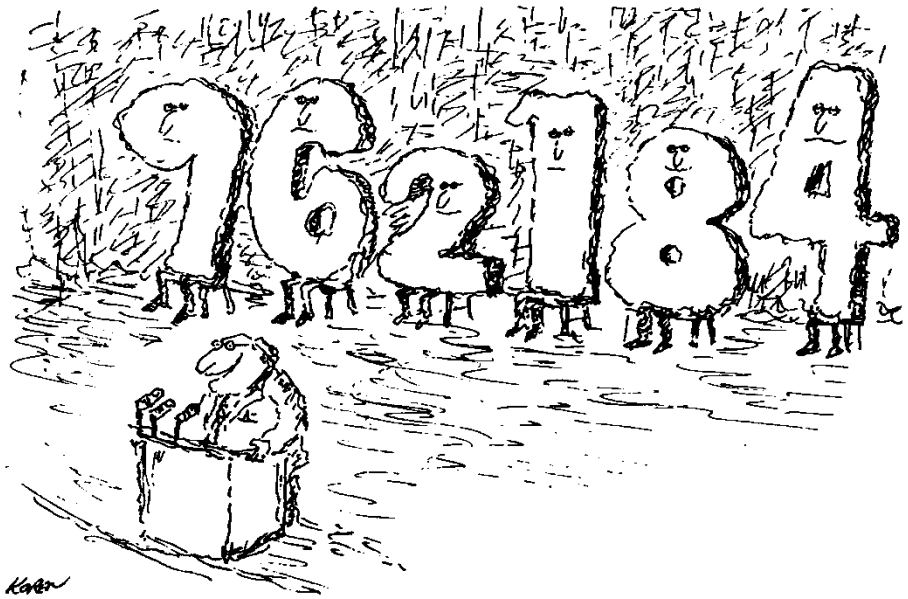
Statistics Review



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

TJHSST

Fall Semester

Review & Self-Study for
Research Statistics 1--KEY

Discussion 1: Measures of Central Tendency

Page 2

Using the previous 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.

   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 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.

   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.

   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.

   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.

   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:

   a) 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 measure of central tendency, even the mean which is affected by the "extremely" high salaries, are lower the $30,000

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

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

Discussion #2: Measures of Spread or Variability

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.

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
variance	33.917	74.515	159.102
standard deviation	5.824	8.632	12.614

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

b) 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 had the most variability, with students performing significantly better and worse than the class average

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

Exercises for Discussion 3

1.

a-c)

FAT CONTENT:

1	**1 2 5 8 8 8 8 8**	***n*=15 mean: 20 g**
2	**0 1 3 3 6 8**	**median: 18 g**
3	**1**	**mode: 18 g**

CARBOHYDRATE CONTENT:

1	**2 5 7 9**	***n*=15 mean: 31.733 g**
2	**8 9**	
3	**3 5 4 6 8**	**median: 34 g**
4	**2 2 6**	
5	**0**	**mode: 42 g**

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

$\mathbf{IQR = 42 - 19 = 23}$

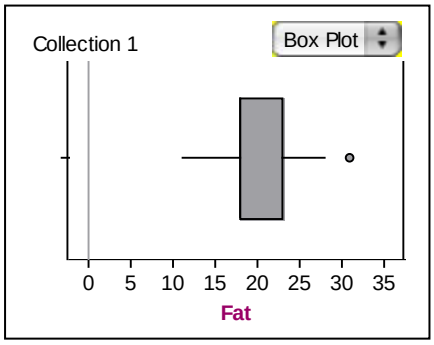
$\mathbf{(1.5)(23) = 34.5}$

$\mathbf{Q1 - 34.5 = -18.5}$

$\mathbf{Q3 - 34.5 = 76.5}$

ok range for outliers: $\mathbf{[-18.5, 76.6]}$ **There are no data points below -18.5 nor above 76.5. Therefore there are no outliers**

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



Note that for the FAT content there is an outlier.
The "ok" interval is [14.5, 26.5], using the formula for outliers.
Therefore the data entry, 31g, is an outlier.

2. Refer to the plots below.

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

b) What was the interquartile range for each class?

IQR_{JR} = 3.5-2.5 = 1.0
IQR_{SR} = 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] Therefore there are no outliers.

Discussion 4: Quantitative Data, Frequency Tables, and Histograms

Exercises for Discussion 4

1. For the following examples, determine whether the data is continuous or discrete:
1. For the following examples, determine whether the data is continuous or discrete:
   a) Population in Fairfax County, Virginia **DISCRETE**
   b) Weight of newspapers collected for recycling on a single day at TJ **CONTINUOUS**
   c) Score on a math test **DISCRETE**
   d) GPA **CONTINUOUS**

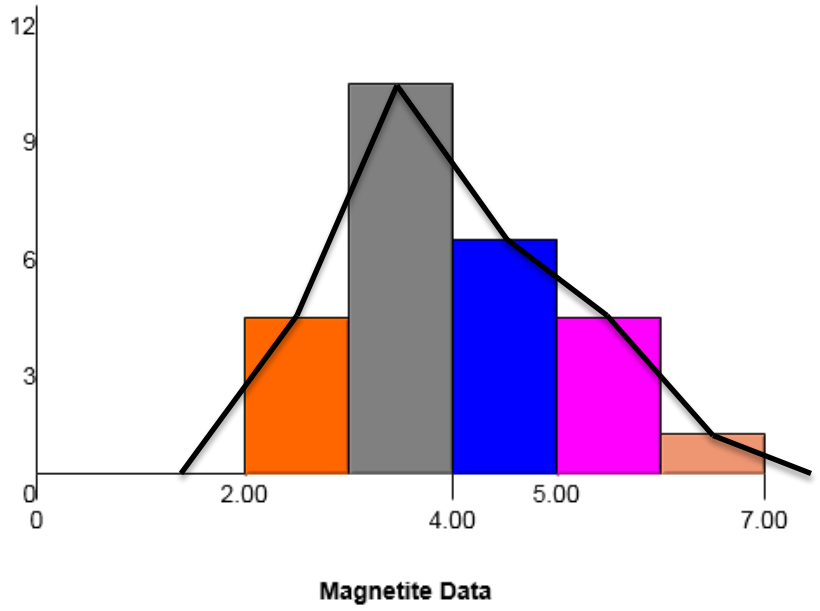
2. For the following set of continuous data, determine an appropriate number of classes and set class limits. Then, set up a frequency table to organize the data.

Low: 2.4
High: 6.2

Answers will vary.

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**

Using the data from the frequency table, design a histogram and then a frequency polygon of the magnetite data.



Magnetite Data

What shape does the frequency polygon appear to have? Explain. **The frequency polygon is skewed slightly to the right.**

Discussion 5: The Scatter Plot

Exercises for Discussion 5:

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.

   a) About how many soft drinks did the average person drink in 1950? **about 105** in 1970? **about 220**

   b) About how many six-packs of soft drinks did the average person drink in 1980? **410/6, approx 68 six packs**

   c) About how many soft drinks did the average person drink per week in 1950? **105/52, approx 2**
   in 1980? **410/52, approx 8**

d) 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. In 200, about 700.**

e) 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.**

2. 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 6: Linear Equations

Exercises for Discussion 6:

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. $m = 3$, $(0, -10)$
$y + 3 = \frac{1}{2}x$	**$y - 5 = -2x$**	**$y - 6 = 0$**	**$y + 10 = 3x$**
$y = \frac{1}{2}x - 3$	**$y = -2x + 5$**	**$y = 6$ (horizontal line)**	**$y = 3x - 10$**

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. $P(-2, 4)$; $m = \frac{2}{3}$
$y - 1 = -3(x + 2)$	**$y + 3 = 4(x + 3)$**	**$y - 4 = \frac{2}{3}(x + 2)$**
$y = -3x - 5$	**$y = 4x + 9$**	**$y = \frac{2}{3}x + \frac{16}{3}$**

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 $(8, 0)$	c. $(-\frac{1}{2}, 2)$ and $(6, 4)$	d. $(-5, 4)$ and $(-5, -2)$
$y + 1 = \frac{2}{5}(x + 3)$ or	**$y - 3 = -\frac{1}{4}(x + 4)$**	**$y - 2 = \frac{4}{13}(x + \frac{1}{2})$**	**Vertical line: $x = -5$**
$y - 1 = \frac{2}{5}(x - 2)$	**$y = -\frac{1}{4}(x - 8)$**	**$y - 4 = \frac{4}{13}(x - 6)$**	**no point-slope or**
$y = \frac{2}{5}x + \frac{1}{5}$	**$y = -\frac{1}{4}x + 2$**	**$y = \frac{4}{13}x + \frac{28}{13}$**	**slope intercept!**

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 $(3, -2)$.

$y + 2 = -\frac{1}{4}(x - 3)$ and $y = -\frac{1}{4}x - \frac{5}{4}$

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

$y - 4 = \frac{1}{3}(x + 3)$ and $y = \frac{1}{3}x + 5$

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.** $y + 6 = -\frac{5}{4}(x + 2)$

m = $\frac{4}{3}$; (2, -6) **m = -4; (6, 1)** **m = $-\frac{5}{4}$; (-2, -6)**

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. $y - 6 = -2(x + 3)$

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

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); $m = 0$

5x + 3y = 31 **x – 2y = 1** **y = 2**

Discussion 7: Linear Equations as Mathematical Models

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.

Let t = number of seconds since pulling the plug and let V = number of gallons of water in tub

$$V - 13 = -\frac{3}{20}(t - 40) \text{ or } V = -\frac{3}{20}t + 19$$

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

i) 16 gallons
ii) 11.5 gallons

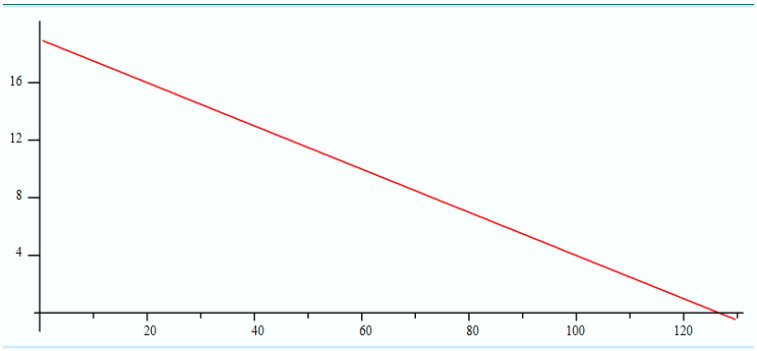
c) Find the gallons-intercept. What does this number represent in the real world?

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?

$126\frac{2}{3}$ seconds; It represents the amount of time needed to empty the tub

e) Plot the graph 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?

$\frac{\text{gallons}}{\text{second}}$ **; This means that water drains out of the tub at a rate of** $\frac{3}{20}$ **gallons per second**

Research 1 Practice Test

Multiple Choice

1. B Adding $30,000 to the total earned by 6 people will crease the mean by $30,000/6 or $5000. Since the owner's salary is already the maximum value in the distribution, increasing it will not change the median.

2. C 110 is the first quartile and 200 is the maximum. This range of weights includes all individuals except the 25% below the first quartile.

Short Answer.

3. a. Describe the shape of this distribution.

This histogram is strongly skewed to the right.

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

This percentage is closer to 10%.

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

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

Mean = **7.309 hours** Standard Deviation = **0.977 hours**

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

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

c. 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.

d. Using an appropriate method, plot the data.

Answers will vary.

Stem (1)	Leaf (0.1)
5	0
6	0000055
7	000000000555555
8	00000555
9	000

e. Using all of your calculations and results from a-d, 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 outliers 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 clearly shows that the center is between 7 and 7.5 hours. The shape of the stem-leaf plot is approximately bell-shaped. The data and graph confirm that on average seniors at this high school got approximately 7.25 hours of sleep the previous night.

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?

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. 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. **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.**

6.

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

Histogram IV displays the variable *number of siblings.* It would be rare for students to have a large number of siblings and very common for students to have one or two siblings.

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

Histogram III displays the variable *price paid for most recent haircut.* Some students will get their hair cut for free by roommates or friends. College students (males particularly) will not be willing to pay a large amount for a haircut. A few females will pay very large amounts. It makes sense to have a large outlier for a particular visit.

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

Histogram I displays the variable *height.* You might expect a more symmetric distribution or one with two peaks (males and females). It's feasible that there were one or two shorter people in the class. Histogram II is also a possibility; perhaps the large peak comes at a value, such as six feet, that many students might round to. It is difficult to explain the extreme outlier in histogram b, unless this is a data entry error or a professional basketball player.

7. Line 1 and Line 2 are perpendicular. Line 1 passes through the points (-8, -4) and (-10, 0). Line 2 passes through the point (x, 9) and (3, 7).

a. Find the missing value of x. **The slope of Line 1 is -2. Perpendicular lines have slopes that are negative reciprocals of one another so the slope of Line 2 is ½. Therefore,** $\frac{7-9}{3-x}=\frac{1}{2}$ **and** $x = 7$.

b. Write the equation of Line 1 in point-slope form. $y = -2x - 20$

c. Write the equation of Line 2 in standard form. $x - 2y = -11$

8. a. **The oldest horse in this sample is female.**
   b. **The most expensive horse in this sample is male.**
   c. **Based on these scatterplots, a 10-year-old male horse is likely to cost more than a 10-year-old female horse.**
   d. **The male gender has a positive association between *price* and *age.***
   e. **The female gender has the stronger association between *price* and *age.***