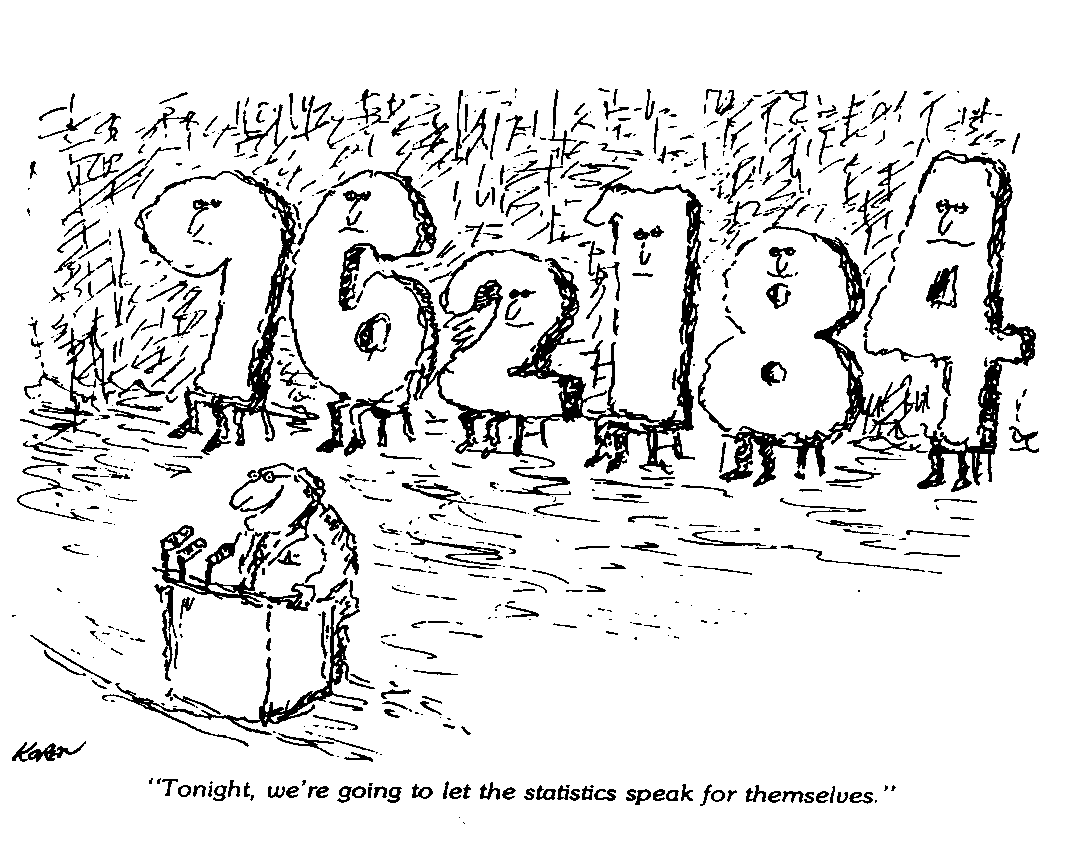
Statistics and

Probability Review Key



**TJHSST**

**Fall Semester Packet 2019**

**Review & Self-Study for**

**Research Statistics 1**

# Part 1: Statistics Review

# Discussion 1: Basic Definitions

Statistics is the field of math used to collect, organize, summarize (descriptive statistics), analyze, and present (inferential statistics) data in a meaningful way.

A variable is any measurable or observable characteristic of a group of objects or people. Data are the actual observations or measurements of a variable. There are two types of data: categorical and numerical. Categorical data can be sorted into groups or categories while numerical data are values or observations that can be measured. For example, types of fish caught would be categorical data while weights of fish caught would be numerical data.

Different types of graphs can be used to display categorical data vs. numerical data. The way data is displayed is dependent on what someone is trying to communicate.

We can also classify data by the number of variables represented. The data set consisting of observations of just one attribute is univariate (hair color or height). The data set consisting of observations of pairs of numbers is bivariate (GPA and SAT for a single student).

Exercises for Discussion 1

Classify the following as categorical or numerical.

a. Number of students in the RS1 classes who turn in their project on the due date. **Numerical**

b. Amount of fluid (in ounces) dispensed by a machine used to fill bottles with Pepsi. **Numerical**

c. Birth order classification (only child, firstborn, middle child, lastborn) in a family. **Categorical**

d. Brand of computer purchased by a customer at Best Buy. **Categorical**

e. Zip code. **Categorical**

f. Puppy weights. **Numerical**

g. Puppy breeds. **Categorical**

h. Numbers on football jerseys. **Categorical**

All of the data above is univariate. You may wish to know if a puppy’s breed is indicative of his weight. These two variables taken together constitute bivariate data.

# Discussion 2: 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  when we are talking about a sample or the symbol  (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 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 five lists: **L1 L2 L3** **L4 L5**

Note: To CLEAR data to make room for new entries, do the following

Press: **[STAT] <ENTER>** and then the up arrow **< > [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.

1. **To Calculate Mean, Median, Mode:**
2. 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>** Continue to press ENTER until the 1-Var Stats are

displayed. The number of times you need to press ENTER varies with your operating system.

At the top of the display is  = MEAN.

1. 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.

1. To find the MODE, scan the ordered data and find which data point occurs most often.

Using the results from above, answer the following questions:

1. What is the mean weight of these fourth graders? **67.68**

What is the median weight? **68** What is the mode? **60**

2. Which do you think is the most representative number for these weights; the mean, median or mode? Explain your choice. **The mean takes into account every single student so it is very representative. In addition, the median is very close to the mean so it confirms that the mean is a good representative number. The mode is so far from the mean and median that it is not a good representative number.**

Exercises for Discussion 2

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 mean, median or mode amount spent by the class on one day? Explain your choice. **MEAN. You could multiply the mean by the number of students to find the total.**

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 your choice. **MEDIAN. If you read more books than the median number, then you know your read more than 50% of the class.**

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 your choice. **MODE. We need to know the skate sizes that are most frequently used.**

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 your choice. **MEDIAN. This would tell you what income half the residents are below. The mean could be skewed by a few very wealthy peoples.**

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 |

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

mean salary: **$26,440** median salary: **$22,000** mode salary: **$18,000**

**The company is not telling the truth. All measures of central tendency, even the mean which is**

**affected by the “extremely” high salaries, are lower than $30,000.**

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 at that salary.

**To Calculate using the data**

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

In the classic operating system: Press: **[2nd] 1 , [2nd] 2** to select L1, L2.

Press: **<ENTER>**

In the new operating system, the List is L1 and the Frequency List is L2.

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 your choice. **The median is the most representative. It is not overly affected by high salaries but recognizes their existence.**

# Discussion #3: 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. You should already be familiar with range and interquartile range. We will study variance and standard deviation in RS1.

The **range** of a set of numbers is the difference between the largest (maxX) and smallest (minX) 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). This information is given in 1-Var Stats as Q1 and Q3.

**Example 1:** In today's world, calculators and computers make the computation of range and interquartile range 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 range is 23 and the interquartile range is 11.

Exercise for Discussion 3

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

|  |  |  |  |
| --- | --- | --- | --- |
| **CLASS** | A | B | **C** |
| mean | 86.286 | 86.357 | 84.429 |
| median | 87 | 87.5 | 87.5 |
| range | 16 | 23 | 44 |
| IQR | 10 | 18 | 17 |

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

b) Discuss the similarities and differences among the three classes. **All three classes had a similar test average. However, class A was the least spread out, with most students performing close to the average. Class C was most spread out (largest range), with students performing significantly better and worse than the class average.**

# Discussion #4: Stem-and-Leaf Plots, Line 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 (directions for sorting a list in your calculator are on page 3).

**Period 1 Test Grades**

|  |  |
| --- | --- |
| **Stem (10)** | **Leaf (1)** |
| **5** | **6** |
| **6** | **89** |
| **7** | **147899** |
| **8** | **014567789** |
| **9** | **012355** |

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 **line plot** or **dotplot**. A line plot or dot plot provides an ordered display of all values in a data set and shows the frequency of data on a number line. These plots are used to show the spread of the data, to include clusters (groups of data points) and gaps (large spaces between data points), and quickly identify the range, mode, and any extreme data values. 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 dot 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. See the example below.

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 Key: 1|2 = 12 grams

4 | 2 2 6

5 | 0

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.

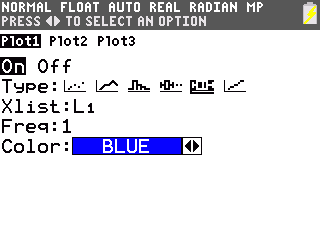
The following box-and-whisker plot illustrates this data:

**Carbohydrate Content**

12 19 34 42 50

| | | | | |

0 10 20 30 40 50 (grams)

Box-and-whisker plots can also be found on the TI-84. First enter the data as described before. Then create the plot:

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 and 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).

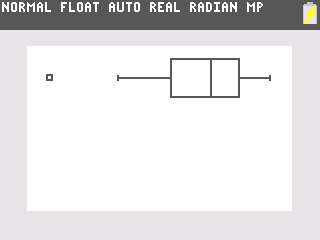
The information found in the example can be used to determine if there are any outliers in the data. An **outlier** is a data point that is very different from the other points and can consequently cause the mean of the data to be overly influenced in one direction. Recall the set of salaries in exercise 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. 19 - 34.5 = -15.5. There are no data points below -15.5, 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.

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

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

The Period 1 Test Grade Data at the beginning of

this discussion is plotted with the modified box-and-whisker plot showing evidence that the grade of 56 is an outlier.

Exercises for Discussion 4

a) Which class had the higher median? **The classes had the same median of 3.**

b) What was the interquartile range for each class? **IQRJR = 3.5-2.5 = 1.0**

**IQRSR = 3.3-2.6 = 0.7**

c) Estimate each of the classes' best and worst grade point averages. Are there any outliers? Explain. **Best JR = 4.0 Worst JR = 0.75**

**Best SR = 4.0 Worst SR = 1.5**

**JR outlier interval [1.0, 5.0], therefore 0.75 is an outlier.**

**SR outlier interval [1.55, 4.35]. Depends on estimate for worst senior GPA; with given estimate there would be an outlier…**

1. Refer to the plots below.

**Average GPA**

Juniors

2.5 3 3.5

Seniors

bo

2.6 3 3.3

| | | | |

0 1 2 3 4 (GPA)

2. 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 carbohydrate content.

**CARBOHYDRATE CONTENT: 1 2 5 7 9 *n*=15**

**2 8 9 Key: stem=10’s; leaves=units**

**3 3 5 4 6 8**

**4 2 2 6**

1. **0**

b) Use this plot to help find the mean, median and mode of the carbohydrate content. **mean: 31.733 g; median: 34 g**; **mode: 42 g**

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

**FAT CONTENT: 1 1 2 5 8 8 8 8 8 *n*=15**

**2 0 1 3 3 6 8 Key: stem=10s; leaves=units**

**3 1**

d**)** Use this plot to help find the mean, median and mode for fat content. Confirm your answers by using your calculator. **mean: 20 g; median: 18 g; mode: 18 g**

e) Are there any outliers for the fat content? Justify your answer using the method above. **IQR = 23 – 18 = 5 so (1.5)(5) = 7.5. Q1 – 7.5 = 10.5 and Q3 + 7.5 = 30.5. So, the range for outliers is outside of {10.5, 30.5]. Thus 31 is an outlier.**

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

This a diagram of the modified box plot. **Note that for the FAT content there is an outlier. The “ok” interval is [10.5, 30.5], using the formula for outliers, therefore the data entry, 31g, is an outlier.**

3. 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. Enter the data in your calculator to find the following:

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

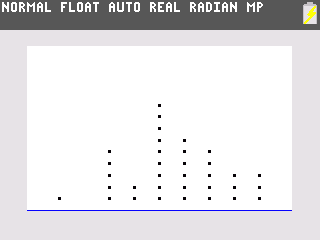
7.5 7 6 7 8 7.5 6 7

1. Find the mean of these data. **Mean** = **7.309 hours**
2. Find and label the five-number summary for these data.

**minimum = 5, lower quartile = 7, median = 7.25, upper quartile= 8, maximum = 9**

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

**1.5 x IQR = 1.5 x 1 = 1.5; 7 – 1.5 = 5.5 so 5 is a low outlier; 8 + 1.5 = 9.5 so no high outliers.**

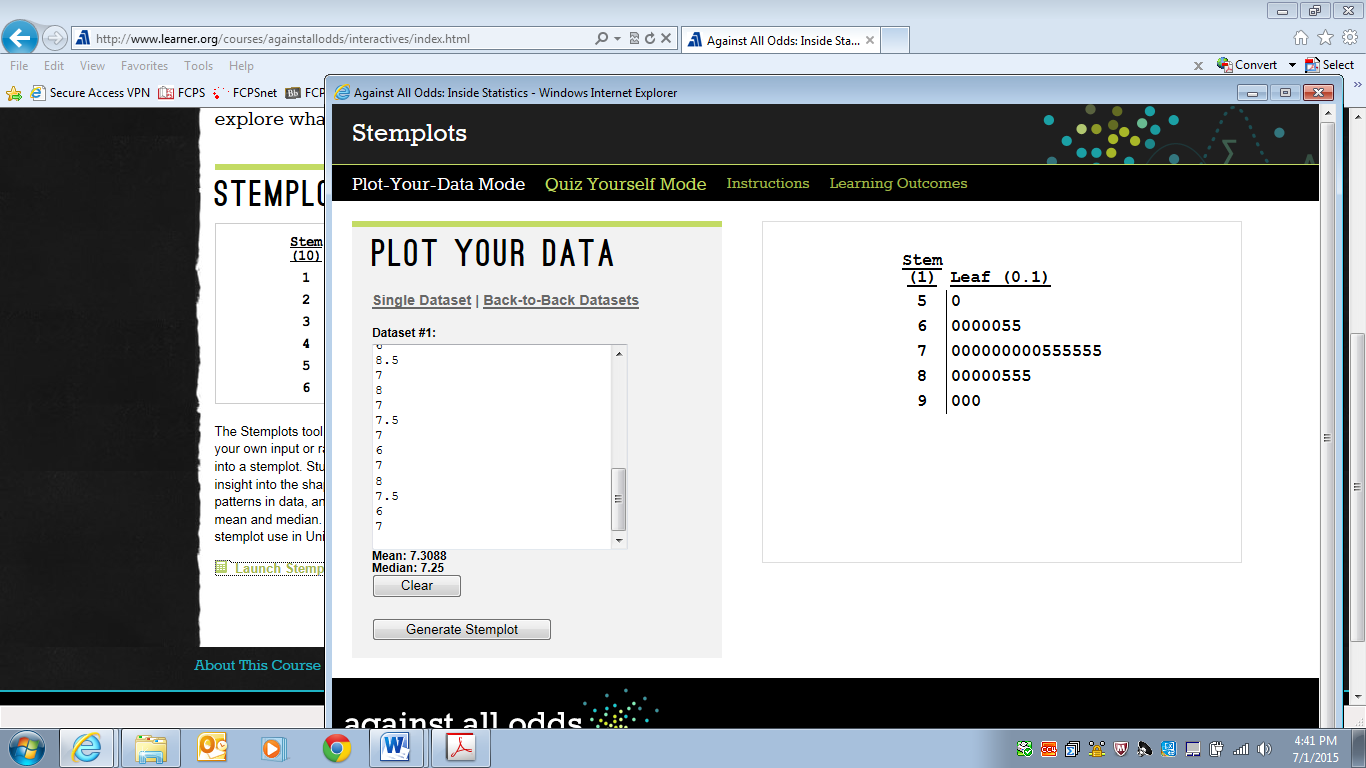


5 6 7 8 9

Hours of sleep

frequency

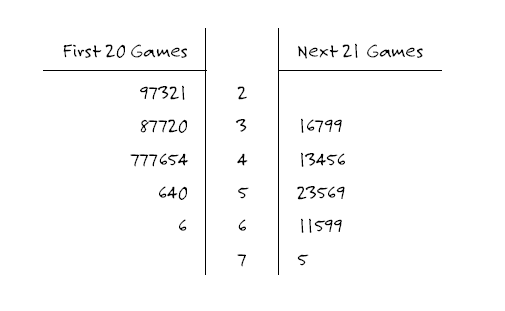
1. Draw a line plot (dot plot) of this data.
2. Draw a stem plot of this data.



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

**The sleep data is tightly grouped, having a small range of 4 hours. There is one low outlier of 5 but there are no high outliers. The center of the data set can be described by either the mean (7.309) or median (7.25) since they are so close to one another. The stem-leaf plot and dot plot clearly show that the center is between 7 and 7.5 hours. The data and graph confirm that on average seniors at this high school got approximately 7.25 hours of sleep the previous night.**

4. 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):

****

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

**No, you cannot tell from this stemplot the order of the Super Bowl games.**

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?

**Yes, you can tell that in 2 of the first 20 and one of the second 21 had 37 points so there were 3 Super Bowls where there were a total of 37 points.**

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.

**This stemplot provides evidence that Super Bowls have become more high-scoring over time because the scores in the last 21 games tend to be slightly higher than the scores in the first 20 games.**

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

**True. The five lowest scores were 21, 22, 23, 27, and 29. 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.

**False. The five highest scores were 75, 69, 69, 66 (in first 20), and 65—and one of these was among the first 20 games.**

# Discussion 5: Quantitative Data, Frequency Tables, and Histograms

Categorical data is most often represented by bar or circle

tally

Black Gray White Red

2

4

6

Colors of car in the parking lot

graphs. Because the data values are separate, the bars as separate.

There is no overlap of the bars.

The graph at the right displays categorical data, car color, in a

bar graph.

In this class, we are most interested in numerical or quantitative data. Generally, quantitative 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. A diagram of quantitative 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 use the rule below 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 exercise in discussion 2 to complete the following example):

1. Enter the data (or remind yourself which list the data is still in from before).

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). Your width should be an integer.

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.)

1. Press [**GRAPH]**.

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

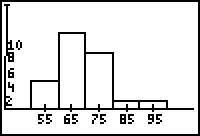
Be sure to:

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 or skip values).

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 2



Frequency

Weights of 4th Graders in lb.

55 65 75 85 95

and use the procedure outlined above to find a histogram. Add labels

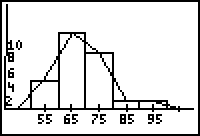
and scales and compare your results to the histogram shown here.

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.



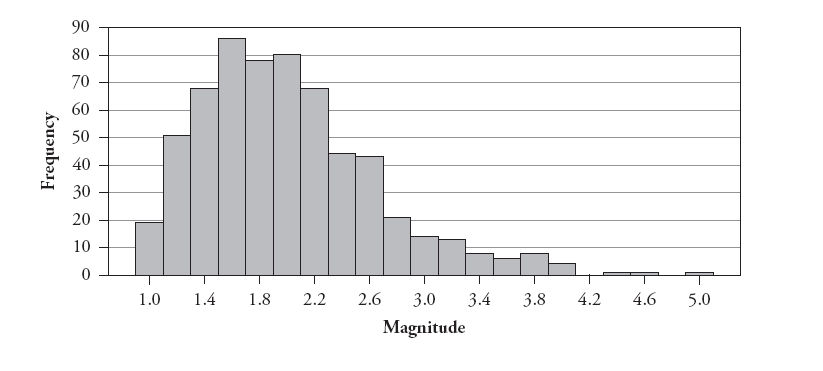
Frequency

Weights of 4th Graders in lb.

55 65 75 85 95

Exercises for Discussion 5

1. 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 this distribution. Include discussion of the measures of central tendency, which is the best and why. Discuss the range and shape of the data. Draw a frequency polygon on top of the histogram. **The earthquake data has a range of 4. It appears that there may be outliers at the high end. Most of the data is in the interval from 1 to 3.3, centered at approximately 2. The mean will be pulled towards the high end because of the few data points in the upper range so the median will be a better measure of the central tendency.**

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

2. 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

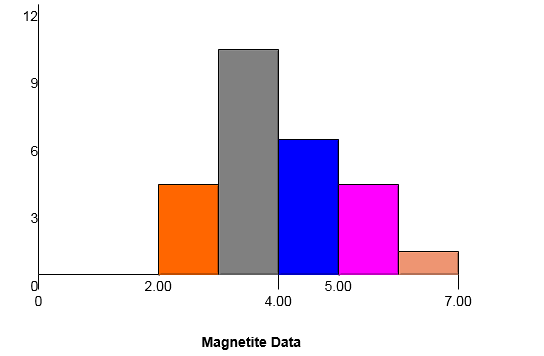
5.6 2.6 3.9 3.4 4.4 3.7 3.7

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:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Interval** | **Class Limits** | **Tally** | **Frequency** | **Relative Frequency** |
| **1** | **2.0-< 3.0** | **|||** | **3** | **3/25** |
| **2** | **3.0-<4.0** | **||||| |||||** | **10** | **10/25** |
| **3** | **4.0-<5.0** | **||||| |** | **6** | **6/25** |
| **4** | **5.0-<6.0** | **|||||** | **5** | **5/25** |
| **5** | **6.0-<7.0** | **|** | **1** | **1/25** |



What does the frequency polygon/histogram reveal about this data? Explain. **The data is centered approximately at 4. The one value will pull the mean towards the right so the median will be a better measure of central tendency. The range is 3.8.**

# Discussion 6: 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 and 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 tend 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].**

Last, sketch the scatterplot on your paper. Label axes and scales clearly.

Exercises for Discussion 6: Put answers on separate sheet of paper.

400

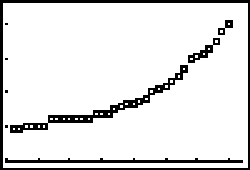
300

200

100

Number of

Soft Drinks



1945 '50 '55 '60 '65 '70 '75 '80

Year

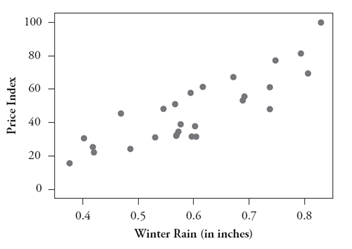
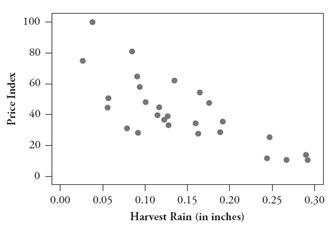
1. The following is a plot over time showing how many 12 ounce soft drinks the average person in the U. S. drank each year from 1945 to 1980.

(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.)

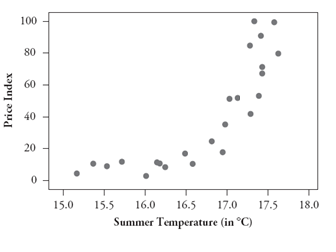
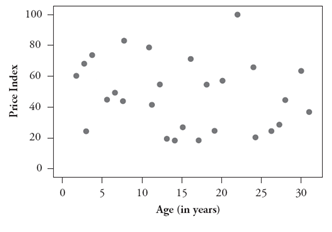
1. About how many soft drinks did the average person drink in 1950? **Approximately 105;** in 1970? **Approximately 220**
2. About how many six-packs of soft drinks did the average person drink in 1980? **410/6, approximately 68 six packs**
3. About how many soft drinks did the average person drink per week in 1950? **105/52, approximately 2;** in 1980? **410/52, approximately 8**
4. If the trend in the plot continued, about how many 12 ounce soft drinks did the average person drink in the year 2000? **Trend: in 10 years, about 150 more; so, in 2000, about 700.**
5. In what year did soft drink consumption start to "take off"? Can you think of any possible reason for this phenomenon? **Around 1961. Diet sodas were introduced, and aluminum cans were starting to be used.**
6. Describe the direction, strength, and form of the association shown in the scatterplot. **Positive, Strong,**

**Non-linear (curved)**

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

a) b)

c) d)



**a) negative, moderate, linear**

**b) positive, strong, linear**

**c) positive, strong, non-linear**

**d) non-linear since there appears to be no pattern, there is no need to comment on strength and direction.**

# 

# Discussion 7: 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.

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)**

The form most often used in upper level math courses is the point-slope form.

Exercises for Discussion 7:

1. Given the slope and *y*-intercept: Write the equations for the lines below in point-slope and slope-

intercept forms.

a. *m* = ½, *b* = –3 b. *m* = –2, *b* = 5 c. *m* = 0, *b* = 6 d. *m = 3*, (0, –10)

***y* + 3 = ½ *x* *y* – 5 = -2*x* *y* – 6 = 0 *y* + 10 = 3*x***

***y* = ½ *x* – 3 *y* = -2*x* + 5 *y* = 6 (horizontal line) *y* = 3*x* – 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* = 

***y* – 1 = -3(*x* + 2) *y* + 3 = 4(*x* + 3) *y* – 4 = (2/3) (*x* + 2)**

***y* = -3*x* – 5 *y* = 4*x* + 9 *y* = (2/3)*x* + (16/3)**

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)

***y* + 1 = (2/5)(*x* + 3) or  *y* – 1 = (2/5)(*x* – 2) *y* – 3 = (-1/4)(*x* + 4) or *y* = (-1/4)(*x* – 8)**

***y* = (2/5)*x* +(1/5) *y* = (-1/4)x + 2**

c. (– ½, 2) and (6, 4) d. (–5, 4) and (–5, –2)

***y* – 2 = (4/13)(*x* + ½ ) or *y* – 4 = (4/13)(*x* – 6) Vertical line: x = –5; no point-slope or slope-**

**y = (4/13)*x* +(28/13) intercept form**

4. Parallel and perpendicular lines:Find the equations of the lines given each of the following:

1. The equation of the line that is parallel to the line *y* = –*x* + 2 through the point (3, –2).

***y* + 2 = - ¼ (*x* – 3) (or y = - ¼ *x* –5/4)**

1. The equation of the line that is perpendicular to the line *y* = –3*x* + 6 through the point (–3, 4).

***y* – 4 = (1/3) (*x* + 3) (or *y* = (1/3)*x* + 5)**

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

a. *y* + 6 = (*x* – 2) b. *y* – 1 = – 4(*x* – 6) c. *y* + 6 = (*x* + 2)

**m = 4/3; (2, -6) m = -4; (6, 1) m = -5/4; (-2, -6)**

6. Write each of the following equations into standard form: (no fractions, A > 0)

a. –5*x* + 11 = ½ *y* b. *y* = *x* + *4* c. *y* – 6 = –2(*x* + 3)

**10*x* + *y* = 22 2*x* – 3*y* = -12 2*x* + *y* = 0**

7. Write the standard form of the linear equation for the line given the following information:

a. (5, 2); *m* =  b. (5, 2); (–3, –2) c. (–5, 2); *m* = 0

**5*x* + 3*y* = 31 *x* – 2*y* = 1 *y* = 2**

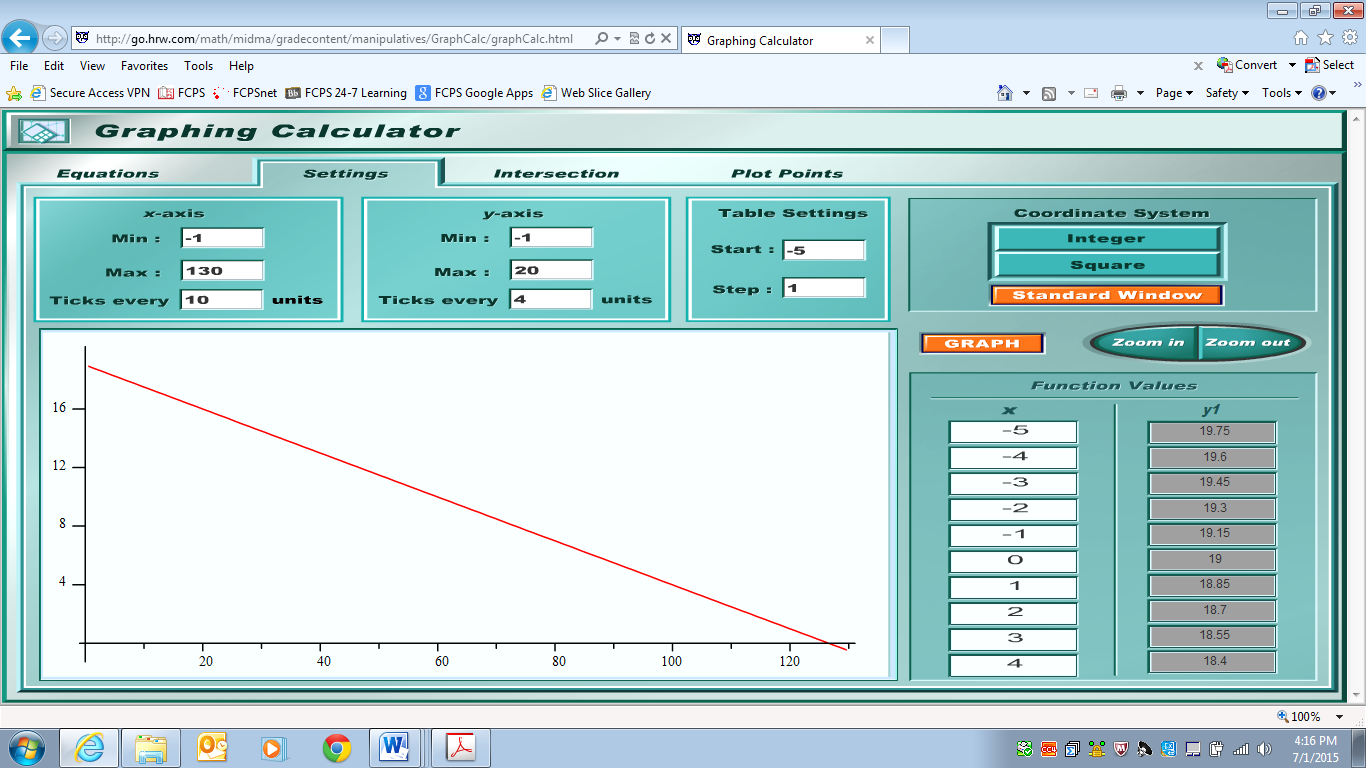
# Discussion 8: Linear Equations as Mathematical Models

Discussion 7 reviewed how to write the equations of lines. Now we can 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 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.

1. Write the linear equation expressing number of gallons left in the tub in terms of the number of seconds since you pulled the plug. **Let *t* = number of seconds since pulling the plug and let *V* = number of gallons of water in tub:** ****
2. How many gallons would be left after: i. 20 seconds? **16 gallons** ii. 50 seconds? **11.5 gallons**
3. Find the gallons-intercept. What does this number represent in the real world? **19 gallons; It represents the amount of water in the tub before the plug is pulled.**

d. Find the time-intercept. What does this number represent in the real world? ** seconds; It represents the amount of time needed to empty the tub.**

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

***V* (number of gallons of water in tub)**

***t* (number of seconds since pulling the plug)**

f. What are the units of the slope? What does this number represent in the real world? **Gallons/second; This means that water drains out of the tub at a rate of 3/20 gallons per second.**

# Part 2: Probability Review

# Discussion 9: The Vocabulary of Probability

Though probability should be taught in 7th and 8th grade math, some students miss these topics. View this free online video that reviews the concepts that you should know using the link below.

**Against All Odds Video 19: Models of Probability:**

<http://www.learner.org/courses/againstallodds/unitpages/unit19.html>

The following are a summary and exercises provided in the Student Guide that accompanies the video. For more information you can go to:

<http://www.learner.org/courses/againstallodds/pdfs/AgainstAllOdds_StudentGuide_Unit19.pdf>

Summary:

Probability is the likelihood that an event, *E*, will occur. *P*(*E*) is a number between 0 and 1, inclusive, such that 0 < *P*(*E*) < 1. In general, if all outcomes of an event are equally likely, the probability of an event occurring is equal to the ratio of desired outcomes to the total number of possible outcomes in the sample space.

A probability of 0 means the event will never occur.

A probability of 1 means the event will always occur.

If all outcomes of an event are equally likely, the theoretical probability of an event =

The experimental probability of an event is determined by carrying out a simulation or an experiment.

The experimental probability of an event =

As the number of trials increases, the experimental probability gets closer to the theoretical probability (Law of Large Numbers).

Experimental probability is based on a calculation which is the result of performing an experiment, conducting a survey, or looking at the history of an event.

Example 1: List the events for each experiment and then find the theoretical probability.

a. Toss one coin. Find the probability of heads.

Solution: T, H. *P*(Head) = ½

b. Toss two coins. Find the probability of at least one head.

Solution: TT, TH, HT, HH *P*(at least one head) = ¾

c. Rolling a die. Probability of a prime number

Solution: 1, 2, 3, 4, 5, 6. *P*(prime number) = ½

Example 2: A bucket contains 15 blue pens, 35 black pens, and 40 red pens. You pick one pen at random. Find each theoretical probability.

a. *P*(black pen) Solution: 35/90 = 7/18

b. *P*(blue pen or red pen) Solution: 15/90 + 40/90 = 55/90 = 11/18

c. *P*(not a blue pen) Solution: 1-15/90 = 75/90 = 5/6

d. *P*(black pen or not a red pen) Solution: This event can be simplified to “not red,” so

*P*(not red) = 50/90=5/9.

Example 3: Jon used a standard deck of 52 cards and selected a card at random. He recorded the suit of the card he chose, and then replaced the card. The results are in the table below.

|  |  |
| --- | --- |
| Diamonds | ||||| || |
| Hearts | ||||| |||| |
| Spades | ||||| ||||| | |
| Clubs | ||| |

a. Based on his results, what is the experimental probability of selecting a heart? Solution: 9/30

b. What is the theoretical probability of selecting a heart? Solution: ¼

c. Based on his results, what is the experimental probability of selecting a diamond or a spade? Solution: 18/30

d. What is the theoretical probability of selecting a diamond or a spade?

Solution: ½

Exercises for Discussion 9

1. Assume that you draw one card from a standard card deck. Find the probability of each of the following.

1. Draw a black card. **1/2**
2. Draw a jack. **4/52 = 1/13**
3. Draw a black card or an ace. **26/52 +4/52 – 2/52 = 28/52 = 7/13**
4. Draw a red card or a face card. **26/52 +12/52 – 6/52 = 32/52 = 8/13**

2. Assume that you roll two dice once. Find the probability of each of the following.

1. Roll a sum of 6 or 7 **5/36 + 6/36 = 11/36**
2. Roll a sum of 5 or an even sum **4/36 + 18/36 = 22/36 = 11/18**
3. Roll a sum of 10 or an odd sum **3/36 + 18/36 = 21/36 = 7/12**
4. Roll a sum of 1 or 12 **0 + 1/36 = 1/36**

3. Assume you have a jar containing 7 brown marbles, 3 blue marbles, and 6 green marbles. You make a single draw from the jar, taking one marble. Find the probability of each of the following.

1. Draw a brown or a blue marble **7/16 + 3/16 = 10/16 = 5/8**
2. Draw a green marble **6/16 = 3/8**
3. Do not draw a blue marble **7/16 + 6/16 = 13/16 or 1 – 3/16 = 13/16**
4. Do not draw either a brown or blue marble. **6/16 = 3/8**

4. Suppose you tossed a number cube (die) 8 times and recorded your results. The recorded data shows you tossed 2 fives.

1. What is the experimental probability of tossing a 5? **2/8 = 1/4**
2. What is the theoretical probability of tossing a 5? **(1/6)(1/6)(1/6)(1/6)(1/6)(1/6)(1/6)(1/6)**

5. Cathy conducted a survey of the students in her RS1 classes to observe the distribution of eye color. The table below is a record of the results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Eye Color | Blue | Brown | Green | Hazel |
| Number | 12 | 58 | 2 | 8 |

1. Find the experimental probability for each of the following:

i*. P*(Blue) = **12/80 = 3/20** ii. *P*(Brown) = **58/80 = 29/40** iii. *P*(Green) = **2/80 = 1/40**

iv. *P*(Hazel) = **8/80 = 1/10**

1. Based on the survey, what is the experimental probability that a student in Cathy’s class has blue or green eyes? **12/80 + 2/80 = 14/80 = 7/40**
2. Based on the survey, what is the experimental probability that a student in Cathy’s class does not have green or hazel eyes? **12/80 + 58/80 = 70/80 = 7/8 OR 1 – 2/80 – 8/80 = 70/80 = 7/8**
3. If the distribution of eye color in Cathy’s class is similar to the distribution in all of the RS1 classes, about how many of the 460 RS1 students would be expected to have brown eyes?

**(29/40)(460) =333.5**

# Discussion 10: Independent and Dependent Events

Two events are *independent* if the fact that one of the events occurs does not affect the probability that the other occurs. If two events are not independent, they are dependent.

If two events are independent, then the probability of the second event does not change regardless of whether the first occurs. For example, the first roll of a number cube does not influence the second roll of the number cube. Other examples of independent events are, but not limited to: flipping two coins; spinning a spinner and rolling a number cube; flipping a coin and selecting a card; and choosing a card from a deck, replacing the card and selecting again.

The probability of two independent events is found by using the following formula:   
*P(A and B) = P(A)∙P(B)*

* Example: When rolling a six-sided number cube and flipping a coin, simultaneously, what is the probability of rolling a 3 on the cube and getting heads on the coin?
* *P(3 and heads)* =

If the outcome of one event has an impact on the outcome of the other event, the events are called dependent. If events are dependent, then the second event is considered only if the first event has already occurred. For example, if you choose a blue card from a set of nine different colored cards that has a total of four blue cards and you do not place that blue card back in the set before selecting a second card, the chance of selecting a blue card the second time is diminished because there are now only three blue cards remaining in the set. Other examples of dependent events include but are not limited to: choosing two marbles from a bag but not replacing the first after selecting it; determining the probability that it will snow and that school will be cancelled.

The probability of two dependent events is found by using the following formula: *P(A and B) = P(A)∙P(B after A)*

* Example: You have a bag holding a blue ball, a red ball, and a yellow ball. What is the probability of picking a blue ball out of the bag on the first pick then *without* replacing the blue ball in the bag, picking a red ball on the second pick?
* *P*(blue *and* red) = *P*(blue)*∙P*(red *after* blue) =

Examples:

1. Is each pair of events dependent or independent?

1. Roll a number cube. Then spin a spinner.

The two events do not affect each other. They are independent.

1. Pick one flash card, then another from a stack of 30 flash cards without replacing the first card.

Picking the first card affects the possible outcomes on the second pick. The events are dependent.

2. Classify each pair of events as *dependent* or *independent*.

a. A member of the junior class is selected as junior class president and a freshman is selected as freshman class president. Solution: Independent

b. Landing on heads after tossing a coin and rolling a 5 on a single 6-sided die. Solution: Independent

c. Choosing a marble from a jar and then choosing a second marble without replacing the first.

Solution: Dependent

Identify each event as dependent or independent and then calculate the probability.

3. What is the probability that a randomly selected number from 20 to 30 is divisible by 3 and then divisible by 12 after replacing the first number?

Solution: Independent;(3/10)(1/10) = 3/100

4. Two numbers are chosen, without replacement, at random from the English alphabet. If *y* is considered to be a consonant, find the probability that

a. both are vowels

b. both are consonants

Both are dependent. For a, (5/26)(4/25) = 20/650. For b, (21/26)(20/25) = 420/650.

Exercises for Discussion 10

1. Four dice are thrown sequentially. What is the probability that the first die shows six and the other dice show an odd number? **(1/6)(1/2)(1/2)(1/2) = 1/48**

2. In a file, there are 4 science papers, 6 English papers, and 5 history papers. If you select two papers at random, without replacement, what is the probability of getting a science paper and then an English paper? **(4/15)(6/14) = 4/35**

3. A box contains 4 white, 3 blue and 6 red marbles. A marble is drawn from the box, kept, and another marble is drawn. Find the probability that

a. Both marbles are red **(6/13)(5/12) = 5/26**

b. Both marbles are blue **(3/13)(2/12) = 1/26**

c. The first marble is red and the second marble is blue **(6/13)(3/12) = 3/26**

d. Neither is red **1 – 5/26 = 21/26**

4. Three people are in an elevator together. Find the probability that all three were born on the same day. **(1/365)(1/365)(1/365) = 1/48627125**

5. Consider the following table giving the contestants in a dog show.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Westies | Schnauzers | Total |
| Females | 12 | 8 | 20 |
| Males | 10 | 14 | 24 |
| Totals | 22 | 22 | 44 |

Find the probability that a randomly selected dog is

1. A Westie **22/44 = 1/2**
2. A female **20/44 = 5/11**
3. A Westie, given it is a female. **12/20 = 3/5**
4. Male, given it is a Schnauzer. **14/22 = 7/11**