says that the correlation between corn prices and soybean prices is $r = 0.7$. This means that

a. when corn prices are above average, soybean prices also tend to be above average.
b. there is almost no relation between corn prices and soybean prices.
c. when corn prices are above average, soybean prices tend to be below average.
d. when soybean prices go up by 1 dollar, corn prices go up by 70 cents.
e. the economist is confused, because correlation makes no sense in this situation.

_____5. The effect of removing the right most point (near the positive x-axis) in the scatter plot shown would be

A. The slope of the linear regression line will increase; r will increase
B. The slope of the linear regression line will increase; r will decrease
C. The slope of the linear regression line will decrease; r will increase
D. The slope of the linear regression line will decrease; r will decrease
E. No change



Short Answer.

6. A student studying the sleeping habits of seniors at his school asked 34 randomly-selected seniors how many hours of sleep they got the previous night. The data, rounded to the nearest half-hour, is given in the table below (and you have already loaded it into your calculator).

8	7.5	9	7.5	9	6	5	9	7.5	7	8	7	6.5
8.5	8	6.5	8.5	6	7	7.5	7	6	8.5	7	8	7
7.5	7	6	7	8	7.5	6	7					

a. Find the mean and standard deviation (this is a sample) of these data.

Mean = _____ Standard Deviation = _____

b. Find and label the five number summary for these data.

_____ _____

_____ _____

_____

c. Determine if there are any outliers in these data. Show your work.

d. Using an appropriate method, plot the data.

e. Using all of your calculations and results from a-d, write a clear, concise paragraph (3-4 sentences) that describes the data set.

7. The following side-by-side stemplot displays the total number of points scored per Super Bowl football game for the first 41 Super Bowls (from 1967–2007), separated according to the first 20 games (1967–1986) and the next 21 games (1987–2007):

First 20 Games		Next 21 Games
97321	2	
87720	3	16799
777654	4	13456
640	5	23569
6	6	11599
	7	5

a. Does this stemplot enable you to determine how many points were scored in the first Super Bowl? If so, what is this number?

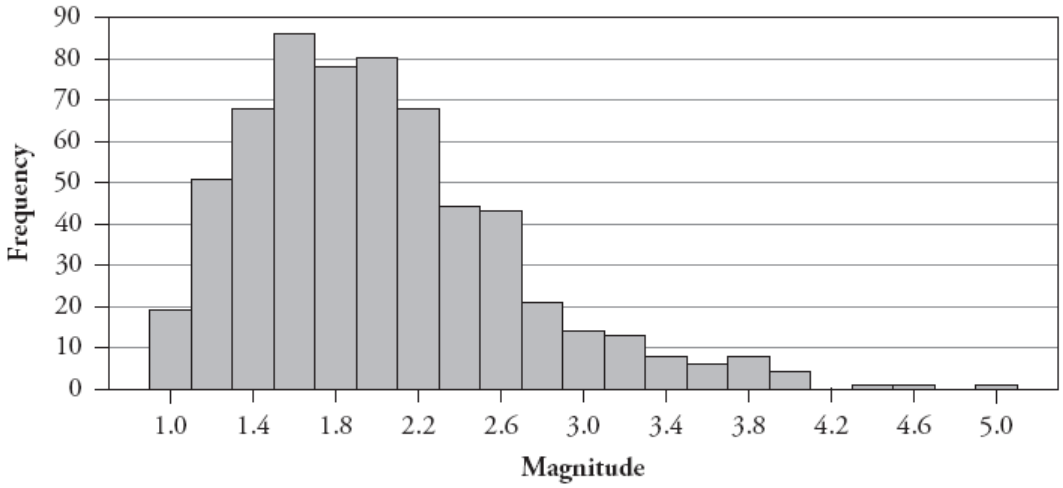
b. Does this stemplot enable you to determine how many of the first 41 Super Bowls had a total of 37 points? If so, what is this number?

c. Does this stemplot provide evidence that Super Bowl games have become more high-scoring over time, more low-scoring over time, or neither? Explain.

d, True or false? (Please circle the appropriate response.) The five lowest-scoring Super Bowls were all played among the first 20 games.

e. True or false? (Please circle the appropriate response.)The five highest-scoring Super Bowls were all played among the next 21 games.

8. The following histogram displays the magnitudes of the 614 earthquakes with Richter scale magnitude greater than 1.0 that occurred in the United States between March 25 and April 1, 2004:

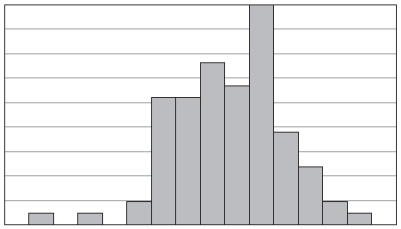


a. Describe the shape of this distribution.

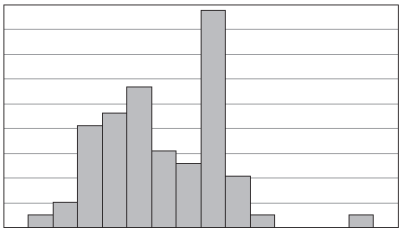
b. Is the percentage of earthquakes of magnitude 3.0 or higher closest to 1%, 10%, or 25%?

9. The following histograms display student responses to several questions on a course survey taken during the first day of a statistics class.

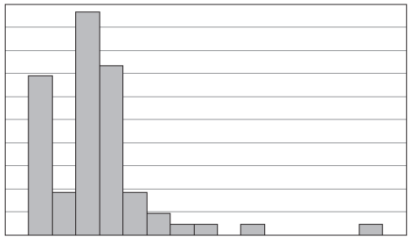
I.



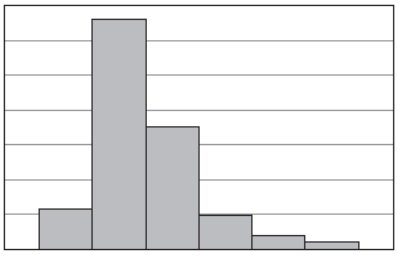
II.



III.



IV.



_____a. Which histogram do you think displays the variable *number of siblings?* Justify your answer.

_____b. Which histogram do you think displays the variable *price paid for most recent haircut?* Justify your answer.

_____c. Which histogram do you think displays the variable *height?* Justify your answer.

10. The following data are monthly rents (in dollars) of studio and one-bedroom apartments in Harrisburg and Philadelphia, Pennsylvania.

Harrisburg (n = 10): 500, 549, 569, 575, 585, 600, 630, 680, 705, 790

Philadelphia (n = 15): 475, 525, 540, 575, 600, 600, 645, 700, 725, 755, 885, 930, 965, 1180, 1300

a. For each city, determine and report the five-number summary of these monthly rents. (Put in order from lowest to highest.)

Harrisburg:

Philadelphia:

b. Construct boxplots of the distributions of rent amounts in these two cities, using the same axis and scale. (Do not bother to check for outliers; there are no outliers in either distribution.) Use a ruler and clearly label.

c. Compare and contrast the distributions of monthly apartment rents in these two cities. Refer to appropriate calculations and displays to support your comments.

11. Suppose that a tire manufacturer believes that the lifetimes of its tires follow a normal distribution with mean 48,000 miles and standard deviation 5,000 miles.

a. Produce a well-labeled sketch of this normal distribution.

b. Determine the *z*-score corresponding to 55,000 miles.

c. Determine the probability that a randomly selected tire lasts for more than 55,000 miles.

d. If the manufacturer wants to issue a guarantee so that 99% of its tires last for longer than the guaranteed lifetime, what *z*-score should it use to determine that guaranteed lifetime?

e. If the manufacturer wants to issue a guarantee so that 99% of its tires last for longer than the guaranteed lifetime, how many miles should it advertise as its guaranteed lifetime?